Title
Optical and Electronic Properties of Nano-Materials from First Principles Computation

Author
Deslippe, Jack

Publication Date
2011

Peer reviewed|Thesis/dissertation

Optical and Electronic Properties of Nano-Materials from First Principles Computation

by

Jack Richard Deslippe

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Physics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Prof. Steven G. Louie, Chair

Prof. Feng Wang

Prof. Daryl Chrzan

Fall 2011

Optical and Electronic Properties of Nano-Materials from First Principles Computation

Abstract

Optical and Electronic Properties of Nano-Materials from First Principles Computation

by

Jack Richard Deslippe

Doctor of Philosophy in Physics

University of California, Berkeley

Prof. Steven G. Louie, Chair

Recent advances in computational physics and chemistry have lead to greater understanding and predictability of the electronic and optical properties of materials. This understanding can be used to impact directly the development of future devices (whose properties depend on the underlying materials) such as light-emitting diodes (LEDs) and photovoltaics. In particular, density functional theory (DFT) has become the standard method for predicting the ground-state properties of solid-state systems, such-as total energies, atomic configurations and phonon frequencies. In the same period, the so called many-body perturbation theory techniques based on the dynamics of the single-particle and two-particle Green's function have become one of the standard methods for predicting the excited state properties associated with the addition of an electron, hole or electron-hole pair into a material. The GW and Bethe-Salpeter equation (GW-BSE) technique is a particularly robust methodology for computing the quasiparticle and excitonic properties of materials.

The challenge over the last several years has been to apply these methods to increasingly complex systems. Nano-materials are materials that are very small (on the order of a nanometer) in at least one dimension (e.g. molecules, tubes/rods and sheets). These materials are of great interest for researchers because they exhibit new and interesting physical and electronic properties compared to those of conventional bulk crystals. These physical properties can often be tuned by controlling the geometry of the materials (for example the chiral angle of a nanotube). Various DFT computer packages have been optimized to compute the ground-state properties of large systems and nano-materials. However, the application of the GW-BSE methodology to large systems and large nano-materials is often thought to be too computationally demanding.

In this work, we will discuss research towards understanding the electronic and optical properties of nano-materials using (and extending) first-principles computational techniques, namely the GW-BSE technique for applications to large systems and nano-materials in particular. While, the GW-BSE approach has, in the past, been prohibitively expensive on systems with more than 50 atoms, in Chapter 2, we show that through a combination methodological and algorithmic improvements, the standard GW-BSE approach can be applied to systems of 500-1000 atoms or 100 AU x 100 AU x 100 AU unit cells. We show that nearly linear parallel scaling of the GW-BSE methodology can be obtained up to tens of thousands (and beyond) of CPUs on current and future high performance supercomputers. In Chapter 3, we will discuss improving the DFT starting point of the GW-BSE

approach through the use of COHSEX exchange-correlations functionals to create a nearly diagonal self-energy matrix. We show applications of this new methodology to molecular systems. In Chapter 4, we discuss the application of the GW-BSE methodology to semiconducting single-walled carbon nanotubes (SWCNTs) and the discovery of novel many-body physics in 1D semiconductors. In Chapter 5, we discuss the application of the GW-BSE methodology to metallic SWCNTs and graphene and the discovery of unexpectedly strong excitonic effects in low-dimensional metals and semi-metals.

To my teachers

Contents

List of Figures

List of Tables

Acknowledgments

I would like to first thank my advisor, Prof. Steven Louie, for his guidance, expertise, patience and support during the years of my doctoral studies at Berkeley. He has been a pleasure to work with and his guidance helped me tremendously in this work. His ability to choose relevant and interesting problems to consider is extremely helpful and his devotion to the field of physics is inspiring.

I would also like to thank all my collaborators in the Louie group, David Prendergast, Catalin Spataru, Li Yang, Cheol-Hwan Park, Georgy Samsonidze, Manish Jain, Peihong Zhang, Johannes Lischner and Rodrigo Capaz for the knowledge they imparted in me and hands on time they gave my questions and problems. I am also grateful to Prof. Marvin Cohen and the Cohen-Louie group for providing a fun and active research environment: Brad Barker, Kevin Chan, Sangkook Choi, Feliciano Giustino, Felipe Jornada, Amy Khoo, Jeff Neaton, Brad Malone, Jonathan Moussa, Jesse Noffsiner, Su-Ying Quek, Filpe Ribeiro, Eric Roman, Jay Sau, Young-Woo Son, David Strubbe, Paul Tangney, Derek Vigil and Jonathan Yates. I also thank Mrs. Katherine Deraadt for all the administrative support and smiles she brings everyday to the office.

It was a great pleasure to collaborate with Prof. Feng Wang and his research groups whose experimental expertise, physical insight and theoretical understanding were extremely helpful in the production of this work. I would also like to thank my other committee member, Prof. Daryl Chrzan, for the useful discussions, past collaborations and the career advice I gained from him.

I would like to acknowledge financial and academic support from the DOE Computational Science Graduate Fellowship (CSGF) for four years. The fellowship gave me the opportunity to learn about other fields of computational science I would not have otherwise been exposed to and helped me choose a career path. I would also like to acknowledge support from the Director, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division, U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Computational resources have been provided by NSF through TeraGrid resources at NICS and by DOE at Lawrence Berkeley National Laboratorys NERSC facility.

Last but not least, I would like to thank my family for their love and support: my two sisters Lisa and Sandy, my parents, my partner Lisa and her family for supporting me these past few years.

Chapter 1

Introduction

Over the past few decades, computational condensed matter physics has benefited tremendously from the advancement of both theoretical and algorithmic methodologies as well as the advancement in computer technology. Condensed matter physicists have used high-performance computers (HPCs) to calculate the properties of solids, liquids and, more recently, nano-materials such as individual molecules, clusters, nanotubes and nanowires. The field has also benefited from the creation of new and more efficient computational techniques that probe materials in novel ways. An example is the now widespread use of the density functional theory (DFT) [60, 76] formalism for computing ground state properties of materials and the so called many-body perturbation theory techniques for excited state material properties [58, 62].

While DFT has been used effectively to study systems with hundreds and thousands of atoms, the many-body perturbation theory techniques, such as the GW method, [58, 62] have been limited to the study of systems whose unit-cell contains only a few or tens of atoms. This severely limits the usefulness of these methods to the study of nanostuctures - one of the classes of materials of greatest interest in current research.

In the following sections of the introduction, we briefly introduce the methods for computing the ground state properties, quasiparticle properties and optical properties of materials that has been developed at Berkeley in the last 25 years. In Chapter 2, we discuss in detail a modern implementation of the GW-Bethe-Salpeter equation (GW-BSE) approach that can compute the electronic and optical properties of large nanostructured materials equivalent to bulk systems with 1000's of atoms in the form of the BerkeleyGW package which scales to tens of thousands of CPUs. In Chapter 3, we discuss the application of this method to molecular systems and other systems that require many empty orbitals. In Chapter 4, we discuss the computation of the quasiparticle and optical properties of semiconducting single walled-carbon nanotubes (SWCNTs) and the unique nature of 1D many-body physics. In the final chapter, Chapter 5, we discuss the application of the GW-BSE method to metallic SWCNTs.

1.1 Basic Electronic Structure Approaches

In principle all the electronic, structural and excited states properties of a material can be determined from the many-body Hamiltonian (for simplicity, we have ommitted relativistic effects):

$$H_{tot} = \sum_j \frac{\mathbf{P}_j^2}{2M_j} + \sum_i \frac{\mathbf{p}_i^2}{2m} + \sum_{j<j'} \frac{Z_j Z_{j'} e^2}{|\mathbf{R}_j - \mathbf{R}_{j'}|} + \sum_{i<i'} \frac{e^2}{|\mathbf{r}_i - \mathbf{r}_{i'}|} + \sum_{i,j} \frac{-Z_j e^2}{|\mathbf{R}_j - \mathbf{r}_i|}. \quad (1.1)$$

Here $\mathbf{R}_j$, $\mathbf{P}_j$, Z_j and M_j refer to nuclei postions, momenta, charge and masses, $\mathbf{r}_i$, $\mathbf{p}_i$ and m_i refer to the positions, momenta and masses of the electrons. This Hamiltonian is expressed in the form of the time-independent Schroedinger equation as the following eigenvalue problem:

$$H\Psi(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N) = E\Psi(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N). \quad (1.2)$$

In this manuscript, we will operate always in the Born-Oppenheimer, or adiabatic, approximation. This approximation assumes that the atomic positions are fixed and can be considered as a set of parameters. This is justified by the fact that the atomic positions are slowly varying compared to motion of the electrons because of the small ratio of masses: (m_e/M). This approximation allows the elimination of two of the terms in H_{tot}, leaving just a Hamiltonian for the electron wavefunction:

$$H_{el} = \sum_i \frac{\mathbf{p}_i^2}{2m} + \sum_{i<i'} \frac{e^2}{|\mathbf{r}_i - \mathbf{r}_{i'}|} + \sum_{i,j} \frac{-Z_j e^2}{|\mathbf{R}_j - \mathbf{r}_i|}. \quad (1.3)$$

We can still determine structural properties, for example the set of atomic positions that minimizes the energy, by considering the change in the total energy as a function of the atomic positions.

An exact eigenvalue/eigenvector decomposition of the Hamiltonian, Eq. 1.3, is intractable in all but the smallest systems. However, such an exact solution is not really desirable. Instead, one is usually interested in computing the properties of a materials that are measurable by experiment such as the response of the material to external probes (optical absorption, photo-emission spectra etc..).

One way to proceed towards getting meaningful information from the Hamiltonian in Eq. 1.3 is to limit the form of the ground state many-electron wavefunction in which one diagonalizes the matrix in Eq. 1.3. This often leads to a set of decoupled single-orbital Schroedinger-like equations in the presence of a mean-field potential. Early attempts at such a formalism were devised by Hartree and Fock, where the self-consistent potential (or mean-field potential) has the form:

$$V_H(\mathbf{r}) = \sum_n \frac{e^2 |\phi_n(\mathbf{r}')|^2}{|\mathbf{r} - \mathbf{r}'|}, \quad (1.4)$$

or

$$V_{HF} = V_H(\mathbf{r}) + V_{ex}, \quad (1.5)$$

where, ϕ_n are the independent electron orbitals and V_{ex} is the nonlocal exchange operator. The first term, V_H on the right side in Eq. 1.5, known as the Hartree potential, describes

the electrostatic interaction of a single electron with the charge density produced by all the electrons in the system in a completely non-correlated way. The second term, known as the exchange-term, results from enforcing the Pauli exclusion principle - i.e. that the many-body wavefunction has to be antisymmetric under particle exchange. The Hartree and Hartree-Fock potentials can be derived from a minimization procedure of the total energy with respect to the ground-state many-body wavefunction in a limited Hilbert space. The many-body wavefunction is restricted to be a simple product of single particle orbitals, in violation of the Pauli-exclusion principle, (the Hartree approximation) or limited to a single Slater determinant of single-particle orbitals (the Hartree-Fock approximation):

$$\psi_H(\mathbf{r}_1, ...\mathbf{r}_N) = \Pi_i \phi_i(\mathbf{r}_i) \tag{1.6}$$

and,

$$\psi_{HF}(\mathbf{r}_1, ...\mathbf{r}_N) = \begin{vmatrix} \phi_1(\mathbf{r}_1 & ... & \phi_N(\mathbf{r}_1) \\ ... & ... & ... \\ \phi_1(\mathbf{r}_N) & ... & \phi_N(\mathbf{r}_N) \end{vmatrix}, \tag{1.7}$$

where ϕ_i are effective single-particle orbitals.

The Hartree-Fock approach for computing the ground-state total energy is inexact because the true many-body wavefunction cannot, in general, be written as a single Slater determinant; rather, it is composed of many Slater determinants that can be formed from a complete set of single-particle orbitals with a given particle number. Minimizing the total-energy in a Hilbert space beyond that of the Hartree-Fock approximation yields corrections that lower the total energy of a many-electron system. The energy difference between the Hartree-Fock ground state and the true many-body ground state is defined as the correlation energy. Techniques exist that systematically improve the many-body wavefunction Hilbert space by including large numbers of Slater-determinants. For example, one technique would be to all include Slater-determinants that can be achieved by swapping the positions of an occupied and unoccupied orbital in the Hartree-fock theory, while another would approach would be to include all Slater determinants that can be created by some small number of single-particle orbitals. These techniques are called multi configuration interaction (CI) techniques, and, their cost generally scales exponentially with the number of single-particle orbitals included. Thus, the formalism becomes quickly prohibitively expensive computationally except when applied to atoms and small molecules typically studied in quantum chemistry.

Before going further it is worth separating physical properties of materials that are associated with all the electrons in the ground state configuration (ground-state properties) vs. properties associated with the addition of a single electron or hole and those associated with the excitation of an electron and a hole simultaneously. Ground state properties include total energies, atomic positions and structural properties like elastic constants and phonon-modes. Measurable quantities related to single-particle excitations in solids include photo-emission and inverse photo-emission as well as transport properties. Two particle (neutral electron + hole) excitations account for most optical properties of solids where an electron, roughly speaking, is promoted to a previously empty orbital, leaving behind a hole in a valence state.

1.2 Ground state properties within density functional theory (DFT)

Density functional theory (DFT) has become a standard theory for computing ground state properties of solid-state systems from first principles. Unlike the the Hartree and Hartree-Fock theories described in the previous section, DFT, in principle (but not necessarily in practice as we discuss below), yields the exact ground state total energies and electron charge density of a many-body interacting system. Also unlike Hartree-Fock theory and CI techniques, DFT does not attempt to solve for the ground state wavefunction of Eq. 1.3 within a restricted basis. Instead, it formulates the total-energy problem in terms of the electron density - a much simpler quantity than the complete many-body wavefunction of the grounds state.

The Hohenberg-Kohn theorem [60] sets up a one-to-one correspondence between the external potential and the electron density of an interacting electron system. Theoretically, if one knows the external potential, one can solve for the eigenstates of Eq. 1.1 and determine all the properties of the system. Therefore, all the relevant physical properties of a material can be known, in principle, if one knows the ground-state density. That is to say, all properties of interest may be written as a functional of the ground-state electron density. In particular, there exists a functional of the density for the total energy of an interacting system:

$$E_V\left[\rho\right] = \int V(\mathbf{r})\,\rho(\mathbf{r})\,d\mathbf{r} \;+\; T_s\left[\rho\right] \;+\; \frac{1}{2}\int \frac{\rho(\mathbf{r})\,\rho(\mathbf{r}')}{|\mathbf{r}\,-\,\mathbf{r}'|}\,d\mathbf{r}d\mathbf{r}' \;+\; E_{xc}\left[\rho\right], \tag{1.8}$$

where $V(r)$ is some external potential (for example the potential from the atomic nuclei) and T_s is the kinetic energy of an equivalent non-interacting system with the same density. The third term of the right hand site of Eq. 1.8 is the Hartree energy of the interacting system. The final term on the right hand side represents the remaining contribution to the total energy not captured in the previous three terms. This term is called the exchange-correlation energy. It includes the exchange energy discussed above with respect to Hartree-Fock theory and contributions to the total energy beyond the exchange (i.e. that derive from the fact that the many-body wavefunction cannot be written as a single Slater determinant) collectively termed "correlation." Included in the DFT correlation energy is the correction to the kinetic energy of the interacting system from that of the non-interacting counterpart with the equivalent density.

From the variational principle, if the exact form of $E_V[\rho]$ was known, one could find the ground-state total energy by minimizing the functional with respect to the density. However, the exact form of $E_{xc}[\rho]$ is not known and is expected to be a non-trivial non-analytic functional of the density. [121] In most practical applications, we do not minimize the total energy with respect to the density directly, but instead we vary the density by constructing the many-particle density from single-particle (non-interacting) orbitals. This scheme is due to Kohn and Sham [76] and maps each interacting system to a non-interacting system in the presence of a potential (the Kohn-Sham potential) where the ground-state density matches that of the interacting system. The non-interacting wavefunctions are taken to be a set of single-particle orbitals (as was the case in the Hartree-Fock approximation

discussed above) subject to the following Kohn-Sham equations:

$$\left\{ \frac{p^2}{2m} \; + \; V(\mathbf{r}) \; + \; V_H\left(\mathbf{r}\right) + V_{xc}(\mathbf{r}) \right\} \; \varphi_i(\mathbf{r}) \; = \; \varepsilon_i \; \varphi_i(\mathbf{r}) \tag{1.9}$$

with

$$\rho(\mathbf{r}) \; = \; \sum_{i}^{occ} \left| \varphi_i\left(\mathbf{r}\right) \right|^2 \tag{1.10}$$

and

$$V_{xc}(\mathbf{r}) = \frac{\delta\, E_{xc}}{\delta\, \rho(\mathbf{r})} \; . \tag{1.11}$$

Here $V_H(r)$ is the Hartree potential, $V_{xc}(r)$ is the exchange-correlation potential (which must be approximated) and ε_i are eigenvalues of the Kohn-Sham equations. In principle, if the exact exchange-correlation functional (and therefore exchange-correlation potential) were known, these equations would yield the exact ground state density of the interacting system. In practice though, one must approximate the exchange-correlation potential. One of the earliest and most successful approximations is known as the local density approximation. [76] Here the exchange correlation energy is approximated as:

$$E_{xc}[\rho] = \int \rho(r)\epsilon_{xc}(\rho(r))dr \tag{1.12}$$

where $\epsilon_{xc}(\rho(r))$ is the exchange-correlation energy per volume of a homogeneous electron gas of density $\rho(r)$ [125, 30, 108]. In the calculations presented in Chapters 3-5, we use the LDA approximation for the exchange-correlation functional unless otherwise noted.

It is important to note that the eigenvalues in the Kohn-Sham equations should not be considered as exact quasiparticle energies. They are formally only Lagrange multipliers in the minimization scheme. It was the misinterpretation of these eigenvalues as quasiparticle energies that led to the well known "band gap problem" - that the Kohn-Sham eigenvalues consistently underestimate the electronic band gap [121]. This underestimation can be by as much as 50% in cases like Si or a qualitative difference in the band topology, such as in the case of Ge where DFT within the local-density approximation (LDA) actually predicts Ge to be a metal (i.e. to have a negative band gap) whereas the quasiparticle gap is finite. See Figure 1.3 for a comparison between DFT gaps, GW gaps and experiment.

1.3 Quasiparticle properties and the GW Method

As mentioned in the previous section, the Kohn-Sham eigenvalues come into the theory as Lagrange multipliers in the variational procedure. With the exception of the eigenvalue of the valence band maximum (VBM) state corresponding to the ionization energy [69], they should not be formally considered as the quasi-electron or quasi-hole energies. Interpreting the Kohn-Sham eigenvalues as the quasiparticle energies leads to a vast underestimation of the band gap in many materials. This is essentially because DFT, within the Kohn-Sham formalism, is a theory for the ground-state. Therefore, we turn

instead to a theory based on the Green's function in order to describe electron excitations in solids and nanostructures.

A theory of the Green's function is more appropriate for describing electron excitations (than a theory of the density for example) because the poles of the single-particle Green's function in frequency space are known to contain the excited state energies of the $N+1$ and $N-1$ interacting electron systems. Here N is the number of electrons in the orginal system with all electrons in the ground-state. The single-particle Green's function describes directly the propagation of single-particle like excitations in the many-body system. It is defined as:

$$G(\mathbf{r},\mathbf{r}',\tau) = -i\langle 0 \left| T\{\psi(\mathbf{r},\tau)\psi^\dagger(\mathbf{r}',0)\} \right| 0\rangle, \tag{1.13}$$

where $\psi(\mathbf{r},t)$ are the second quantized operators for creating a particle at position $\mathbf{r}$ and time t in the Heisenberg picture and $|0\rangle$ is the N-particle ground state. The physical meaning of the Green's function is the probability of finding a particle at time τ and position $\mathbf{r}$ if one was added to the ground-state of the N-particle system at time 0 and position $\mathbf{r}'$. One may transform this expression to the basis of some appropriately chosen mean-field single-particle orbitals and arrive at:

$$G(p,\tau) = -i\langle 0 \left| T\{c_p(\tau)c_p^\dagger(0)\} \right| 0\rangle, \tag{1.14}$$

where $c_p(t)$ are the second quantized operators for creating a particle at quantum number p (for example, band index n and crystal momentum $\mathbf{k}$) and time t.

For system with periodic translation symmetry (such as bulk crystals), one can express the Green's function in another representation, the Lehmann representation (described here for the interacting electron gas for simplicity): [43]

$$G(\mathbf{k},\omega) = V\sum_i \left[\frac{\langle\Psi_0|\psi(0)|\Psi_{i\mathbf{k}}\rangle\langle\Psi_{i\mathbf{k}}|\psi^+(0)|\Psi_0\rangle}{\omega-\mu-\varepsilon_i(N+1)+i\eta} + \frac{\langle\Psi_0|\psi(0)|\Psi_{i-\mathbf{k}}\rangle\langle\Psi_{i-\mathbf{k}}|\psi^+(0)|\Psi_0\rangle}{\omega-\mu+\varepsilon_i(N-1)-i\eta} \right] \tag{1.15}$$

where, Ψ is the many-body states (0 represents the N-particle system ground state and $i\mathbf{k}$ represents the ith many-body excited state with wavevector $\mathbf{k}$ of the $N+1/N-1$ particle system), μ is the chemical potential and ε_i is the energy of the ith excited state of the $N+1/N-1$ particle system above the $N+1/N-1$ ground state energy, and the second quantized destruction operator is $\psi(\mathbf{r}) = e^{-i\mathbf{p}\cdot\mathbf{r}}\psi(\mathbf{0})e^{i\mathbf{p}\cdot\mathbf{r}}$. [43, 85] For a *non-interacting* system, the Lehmann representations yields a Green's function with simple poles at the independent electron excitation energies of the $N+1$ and $N-1$ particle systems of a give wavevector. For a non-interacting system, the $i\mathbf{k}$ state represents an addition or subtraction from an electron from a single particle state that contributes to a single isolated pole in the Green's function. For a system with moderate electron-electron interactions, $G(\mathbf{k},\omega)$ along the real ω axis consists of well-defined peaks, similar to the sharp poles in the case of the noninteracting spectrum, but each peak now has a finite width corresponding to a pole position in the analytical continuation of G off the real axis.

Each pole in G corresponds to pole in the spectral function, $A(\omega) = (1/\pi)|\mathrm{Im}G(\omega)|$, of the form:

$$A(p,\omega) = \frac{\frac{i}{2\pi}Z_p}{\omega-[E_p-\mu]} + c.c. + \text{ correction terms.} \tag{1.16}$$

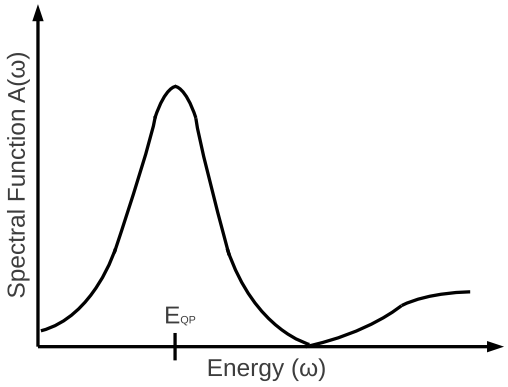

Figure 1.1: Diagram of the spectral function around a pole in the Green's function. The spectral function contains a quasiparticle peaks and a background. The main peak has a Lorentzian shape and represents the quasiparticle. The width of the peak is related to the quasiparticle lifetime.

See Fig. 1.1 for qualitative picture. Since, G can also be written in terms of the spectral function as:

$$G(p,\omega) = \int_C \frac{A(p,\omega')}{\omega - \omega'} d\omega', \tag{1.17}$$

(where C is an appropriate contour) we can approximate G around a pole as:

$$G(p,\tau) = -iZ_p e^{-iRe(E_p)\tau}\, e^{-\Gamma_p \tau} + \text{ correction terms}, \tag{1.18}$$

where Z_p is the renormalization factor, and E_p is the real part of the complex energy whose real-part gives the single-particle like time oscillation and whose imaginary part, Γ_p, gives rise to a damping in time. The usual interpretation of this structure is that it represents a quasiparticle (single-electron like) state. Since the quasiparticle is not a true eigenstate of the $N+1$ or $N-1$ interacting electron system, it acquires a finite lifetime, $1/\Gamma_p$.

It can be shown that the time evolution of the Green's function obeys the following Dyson's equation:

$$G(\mathbf{r},\mathbf{r}';\omega) = G_0(\mathbf{r},\mathbf{r}';\omega) + \int G_0(\mathbf{r},\mathbf{r}_1;\omega)\Sigma(\mathbf{r}_1,\mathbf{r}_2;\omega)G(\mathbf{r}_2,\mathbf{r}';\omega)d\mathbf{r}_1 d\mathbf{r}_2 \tag{1.19}$$

or equivalently,

$$(\hbar\omega - H_o - V_H)G(\mathbf{r},\mathbf{r}';\omega) - \int \Sigma(\mathbf{r},\mathbf{r}'',\omega)G(\mathbf{r}'',\mathbf{r}';\omega) = \delta(\mathbf{r},\mathbf{r}'). \tag{1.20}$$

This equation is shown diagrammatically in Fig. 1.2. Here, Σ is a non-Hermitian, non-local, energy dependent operator that includes the effects of exchange and correlation. If

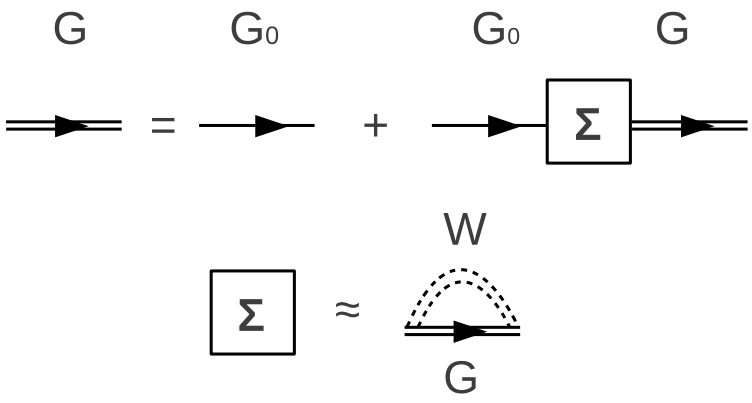

Figure 1.2: Diagramatic representation of the Dyson equation and the GW approximation for Σ.

we express G in the spectral representation, as in Eq. 1.17,

$$G(\mathbf{r},\mathbf{r}';\omega) = \sum_{n\mathbf{k}} \frac{\psi_{n\mathbf{k}}(\mathbf{r})\psi^*_{n\mathbf{k}}(\mathbf{r}')}{\omega - E_{n\mathbf{k}} - i\delta_{n\mathbf{k}}}, \tag{1.21}$$

where $E_{n\mathbf{k}}$ is a complex energy, we arrive at a reworked Dyson's equation for the quasiparticle wavefunctions and energies:

$$[E_{n\mathbf{k}} - H_o(\mathbf{r}) - V_H(\mathbf{r})]\ \psi_{n\mathbf{k}}(r) - \int \Sigma(\mathbf{r},\mathbf{r}';E_{n\mathbf{k}})\psi_{n\mathbf{k}}(\mathbf{r}')d\mathbf{r}' = 0. \tag{1.22}$$

Here $\psi_{n\mathbf{k}}$ and $E_{n\mathbf{k}}$ are the quasiparticle wavefunctions and complex energies respectively. This set of equations looks similar to the Kohn-Sham equations with the exception that the exchange-correlation potential, V_{xc}, is replaced by a non-local, non-Hermitian and energy dependent operator, Σ.

Both the energy dependence and non-locality of Σ make solving the Dyson's equation considerably more complex than solving the Kohn-Sham equations. As discussed in Chapter 2, we typically do not construct Eq. 1.22 as a matrix equation and diagonalize to find right and left eigenstates. Instead, we usually start with a suitable mean-field approximation for Σ, such as the V_{xc} from DFT in a suitable approximation (such as LDA) and treat the self energy operator in Eq. 1.22 as a perturbation. In other words we consider the self-energy in terms of the perturbation $\Sigma - V_{xc}$. The validity of this approach requires that the off-diagonal elements of $\Sigma - V_{xc}$ are small. In Chapter 3, we discuss that in the case of

molecular systems, this assumption often fails if one uses V_{xc} within DFT as the mean-field starting point. In that chapter, we propose two more appropriate mean-field choices.

In order to proceed, we must have an approximation for the non-local, energy dependent self energy operator, Σ, in Eq. 1.22. A systematic expansion of Σ in terms of the screened-coulomb interaction, W, was first worked out by Hedin [58]. It leads to a series of coupled equations for the Green's function, G, the polarizability (or dielectric) matrix P (used to screen the Coulomb interaction), and the vertex function Γ. The self energy operator Σ is given by:

$$\Sigma(1,2) = i \int G(1,3^+)W(1,4)\Gamma(3,2,4)d(34) \tag{1.23}$$

where each number corresponds to a composite space, time coordinate, $(1) \rightarrow (\mathbf{r}_1, t_1)$, and the + indicates an infinitesimal is added to the time coordinate. The vertex function, Γ, is defined through:

$$\Gamma(1,2,3) = \delta(1,2)\delta(2,3) + \int \frac{\delta\Sigma(1,2)}{\delta G(4,5)} G(4,6)G(7,5)\Gamma(6,7,3)d(4567), \tag{1.24}$$

and the screened Coulomb interaction W is defined as:

$$W(1,2) = v(1,2) + \int v(1,3)\chi(3,4)W(4,2)d(34). \tag{1.25}$$

Here $v(1,2)$ is the bare Coulomb interaction, $1/|\mathbf{r}_1 - \mathbf{r}_2|$, in atomic units and χ is the polarizability matrix that can be related to G as:

$$\chi(1,2) = -i \int G(1,3)\Gamma(3,4,2)G(4,1^+)d(34). \tag{1.26}$$

The four equations 1.23 - 1.26 in combination with the Dyson's equation are collectively known as Hedin's equations. [58, 59]

The prescription proposed by Hedin for evaluating this set of equations is to first take the simplest approximation for the vertex function: $\Gamma(1,2,3) \approx \delta(1,2)\delta(2,3)$. In this approximation we arrive at the following equations:

$$\Sigma(1,2) = iG(1,2)W(1^+,2), \tag{1.27}$$

$$\chi(1,2) = iG(1,2^+)G(2,1) \tag{1.28}$$

and,

$$W(1,2) = v(1,2) + \int v(1,3)\chi(3,4)W(4,2)d(34). \tag{1.29}$$

This level of approximation is the one typically used in first-principles GW implementations. [63, 62]

Let us now discuss the mean-field starting point of GW calculations in more detail. In principle, the GW formalism is not dependent on the density-functional formalism and it can be solved in any basis, so long as the basis is complete and the Dyson's equation is solved in full within the basis in a self-consistent fashion. In practice, though, the GW

calculations presented in this manuscript, with the exception of Chapter 3, take as input the Kohn-Sham wavefunctions and eigenvalues. For the case of most bulk semiconductors, and even the nanostructures considered presently, it is found that the Dyson's equation, expressed as a matrix equation in the Kohn-Sham basis is nearly diagonal. This is not to say that DFT does an accurate job computing the quasiparticle energies in the system; it only means that the Schroedinger-like equation with either the DFT V_{xc} or $\Sigma = iGW$ often have similar eigenfunctions. The eigenvalues may still differ significantly as was the case in the band gap problem discussed above. In the case of silicon, the Kohn-Sham wavefunctions within LDA differ from the fully diagonalized GW quasiparticle wavefunctions by less than 0.1% [63]. However, there are important cases, particularly in molecules, where expressing Σ as a matrix within the LDA orbital basis does not yield a diagonal representation and a more preferred basis set can be used. We will discuss such cases in Chapter 3.

Since in many cases, as discussed above, only the diagonal elements are sizable within the Kohn-Sham orbital basis used throughout the manuscript, for the purposes of the nanostructures studied, the effects of Σ can be treated perturbatively. Thus, for the systems considered, we treat Σ as $\Sigma = V_{xc} + (\Sigma - V_{xc})$, where V_{xc} is an approximation to the diagonal elements of the Kohn-Sham exchange-correlation potential. In principle, the process of correcting the eigenfunctions and eigenvalues that are used to construct Σ could be repeated until self-consistency is reached; however, in practice, it is found that the first-order perturbation theory approach for a given Σ is sufficient. From practice, self-consistency in Σ is found to only improve the accuracy of the GW result if it is coupled with the inclusion of a vertex function in the Σ operator itself. [61]

The most computationally demanding ingredient of a GW calculation is the dynamic dielectric matrix. As we will discuss in the next chapter, the generation of this matrix usually requires a sum over Kohn-Sham empty orbitals. Both the generation of the required empty orbitals and the computation of the matrix elements are computationally expensive. As described in Chapter 3, requiring an even larger number of empty orbitals is the expression for Σ itself. Σ, within the GW approximation, can be broken into two terms (see the next chapter for details): $\Sigma_{\mathrm{COH}} + \Sigma_{\mathrm{SEX}}$. Σ_{SEX}, called the screened-exchange term, is simply the exchange, or Fock, operator from Hartee-Fock screened with the dielectric matrix. Σ_{COH} is a term absent in Hartree-Fock theory describing the interaction of an electron with the induced-charge created around the charged quasiparticle. Σ_{COH}, in practice, involves a sum over an even a larger number of empty Kohn-Sham orbitals than required in the polarizability sum. In practice, the dependence of Σ_{COH} on empty orbitals significantly increases the computational time needed to generate the unoccupied Kohn-Sham orbitals - particularly in nanosystems such as molecules, where absolute quasiparticle energies are required. As lot of recent research effort has been spent on approximating or eliminating the required sum over empty states. [136, 49, 140, 141, 25, 113] We discuss in Chapter 3 a new method to reduce the number of empty orbitals required for computing Σ.

The GW methodology has been successfully applied over the last two decades to a variety of systems from bulk crystalline semiconductors, insulators and metals to molecules, and more recently nanostructures of increasing complexity and size. Figure 1.3 shows electronic band gaps of the GW methodology compared to DFT and experiment. For more implementation details of the GW methodology with a particular emphasis towards the

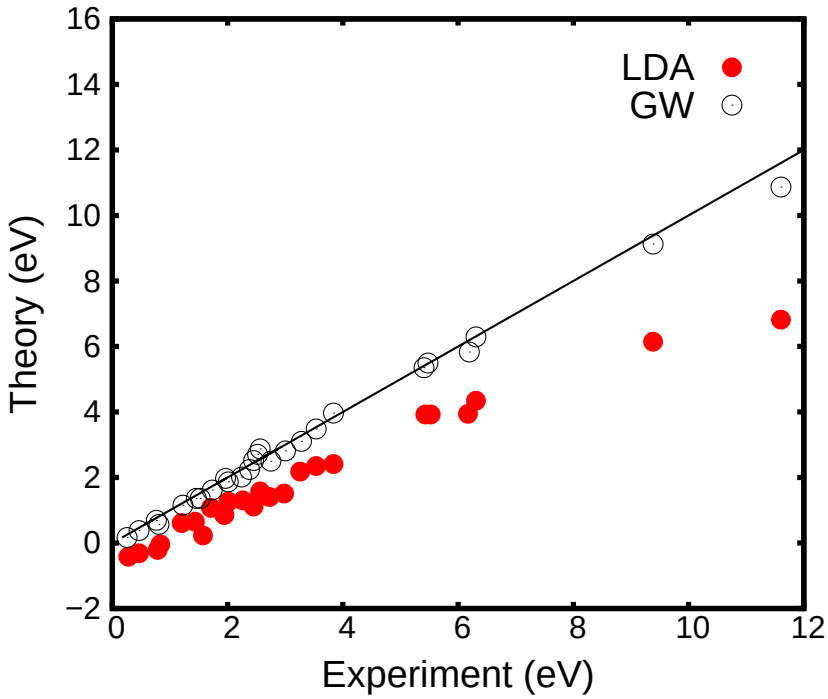

Figure 1.3: Comparison of the computed LDA (solid circles) and GW (empty circles) energy gaps for a variety of sytems to experiment. The black line, y=x, represents the experimental gaps. Data from [84].

extension of the method to the study of large systems and nanostructures, see Chapter 2.

1.4 Optical properties and the Bethe-Salpeter equation (BSE)

The single-particle Green's function described in the previous section has poles at complex energies with real parts related to the single particle excitation energies - the energies associated with creating a single electron or single hole on top of the N-particle ground state. However, since the GW methodology is a one-particle theory, it is not intended to yield accurate optical properties. This is because optical properties are related to neutral excitations: that is, the excitation of both an electron and a hole. In many systems (with the exception perhaps of bulk metals), the electron and the hole interact strongly. This interaction leads to a qualitative difference in optical spectra. [115]

The approach we take, is to consider the two-particle Green's function and the Bethe-Salpeter equation method for its evaluation. [132, 115] In this scheme, we approximate the neutral excited states of an N-particle system as a superposition of quasi-electron

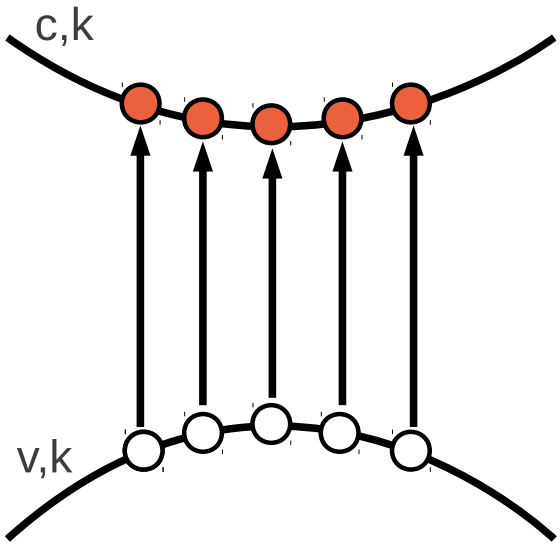

Figure 1.4: Schematic of the single-particle like electron-hole transitions that make up the basis set for which the interacting wavefunction A^S_{vc} is expanded.

and quasi-hole states plus some correction (assumed to be small):

$$|N,S\rangle = \sum_{v}^{hole}\sum_{c}^{elec} A^S_{vc}\, a^\dagger_v\, b^\dagger_c\ |N,0\rangle \quad + \quad ... \tag{1.30}$$

where S is an index labeling a particular neutral two-particle excited state, called an exciton, of the N-interacting electron system, and $a^\dagger$ and $b^\dagger$ are the creation operators for holes and electrons respectively. The origin of this expression can be found from defining a quasi-two-particle wavefunction analagous to the quasi-particle wavefunctions of the previous section as:

$$\chi_S(\mathbf{x},\mathbf{x}') = -<N,0|\psi_v(\mathbf{x}')\psi^\dagger_c(\mathbf{x})|N,S> . \tag{1.31}$$

Here, $|N,0>$ and $|N,S>$ refer to the N-particle ground state and two-particle excited states respectively. $\mathbf{x}$ refers to a joint space, time coordinate $(\mathbf{r},t)$. We can expand this function in the basis of the quasi-electron and quasi-hole wavefunctions as:

$$\chi_S(\mathbf{r},\mathbf{r}')=\sum_{cv} A^S_{cv}\ \psi_c(\mathbf{r})\psi^*_v(\mathbf{r}'), \tag{1.32}$$

This is illustrated in Fig. 1.4.

In the works of Strinati [132] and Rohlfing and Louie [115], it is shown that χ_S satisfies the following equation (shown within the Tamm-Dancoff approximation [115] for simplicity):

$$(E_c - E_v)\, A^S_{cv} + \sum_{cv,c'v'} K_{cv,c'v'}\,(\Omega_S)\, A^S_{c'v'} = \Omega_S\, A^S_{cv} \quad . \tag{1.33}$$

Here, Ω_S is the excitation energy of the two-particle excited state, $E_S - E_0$. For bound-exciton states, i.e. states where the hole amplitude is localized around the electron position, the exciton binding can be inferred from the difference in energy between Ω_S and the minima of all $E_c - E_v$ non-interacting energies that contribute to the exciton state (i.e. have non-zero A^S_{cv}). This equation is termed the Bethe-Salpeter equation since it derives from the Bethe-Salpeter equation for the two particle Green's function in much the same way the Dyson's equation 1.22 derives from the Dyson's equation for the single particle Green's function. K, termed the electron-hole interaction Kernel, contains the effective interaction between the quasi-electron and quasi-hole and is given by:

$$K(12,34)=\frac{\delta[V_H\delta(1,3)+\Sigma(1,3)]}{\delta G(4,2)}, \tag{1.34}$$

and often approximated as: [115]

$$\begin{aligned} K(12,34) &= -i\delta(1,3)\delta(2^-,4)V_c(1,4) + i\delta(1,4)\delta(3,2)W(1^+,3) \\ &= K_x + K_d. \end{aligned} \tag{1.35}$$

Here, K_x is a bare-exchange interaction similar to that discussed above in reference to Hartree-Fock theory. K_d is a screened-direct interaction that can be physically understood as the Coulomb interaction between the charged electron and charged hole and the induced screening electrons of the background system.

One may compute directly the optical spectra from the two-particle excited states $|N,S>$. For example the imaginary part of the macroscopic dielectric matrix (related to many optics phenomena such as absorption and scattering spectroscopy) is:

$$\varepsilon_2(\omega) = \frac{16\pi^2 e^2}{\omega^2} \sum_s \left|\langle N,0|\,\hat{\mathbf{e}}\cdot\mathbf{v}\,|N,S\rangle\right|^2 \delta(\Omega_S - \hbar\omega), \tag{1.36}$$

where the matrix elements defining the oscillator strength for each transition are:

$$\langle N,0|\,\hat{\mathbf{e}}\cdot\mathbf{v}\,|N,S\rangle = \sum_{cv} A^S_{cv}\ \langle c|\,\hat{\mathbf{e}}\cdot\mathbf{v}\,|v\rangle\,. \tag{1.37}$$

With the excitation energies and amplitudes of the electron-hole pairs, A, one can also obtain higher order optical effects such as multi-photon absorption and phonon-assisted absorption spectra. When applying this method to isolated nanosystems in supercell calculations, it is important, as it is in the GW calculation, to replace W with the appropriate screened truncated interaction. As will be illustrated in the final Chapter, even with well-separated systems that are considered reasonable in DFT calculations, the nature of the interaction between the electron and hole in an untruncated interaction calculation can be very different from the isolated case owing to the unwanted influence of neighboring replicas. Because of the generally reduced screening and confinement effects, one expects stronger excitonic effects in reduced dimensional systems, which as we will see in the following chapters, is indeed the case.

Chapter 2

A Modern Implementation of the GW-BSE Method for Complex Materials and Nanostructures:

The GW formalism [58] laid out in Chapter 1 was first developed as an *ab inito* methodology in computational research codes in the late 1980's [63, 62] with mainly traditional bulk systems in mind. Over the last few decades, the methodology has been successfully applied to the study of the quasiparticle properties of a large range of material systems from traditional bulk semiconductors, insulators and metals to, more recently, nanosystems like polymers, nano-wires and molecules [62, 83, 127, 128, 37, 111]. The GW approach has proven to yield quantitatively accurate quasiparticle band gaps and dispersion relations from first principles.

Additionally, the Bethe-Salpeter equation (BSE) approach to the optical properties of materials has proven exceptionally accurate in predicting the optical response of a similarly large class of materials within the same approximations as GW for Σ [132, 115, 7, 20].

The combined GW-BSE approach is now arguably regarded as the most accurate methodology commonly used for computing the quasiparticle and optical properties within computational physics of condensed-matter systems from first principles. However, a percieved drawback of the GW methodology is its computational cost; a GW-BSE calculation is usually thought to be an order of magnitude (or worse) more than a typical DFT calculation for the same system. Since the pioneering work of Ref [62] many GW implementations have been made, but all are limited to small systems of the size of 10's of atoms, and scaling to only small numbers of CPUs on the order of 100. Thus, there is a great need in the community for a modern implementation of the GW-BSE methodology for use on large and complex materials. We have developed such a modern implementation, in the form of the BerkeleyGW package, over the past several years in order to meet this need.

BerkeleyGW is a massively parallel computer package that implements the *ab initio* GW methodology of Hybertsen and Louie [62] and includes many more recent advances, such as the Bethe-Salpeter equation approach for optical properties [115]. It alleviates the restriction to small numbers of atoms and scales beyond thousands of CPUs. The package is intended to be used on top of a number of mean-field (DFT and other) codes that focus

on ground-state properties such as `PARATEC` [3], `Quantum ESPRESSO` [48], `SIESTA` [126, 104] and an empirical pseudopotential code (EPM) included in the package (based on TBPW [89]).

In this chapter, we will summarize some of the major sections, advances and implementation details of the BerkeleyGW package that are relevant for the study of nanostructured materials presented in the subsequent chapters. We present a breakdown of a typical calculation on a large nano-system utilizing the package. We save for the appendix many of the detailed implementation issues and performance details, including the parallel processor scaling - an issue of great importance to utilizing the package on complex materials. For the interested reader, the latest source code and help forums can be found by visiting the website at `http://berkeleygw.org/`.

2.1 Theoretical Framework

Let us first summarize the relevant parts of the *ab initio* GW-BSE approach whose theory was laid out in Chapter 1 as it is implemented in the BerkeleyGW package. The *ab initio* GW-BSE approach is a many-body Green's-function methodology whose only input parameters are the constituent atoms and the approximate structure of the system [62, 115]. Typical calculations on the ground- and excited-state properties using the GW-BSE method can be broken down into three steps: 1) the solution of the ground-state structural and electronic properties within a suitable ground-state theory such as the pseudopotential density-functional theory (DFT), 2) the calculation of the quasiparticle energy values and wavefunctions within the GW approximation for the electron self-energy operator, and 3) the calculation of the two-particle correlated electron-hole excited states through the solution of a Bethe-Salpeter equation.

DFT calculations, often the chosen starting point for GW, are performed by solving the self-consistent Kohn-Sham equations with an approximate functional for the exchange-correlation potential, $V_{\rm xc}$ – one common approximation being the local density approximation (LDA) [76]:

$$\left[-\frac{1}{2}\nabla^2 + V_{\rm ion} + V_{\rm H} + V_{\rm xc}^{\rm DFT}\right]\psi_{n\mathbf{k}}^{\rm DFT} = E_{n\mathbf{k}}^{\rm DFT}\psi_{n\mathbf{k}}^{\rm DFT} \tag{2.1}$$

where $E_{n\mathbf{k}}^{\rm DFT}$ and $\psi_{n\mathbf{k}}^{\rm DFT}$ are the Kohn-Sham eigenvalues and eigenfunctions respectively, $V_{\rm ion}$ is the ionic potential, $V_{\rm H}$ is the Hartree potential and $V_{\rm xc}$ is the exchange-correlation potential within a suitable approximation. When DFT is chosen as the starting point for GW, the Kohn-Sham wavefunctions and eigenvalues are used here as a first guess for their quasiparticle counterparts. The quasiparticle energies and wavefunctions (*i.e.* the one-particle excitations) are computed by solving the following Dyson equation [59, 62]:

$$\left[-\frac{1}{2}\nabla^2 + V_{\rm ion} + V_{\rm H} + \Sigma(E_{n\mathbf{k}}^{\rm QP})\right]\psi_{n\mathbf{k}}^{\rm QP} = E_{n\mathbf{k}}^{\rm QP}\psi_{n\mathbf{k}}^{\rm QP} \tag{2.2}$$

where Σ is the self-energy operator within the GW approximation, and $E_{n\mathbf{k}}^{\rm QP}$ and $\psi_{n\mathbf{k}}^{\rm QP}$ are the quasiparticle energies and wavefunctions, respectively. For systems of dimension less

than three, the Coulomb interaction may be replaced by a truncated interaction. The interaction is set to zero for particle separation beyond the size of the material in order to avoid unphysical interaction between the material and its periodic images in a super-cell [32] calculation. The electron-hole excitation states (probed in optical or other measurements) are calculated through the solution of a Bethe-Salpeter equation [115, 132] for each exciton state S:

$$\left(E^{\mathrm{QP}}_{c\mathbf{k}} - E^{\mathrm{QP}}_{v\mathbf{k}}\right) A^S_{vc\mathbf{k}} + \sum_{v'c'\mathbf{k}'} \left\langle vc\mathbf{k}|K^{\mathrm{eh}}|v'c'\mathbf{k}'\right\rangle = \Omega^S A^S_{vc\mathbf{k}} \tag{2.3}$$

where $A^S_{vc\mathbf{k}}$ is the exciton wavefunction (in the Bloch representation), Ω^S is the excitation energy, and K^{eh} is the electron-hole interaction kernel. The exciton wavefunction can be expressed in real space as:

$$\Psi(\mathbf{r}_e, \mathbf{r}_h) = \sum_{\mathbf{k},c,v} A^S_{vc\mathbf{k}} \psi_{\mathbf{k},c}(\mathbf{r}_e) \psi^*_{\mathbf{k},v}(\mathbf{r}_h), \tag{2.4}$$

and the imaginary part of the dielectric function, if one is interested in the optical response, can be expressed as

$$\epsilon_2(\omega) = \frac{16\pi^2 e^2}{\omega^2} \sum_S \left|\mathbf{e} \cdot \langle 0|\mathbf{v}|S\rangle\right|^2 \delta\left(\omega - \Omega^S\right) \tag{2.5}$$

where

$$\langle 0|\mathbf{v}|S\rangle = \sum_{vc\mathbf{k}} A^S_{vc\mathbf{k}} \langle v\mathbf{k}|\mathbf{v}|c\mathbf{k}\rangle , \tag{2.6}$$

and where $\mathbf{v}$ is the velocity operator along the direction of the polarization of light, $\mathbf{e}$. One may compare this to the non-interacting absorption spectra:

$$\epsilon_2(\omega) = \frac{16\pi^2 e^2}{\omega^2} \sum_{vc\mathbf{k}} \left|\mathbf{e} \cdot \langle v\mathbf{k}|\mathbf{v}|c\mathbf{k}\rangle\right|^2 \delta\left(\omega - E^{\mathrm{QP}}_{c\mathbf{k}} + E^{\mathrm{QP}}_{v\mathbf{k}}\right). \tag{2.7}$$

An example absorption spectrum for silicon computed with the BerkeleyGW package at the GW and the GW-BSE levels is shown in Fig. 2.1. Only when both the quasiparticle effects within the GW approximation and the excitonic effects through the solution of the Bethe-Salpeter equation are included is good agreement with experiment reached.

2.2 Computational Layout

2.2.1 Major Components of a GW-BSE Calculation for Complex Materials

Figure 2.2 illustrates the procedure for carrying out an *ab initio* GW-BSE calculation to obtain quasiparticle and optical properties using the BerkeleyGW code. First, one obtains the mean-field electronic orbitals and eigenvalues as well as the charge density. One can utilize one of the many supported DFT codes [3, 48, 126] to construct this mean-field starting point and convert it to the BerkeleyGW format using the included wrappers.

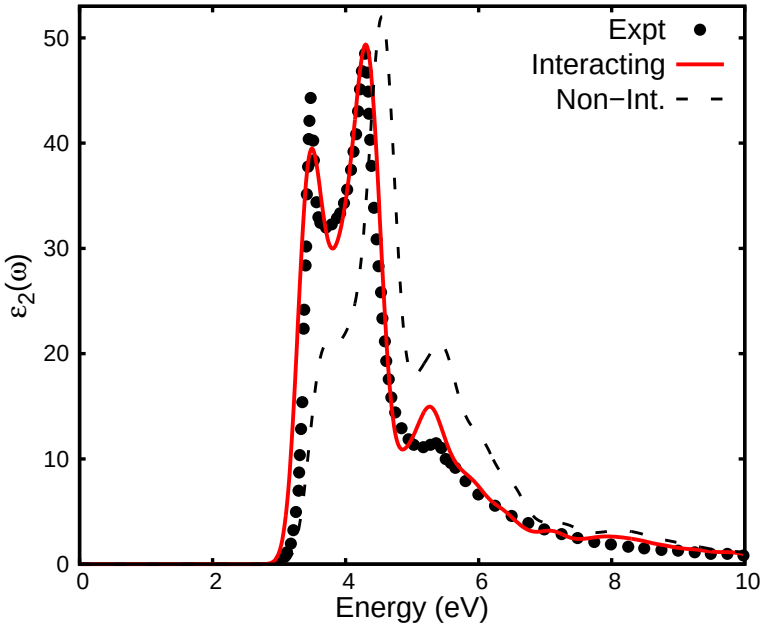

Figure 2.1: The absorption spectra for silicon calculated at the GW (black-dashed) and GW-BSE (red-solid) levels using the BerkeleyGW package. Experimental data from [46].

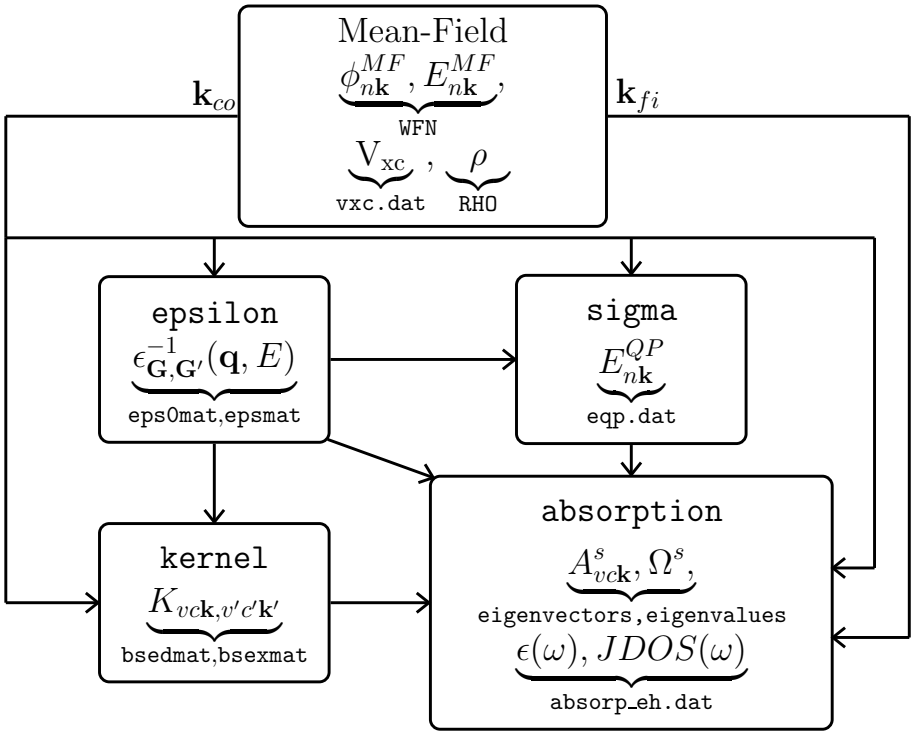

Figure 2.2: Flow chart of a GW-BSE calculation performed using the BerkeleyGW package described in this chapter.

The `epsilon` executable produces the polarizability and inverse dielectric matrices. In the `epsilon` executable, the static or frequency-dependent polarizability and dielectric function are calculated within the random phase approximation (RPA) using the electronic eigenvalues and eigenfunctions from a mean-field reference system. The main output is the file `epsmat` that contains the inverse-dielectric matrix.

In the `sigma` executable, the screened Coulomb interaction, W, is constructed from the inverse dielectric matrix and the one-particle Green's function, G, is constructed from the mean-field eigenvalues and eigenfunctions. We then calculate the diagonal and (optionally) off-diagonal elements of the self-energy operator, $\Sigma = iGW$, as a matrix in the mean-field basis. In many cases, only the diagonal elements are sizable within the chosen mean-field orbital basis; in such cases, in applications to real materials, the effects of Σ can be treated within first-order perturbation theory. If off-diagonal terms are not requested, the `sigma` executable considers Σ in the form $\Sigma = V_{\rm xc} + (\Sigma - V_{\rm xc})$, where $V_{\rm xc}$ is the independent-particle exchange-correlation potential of the chosen mean-field system. For moderately correlated electron systems, the best available mean-field Hamiltonian may often be taken to be the Kohn-Sham Hamiltonian [76]. However, many mean-field starting points are consistent with the BerkeleyGW package such as Hartree-Fock, static COHSEX and hybrid functionals. In principle, the process of correcting the eigenfunctions and eigenvalues (which determine W and G) could be repeated until self-consistency is reached or the Σ matrix is diagonalized in full; however, in practice, it is found that an adequate solution is often obtained within first-order perturbation theory on the Dyson's equation for a given Σ. Comparison of calculated energies with experiment shows that this level of approximation is very accurate for semiconductors and insulators and for most conventional metals. The output of the `sigma` executable are $E^{\rm QP}$, the quasiparticle energies, which are written to the file `eqp.dat` using the `eqp.py` post-processing utility on the generated `sigma.log` files for each `sigma` run.

The BSE executable, `kernel`, takes as input the full dielectric matrix calculated in the `epsilon` executable, which is used to screen the attractive direct electron-hole interaction, and the quasiparticle wavefunctions, which often times are taken to be the same as the mean-field wavefunctions. The direct and exchange part of the electron-hole kernel are calculated and output into the `bsedmat` and `bsexmat` files respectively. The `absorption` executable uses these matrices, the quasiparticle energies and wavefunctions from a coarse **k**-point grid GW calculation, as well as the wavefunctions from a fine **k**-point grid. The quasiparticle energy corrections and the kernel matrix elements are interpolated onto the fine grid. The Bethe-Salpeter Hamiltonian, consisting of the electron-hole kernel with the addition of the kinetic-energy term, is constructed in the quasiparticle electron-hole pair basis and diagonalized yielding the exciton wavefunctions and excitation energies, printed in file `eigenvectors`. Exciton binding energies can be inferred from the energy of the correlated exciton states relative to the interband-transition continuum edge. With the excitation energies and amplitudes of the electron-hole pairs, one can then calculate the macroscopic dielectric function for various light polarizations which is written to the file `absorption_eh.dat`. This may be compared to the absorption spectra without the electron-hole interaction included, printed in file `absorption_noeh.dat`.

Example input files for each executable are contained within the source code for

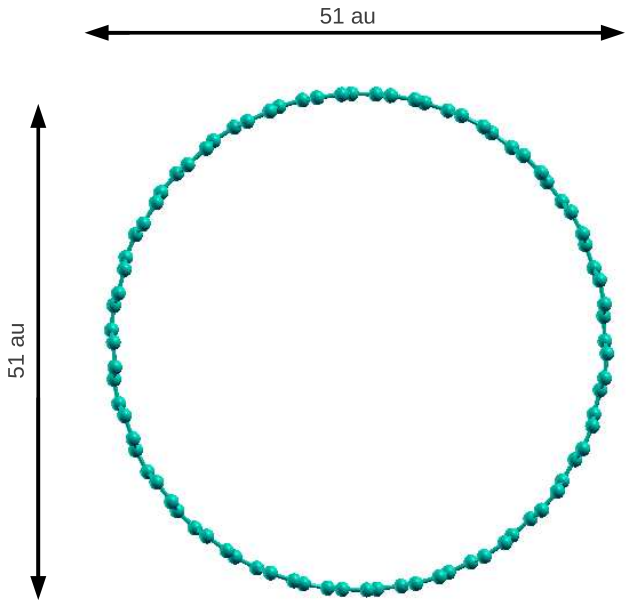

Figure 2.3: The cross-section of the (20,20) SWCNT used throughout the chapter as a benchmark system.

the package, as well as complete example calculations for silicon, the (8,0) and (5,5) single-walled carbon nanotubes (SWCNTs), the CO molecule, and sodium metal.

Throughout the chapter, atomic units are used. Additionally, sums over $\mathbf{k}$ and $\mathbf{q}$ are accompanied by an implicit division by the volume of the supercell, $V_{\text{sc}} = N_k V_{\text{uc}}$, where N_k is the number of points in the $\mathbf{k}$-grid and V_{uc} is the volume of the unit cell.

Throughout the chapter and appendices, we refer to benchmark numbers from calculations on the (20,20) SWCNT. This system has 80 carbon atoms and 160 occupied bands. We use 800 unoccupied bands in all sums requiring empty orbitals. We use a supercell of size $80 \times 80 \times 4.6$ au^3 equivalent to a bulk system of greater than 500 atoms. We use a $1 \times 1 \times 32$ coarse $\mathbf{k}$-grid and a $1 \times 1 \times 256$ fine $\mathbf{k}$-grid. We calculate the self-energy corrections within the diagonal approximation for 8 conduction and 8 valence bands. The Bethe-Salpeter equation is solved with 8 conduction and 8 valence bands. The relative costs of the various steps in the GW-BSE calculation using the BerkeleyGW package is shown in Table 2.1. As can be seen from the table, the actual time to solution for the GW-BSE part of the calculation is smaller than that of the DFT parts.

2.2.2 Dielectric Matrix: `epsilon`

`epsilon` is a standalone executable that computes either the static or dynamic RPA polarizability and corresponding inverse dielectric function from input electronic eigen-

Step	# CPUs	CPU Hours	Wall Hours
DFT Coarse	64 × 32	19000	9.1
DFT Fine	64 × 256	29000	1.8
`epsilon`	1600 × 32	61000	1.2
`sigma`	960 × 16	46000	3.0
`kernel`	1024	600	0.6
`absorption`	256	500	2.0

Table 2.1: Breakdown of the CPU and wall time spent on the calculation of the (20,20) SWCNT with parameters described in the text. The × indicates an additional level of trivial parallelization over the 256 or 32 (16 if time-reversal symmetry is utilized) **k**- or **q**-points.

values and eigenvectors computed in a suitable mean-field code. As we discuss in detail below, the input electronic eigenvalues and eigenvectors can come from a variety of different mean-field approximations including DFT within LDA/GGA [76, 107], generalized Kohn-Sham hybrid-functional approximations as well as direct approximations to the GW Dyson's equation such as the static-COHSEX [59, 68] approximation and the Hartree-Fock approximation.

We will first discuss the computation of the static polarizability and the inverse dielectric matrix. The `epsilon` executable computes the static RPA polarizability using the following expression [62]:

$$\chi_{\mathbf{GG'}}(\mathbf{q};0) = \sum_{n}^{\text{occ}}\sum_{n'}^{\text{emp}}\sum_{\mathbf{k}} M_{nn'}(\mathbf{k},\mathbf{q},\mathbf{G})M^{*}_{nn'}(\mathbf{k},\mathbf{q},\mathbf{G'})\frac{1}{E_{n\mathbf{k+q}}-E_{n'\mathbf{k}}}. \tag{2.8}$$

where

$$M_{nn'}(\mathbf{k},\mathbf{q},\mathbf{G}) = \langle n\mathbf{k}{+}\mathbf{q}|\, e^{i(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}}\,|n'\mathbf{k}\rangle \tag{2.9}$$

are the plane-wave matrix elements. Here, $\mathbf{q}$ is a vector in the first Brillouin zone, $\mathbf{G}$ is a reciprocal lattice vector, $\langle n\mathbf{k}|$ and $E_{n\mathbf{k}}$ are the meanfield electronic eigenvectors and eigenvalues. The matrix in Eq. 2.8 is to be evaluated up to $|G^2| < |E_{\text{cut}}|$ for both $\mathbf{G}$ and $\mathbf{G'}$ where E_{cut} defines the dielectric energy cutoff. The number of empty states, n', included in the summation must be such that the highest empty state included has an energy corresponding to E_{cut}. There is therefore one, rather than two, convergence parameter in evaluating Eq. 2.8 – one must either choose to converge with empty states or with the dielectric energy cutoff and set the remaining parameter to match the chosen convergence parameter. The `epsilon` code itself reports the convergence of Eq. 2.8 in an output file called `chi_converge.dat` (plotted in Fig. 2.4), that presents the computed value of $\chi_{\mathbf{GG'}=0}(\mathbf{q};0)$ and $\chi_{\mathbf{GG'}=\mathbf{G}_{max}}(\mathbf{q};0)$ using partial sums in Eq. 2.8 where $\mathbf{G}_{\text{max}}$ is the largest reciprocal-lattice vector included, and the number of empty states is varied between 1 and the maximum number requested in the input file, `epsilon.inp`.

With the expression for χ above, we can obtain the dielectric matrix as

$$\epsilon_{\mathbf{GG'}}(\mathbf{q};0) = \delta_{\mathbf{GG'}} - v(\mathbf{q}{+}\mathbf{G})\chi_{\mathbf{GG'}}(\mathbf{q};0) \tag{2.10}$$

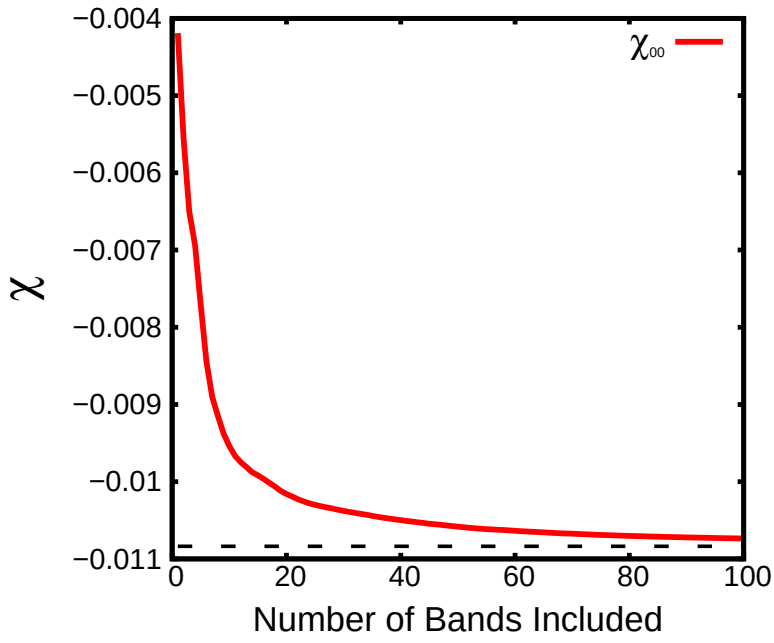

Figure 2.4: Example convergence output plotted from `chi_converge.dat` showing the convergence of the sum in Eq. 2.8 for the $\mathbf{G}, \mathbf{G}' = 0$ and $\mathbf{q} = (0, 0, 0.5)$ component of χ in ZnO.

where $v(\mathbf{q}+\mathbf{G})$ is the bare Coulomb interaction defined as:

$$v(\mathbf{q}+\mathbf{G}) = \frac{4\pi}{|\mathbf{q}+\mathbf{G}|^2} \tag{2.11}$$

in the case of bulk crystals where no truncation is necessary. We discuss in Sec. 2.3 how to generalize this expression for the case of nano-systems where truncating the interaction in non-periodic directions greatly improves the convergence with supercell size.

It should be noted that we use an asymmetric definition of the Coulomb interaction, as opposed to symmetric expressions such as

$$v(\mathbf{G},\mathbf{G}') = \frac{4\pi}{|\mathbf{q}+\mathbf{G}|\,|\mathbf{q}+\mathbf{G}'|}. \tag{2.12}$$

This causes $\epsilon_{\mathbf{G}\mathbf{G}'}(\mathbf{q};0)$ and $\chi_{\mathbf{G}\mathbf{G}'}(\mathbf{q};0)$ to be also asymmetric in $\mathbf{G}$ and $\mathbf{G}'$. This asymmetry is resolved when constructing the static screened Coulomb interaction using the expression:

$$W_{\mathbf{G}\mathbf{G}'}(\mathbf{q};0) = \epsilon^{-1}_{\mathbf{G}\mathbf{G}'}(\mathbf{q};0)v\big(\mathbf{q}+\mathbf{G}'\big). \tag{2.13}$$

Here W is symmetric in $\mathbf{G}$ and $\mathbf{G}'$ even though both v and ϵ^{-1} individually are not.

The computation of $\epsilon^{-1}_{\mathbf{G}\mathbf{G}'}(\mathbf{q};0)$ in the `epsilon` code involves three computationally intensive steps: the computation of the matrix elements needed for the summation in Eq. 2.8, the summation itself and the inversion of the dielectric matrix to yield $\epsilon^{-1}_{\mathbf{G}\mathbf{G}'}(\mathbf{q};0)$.

The `epsilon` code first computes all the matrix elements $M_{nn'}$ required in the summation for Eq. 2.8. This step is generally the most time-consuming step in the execution of the `epsilon` code. Naively, one might think this process scales as N^4, where N is the number of atoms in the system. This is because both the number of valence and conduction bands needed scales linearly with N and the number of $\mathbf{G}$ vectors scales linearly with the cell volume which itself scales linearly with the number of atoms. Thus, we must calculate N^3 matrix elements each of which involves a sum over the plane-wave basis set for the eigenfunctions. We therefore have an N^4 scaling. However, we can achieve $N^3 \log N$ scaling by using fast Fourier transforms (FFTs), noting that the expression in Eq. 2.9 is a convolution in Fourier space [44]. Therefore, Eq. 2.9 can be written as the Fourier transform of a direct product of the wavefunctions in real space:

$$M_{nn'}(\mathbf{k},\mathbf{q},\{\mathbf{G}\}) = FFT^{-1}\left(\phi_{n,\mathbf{k}+\mathbf{q}}(\mathbf{r}) * \phi^*_{n',\mathbf{k}}(\mathbf{r})\right). \tag{2.14}$$

The FFTs are implemented with FFTW [45] and scale as $N \log N$. The computation of all the matrix elements needed for Eq. 2.8 therefore scales as $N^3 \log N$. We discuss in the following sections that the computation of these matrix elements can be very trivially parallelized up to tens of thousands of CPUs. Given an infinite resource of CPUs, our implementation would have a wall-time scaling of $N \log N$, nearly linear in the number of atoms.

Having computed the individual matrix elements required in Eq. 2.8, we now turn our attention to the summation involved in the same expression. It should be noted that the formal scaling of this step with the number of atoms is N^4 since one must sum over the number of occupied bands and the number of unoccupied bands for every $\mathbf{G}$ and $\mathbf{G}'$

pair – each one of these quantities scaling linearly with the number of atoms. This step therefore has formally the worst scaling of the entire GW process – leading many to claim that GW as a whole scales like N^4. However, in practice for most systems currently under study within a generalized plasmon-pole (GPP) [62] or other approximation where this sum is done only once for the static polarizability, this step represents less than 10 percent of a typical calculation time even for systems of 100's of atoms because it can be greatly optimized and parallelized. In particular, Eq. 2.8 can be written very compactly as a single matrix-matrix product for each $\mathbf{q}$:

$$\chi_{\mathbf{GG'}}(\mathbf{q};0) = \mathbf{M}(\mathbf{G},\mathbf{q},(n,n',\mathbf{k})) \cdot \mathbf{M}^T(\mathbf{G'},\mathbf{q}(n,n',\mathbf{k})) \tag{2.15}$$

where $(n,n',\mathbf{k})$ represents a single composite index that is summed over as the inner dimension in the matrix-matrix product. The matrices $\mathbf{M}$ can be expressed in terms of the matrix elements M as:

$$\mathbf{M}(\mathbf{G},\mathbf{q},(n,n',\mathbf{k})) = M_{nn'}(\mathbf{k},\mathbf{q},\mathbf{G}) \cdot \frac{1}{\sqrt{E_{n\mathbf{k}+\mathbf{q}} - E_{n'\mathbf{k}}}}. \tag{2.16}$$

The single dense matrix-matrix product required in Eq. 2.15 still scales as N^4 since the inner dimension, $(n,n',\mathbf{k})$, scales as N^2 and dense matrix multiplication itself scales as N^2. However, in the BerkeleyGW package, this single step is still made quite rapid for even systems as large as 100's of atoms. The LEVEL 3 BLAS [8] libraries `DGEMM` and `ZGEMM` and their parallel analogues are used to compute the single matrix product in Eq. 2.15. As we discuss further in Sec. A.1.1, in the evaluation of Eq. 2.9, the parallel wall-time scaling is N^2 with the number of atoms.

Finally, once we have constructed $\chi_{\mathbf{GG'}}(\mathbf{q};0)$ we can construct the RPA dielectric matrix and inverse dielectric matrix required for the computation of the screened Coulomb interaction, W. The dielectric matrix as implemented in the code is expressed in Eq. 2.10.

Here, we require for the first time the Coulomb interaction in reciprocal space $v(\mathbf{q}+\mathbf{G})$, which can be computed trivially from Eq. 2.11 for the case of bulk crystals, but requires an FFT for the case of nanostructured materials. We discuss this more in Sec. 2.3.

There is a clear problem in directly computing $\epsilon_{\mathbf{00}}(\mathbf{q}=\mathbf{0})$ due to the fact the Coulomb interaction, Eq. 2.11, diverges as $\mathbf{q} \to 0$ except in the case of box-type truncation schemes (see Sec. 2.3). For semiconducting systems, due to orthogonality, the matrix elements (Eq. 2.9) themselves go to 0 with the form $|M_{nn'}(\mathbf{k},\mathbf{q},\mathbf{G}=\mathbf{0})| \propto |q|$. So, $\epsilon(\mathbf{q}\to 0)$ contains a non-trivial $\frac{q^2}{q^2}$ limit. There are multiple ways to handle this limit, including replacing the "velocity" operator in Eq. 2.9 with the momentum operator, plus the commutators with the non-local part of the mean-field Hamiltonian. [15, 62]. The `epsilon` code has implemented a simpler scheme, however, in which we numerically take the limit as $\mathbf{q} \to 0$ by evaluating $\epsilon_{\mathbf{00}}(\mathbf{q}_0)$ at a small but finite $\mathbf{q}_0$ usually taken as approximately 1/1000th of the Brillouin zone. For semiconducting systems, where $\epsilon_{\mathbf{00}}(\mathbf{q}=\mathbf{0}) \to C$, it is sufficient to construct a separate $\mathbf{k}$-grid for the conduction and valence bands shifted by the small vector $\mathbf{q}_0$ in order to compute $M_{nn'}(\mathbf{k},\mathbf{q}_0,\mathbf{G}=\mathbf{0})$, where n is a valence and n' a conduction band, and to evaluate the correct limiting $\frac{q^2}{q^2}$ ratio. For metals, however, intraband transitions have $|M_{nn'}(\mathbf{k},\mathbf{q},\mathbf{G}=\mathbf{0})| \propto C$, yielding $\epsilon_{\mathbf{00}}(\mathbf{q}\to 0) \propto \frac{C'}{q^2}$. In this case,

the two-**k**-grid treatment is insufficient, because the proportionality coefficient C' depends sensitively on the density of states (DOS) at the Fermi energy. Therefore a **k**-grid sampling of the same spacing as $\mathbf{q}_0$ is required, although fewer conduction bands are necessary in the sum since $\epsilon(\mathbf{q} \to 0)$ is dominated by intra-band transitions. Note that this treatment of intraband transitions is still the zero-temperature limit in our code, as the effect of thermal occupations is small in GW except at very large temperatures [21]. Effectively occupations are taken as one below the Fermi level, zero above the Fermi level, and 1/2 at the Fermi level (as needed for graphene at the Dirac point). This is despite any smearing that may have been used in the underlying mean-field calculation.

The inversion of the dielectric matrix required to compute W, Eq. 2.13, is done with LAPACK and ScaLAPACK (for parallel calculations) using `ZGESV`, `DGESV` and their parallel counterparts. The inversion scales like N^3 with the number of atoms and, as we discuss below, scales well up to 100's of processors with ScaLAPACK. In general, for systems of up to 100's of atoms, the inversion step represents less than 10 percent of the total computation time for `epsilon`.

We have so far limited ourselves to situations in which only a direct calculation of the static polarizability, Eq. 2.8, is required, such as in the static-COHSEX approximation [68] or when utilizing a GPP model [62] to extend the dielectric response to non-zero frequencies. However, in order to do a more refined calculation, the dielectric matrix can be computed directly at real frequencies without extrapolation, as is formally required in the Dyson equation. We use in the package the advanced and retarded dielectric functions, defined as:

$$\begin{aligned} \epsilon^{\mathrm{r/a}}_{\mathbf{GG'}}(\mathbf{q};E) &= \delta_{\mathbf{GG'}} - v(\mathbf{q}+\mathbf{G}) \sum_{n}^{\mathrm{occ}} \sum_{n'}^{\mathrm{emp}} \sum_{\mathbf{k}} M_{nn'}(\mathbf{k},\mathbf{q},\mathbf{G}) M^*_{nn'}(\mathbf{k},\mathbf{q},\mathbf{G'}) \qquad (2.17) \\ &\times \frac{1}{2}\left[\frac{1}{E_{n\mathbf{k+q}} - E_{n'\mathbf{k}} - E \mp i\delta} + \frac{1}{E_{n\mathbf{k+q}} - E_{n'\mathbf{k}} + E \pm i\delta} \right] \end{aligned}$$

where E is the evaluation frequency and δ is a broadening parameter chosen to be consistent with the energy spacing afforded by the **k**-point sampling of the calculation, using the upper (lower) signs for the retarded (advanced) function. In principle, one must converge the calculation with respect to increasing the **k**-point sampling and decreasing this broadening parameter.

In the `epsilon` code, we compute Eq. 2.17 on a grid of real frequencies, E, specified by a frequency spacing, a low-frequency cutoff, a high-frequency cutoff and a frequency-spacing increment. We sample the frequency on the real axis uniformly from 0 to the low-frequency cutoff with a sampling rate given by the frequency spacing. We then increase the frequency spacing by the step increment until we reach the high-frequency cutoff (Fig. 2.6). In general, one must also refine this frequency grid until convergence is reached, but we find that for the purpose of calculating band gaps of typical semiconductors, a frequency spacing of a few hundred meV and a high-frequency cutoff of twice the dielectric energy cutoff is sufficient, though it should be noted this energy can be quite high (*e.g.* the case of ZnO [122]).

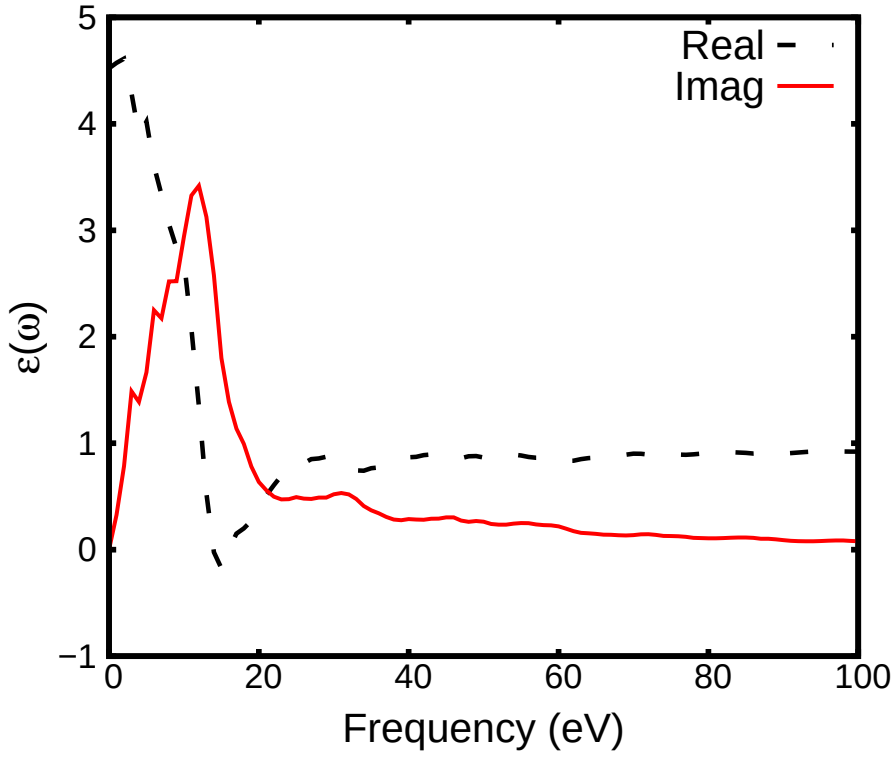

Figure 2.5: Example output plotted from `EpsDyn` showing the computed $\epsilon_{\mathbf{00}}(\omega)$ in ZnO.

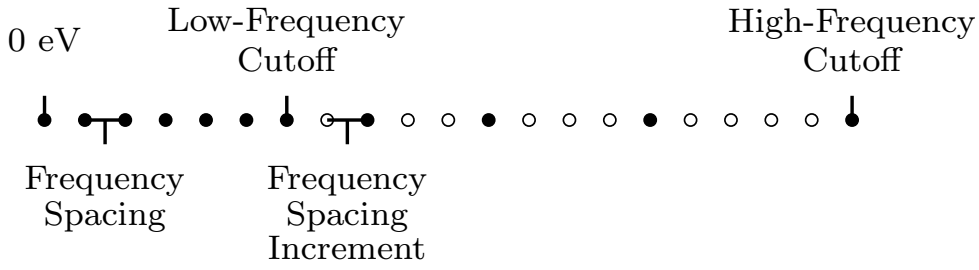

Figure 2.6: Schematic of the frequency-grid parameters for a full-frequency calculation in `epsilon`. The open circles are a continuation of the uniform grid that are omitted above the low-frequency cutoff.

In cases where the computation of the "full-frequency" dielectric response function is required, the bottleneck of the calculation often does become the N^4 summation step of Eq. 2.15. This is because the computation of the matrix elements, Eq. 2.9, needs only be done once, whereas the summation must be done for all frequencies separately. Because of this, a full-frequency `epsilon` calculation of between 10-50 frequencies costs only twice the time of a static `epsilon` calculation, but the cost scales linearly with frequencies after this point.

2.2.3 Computation of the Self-Energy: `sigma`

The `sigma` executable takes as input the inverse epsilon matrix calculated from the `epsilon` executable and a suitable set of mean-field electronic energies and wavefunctions. It computes a representation of the Dyson's equation, Eq. 2.2, in the basis of the mean-field eigenfunctions through the computation of the diagonal and off-diagonal elements of Σ:

$$\begin{aligned}\langle\psi_{n\mathbf{k}}|H^{\mathrm{QP}}(E)|\psi_{m\mathbf{k}}\rangle = \\ E^{\mathrm{MF}}_{n\mathbf{k}}\delta_{n,m} + \langle\psi_{n\mathbf{k}}|\Sigma\left(E\right) - \Sigma^{\mathrm{MF}}\left(E\right)|\psi_{m\mathbf{k}}\rangle\end{aligned} \tag{2.18}$$

where E is an energy parameter that should be self-consistently set to the quasiparticle eigenvalues, $E^{\mathrm{MF}}_{n\mathbf{k}}$ and $\psi_{n\mathbf{k}}$ are the mean-field eigenvalues and eigenvectors and Σ^{MF} is a mean-field approximation to the electronic self energy operator, such as V_{xc} in the case of a DFT starting point.

It is often the case that the mean-field wavefunctions are sufficiently close to the quasiparticle wavefunctions [62] that one may reduce Eq. 2.18 to include only diagonal matrix elements. In this case the user may ask for only diagonal elements, and the quasiparticle energies will be updated in the following way:

$$E^{\mathrm{QP}}_{n\mathbf{k}} = E^{\mathrm{MF}}_{n\mathbf{k}} + \langle\psi_{n\mathbf{k}}|\Sigma\left(E\right) - \Sigma^{\mathrm{MF}}\left(E\right)|\psi_{n\mathbf{k}}\rangle. \tag{2.19}$$

The mean field in Eq. 2.19 and Eq. 2.18 can be DFT within the LDA or GGA schemes as well as within a hybrid-functional approach, in which case $\Sigma^{\mathrm{MF}}\left(E\right) = V_{\mathrm{xc}}$, which is local and energy-independent in the case of LDA. The starting mean-field can also be an approximation to the Dyson's equation, Eq. 2.2, such as Hartree-Fock (the zero-screening limit) or static COHSEX (the static-screening limit) [114, 26, 68] - see chapter 3 for a discussion of COHSEX as a starting point. The use of these mean-field starting points for construction of Eq. 2.18 and Eq. 2.19 is classified as a one-shot G_0W_0 calculation (the 0 here means that both G and W are constructed from the mean-field eigenvalues and eigenvectors). One can also start from a previous iteration of GW in an eigenvalue or eigenvector self-consistency scheme [26, 144]. In this case, the 'MF' superscripts in Eq. 2.19 and 2.18 should be renamed "previous" to designate the self-consistency.

The `sigma` executable itself can evaluate the matrix elements of Σ in Eq. 2.19 and Eq. 2.18 within various approximations: Hartree-Fock, static COHSEX, GW within a GPP model and full-frequency GW.

For GW and static-COHSEX calculations, Σ can be broken into two parts, $\Sigma = \Sigma_{\mathrm{SX}} + \Sigma_{\mathrm{CH}}$, where Σ_{SX} is the screened exchange operator and Σ_{CH} is the Coulomb-hole

operator. [62, 58, 59] These are implemented in the `sigma` executable in the following way for a full-frequency calculation:

$$\langle n\mathbf{k}|\,\Sigma_{\rm SX}(E)\,|n'\mathbf{k}\rangle = -\sum_{n''}^{\rm occ}\sum_{\mathbf{q}\mathbf{G}\mathbf{G}'} M^*_{n''n}(\mathbf{k},-\mathbf{q},-\mathbf{G})M_{n''n'}(\mathbf{k},-\mathbf{q},-\mathbf{G}') \times \left[\epsilon^{\rm r}_{\mathbf{G}\mathbf{G}'}\right]^{-1}\left(\mathbf{q};E-E_{n''\mathbf{k}-\mathbf{q}}\right)v\left(\mathbf{q}+\mathbf{G}'\right) \tag{2.20}$$

and

$$\langle n\mathbf{k}|\,\Sigma_{\rm CH}(E)\,|n'\mathbf{k}\rangle = \frac{i}{2\pi}\sum_{n''}\sum_{\mathbf{q}\mathbf{G}\mathbf{G}'} M^*_{n''n}(\mathbf{k},-\mathbf{q},-\mathbf{G})M_{n''n'}(\mathbf{k},-\mathbf{q},-\mathbf{G}') \times \int_0^\infty dE'\,\frac{\left[\epsilon^{\rm r}_{\mathbf{G}\mathbf{G}'}\right]^{-1}(\mathbf{q};E') - \left[\epsilon^{\rm a}_{\mathbf{G}\mathbf{G}'}\right]^{-1}(\mathbf{q};E')}{E-E_{n''\mathbf{k}-\mathbf{q}}-E'+i\delta}\;v\left(\mathbf{q}+\mathbf{G}'\right) \tag{2.21}$$

where M is defined in Eq. 2.6 and $\epsilon^{\rm r}$ and $\epsilon^{\rm a}$ are the retarded and advanced dielectric matrices defined in Eq. 2.17.[129] In practice the `sigma` executable computes the matrix elements of bare exchange, $\Sigma_{\rm X}$ and of $\Sigma_{\rm SX}-\Sigma_{\rm X}$, where the matrix elements of $\Sigma_{\rm X}$ are obtained by replacing $\left[\epsilon^{\rm r}_{\mathbf{G}\mathbf{G}'}\right]^{-1}\left(\mathbf{q};E-E_{n''\mathbf{k}-\mathbf{q}}\right)$ with $\delta_{\mathbf{G},\mathbf{G}'}$ in Eq. 2.20 (as given by Eq. 2.29 below). The integral in Eq. 2.21 over frequency is done numerically on the frequency grid used in the `epsilon` executable (Fig. 2.6).

For GPP calculations, the corresponding expressions used in the code are:

$$\langle n\mathbf{k}|\,\Sigma_{\rm SX}(E)\,|n'\mathbf{k}\rangle = -\sum_{n''}^{\rm occ}\sum_{\mathbf{q}\mathbf{G}\mathbf{G}'} M^*_{n''n}(\mathbf{k},-\mathbf{q},-\mathbf{G})M_{n''n'}(\mathbf{k},-\mathbf{q},-\mathbf{G}') \times \left[\delta_{\mathbf{G}\mathbf{G}'}+\frac{\Omega^2_{\mathbf{G}\mathbf{G}'}(\mathbf{q})}{\left(E-E_{n''\mathbf{k}-\mathbf{q}}\right)^2-\tilde{\omega}^2_{\mathbf{G}\mathbf{G}'}(\mathbf{q})}\right]v\left(\mathbf{q}+\mathbf{G}'\right) \tag{2.22}$$

and

$$\langle n\mathbf{k}|\,\Sigma_{\rm CH}(E)\,|n'\mathbf{k}\rangle = \frac{1}{2}\sum_{n''}\sum_{\mathbf{q}\mathbf{G}\mathbf{G}'} M^*_{n''n}(\mathbf{k},-\mathbf{q},-\mathbf{G})M_{n''n'}(\mathbf{k},-\mathbf{q},-\mathbf{G}') \times \frac{\Omega^2_{\mathbf{G}\mathbf{G}'}(\mathbf{q})}{\tilde{\omega}_{\mathbf{G}\mathbf{G}'}(\mathbf{q})\left[E-E_{n''\mathbf{k}-\mathbf{q}}-\tilde{\omega}_{\mathbf{G}\mathbf{G}'}(\mathbf{q})\right]}\;v\left(\mathbf{q}+\mathbf{G}'\right) \tag{2.23}$$

where Ω and $\tilde{\omega}$ are the effective bare plasma frequency and the GPP mode frequency [62] defined as:

$$\Omega^2_{\mathbf{G}\mathbf{G}'}(\mathbf{q}) = \omega_{\rm p}^2\frac{(\mathbf{q}+\mathbf{G})\cdot(\mathbf{q}+\mathbf{G}')}{|\mathbf{q}+\mathbf{G}|^2}\frac{\rho(\mathbf{G}-\mathbf{G}')}{\rho(\mathbf{0})} \tag{2.24}$$

and

$$\tilde{\omega}^2_{\mathbf{G}\mathbf{G}'}(\mathbf{q}) = \frac{\Omega^2_{\mathbf{G}\mathbf{G}'}(\mathbf{q})}{\delta_{\mathbf{G}\mathbf{G}'}-\epsilon^{-1}_{\mathbf{G}\mathbf{G}'}(\mathbf{q};0)} \tag{2.25}$$

Here, ρ is the charge density in reciprocal space and $\omega_{\rm p}^2 = 4\pi\rho(\mathbf{0})e^2/m$ is the classical plasma frequency. In this case, the integral over energy that is necessary in the full-frequency

expression, Eq. 2.21, is reduced to a single term using an analytical approximation to the frequency dependence of the dielectric matrix requiring only the static dielectric matrix $\epsilon^{-1}_{\mathbf{GG'}}(\mathbf{q};0)$ in Eq. 2.25. The analytical approximation is done using the f-sum rule for each $\mathbf{GG'}$ pair as described in Ref. [62]. This reduces the computational cost of evaluating the Σ matrix elements by a factor of the number of frequencies. It is important to note that for systems without inversion symmetry, ρ in Eq. 2.24 and V_{xc} in Eqs. 2.18 and 2.19 are complex functions in reciprocal space (even though these are real functions when transformed to real space).

For static COHSEX calculations, the expressions used in the code are:

$$\langle n\mathbf{k}|\,\Sigma_{\mathrm{SX}}(0)\,|n'\mathbf{k}\rangle = -\sum_{n''}^{\mathrm{occ}}\sum_{\mathbf{qGG'}} M^*_{n''n}(\mathbf{k},-\mathbf{q},-\mathbf{G})M_{n''n'}(\mathbf{k},-\mathbf{q},-\mathbf{G'})\epsilon^{-1}_{\mathbf{GG'}}(\mathbf{q};0)v\big(\mathbf{q}+\mathbf{G'}\big) \tag{2.26}$$

and

$$\begin{aligned}\langle n\mathbf{k}|\,\Sigma_{\mathrm{CH}}(0)\,|n'\mathbf{k}\rangle &= \frac{1}{2}\sum_{n''}\sum_{\mathbf{qGG'}} M^*_{n''n}(\mathbf{k},-\mathbf{q},-\mathbf{G})M_{n''n'}(\mathbf{k},-\mathbf{q},-\mathbf{G'}) && (2.27)\\ &\quad\times\left[\epsilon^{-1}_{\mathbf{GG'}}(\mathbf{q};0)-\delta_{\mathbf{GG'}}\right]v\big(\mathbf{q}+\mathbf{G'}\big)\\ &= \frac{1}{2}\sum_{\mathbf{qGG'}} M_{nn'}(\mathbf{k},\mathbf{q}=\mathbf{0},\mathbf{G'}-\mathbf{G})\left[\epsilon^{-1}_{\mathbf{GG'}}(\mathbf{q};0)-\delta_{\mathbf{GG'}}\right]v\big(\mathbf{q}+\mathbf{G'}\big) && (2.28)\end{aligned}$$

where Eqs. 2.26 and 2.27 can formally be derived from Eqs. 2.22 and 2.23 by setting $\big(E-E_{n''\mathbf{k}-\mathbf{q}}\big)$ to zero. Using the completeness relation for the sum over empty states, Eq. 2.27 can be written in a closed form given by Eq. 2.28, which now does not involve the empty orbitals.

For Hartree-Fock calculations, we compute the matrix elements of bare exchange:

$$\langle n\mathbf{k}|\,\Sigma_{\mathrm{X}}\,|n'\mathbf{k}\rangle = -\sum_{n''}^{\mathrm{occ}}\sum_{\mathbf{qGG'}} M^*_{n''n}(\mathbf{k},-\mathbf{q},-\mathbf{G})M_{n''n'}(\mathbf{k},-\mathbf{q},-\mathbf{G'})\delta_{\mathbf{GG'}}v\big(\mathbf{q}+\mathbf{G'}\big) \tag{2.29}$$

In principle, the inner and outer orbitals used in Eqs. 2.20 – 2.29 originate from the same mean-field solution. However, there is an option in the `sigma` executable to use a different mean-field solution for the inner and outer states. This is useful if one wishes to construct the Σ operator within one mean field but expand the Σ matrix using different orbitals, i.e. in order to evaluate matrix elements in a different basis than the mean-field wavefunctions when the quasiparticle wavefunctions are significantly different. This is also useful for verifying the accuracy of the linearization approximation as given by Eq. 2.30 below.

Eq. 2.19 depends on the evaluation energy parameter E. This parameter should be determined self-consistently to the quasiparticle energy $E^{\mathrm{QP}}_{n\mathbf{k}}$. In principle, what one may do is start by setting $E = E^{\mathrm{MF}}_{n\mathbf{k}}$ and find $E^0_{n\mathbf{k}}$ using Eq. 2.19. One can then set $E = E^0_{n\mathbf{k}}$ and resolve Eq. 2.19 arriving at a new quasiparticle energy $E^1_{n\mathbf{k}}$. One can then repeat this process until convergence is reached. This process can be achieved using the different set

of inner and outer states as described in the previous paragraph – where the outer-state eigenvalues are updated after each step, and the eigenfunctions are left unchanged. In many cases, one can avoid this process by computing $\Sigma(E)$ on a grid of energies and interpolating or extrapolating to $E^{\mathrm{QP}}_{n\mathbf{k}}$. In particular, in many systems, $\Sigma(E)$ is a nearly linear function of E so one may compute $\Sigma(E)$ for two grid points and evaluate the self-consistent $E^{\mathrm{QP}}_{n\mathbf{k}}$ using Newton's method [62]:

$$E^{\mathrm{QP}}_{n\mathbf{k}} = E^{0}_{n\mathbf{k}} + \frac{d\Sigma/dE}{1 - d\Sigma/dE}(E^{0}_{n\mathbf{k}} - E^{\mathrm{MF}}_{n\mathbf{k}}) \tag{2.30}$$

The derivative that appears here is also related to the quasiparticle renormalization factor:

$$Z = \frac{d\Sigma/dE}{1 - d\Sigma/dE} \tag{2.31}$$

For full-frequency calculations, Eq. 2.20 and Eq. 2.21 are evaluated on a frequency grid, E, (not to be confused with the frequency grid over which the integrals are carried out) specified by the user. One then has access directly to Re $\Sigma(\omega)$ and to Im $\Sigma(\omega)$, printed in the file `spectrum.dat`, which can be used to construct the spectral function:

$$A_{\mathbf{k}}(\omega) = \frac{1}{\pi} \cdot \sum_{n} \frac{|\text{Im } \Sigma_{n\mathbf{k}}(\omega)|}{(\omega - E^{\mathrm{MF}}_{n\mathbf{k}} - \text{Re } \Sigma_{n\mathbf{k}}(\omega) + V^{n\mathbf{k}}_{\mathrm{xc}})^2 + \text{Im } \Sigma_{n\mathbf{k}}(\omega)^2}, \tag{2.32}$$

where we are using the mean-field exchange-correlation matrix element $V^{n\mathbf{k}}_{\mathrm{xc}} = \langle n\mathbf{k}\,|V_{\mathrm{xc}}|\,n\mathbf{k}\rangle$. This quantity can be used to compare directly with the quasiparticle spectrum from photoemission experiments and various other measurements of the bandstructure.

The plane-wave matrix elements required in Eqs. 2.20 – 2.29 are similar to those of Eq. 2.6 required for the construction of the irreducible polarizability matrix. In the current case, however, we require additional matrix elements between valence-valence band pairs as well as conduction-conduction band pairs. As was the case in the `epsilon` executable, the matrix elements are computed using FFTs utilizing the FFTW library.[45] For each pair of outer states, n and n', we sum over all occupied and unoccupied inner states, n'', included in the calculation (typically states of energy up to the dielectric energy cutoff). Therefore, the computational cost of computing all the necessary matrix elements scales as $N^2 \log N$, where N is the number of atoms (a factor of $N \log N$ comes from the FFTs). If one is interested in all the diagonal matrix elements, Eq. 2.19, in a given energy range (as opposed to just a fixed small number of states – *e.g.* VBM and CBM) then an additional factor of N is included in the scaling which becomes $N^3 \log N$. If one requires both diagonal and off-diagonal elements within a given energy window (such as in a self-consistent GW scheme), then the scaling becomes $N^4 \log N$.

Once the plane-wave matrix elements have been computed, the summations in the Coulomb-hole terms of Eqs. 2.20 – 2.29 for a particular n, n' pair scale individually as N^3. Again, if all diagonal or off-diagonal matrix elements of Σ in a given energy window must be computed an additional factor of N or N^2 respectively is added to the scaling.

It is important to point out that the Coulomb-hole summations in terms of Eqs. 2.20 – 2.29 converge exceptionally slowly with respect to the number of empty states included in the sums. The highest empty state included should have energy at least the dielectric energy cutoff. Additionally, the convergence of the sums in 2.20 – 2.29 should be tested with respect to the dielectric energy cutoff. As shown in Figure 2.7, the convergence with respect to the dielectric energy cutoff, and the corresponding number of empty states, is very slow in many cases. This problem is similar to convergence issues with respect to empty states in the `epsilon` executable, as been discussed above. However, one finds that in many cases (particularly for bulk systems) the final E^{QP} converges much more slowly with respect to the number of empty states in the Coulomb-hole expression than in the polarizability expression [122] – see Fig. 2.7 for a comparison of these two rates in ZnO. The partial sums of the Coulomb-hole matrix elements with respect to number of states included in the sum is written to the file `ch_converge.dat`. Example output from `ch_converge.dat` is plotted in Fig. 2.7.

2.2.4 Optical Properties: `BSE`

The optical properties of materials are computed in the Bethe-Salpeter equation (BSE) executables. Here the eigenvalue equation represented by the BSE, Eq. 2.3, is constructed and diagonalized yielding the excitation energies and wavefunctions of the correlated electron-hole excited states. There are two main executables: `kernel` and `absorption`. In the first, the electron-hole interaction kernel is constructed on a coarse **k**-point grid, and in the second the kernel is (optionally) interpolated to a fine **k**-point grid and diagonalized.

The `kernel` executable constructs the second term of the left-hand side of Eq. 2.3 which is referred to as the electron-hole kernel. The kernel, K, as implemented in the package, is limited to the static approximation, and contains two terms, a screened direct interaction and a bare exchange interaction, $K^{\mathrm{eh}} = K^{\mathrm{d}} + K^{\mathrm{x}}$, defined in the following way [115]:

$$\langle vc\mathbf{k}|K^{\mathrm{d}}|v'c'\mathbf{k}'\rangle = - \int d\mathbf{r}d\mathbf{r}'\psi_c^*(\mathbf{r})\psi_{c'}(\mathbf{r})W(\mathbf{r},\mathbf{r}')\psi_{v'}^*(\mathbf{r}')\psi_v(\mathbf{r}') \tag{2.33}$$

and

$$\langle vc\mathbf{k}|K^{\mathrm{x}}|v'c'\mathbf{k}'\rangle = \int d\mathbf{r}d\mathbf{r}'\psi_c^*(\mathbf{r})\psi_v(\mathbf{r})v(\mathbf{r},\mathbf{r}')\psi_{v'}^*(\mathbf{r}')\psi_{c'}(\mathbf{r}'). \tag{2.34}$$

These matrices are constructed on a coarse grid of **k**-points, in most cases the same grid used within the GW calculation because one must have previously constructed the dielectric matrix $\epsilon^{-1}(\mathbf{q})$ for $\mathbf{q} = \mathbf{k} - \mathbf{k}'$. We calculate these matrices in **G**-space using the prescription of Rohlfing and Louie [115]:

$$\langle vc\mathbf{k}|K^{\mathrm{d}}|v'c'\mathbf{k}'\rangle = \sum_{\mathbf{G}\mathbf{G}'} M_{cc'}(\mathbf{k},\mathbf{q},\mathbf{G})W_{\mathbf{G}\mathbf{G}'}(\mathbf{q};0)M_{vv'}^*(\mathbf{k},\mathbf{q},\mathbf{G}') \tag{2.35}$$

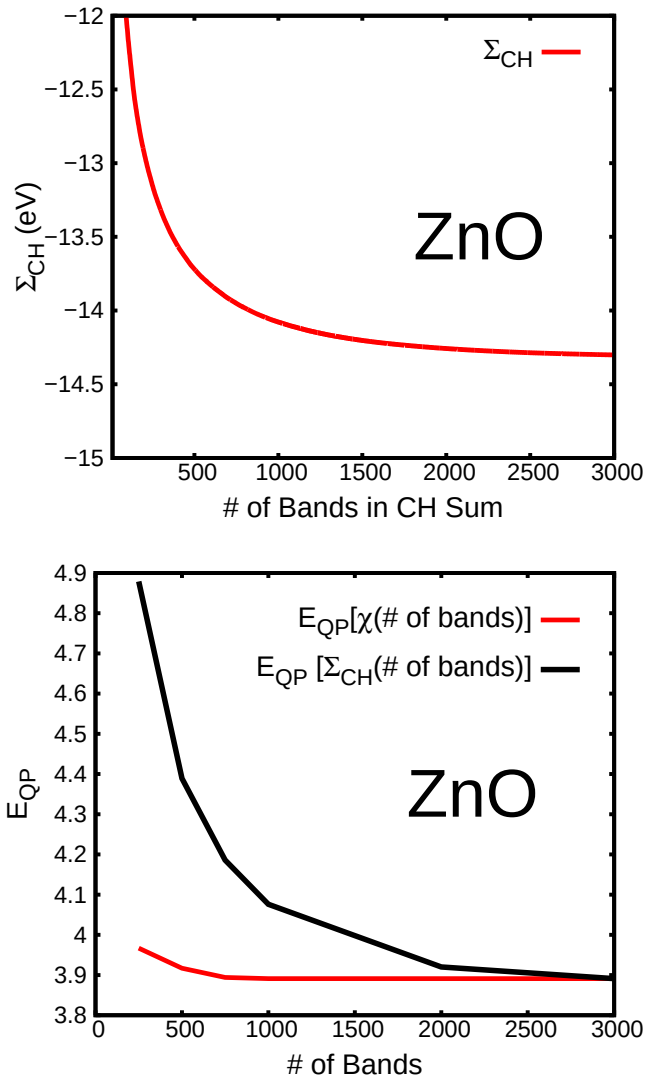

Figure 2.7: ZnO Convergence of the VBM. Top: Example convergence output from file `ch_convergence.dat` showing the Coulomb-hole sum value *vs.* the number of bands included in the sum. The black line is the best-guess converged value using the modified static-remainder approach [35]. Bottom: The convergence of E_{QP} with respect to empty states in the polarizability sum, Eq. 2.8, and with respect to empty states in the Coulomb-hole sum, Eq. 2.23. The red curve shows the VBM E_{QP} in ZnO using a fixed 3,000 bands in the Coulomb-hole summation and varying the number of bands included in the polarizability summation. The black curve shows the VBM E_{QP} in ZnO using a fixed 1,000 bands in the polarizability summation and varying the number of bands included in the Coulomb-hole summation.

and

$$\langle vc\mathbf{k}|K^{\mathrm{x}}|v'c'\mathbf{k}'\rangle = \sum_{\mathbf{G}\mathbf{G}'} M_{cv}(\mathbf{k},\mathbf{q},\mathbf{G})v(\mathbf{q}+\mathbf{G})\delta_{GG'}M^*_{c'v'}(\mathbf{k},\mathbf{q},\mathbf{G}') \tag{2.36}$$

where M is defined in Eq. 2.9 and calculated using FFTs as described above in Sec. 2.2.2.

For each $\mathbf{k}$ and $\mathbf{k}'$, we must therefore calculate all the matrix elements $M_{vv'}$, $M_{cc'}$, and M_{vc}. The number of valence and conduction bands required to calculate the absorption spectra within a given energy window each scales linearly with the number of atoms N. So, formally we again have $N^3 \log N$ scaling with the use of FFTs. The summations involved in Eq. 2.35 and Eq. 2.36, however, formally scale as N^6 since there are N^4 terms to compute and each involves a sum over $\mathbf{G}\mathbf{G}'$. In practice, though, except for the largest systems considered, the summations require less time than the matrix elements, and N_v and N_c remain small compared to the values required in the GW step, for example where states with energy up to the dielectric cutoff were required. Usually the energy window used in solving the BSE is approximately 10 eV, giving a spectra converged beyond the visible region. As we discuss below, within the BerkeleyGW package, the parallel wall-time scales as N^2 for this step. However, the N^6 scaling will present a considerable challenge when applying the code to systems of size greater than 100's of atoms.

As was the case for GW code, the $\mathbf{q} \to 0$ limit must be handled carefully and differently depending on the type of screening in the system. For the exchange kernel, we zero out all $\mathbf{G} = \mathbf{G}' = 0$ contributions to the kernel matrix elements, as was discussed in Ref. [55] which gives directly Im ϵ_{M} where ϵ_{M} is the macroscopic dielectric constant. For the direct term, however, we must handle the $\mathbf{G} = 0$ case specially. For these purposes, the $\mathbf{G} = 0$ and $\mathbf{G}' = 0$ terms are removed from Eq. 2.35 and treated separately. For each $(\mathbf{k}cv, \mathbf{k}'c'v')$ we save three terms: the body term, which contains the result of the sum in Eq. 2.35 with the $\mathbf{G} = 0$ or $\mathbf{G}' = 0$ terms removed; the wing term, which contains all the sum of all the remaining terms in the sum with the exception of the single term where $\mathbf{G} = \mathbf{G}' = 0$; the head term, which contains the remaining term from the sum where $\mathbf{G} = \mathbf{G}' = 0$. For metallic systems, as we discussed above, $\epsilon^{-1}(\mathbf{q}, \mathbf{G} = \mathbf{G}' = 0) \propto 1/v(\mathbf{q})$ so that $W(\mathbf{q}, \mathbf{G} = \mathbf{G}' = 0) \propto C$, thus the head term in the kernel remains well behaved. For semiconductors however, $W(\mathbf{q}, \mathbf{G} = \mathbf{G}' = 0) \propto 1/q^2$ so that the head term in the kernel actually diverges as $1/q^2$ when $\mathbf{q} \to 0$. Similarly, the wing term diverges as $1/q$ when $\mathbf{q} \to 0$ for semiconductors, while it again remains well behaved for metals. These limits are summarized in Table 2.2.

Because exciton binding energies and absorption spectra depend sensitively on quantities like the effective mass and joint density of states, it is essential in periodic systems to sample the $\mathbf{k}$-points on a very fine grid. Directly calculating the kernel on this grid in the `kernel` executable would be prohibitively expensive; so instead we interpolate the kernel in the `absorption` executable before diagonalization. For semiconductors, the head and wing kernel terms are not smooth functions of $\mathbf{k}$ and $\mathbf{k}'$ (as we have shown above, they diverge for $\mathbf{q} = \mathbf{k} - \mathbf{k}' \to 0$). Therefore, the quantities that we interpolate are $q^2 \cdot K^{\mathrm{d}}_{\mathrm{head}}$, $q \cdot K^{\mathrm{d}}_{\mathrm{wing}}$ and the body term directly as they are now smooth quantities [115]. For metals, we directly interpolate the kernel without any caveats because all the contributing terms are smooth

functions of $\mathbf{k}$ and $\mathbf{k}'$. As in GW, we treat metals with zero-temperature occupations.

The `absorption` executable requires as input both coarse- and fine-grid wavefunctions. The interpolation is done through a simple expansion of the fine-grid wavefunction in terms of nearest coarse-grid wavefunction:

$$u_{n\mathbf{k}_{\rm fi}} = \sum_{n'} C^{\mathbf{k}_{\rm co}}_{n,n'} u_{n'\mathbf{k}_{\rm co}} \tag{2.37}$$

where $\mathbf{k}_{\rm co}$ is the closest coarse-grid point to the fine-grid point, $\mathbf{k}_{\rm fi}$, and the coefficients $C^{\mathbf{k}_{\rm fi}}_{n,n'}$ are defined as the overlaps between the coarse-grid and fine-grid wavefunctions:

$$C^{\mathbf{k}_{\rm co}}_{n,n'} = \int d\mathbf{r}\, u_{n\mathbf{k}_{\rm fi}}(\mathbf{r}) u^*_{n'\mathbf{k}_{\rm co}}(\mathbf{r}). \tag{2.38}$$

The coefficients $C^{\mathbf{k}_{\rm co}}_{n,n'}$ are normalized so that $\sum_{n'} |C^{\mathbf{k}_{\rm co}}_{n,n'}|^2 = 1$. It should be noted that for a given set of fine bands one can systematically improve the interpolation by including more valence and conduction bands in the coarse grid due to the completeness of the Hilbert space at each $\mathbf{k}$. It should also be noted that we do restrict n and n' to be either both valence or both conduction bands – this is acceptable due to the different character of the conduction and valence bands in most systems.

Using these coefficients we interpolate the kernel with the following formula:

$$\langle vc\mathbf{k}_{\rm fi}|K|v'c'\mathbf{k}'_{\rm fi}\rangle = \sum_{n_1,n_2,n_3,n_4} C^{\mathbf{k}_{\rm co}}_{c,n_1} C^{*\mathbf{k}_{\rm co}}_{v,n_2} C^{*\mathbf{k}'_{\rm co}}_{c',n_3} C^{\mathbf{k}'_{\rm co}}_{v',n_4} \langle n_2 n_1 \mathbf{k}_{\rm co}|K|n_4 n_3 \mathbf{k}'_{\rm co}\rangle \tag{2.39}$$

where K is one of the head, wing, body or exchange kernel terms. As in the case of `epsilon`, this summation can be performed compactly as a set of matrix-matrix multiplications. We utilize the Level 3 BLAS calls `DGEMM` and `ZGEMM` to optimize the performance.

One can systematically improve on the interpolation by using the closest four coarse-grid points to each fine point and using a linear interpolation layer in addition to the wavefunction-based interpolation described above. This is done by default for the interpolation of the first term of Eq. 2.3 for the quasiparticle self-energy corrections $E^{\rm QP} - E^{\rm MF}$:

$$E^{\rm QP}_n(\mathbf{k}_{\rm fi}) = E^{\rm MF}_n(\mathbf{k}_{\rm fi}) + \langle \sum_{n'} |C^{\mathbf{k}_{\rm co}}_{n,n'}|^2 (E^{\rm QP}_{n'}(\mathbf{k}_{\rm co}) - E^{\rm MF}_{n'}(\mathbf{k}_{\rm co}))\rangle_{\mathbf{k}_{\rm co}} \tag{2.40}$$

where the brackets indicate linear interpolation using the tetrahedron method. In this case, the wavefunction-based interpolation layer guarantees that the band crossings are properly handled, and the linear interpolation layer ensures that we correctly capture the energy dependence of the self-energy corrections. In this way, we can construct $E^{\rm QP}$ on the fine grid, or any arbitrary point, given $E^{\rm MF}$ on the fine grid and $E^{\rm QP}$ and $E^{\rm MF}$ on the coarse grid (Fig. 2.8).

As an alternative to calculating the quasiparticle corrections on the coarse grid and interpolating them to the fine grid, the user may choose a less refined method of specifying

the corrections using a three-parameter model involving a scissor-shift parameter ΔE to open the energy gap at the Fermi energy, a zero energy E_0 (typically the band edge), and an energy-scaling parameter C changing the bandwidth (the parameters are specified separately for valence and conduction bands):

$$E^{\mathrm{QP}} = E^{\mathrm{MF}} + \Delta E + C\left(E^{\mathrm{MF}} - E_0\right). \tag{2.41}$$

Having constructed the kernel on the fine grid, we now consider the diagonalization of the kernel. The kernel matrix is of dimension $N_c \cdot N_v \cdot N_k$ where N_k is the number of **k**-points on the fine grid. Formally, the kernel dimension scales as N for periodic systems with small unit cells and N^2 for large systems, where N is the number of atoms. For bulk systems with small unit cells, $N_k \propto 1/N$, but the reduction with increased cell size quickly saturates for large systems and large molecules where $N_k = O(1)$ (to compute a smooth continuum absorption onset, it is necessary to include some level of **k**-point sampling even for isolated systems). The matrix can be diagonalized exactly within LAPACK (`zheevx`) or ScaLAPACK (`pzheevx`). The diagonalization therefore scales as N^3 for periodic systems with small unit cells and N^6 for large systems.

The result of the diagonalization is the set of exciton eigenvalues Ω^S and eigenfunctions $A^S_{cv\mathbf{k}}$ which can be used to construct the absorption spectra (or Im $\epsilon_2(\omega)$) using Eq. 2.5. There are a number of post-processing tools in the package, such as `PlotXct`, which plots the exciton wavefunction in real space, to analyze the exciton states.

The N^6 scaling for large systems in the diagonalization is, in practice, more limiting than the N^6 step in the construction of the kernel. This is because the latter step can be very efficiently parallelized while diagonalization, even with the use of ScaLAPACK, typically saturates at $O(1000)$ CPUs. Often one is only interested in the absorption spectrum, and not all of the correlated exciton eigenfunctions and eigenvalues. For such systems, we use the Haydock recursion iteration method [57, 19]. This is an iterative method based on spectral decomposition and requires only matrix-vector products, which can be efficiently parallelized. This method gives directly the absorption spectrum, the equivalent of Eq. 2.5. In principle, one can get eigenvalues and eigenvectors for a small energy range of interest using iterative Lanczos algorithms.

As mentioned above, the electron-hole kernel should be constructed with a sufficient number of valence and conduction bands to cover the energy window of interest – typically all bands within the desired energy window from the Fermi-energy should be included so that the energy window of the bands included in the calculation is at least twice that of the desired absorption energy window. The `absorption` executable computes the percent deviation from the f-sum rule [62]:

$$\int_0^\infty \epsilon_2(\omega)\omega d\omega = -\frac{\pi\omega_{\mathrm{p}}^2}{2}. \tag{2.42}$$

One should converge this quantity with both the number of valence and conduction bands included. The absorption spectra (or ϵ_2) in the energy window of interest converges much more quickly than ϵ_1 if high-energy transitions outside of the window of interest contribute greatly to the sum rule, since $\epsilon_1(\omega)$ is related to an integration over all frequencies of $\epsilon_2(\omega)$, since ϵ via the Kramers-Kronig relation.

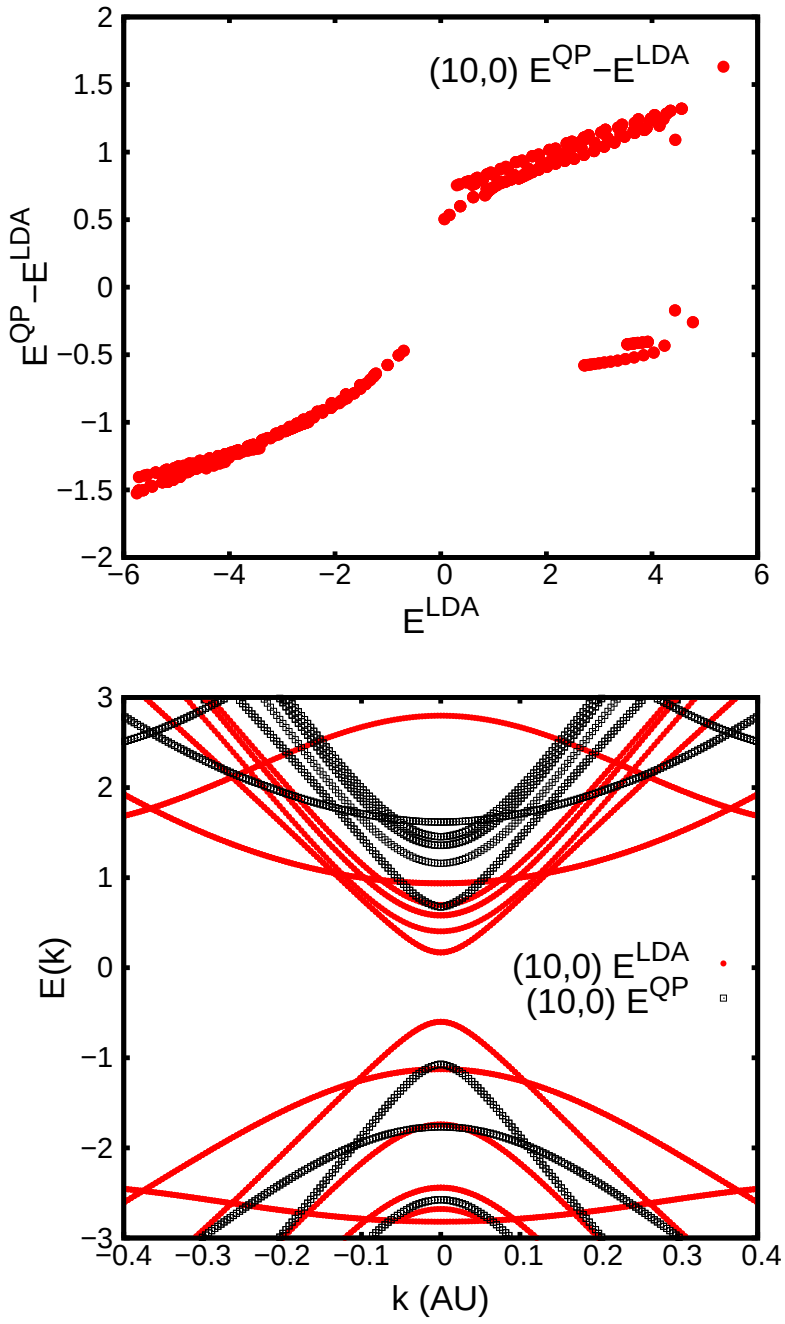

Figure 2.8: Top: GW quasiparticle self-energy corrections, $E^{\mathrm{QP}} - E^{\mathrm{LDA}}$ vs. the LDA energy for (10,0) SWCNT. Both a rigid opening of the band gap and a non-linear energy scaling are present. Bottom: The fine-grid quasiparticle bandstructure using the interpolated self-energy corrections (black-open) and the LDA uninterpolated bandstructure (red-closed). 256 points are used to sample the Brillouin zone.

2.3 Coulomb Interaction

The bare Coulomb interaction is used in many places throughout the code. In all cases, the 1D array $v(\mathbf{q}+\mathbf{G})$ is generated from a single `vcoul_generator` routine in the `Common` directory. However, there is a lot that can be specified about the Coulomb interaction in the code. The Coulomb interaction can be truncated to eliminate the spurious interaction between periodic images of nanosystems in a supercell calculation. One can implement a cell-averaging technique whereby the value of the interaction at each $\mathbf{q}$-point (or $\mathbf{q} \to 0$ in particular) can be replaced by the average of $v(\mathbf{q}+\mathbf{G})$ in the volume the $\mathbf{q}$-point represents. Finally, this average can be made to also include the $\mathbf{q}$-dependence of the inverse dielectric function if W is the final quantity of relevance for the application – such as in the evaluation of W for the self energy.

The BerkeleyGW package contains five general choices for the Coulomb interaction. Firstly, one may choose to use the bulk, untruncated value expressed in Eq. 2.11. There are, in addition, 4 choices of Coulomb interaction that truncate the interaction beyond a certain cutoff in real space, of generic form

$$v_{\mathrm{t}}(\mathbf{r}) = \frac{\Theta(f(\mathbf{r}))}{r} \tag{2.43}$$

where f is some function that describes the geometry in which the interaction is truncated. The four choices available implement the methods of Ismail-Beigi [66]: Wigner-Seitz slab truncation, Wigner-Seitz wire truncation, Wigner-Seitz box truncation, and spherical truncation. The Wigner-Seitz box truncation and the spherical truncation truncate the Coulomb interaction in all three spatial directions, yielding a finite value of $v_{\mathrm{t}}(\mathbf{q}=0)$. All the Wigner-Seitz truncation schemes truncate the interaction at the edges of the Wigner-Seitz cell in the non-periodic directions. Slab truncation is intended for nano-systems with slab-like geometry; the Coulomb interaction is truncated at the edges of the unit cell in the direction (z) perpendicular to the slab plane (xy). Wire truncation is intended for nano-systems with wire-like geometry; the Coulomb interaction is truncated at the edges of the unit cell in the two directions (xy) perpendicular to the wire axis (z). Spherical truncation allows the user to manually specify a spherical truncation radius outside of which the Coulomb interaction will be truncated.

Like the untruncated interaction, the slab-truncation and spherical-truncation schemes have the benefit that $v_{\mathrm{t}}(\mathbf{q}+\mathbf{G})$ can be constructed analytically:

$$v_{\mathrm{t}}^{\mathrm{sph}}(\mathbf{q}) = \frac{4\pi}{q^2} \cdot (1 - r_{\mathrm{c}} \cdot q) \tag{2.44}$$

$$v_{\mathrm{t}}^{\mathrm{slab}}(\mathbf{q}) = \frac{4\pi}{q^2} \cdot (1 - e^{-q_{xy} \cdot z_{\mathrm{c}}} \cos(q_z \cdot z_{\mathrm{c}})) \tag{2.45}$$

where r_{c} and z_{c} are the truncation distances in the radial and perpendicular directions, respectively. The wire-truncation and box-truncation on the other hand are computed numerically through the use of FFTs. First, the truncated interaction, ($2K_0(|q_z|\rho)$ for wire where $\rho = \sqrt{x^2+y^2}$ and K_0 is the modified Bessel function and $1/r$ for box), is constructed on a real-space grid in the Wigner-Seitz cell and folded into the traditional unit

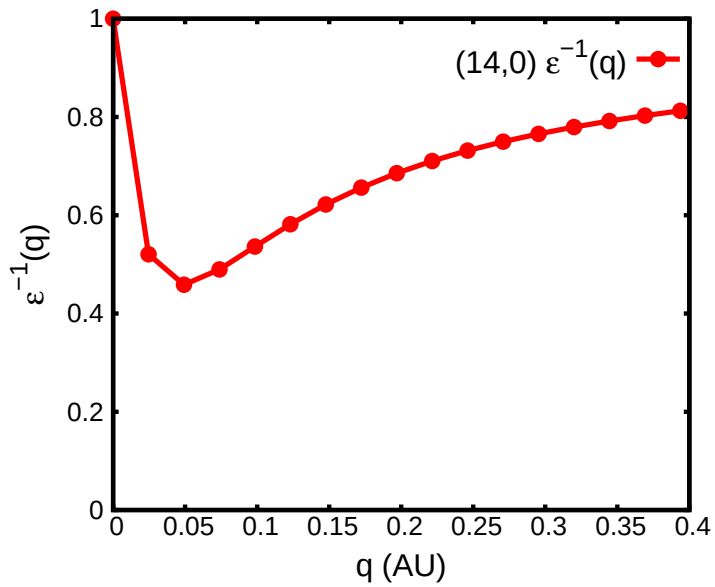

Figure 2.9: $\epsilon^{-1}(q)$ in the (14,0) single-walled carbon nanotube for $\mathbf{q}$ along the tube axis as reported in the `epsilon.log` output file. The circles represent $\mathbf{q}$-grid points included in a $1 \times 1 \times 32$ sampling of the first Brillouin zone.

cell. The real-space grid is typically more dense than the charge density grid used in DFT calculations. The density of points on the real-space grid relative to the charge density grid in xy directions for wire-truncation and in xyz directions for box-truncation is specified by parameters `n_in_wire = 4` and `n_in_box = 2`, respectively. The origin of the Coulomb potential is offset from the origin of the coordinate system by half a grid step to avoid the singularity. We then FFT to yield the $v_{\rm t}(\mathbf{q})$ directly. The FFT is done in parallel. For wire-truncation, xy-planes are evenly distributed among processors and each processor performs 2D-FFTs in xy-planes it owns. For box-truncation, the parallel 3D-FFT is performed as follows: xy-planes are distributed and 2D-FFTs in xy-planes are carried out the same way as for wire-truncation; the data is then transferred from xy-planes to z-rods which are also evenly distributed among processors; finally, each processor performs 1D-FFTs in z-rods it owns. After the FFT, the origin of the Coulomb potential is shifted back to the origin of the coordinate system by multiplying $v_{\rm t}(\mathbf{G})$ with $\exp(i2\pi\mathbf{G}\cdot\frac{1}{2}\mathbf{d})$ where $\mathbf{d}$ is the real-space grid displacement vector.

As mentioned above, for all interaction choices with the exception of cell-box and spherical truncation, the Coulomb interaction diverges as $\mathbf{q} \to 0$. For the case of no truncation, $v(\mathbf{q} \to 0) \propto 1/q^2$; for slab truncation, $v_{\rm t}^{\rm slab}(\mathbf{q} \to 0) \propto 1/q$; for wire truncation, $v_{\rm t}^{\rm wire}(\mathbf{q} \to 0) \propto -\ln(q)$. As we mentioned in Sec. 2.2.2, this divergence is handled in `epsilon` by taking a numerical limit – that is, evaluating ϵ at a small but finite $\mathbf{q}_0$. For `sigma` and `absorption` on the other hand, we are interested in directly evaluating $W(\mathbf{q}, \mathbf{G}\mathbf{G}')$ matrix elements and the appropriate treatment is to replace the divergent (for non-metals)

$W(\mathbf{q} \to 0)$ with an average over the volume in reciprocal space that $\mathbf{q} = 0$ represents:

$$W^{\text{avg}}(\mathbf{q} = 0, \mathbf{G}\mathbf{G}' = 0) = \frac{N_{\mathbf{q}} \cdot \text{Vol}}{2\pi} \int_{\text{cell}} d^n q \; W(\mathbf{q}) \tag{2.46}$$

where "cell" represents the volume in reciprocal space closer to $\mathbf{q} = 0$ than any other $\mathbf{q}$-points, and $d^n q$ represents the appropriate dimensional differential for the truncation scheme (e.g., $d^2 q$ for slab truncation). Equation 2.46 yields a finite number in all truncation schemes even when $W(\mathbf{q} \to 0)$ itself is divergent because of the reduced phase-space around $\mathbf{q} \to 0$.

For metallic systems, it is particularly important to use Eq. 2.46 to average W, as opposed to separately averaging $v(\mathbf{q})$ and $\epsilon^{-1}(\mathbf{q})$, since for metals $\epsilon(\mathbf{q}) \propto v(\mathbf{q})$ at small $\mathbf{q}$, yielding a constant limit: $W^{\text{metal}}(\mathbf{q} \to 0) = C$. The user can tell the code which model to use for the $\mathbf{q}$-dependence of ϵ^{-1} by specifying one of the screening flags in the input files. The $\mathbf{q} \to \mathbf{0}$ limits for the inverse dielectric function and screened Coulomb interaction are enumerated in Table 2.2 for the semiconductor and metallic screening types.

As shown in Fig. 2.9, including ϵ^{-1} in the cell-averaging scheme is also important when using a truncated interaction where $\epsilon(\mathbf{q} = 0) = 1$ but quickly rises by the first non-zero $\mathbf{q}$-point [66]. The figure shows $\epsilon^{-1}(q)$ along the tube axis in the (14,0) SWCNT. $\epsilon^{-1}(0) = 1$ but decreases nearly by half by the first non-zero grid point included in a $1 \times 1 \times 32$ sampling of the first Brillouin zone. $\epsilon^{-1}(0) = 1$ is a general property of truncated systems with semiconductor-type screening since the q^2 dependence of the polarizability approaches 0 faster than the Coulomb interaction diverges at $q \to 0$. BerkeleyGW uses this W-averaging procedure by default for cases with truncated Coulomb interaction.

In general, using an extension of Eq. 2.46 even for $\mathbf{q}, \mathbf{G}\mathbf{G}' \neq 0$ can speed up the convergence of a `sigma` or `absorption` calculation with respect to the number of $\mathbf{q}$-points required in the calculation. This can be easily explained by the fact that one is replacing a finite sum over $\mathbf{q}$-points with an integral – mimicking a calculation on a much larger set of $\mathbf{q}$-points or a much larger unit cell. The user can ask that BerkeleyGW use the cell-averaged W for all $\mathbf{q}$ and $\mathbf{G}$ below an energy cutoff specified in the input file.

The averaging is implemented in the code using a Monte Carlo integration method with 2,500,000 random points in each cell.

To conclude this chapter, we want to point the interested reader to the appendices of this work and www.berkeleygw.org to download the BerkeleyGW code and to find more details of usage, methodology etc... The package may be used an example of a modern GW-BSE implementation emphasizing the study of large and complex materials. The extension of the GW-BSE methodology to the study of large and complex materials within the BerkeleyGW package is the basis for the work presented in the next three chapters.

$\epsilon^{-1}_{\mathbf{GG'}}$	Semiconductor	Semiconductor Truncation	Metal	Metal Truncation
Head	Constant	1	q^2	$\frac{1}{V_0^T(q)}$
Wing	q	$q\epsilon^{-1}_{00q}$	q^2	$\frac{1}{V_0^T(q)}$
Wing′	$\frac{1}{q}$	$q\epsilon^{-1}_{00q}V_0^T(q)$	Constant	Constant
$W_{\mathbf{GG'}}$	Semiconductor	Semiconductor Truncation	Metal	Metal Truncation
Head	$\frac{1}{q^2}$	$V_0^T(q)\epsilon^{-1}_{00q}$	Constant	Constant
Wing	$\frac{1}{q}$	$q\epsilon^{-1}_{00q}V_0^T(q)$	Constant	Constant
Wing′	$\frac{1}{q}$	$q\epsilon^{-1}_{00q}V_0^T(q)$	Constant	Constant

Table 2.2: Top: $\mathbf{q} \to \mathbf{0}$ limits of the head, $\epsilon^{-1}_{\mathbf{00}}(\mathbf{q})$, wing, $\epsilon^{-1}_{\mathbf{G0}}(\mathbf{q})$ and wing′, $\epsilon^{-1}_{\mathbf{0G'}}(\mathbf{q})$, of the inverse dielectric matrix. Bottom: $\mathbf{q} \to \mathbf{0}$ limits of the head and wings of the screened Coulomb interaction, $W_{\mathbf{GG'}}(\mathbf{q})$.

Chapter 3

Applications to molecules and systems requiring many empty states

As described in the previous chapter, the GW methodology has been successfully applied to the study of the quasiparticle properties of a wide range of systems from traditional bulk semiconductors, insulators and metals [62, 83] to nanosystems like polymers, nanotubes and molecules [127, 128, 37]. The approach yields quantitatively accurate quasiparticle band gaps, dispersion relations and optical spectra from first principles. A perceived drawback of the GW methodology is its computational cost; usually thought to be an order of magnitude more than a typical DFT calculation. As discussed in the previous chapter, the most expensive part of a tyical GW computation is actually the DFT step.

One of the main computational bottlenecks of the traditional GW method [62] is the cost to generate the large number of empty orbitals needed to converge the Coulomb-hole summation term of the self-energy. Even for relatively simple materials like ZnO [122], the required number of empty bands to converge the Coulomb-hole summation is on the order of several thousands.

For molecules the computational requirements can be even higher. The number of empty orbitals required in the Coulomb-hole summation is fixed by the dielectric energy cutoff $\mathrm{E_{cut}}$ and the supercell volume, $\mathrm{Vol_{SC}}$:

$$N_b \propto \mathrm{E_{cut}^{3/2}} * \mathrm{Vol_{SC}}. \tag{3.1}$$

For molecular systems, studied with plane-wave basis set and periodic boundary conditions (as in the BerkeleyGW package described above) even when using a truncated Coulomb interaction, the supercell volume must be taken as at least eight times the volume taken up by the electronic charge-density. Typically, the volume containing 99% of the integrated charge density is chosen. The factor of eight comes from the requirement that an equivalent amount of vacuum is included in each direction; so that even if the Coulomb interaction is truncated at a separation greater than that of the greatest separation of charge on the molecule, there will be no interaction between charge on one molecule and it's periodic images.

It should be pointed out that, the use of the plane-wave basis set over localized basis set alternatives comes with significant advantages despite the problems above. Predominantly, the plane-wave basis sets describes with ease the free-electron like states above the vacuum, whose inclusion is critical for computing a converged polarizability χ and self-energy, Σ. Localized orbitals sufficiently describe the occupied molecular orbitals, but fail to describe most of the empty orbitals above the vacuum. Therefore, one cannot greatly reduce the cost of GW calculations on molecules by resorting to a localized-basis set alone.

Molecular systems also often require the costly computation of off-diagonal matrix elements of the self-energy, Σ, operator. This is because within the G_0W_0 approximation using the common starting point of DFT, the lowest unoccupied molecular orbital (LUMO) is often bound (having energy below the vacuum levels) at the mean-field level, whereas the true quasiparticle state is a resonant orbital with energy above the vacuum and wavefunction including some free-electron like contribution. In such a case the diagonal approximation of Σ in the DFT basis breaks down and one must in general compute off-diagonal elements within a large energy window. This can increase the cost of such calculations by multiple orders of magnitude.

Both the issue of empty orbital generation and off-diagonal elements of Σ are discussed in the following two sections.

3.1 Empty States

There has been much research effort invested in recent years to reduce the need for empty orbitals in the GW formalism [113, 140, 49, 25, 136]. In particular, it was proposed by Tiago et. al. [136] that a truncation of the sum over empty orbitals in Eq. (3.2), below, can be achieved with minimal loss of accuracy by adding the contribution of the remaining orbitals within the static (COHSEX) approximation [59, 62]. This approach, however, was shown to be of limited use by Bruneval et. al. [25], where instead of using the static approximation for the remaining sum, the authors proposed using an approach based on a common non-zero energy denominator in Eq. (3.2). The main drawback of this common energy denominator approximation (CEDA) approach (known also as the extrapolar method) is that the energy denominator is not uniquely defined and can only be treated as a somewhat ad-hoc parameter, and the quasiparticle energy convergence is not monotonic with this parameter. Recently, however, studies by Kang and Hybertsen [72] have shown that a modified static COHSEX approach can be used to accurately minimize the empty orbitals problem in the Coulomb-hole summation of Eq 3.2. Therefore, we propose a modified static remainder approach based on Tiago's results [136] that is more fully justified by the recent Kang-Hybertsen result [72]. The new approach yields accurate GW Coulomb-hole absolute energies with less than 10% of the traditionally necessary empty orbitals. Furthermore, unlike the extrapolar method of Bruneval [25], this approach yields an easy to implement procedure with no adjustable parameters. For simplicity of presentation, we shall discuss our approach within the generalized plasmon pole model for the dielectric matrix. However the approach can be applied straightforwardly to full frequency calculations.

Within the GW and static-COHSEX (the zero-frequency limit of GW) approximations for the self energy, the self-energy operator, Σ, can be broken into two parts, [62, 59]

$\Sigma = \Sigma_{\rm SX} + \Sigma_{\rm CH}$ where $\Sigma_{\rm SX}$ is the screened-exchange operator and $\Sigma_{\rm CH}$ is the Coulomb-hole operator. In a conventional GW calculation within the generalized plasmon-pole approximation, both the calculation of the Coulomb-hole self energy term:

$$\langle n\mathbf{k}|\,\Sigma^N_{\rm CH}(\mathbf{r},\mathbf{r}';E)\,|n'\mathbf{k}\rangle = \frac{1}{2}\sum_{n''}^{N}\sum_{\mathbf{q}\mathbf{G}\mathbf{G}'}\langle n\mathbf{k}|\,e^{i(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}}\,|n''\mathbf{k}-\mathbf{q}\rangle\,\langle n''\mathbf{k}-\mathbf{q}|\,e^{-i(\mathbf{q}+\mathbf{G}')\cdot\mathbf{r}'}\,|n'\mathbf{k}\rangle \quad (3.2)$$
$$\times\{\frac{\Omega^2_{\mathbf{G}\mathbf{G}'}(\mathbf{q})}{\tilde{\omega}_{\mathbf{G}\mathbf{G}'}(\mathbf{q})\left[E-E_{n''\mathbf{k}-\mathbf{q}}-\tilde{\omega}_{\mathbf{G}\mathbf{G}'}(\mathbf{q})\right]}\,v(\mathbf{q}+\mathbf{G}')\}$$

and the calculation of the dielectric screening matrix, $\epsilon = 1 + 4\pi\chi$, at $\omega = 0$:

$$\epsilon_{\mathbf{G}\mathbf{G}'}(\mathbf{q};0) = \delta_{\mathbf{G}\mathbf{G}'} \quad (3.3)$$
$$-\,v(\mathbf{q}+\mathbf{G})\sum_{n}^{\rm occ}\sum_{n'}^{N}\sum_{\mathbf{k}}\langle n\mathbf{k}+\mathbf{q}|\,e^{i(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}}\,|n'\mathbf{k}\rangle$$
$$\langle n'\mathbf{k}|\,e^{-i(\mathbf{q}+\mathbf{G}')\cdot\mathbf{r}'}\,|n\mathbf{k}+\mathbf{q}\rangle\times\frac{1}{E_{n\mathbf{k}+\mathbf{q}}-E_{n'\mathbf{k}}},$$

involve a summation over empty orbitals. Here N is the number of empty orbitals in the truncated sum, $n\mathbf{k}$ is a Bloch orbital with a given crystal momentum $\mathbf{k}$, band index n and energy $E_{n\mathbf{k}}$, $v(\mathbf{q}+\mathbf{G})$ is the bare Coulomb interaction in reciprocal space and $\Omega_{\mathbf{G}\mathbf{G}'}(\mathbf{q})$ and $\tilde{\omega}_{\mathbf{G}\mathbf{G}'}(\mathbf{q})$ are plasmon-pole parameters [62]. The dielectric matrix $\epsilon_{\mathbf{G}\mathbf{G}'}(\mathbf{q};\omega)$ is required to construct the screened Coulomb interaction $W_{\mathbf{G}\mathbf{G}'}(\mathbf{q};\omega) = \epsilon^{-1}_{\mathbf{G}\mathbf{G}'}(\mathbf{q};\omega)v(\mathbf{q}+\mathbf{G}')$ where v is the bare Coulomb interaction.

In practice, the band convergence of absolute energy levels in $\Sigma_{\rm CH}$ with the number of empty orbitals is extremely slow. For many systems, the quasiparticle energy dependence on the number of empty orbitals in the dielectric screening (e.g. Eq. (3.3)) converges much faster than with the number of empty orbitals in Eq. (3.2). For example, recent calculations for ZnO show that the Coulomb-hole contribution to the electronic band gap does not converge until 3,000+ empty orbitals are included in the summation in Eq. (3.2) [122]. In Fig. 3.1, we demonstrate the slow convergence of Eq. (3.2) in ZnO. The situation is even worse for nanosystems where absolute energies are often required for applications involving interfaces over which absolute energy level alignment is needed such as the cases for molecular electronics or photovoltaic applications.

From the bottom panel of Fig. 3.1, it is immediately evident that the quasiparticle energy converges much more slowly with respect to the number of empty orbitals included in the Coulomb-hole summation, Eq. (3.2) than from the ϵ summation, Eq. (3.3). Additionally, one may compute Eq. (3.3) in an alternative density functional perturbation theory approach that avoids the sum over empty orbitals. Similar techniques [49, 141] to avoid the empty orbitals for Eq. (3.2) have been proposed but they are more difficult to implement and use. Therefore, any reduction in the summation in Eq. (3.2) over large numbers of empty orbitals can greatly reduce the cost of calculation for standard GW approaches.

The Coulomb-hole self-energy contribution to the convergence of energy levels and electronic gaps in bulk silicon (using a 5x5x5 $\mathbf{k}$-point grid) and the silane (SiH_4) molecule (in a supercell calculation) are shown in Fig. 3.2 as a function of the band cutoff, N, in

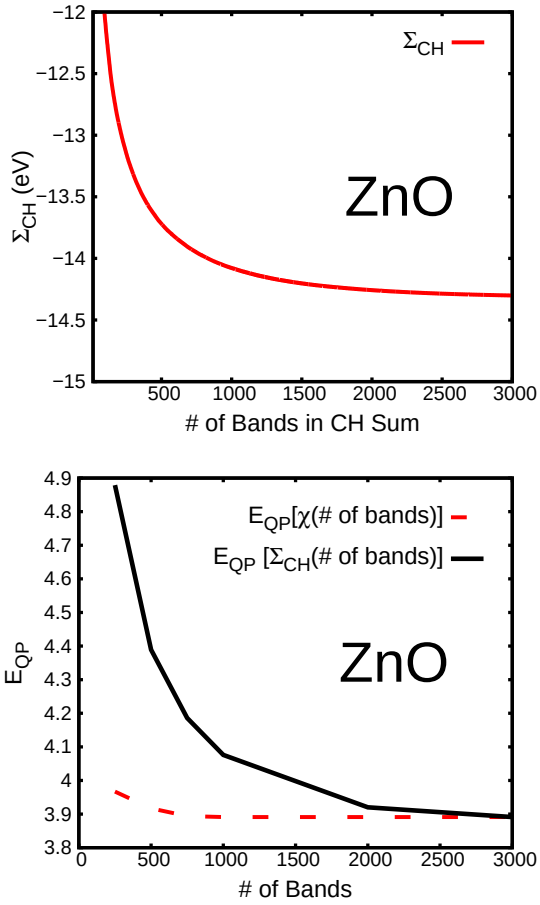

Figure 3.1: Left: The convergence of the Coulomb hole contribution to the self-energy, Eq. (3.2), with respect to the number of orbitals included in the summation, N, using a dielectric matrix calculated with 1000 empty bands. For all calculations on ZnO, a 5x5x4 **k**-point grid is used. Right: The convergence of the quasiparticle energy, E_{QP}, with respect to empty states in the polarizability sum Eq. (3.3) and with respect to empty states in the Coulomb-hole sum Eq. (3.2). The red curve shows the VBM E_{QP} in ZnO using a fixed 3,000 bands in the Coulomb-hole summation and varying the number of bands included in the polarizability summation. The black curve shows the VBM E_{QP} in ZnO using a fixed 1,000 bands in the polarizability summation and varying the number of bands included in the Coulomb-hole summation.

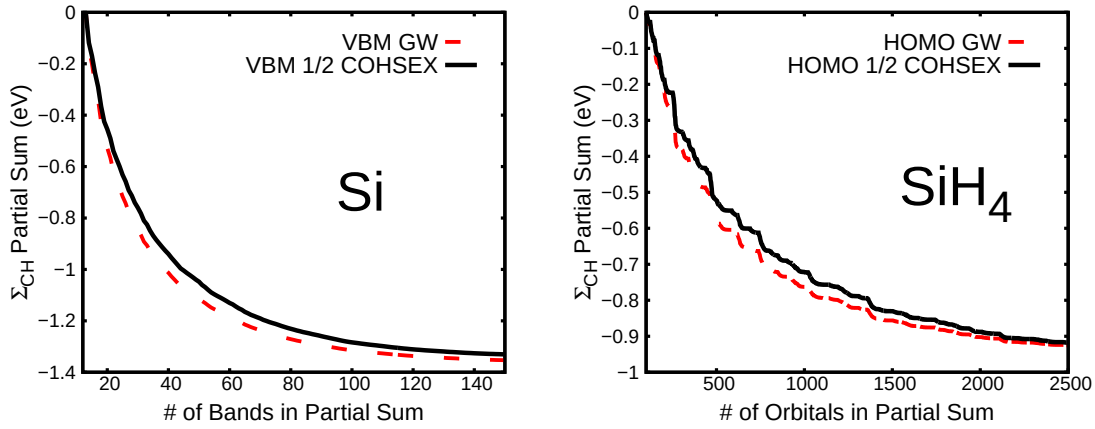

Figure 3.2: Comparison between the contributions to the Coulomb-hole sum for the full GW operator vs. results from the 1/2 the static COHSEX Coulomb-hole operator for orbitals beyond the number of real DFT bands/orbitals used: 12 in silicon and 100 in Silane. A 5x5x5 $\mathbf{k}$ grid is used in Si. The plotted quantity is $\sum_{n''=n_{DFT}+1}^{N} \sum_{\mathbf{qGG'}} \langle n\mathbf{k}| e^{i(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}} |n''\mathbf{k}-\mathbf{q}\rangle \langle n''\mathbf{k}-\mathbf{q}| e^{-i(\mathbf{q}+\mathbf{G'})\cdot\mathbf{r'}} |n'\mathbf{k}\rangle \times \mathrm{I}^{\mathrm{CH}}_{\mathbf{GG'}}(\mathbf{q}, n, n', n'')$ where I^{CH} is the term in $\{\}$ in Eqs. (3.2) and (3.4) respectively.

Eq. (3.2). The convergence on energy levels in silane is significantly slower than that in silicon [115] because of the large number of free-electron like vacuum states. The silicon calculations were done with a 25 Rydberg wavefunction cutoff and a 10 Rydberg dielectric matrix cutoff. The silane calculations were done with a 75 Rydberg wavefunction cutoff and a 6 Rydberg dielectric matrix cutoff. The needed volume of the supercell used, $(25au)^3$, and the corresponding number of vacuum states, is minimized by using a truncated Coulomb interaction [66]. Despite this, the largest computational cost in the GW calculation on silane is the DFT generation of the empty orbitals, representing more than 50% of the total computational expense. The calculation of the polarizability and the evaluation of the self energy require less computational time, and they scale nearly linearly to thousands of CPUs.

The static COHSEX method is the static limit of the GW approximation for the self energy – where everywhere $\epsilon(\mathbf{G}, \mathbf{G'}, \omega)$ is replaced by $\epsilon(\mathbf{G}, \mathbf{G'}, \mathbf{0})$. The static remainder approach is based on the fact that the expectation value of the static COHSEX Coulomb-hole operator can be expressed either in a closed form or as a sum over empty orbitals:

$$\Sigma_{\mathrm{CH}}^{Stat/N}(n,\mathbf{k}) = \tag{3.4}$$
$$\frac{1}{2}\sum_{n''}^{N}\sum_{\mathbf{qGG'}} \langle n\mathbf{k}|\, e^{i(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}}\, |n''\mathbf{k}-\mathbf{q}\rangle \langle n''\mathbf{k}-\mathbf{q}|\, e^{-i(\mathbf{q}+\mathbf{G'})\cdot\mathbf{r'}}\, |n\mathbf{k}\rangle$$
$$\times \{\left[\epsilon^{-1}_{\mathbf{GG'}}(\mathbf{q};0)-\delta_{\mathbf{GG'}}\right] v\big(\mathbf{q}+\mathbf{G'}\big)\}$$

and,

$$\Sigma_{\mathrm{CH}}^{Coh/\infty}(n,\mathbf{k}) = \frac{1}{2}\sum_{\mathbf{qGG'}} \langle n\mathbf{k}|\, e^{i(\mathbf{G}-\mathbf{G'})\cdot\mathbf{r}}\, |n\mathbf{k}\rangle \left[\epsilon^{-1}_{\mathbf{GG'}}(\mathbf{q};0)-\delta_{\mathbf{GG'}}\right] v\big(\mathbf{q}+\mathbf{G'}\big) \tag{3.5}$$

where N and ∞ denote a truncated empty state summation and closed form expression, respectively. Equation (3.4) is equal to Eq. (3.2) in the limit of static dielectric screening. In our modified static remainder approach, we calculate both the GW Σ_{CH} partial sum (Eq. (3.2)) and the COHSEX Σ_{CH} partial sum (Eq. (3.4)) up to the number of DFT bands available. We then add a modified static correction to the GW Coulomb-hole energies:

$$\langle n\mathbf{k}|\Sigma^{\infty}_{\mathrm{CH}}\big(\mathbf{r},\mathbf{r'};E\big)\,\big|n'\mathbf{k}\big\rangle = \tag{3.6}$$
$$\langle n\mathbf{k}|\,\Sigma^{N}_{\mathrm{CH}}\big(\mathbf{r},\mathbf{r'};E\big)\,\big|n'\mathbf{k}\big\rangle + \frac{1}{2}\Bigg(\,\langle n\mathbf{k}|\,\Sigma^{Stat/\infty}_{\mathrm{CH}}\big(\mathbf{r},\mathbf{r'}\big)\,\big|n'\mathbf{k}\big\rangle - \langle n\mathbf{k}|\,\Sigma^{Stat/N}_{\mathrm{CH}}\big(\mathbf{r},\mathbf{r'}\big)\,\big|n'\mathbf{k}\big\rangle\,\Bigg).$$

The factor of 1/2 in Eq. (3.6) is justified from the recent work of Kang and Hybertsen [72], where the authors show that the GW contribution of high energy bands (corresponding to large **G**-vectors) to the Coulomb-hole self energy asymptotes to 1/2 of the equivalent static COHSEX band contribution. In that paper, the authors propose using a specific modified static-COHSEX operator to entirely remove the need for empty orbitals. They alter the static-COHSEX operator to mimic the full-dynamical behavior even for contributions from the low energy orbitals. In our current approach, we include the full GW contribution from the low energy orbitals and add a single correction for the high-energy orbitals, where the static approximation is expected to perform well. One advantage of the present approach is that it can be used in conjunction with a full-frequency (as opposed to GPP model) screening approach to both calculate the fine structure of energy dependence of the self energy, $\Sigma(\omega)$, as well as converging the absolute value with respect to empty orbitals.

A comparison between the convergence of the residual value of the GW expression (Eq. (3.2)) and 1/2 of the static COHSEX approximation (Eq. (3.4)) for the Coulomb-hole contribution to the electron self-energy starting at some n_{DFT} is shown in Fig. 3.2. The figure shows the cumulative contributions of the high-energy orbitals to Σ_{CH} for both the GW operator and the 1/2 static COHSEX operator for orbitals above 12 and 100 for silicon and silane, respectively. The residual value of the 1/2 static COHSEX results reproduce the equivalent GW curves extremely well. Therefore, replacing the GW operator with the modified static remainder in Eq. (3.6) yields very good agreement with a fully converged GW calculation. This justifies the truncation of the partial sum in Eq. (3.2) and the

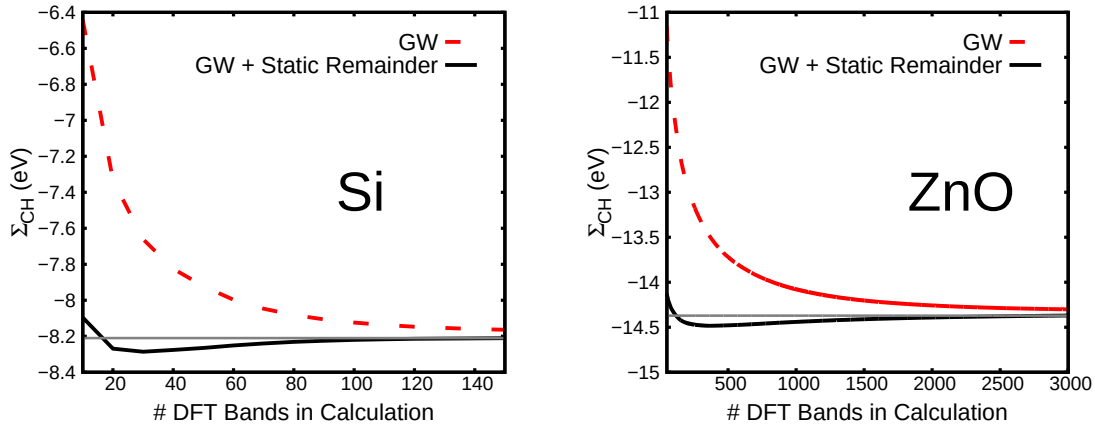

Figure 3.3: Coulomb-hole energies of the valence band maximum in Si (left) and ZnO (right) in the modified static-remainder approach compared to the energies from the standard approach of truncating the Coulomb-hole summation in Eq. (3.2) as a function of the number of DFT bands. In the static-remainder approach the summation is also truncated at the same number of bands but the modified static remainder is added to the sum. A 5x5x5 and 5x5x4 **k**-point grid is used in Si and ZnO respectively.

addition of the modified static remainder correction. For both silicon and silane, one can get a converged Σ_{CH} to within 10's of meV with less than 10% of the original number of empty orbitals required. This is a very high level of accuracy considering the modified static correction in both cases is greater than 1 eV. Even higher accuracy may be reached if one increases the number of actual DFT empty orbitals used. In the case of Si, a converged Σ_{CH} can be reached with the use of only 10 empty bands. A 5x5x5 **k**-point grid was used for Si. Furthermore, the convergence for this approach is nearly monotonic in terms of the number of empty Kohn-Sham orbitals employed in the calculation. Figure 3.3 shows the convergence of the modified static remainder corrected Σ_{CH} and the uncorrected GW Σ_{CH} as a function of DFT empty bands used for Si an ZnO.

To test the modified static reminder approach on a large molecular system, we compute the Coulomb-hole contribution to the self-energy for the BND (bithiophene naphthalene diimide) molecule containing 46 atoms [134]. The supercell was set to 76.93 x 36.31 x 20.18 atomic units. The calculations were done with a 60 Rydberg wavefunction cutoff and a 6 Rydberg dielectric matrix cutoff. The polarizability was computed with 953 orbitals (78 occupied + 875 empty orbitals up to 1 Rydberg cutoff in DFT eigenvalues), and the Coulomb-hole part of the self-energy was evaluated as a function of the number of orbitals, as shown in Fig. 3.4. One can see that the Coulomb-hole term computed with 953 orbitals without the addition of the remainder is only converged to within 1 eV. Including the static remainder correction improves the convergence to better than 0.1 eV.

In conclusion, a modified static remainder approach that reduces the number of empty states involved in evaluating Σ_{CH} by over an order of magnitude has been presented.

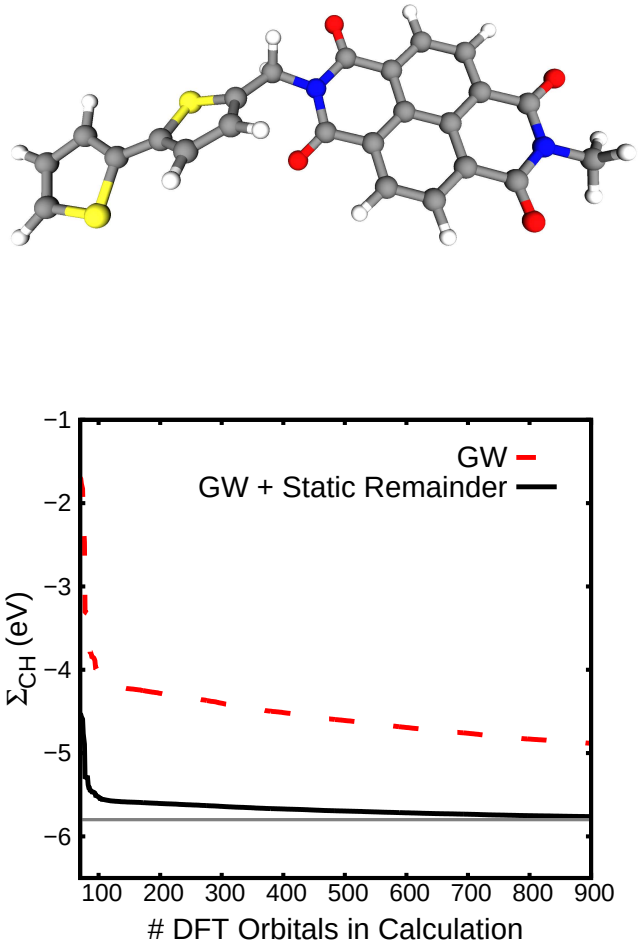

Figure 3.4: (left) Model of the BND (bithiophene naphthalene diimide) molecule. (right) Coulomb-hole part of the self-energy, with and without the static-remainder, for the highest occupied molecular orbital (HOMO) of the BND molecule as a function of the number of DFT orbitals included in the Coulomb-hole sum.

This approach is particularly useful when applying the GW method to molecules and other nanostructures where absolute energies, as opposed to just energy gaps, are desired. A limitation of this method is that it does not address the problem of the sum over empty states required in evaluating the dielectric matrix (Eq. (3.3)). However, the dielectric matrix converges faster than the absolute energies of Σ_{CH} for many solids [122] and can more easily be replaced by calculation using the density functional perturbation theory approaches. Our approach here shows nearly monotonic convergence towards the converged GW Σ_{CH} values and can be implemented in a simple and automatic way in standard GW computer codes.

3.2 Off-diagonal Elements of Σ

As mentioned in the previous chapters and section, the GW approximation to the electron self energy has become the method of choice for treating the excited-state properties of solids from first principles [59, 62]. This approach is typically implemented starting from a DFT mean-field within the plane wave pseudo-potential formalism [63, 62, 50], and has been shown to work extremely well for a wide variety of condensed matter systems – metals [99, 100], semiconductors and insulators, [63, 62, 50] and for nanostructures [127, 37]. However, one of the commonly used approximations, that the DFT Kohn-Sham orbitals are the same as the quasiparticle wavefunctions, can sometimes break-down, leading to errors.

Some notable examples where such a break-down may occur are in the calculation of electron affinity in molecular systems and defect levels in solids. In molecular systems, quasiparticle states of interest could have a mean-field energy below the vacuum level whereas the actual quasiparticle level (after the self-energy correction) may be above the vacuum level. The former is a localized bound state; the latter is a resonant state [51]. A similar problem can occur with defect states in solids. The defect level within DFT can be within the conduction band continuum (a resonant state), however after the GW self-energy correction within the band gap of the solid (a localized state).

There have been several attempts to address this problem [41, 26, 53]. One method would be to expand the quasiparticle wavefunctions in terms of the Kohn-Sham orbitals. But the approach involves the calculation of the off-diagonal elements of Σ on a grid of frequencies which is conceptually and numerically difficult. Another method is the QPscGW approach of Faleev et al. [41] that iteratively constructs a mean-field starting point such that, by construction, the quasiparticle wavefunctions and mean-field orbitals are close. However, as we discuss below, the utility of this approach is limited due to its high computational cost. Alternatively, several groups [26, 53] have constructed the full Hamiltonian in so-called static-COHSEX and carried the calculation to varying levels of self-consistency as a mean-field starting point for a subsequent GW calculation. Unfortunately, again the off-diagonal matrix elements of the GW Hamiltonian in the Kohn-Sham basis can sometimes converge slowly with respect to Hamiltonian size. As a result, these approaches often suffer due to a small basis set used for constructing the static-COHSEX Hamiltonian.

Below, two alternative methods are presented based on the static-COHSEX approximation that allow us to efficiently construct the improved quasiparticle wavefunctions and subsequently perform a GW calculation with Σ nearly diagonal in the new basis. The first method is a fully self-consistent static-COHSEX followed by GW (sc-

COHSEX+GW) method, where approximate quasiparticle wavefunctions are obtained from a self-consistent solution to the static-COHSEX Hamiltonian. The second, the static-offdiagonal GW method, allows for a computationally less intensive treatment of effectively just the off-diagonal matrix elements (in the Kohn-Sham basis) within the static-COHSEX Hamiltonian to obtain approximate quasiparticle wavefunctions. Both these approaches have been implemented within a plane-wave basis set. The main advantage of these methods over the previous ones is that we work completely in a plane-wave basis. Both approaches do not require an explicit construction of the Hamiltonian, as in a typical DFT calculation, where only the Hamiltonian times the wavefunction is required. We also apply these methods to molecular and bulk solid examples of silane and silicon respectively. These methods make significant improvement to the electron affinity of silane. In silicon, the static-offdiagonal GW gives virtually identical quasiparticle energies to a conventional GW calculation. The sc-COHSEX+GW on the other hand overestimates the band gaps for reasons to be discussed below.

The quasiparticle energies within a conventional GW approach are typically calculated within a first-order perturbation theory approximation:

$$\epsilon^{\mathrm{QP}} = \epsilon^{\mathrm{DFT}} + \langle \psi^{\mathrm{DFT}} | \Sigma_{\mathrm{GW}}(\psi^{\mathrm{DFT}}, \varepsilon^{-1}_{\mathrm{DFT}}) - V_{\mathrm{XC}} | \psi^{\mathrm{DFT}} \rangle \tag{3.7}$$

which is known as the diagonal $\mathrm{G_0W_0}$ approximation. In Eq. (3.7), ϵ^{QP} is the quasiparticle energy, ψ^{DFT} and ϵ^{DFT} are DFT eigenfunctions and eigenvalues. $\Sigma_{\mathrm{GW}}(\psi^{\mathrm{DFT}}, \varepsilon^{-1}_{\mathrm{DFT}})$ is the self energy operator constructed with DFT eigenvalues, eigenfunctions and dielectric matrix $\varepsilon^{-1}_{\mathrm{DFT}}$, and V_{XC} is the DFT exchange-correlation potential. Within this approach, the dynamic Σ_{GW} is to be evaluated at ϵ^{QP} in a self-consistent way.

As seen from Eq. (3.7), the diagonal $\mathrm{G_0W_0}$ approach assumes that the DFT (often LDA or GGA) eigenfunctions are a good approximation to the quasiparticle wavefunctions. As discussed previously, there are known limitations of this approximation (where $\psi^{\mathrm{QP}} \not\approx \psi^{\mathrm{DFT}}$). One way to solve this problem is to diagonalize the full $\mathrm{G_0W_0}$ matrix [115, 53], $\epsilon^{\mathrm{DFT}}\delta_{ij} + \langle \psi_i^{\mathrm{DFT}} | \Sigma_{\mathrm{GW}}(\epsilon^{\mathrm{QP}}) - \mathrm{V_{XC}} | \psi_j^{\mathrm{DFT}} \rangle$, constructed in the DFT eigenfunction basis ψ_j^{DFT}. However, in practice, the matrix is limited to a small number of states (rows/columns) due to computational cost. Additionally, all the matrix elements should be evaluated at ϵ^{QP} for each separate quasiparticle level, which is challenging to evaluate in a self-consistent fashion. Thus, diagonalizing the full $\mathrm{G_0W_0}$ matrix with sufficient rows and columns is extremely difficult.

Instead of constructing and diagonalizing the full $\mathrm{G_0W_0}$ matrix in the ψ^{DFT} basis, we propose the static-offdiagonal GW approach as shown in the left track of Figure 3.5. In this approach, using the DFT eigenvalues and eigenfunctions, we construct the static-COHSEX operator *in a plane-wave basis set* up to the convergent plane-wave DFT wavefunction cutoff. In particular, in the static-COHSEX operator, the screened exchange (SEX) and Coloumb hole (COH) terms are computed from the DFT eigenfunctions and eigenvalues, but expressed as matrices in the plane-wave basis. We then diagonalize the COHSEX Hamiltonian, $(H_0^{\mathrm{DFT}} + \Sigma_{\mathrm{COHSEX}})$ using an iterative alogrithm. Here H_0^{DFT} is defined as the DFT Hamiltonian without the exchange-correlation term, V_{xc}. It is worthwhile to point out that solving this eigensystem iteratively only requires one to compute $(H_0^{\mathrm{DFT}} + \Sigma_{\mathrm{COHSEX}})\phi$ products, where ϕ is some trial quasiparticle wavefunction. This is

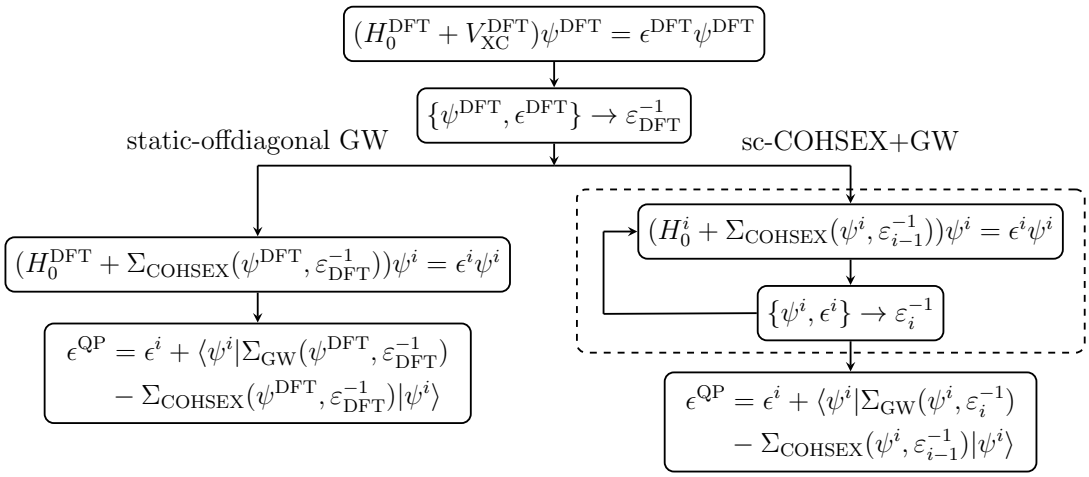

Figure 3.5: Outline of the static-offdiagonal GW and the sc-COHSEX+GW methodologies. The H_0^i refers to the kinetic, ionic and hartree potentials constructed with density from ψ^i. See text for details.

similar to any non self-consistent DFT hybrid functional calculation. Having solved this eigensytem, one then does a diagonal G_0W_0 calculation as shown in the left track of Figure 3.5, but now in the basis of the quasiparticle wavefunctions. Note that this procedure does not change the mean-field in the GW calculation. This approach, which is equivalent to diagonalizing the G_0W_0 matrix in the static limit in a complete plane-wave basis, is an effective scheme for the inclusion of the offdiagonal G_0W_0 matrix elements of the Kohn-Sham basis. It can also be seen as a transformation to a basis within which the G_0W_0 matrix (still constructed from G_0 and W_0 using the DFT eigenvalues and eigenfunctions) is nearly diagonal.

Alternatively, one could replace the DFT mean-field starting point completely by replacing the DFT mean-field Hamiltonian with a self-consistent static-COHSEX (sc-COHSEX) mean-field Hamiltonian. This approach is outlined on the right track of Figure 3.5. As before, we use the DFT eigenfunctions and eigenvalues to construct an initial polarizabilty. However, in this second approach, the SEX operator (with fixed screening) is updated self-consistently as we diagonalize the static-COHSEX Hamiltonian. The eigenvalues and eigenfunctions from this diagonalization are used to construct a new polarizabilty and dielectric matrix. This process is repeated to reach self-consistency in the dielectric matrix. In practice, for the systems considered, we find that one update of the polarizability is sufficient. We then do a standard diagonal G_0W_0 calculation, in the basis of the SC-COHSEX orbitals, using the sc-COHSEX eigenvalues, eigenfunctions and updated polarizabilty as our mean-field starting point.

Lets now compare our sc-COHSEX method with previous self-consistent quasiparticle methods of Bruneval et al [26] and QPscGW [41]. In the work of Bruneval et al [26], a similar self-consistent COHSEX approach is used, with the important difference that they work in the DFT Kohn-Sham orbital basis. In particular, they construct the offdiagonal

matrix elements of the static-COHSEX operator only for valence band and low-energy conduction band states (a small fraction of the DFT orbitals needed to construct Σ_{COH}). This restricts the freedom that the quasiparticle wavefunctions have. We avoid this problem by working directly in plane-wave basis to construct and diagonalize the sc-COHSEX Hamiltonian operator. Using this complete basis removes any bias on the low-energy DFT orbitals. The QPscGW approach [41] does not make use of static-COHSEX approximation. It seeks a mean field that gives eigenvalues closest to the quasiparticle energies iteratively. However, the QPscGW approach also suffers from the same problem of working in a restricted basis. In this case, the restricted basis is required due to the extremely high computational costs of constructing the Σ matrix that includes some dynamical effects, because one must sum over a large number of empty states as well as integrate over frequencies when constructing each matrix element of Σ. Additionally, this method (as well as the sc-COHSEX+GW methods described above) tends to over-estimate band gaps because the gap in the self-consistent mean-field used to construct the polarizability is higher than the optical gap of the system. It is well known [61, 12] that self-consistency in the polarizability cancels vertex (or excitonic) effects, and so including only self-consistency without higher order corrections leads to larger gaps. The static-offdiagonal GW approach, on the other hand, does not suffer from this problem. In the static-offdiagonal GW approach, we continue to use the DFT (LDA or GGA) polarizability and Σ, but include the important off-diagonal effects in the quasiparticle wavefunctions in the static approximation.

An illustrative example for our methods is the silane molecule. It has been shown [115, 53] that the LUMO level is below the vacuum level in DFT – but the correct quasiparticle LUMO level is above the vacuum level. This leads to a qualitative difference in the DFT and quasiparticle wavefuntion – the Kohn-Sham wavefunction is too localized, while the quasiparticle one mixes with continuum states and is a resonant state.

Table 3.1 shows the calculated HOMO and LUMO energies from different methods and experiment. In particular, with the traditional diagonal only G_0W_0 method, the quasiparticle LUMO levels range from 0.63 – 1.1 eV. In the full-Σ approaches of [115, 51] and [53] the LUMO is found to be nearly 1 eV lower than the respective diagonal G_0W_0 energies and in much better agreement with the quantum monte carlo (QMC) results. The results with our new methods for the LUMO level agrees well with the QMC result.

Our DFT calculations were performed using plane waves and pseudopotentials as implemented in PARATEC [3]. We expanded the wavefunctions in plane waves up to an energy cutoff of 75 Ry. We used the Γ point sampling of the Brillouin zone and spherical truncation of the coulomb interaction to avoid silane-silane interactions. For the GW calculations, we used the BerkeleyGW [1] package. We used a dielectric matrix energy cutoff of 6 Ry. The dynamical contributions to the self-energy were treated within a generalized plasmon pole model [63, 62]. We performed all calculations at two volumes in a simple cubic lattice corresponding to lattice constants of 22.5 au and 25 au. All the results presented were extrapolated to infinite volume limit.

The HOMO and LUMO charge distributions within LDA and within our sc-COHSEX and static-offdiagonal GW approaches are plotted in Figure 3.6. While the HOMO wavefunctions do not change between all three methods, the LUMO quasiparticle wavefunctions are much more delocalized in sc-COHSEX and static-offdiagonal GW.

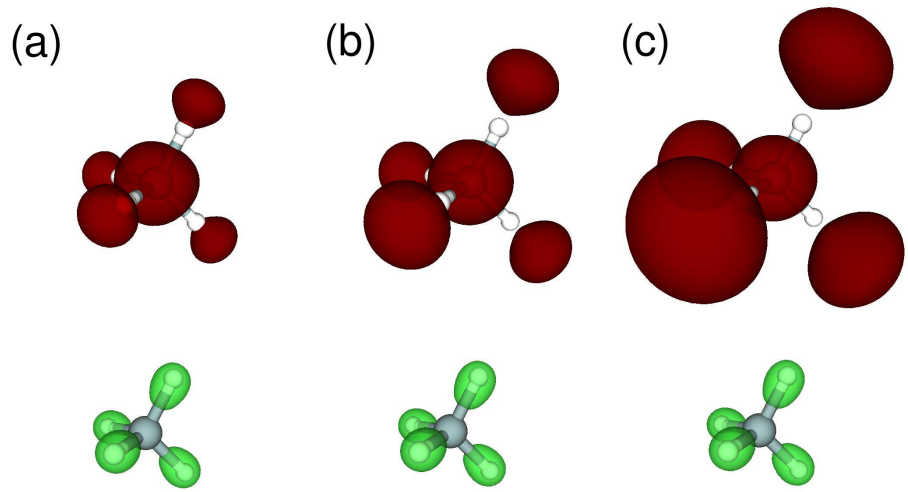

Figure 3.6: HOMO (bottom) and LUMO (top) quasiparticle wavefunction of the silane molecule within (a) LDA+GW, (b) static-offdiagonal GW and (c) sc-COHSEX + GW. The plotted quantity is an iso-surface of $|\psi_5(\vec{r})|^2$ for the LUMO and $\sum_{n=2,3,4}|\psi_n(\vec{r})|^2$ for the HOMO at iso-value of 1/3 of the maximum for the wavefunction amplitude.

This is consistent with the results shown in table 3.1, that shows the electron affinity within sc-COHSEX+GW and static-offdiagonal-GW approaches to be ~ 0 eV. The ionization potential does not get affected in these approaches.

Figure 3.7 shows the contribution to the second-order perturbation correction to the LUMO energy, E^{QP}_{LUMO}, from $\Sigma_{GW}(E_{LUMO}) - \Sigma_{mf}$ (where mf stands for mean-field) from intermediate states 1 to 32. The LDA mean-field ($\Sigma_{mf} = V_{XC}$) starting point shows large corrections to the quasiparticle energy coming from states 9, 17 and 29. This corresponds to large offdiagonal elements in the Σ matrix illustrating a failure of LDA to correctly describe the LUMO quasiparticle orbital. If one accounts for these second order corrections, the electron affinity becomes close to those from more accurate approaches. However, this comes at an additional cost of evaluating offdiagonal Σ matrix elements. Also seen in Figure 3.7, the contributions in both the sc-COHSEX+GW approach and the static-offdiagonal GW approach are small. This shows that, in both new approaches, the off-diagonal elements of Σ are effectively included in the mean-field starting point (sc-COHSEX) or treated adequately within the static approximation (static-offdiagonal approach). In other words, this means that the quasiparticle wavefunctions are well described by the static-offdiagonal and sc-COHSEX wavefunctions respectively. It is worth pointing out that the reason the quasiparticle wavefunctions in Figure 3.6 for both the sc-COHSEX and static-offdiagonal approaches are not the same is because the corresponding Σ_{GW} operators are not the same.

Table 3.2 shows the result of application of these approaches to silicon. These calculations were done with a $6 \times 6 \times 6$ k-point sampling of the Brillouin zone, 35 Ry cutoff for the wavefunctions and 12 Ry cutoff for the dielectric matrix. The generalized plasmon

Starting	HOMO		LUMO	
Mean-Field	mf	mf+G_0W_0	mf	mf+G_0W_0
LDA	–8.53	–12.49	–0.46	0.79
LDA [51]	–8.4	–12.7	–0.6	1.1
LDA [53]	–8.42	–12.67	–0.50	0.63
Full-Σ [53]	—	–12.66	—	–0.42
Full-Σ [51]	—	–12.7	—	0.3
static-offdiagonal GW	–14.49	–12.50	–0.02	–0.02
sc-COHSEX+GW	–14.13	–12.86	–0.02	0.00
QMC [51]	—	–12.6	—	0.2
Experiment [67]	—	–12.6	—	—

Table 3.1: HOMO and LUMO quasiparticle energies calculated in the present and previous approaches. All values are in eV.

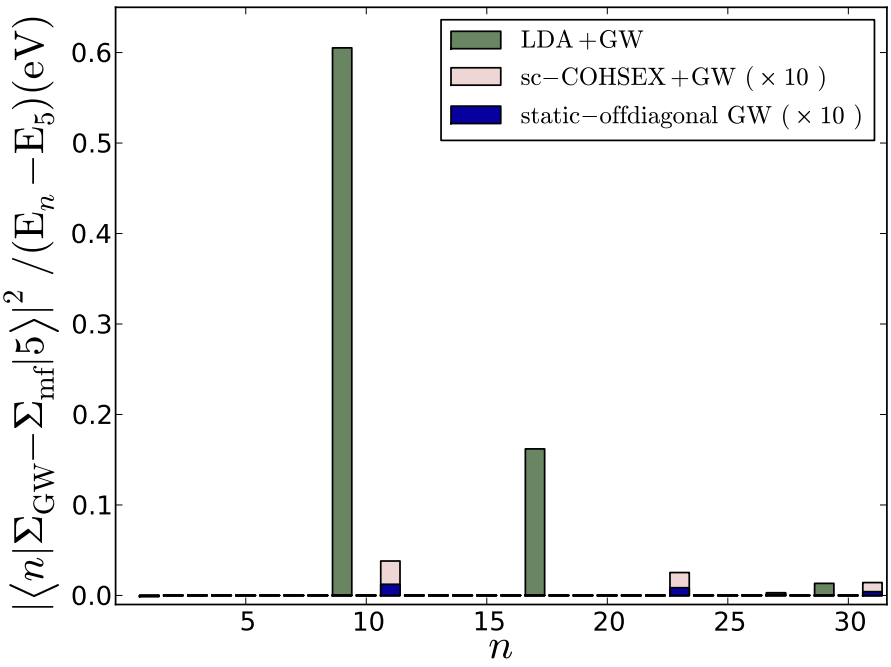

Figure 3.7: Contributions to the second order perturbation correction in the quasiparticle energy of the LUMO, state 5, in silane within the LDA+GW, sc-COHSEX+GW and static-offdiagonal GW approaches. As indicated in the legend, the corrections in the latter two approaches are multiplied by a factor of 10 for clarity.

	Direct Gap		Band Gap	
Mean-Field	mf	mf+G_0W_0	mf	mf+G_0W_0
LDA	2.56	3.29	0.53	1.29
LDA [62]	2.57	3.35	0.52	1.29
LDA [26]	2.57	3.20	0.51	1.14
sc-COHSEX+GW [26]	—	3.69	—	1.56
static-offdiagonal GW	3.79	3.32	1.82	1.29
sc-COHSEX+GW	3.74	3.69	1.72	1.63
Experiment [78]	—	3.40	—	1.17

Table 3.2: Direct gap at Γ and indirect band gap for silicon calculated within various approximations. All values are in eV.

pole model [63, 62] was used to extend the static dielectric matrix to finite frequencies. Table 3.2 shows our calculated values of the direct and indirect band gaps in silicon. As can be seen in the table the static-offdiagonal GW approach gives the same gaps as previous calculations using the diagonal Σ approximation within the Kohn-Sham basis. [63] The sc-COHSEX [26] approach tends to overestimate the gaps slightly due to the aforementioned reasons.

In summary, we presented two approaches for going beyond the diagonal Σ constructed within G_0W_0 and the DFT mean-field. The sc-COHSEX+GW approach, can be viewed as a diagonal G_0W_0 approach with an improved mean-field starting point where the offdiagonal matrix elements of $\Sigma - \Sigma_{\mathrm{mf}}$ are small. The static-offdiagonal-GW approach does not change the mean-field starting point of a typical DFT+GW calculation but constructs and diagonalizes the $\Sigma - V_{\mathrm{xc}}$ in the static approximation. The latter approach is significantly less computationally expensive than the former. We showed that both methods give good quasiparticle wavefunctions and energies for the electron affinity of silane and that with both approaches the offdiagonal elements of Σ are small. In silicon, static-offdiagonal GW gives band gaps in good agreement with experiment, while sc-COHSEX+GW slightly overestimates them as in other self-consistent GW methods.

Combined with the modified static-remainder method of the previous section and the advancements in the GW-BSE methodology presented in the previous chapter, the methods here provide a robust system for calculating the excited electronic and optical properties of large isolated molecules and other nanostuctures.

Chapter 4

Semiconducting single-walled carbon nanotubes

Carbon nanotubes are sp2-bonded tubular structures with a diameter on the order of a nanometer but with length which can be up to centimeters long. These structures possess unique structural and electronic properties. [117] Single-walled carbon nanotubes (SWCNTs) can be metals or semiconductors depending sensitively on their geometric structure, which is indexed by a pair of numbers (m, n) where m and n are the two integers specifying the circumferential vector in units of the two primitive translation vectors of graphene. [118, 91, 54, 9, 127, 128] Because of the reduced dimensionality of carbon nanotubes, many-electron (both self-energy and excitonic) effects are shown to be extraordinarily important in the optical properties of these systems. [127, 37] The strong excitonic effects (the result of the correlation between an excited quasi-electron in a conduction state and the quasi-hole in the valence state) in their optical properties have been predicted by first-principles theory [127, 37] and have subsequently been confirmed by experiment. [146, 90] Other experimental advances have allowed the measurement and characterization of the excitation features in the optical response of individual isolated SWCNTs. [80, 81, 39, 145] In this chapter, we discuss several novel theoretical results that have been obtained from first-principles calculations on the SWCNTs and related 1D nanostructures. Because of the importance of the electron-electron interactions, an accurate description of the quasiparticle and optical properties of nanotubes and other quasi-one dimensional materials requires the use of many-body perturbation theory methods which accurately account for the electron self energy and the electron-hole interaction in optically excited states. The first-principles GW-Bethe-Salpeter equation (GW-BSE) methodology [62, 115, 127] has proven to be the ideal tool in describing these properties of nanostructures.

The basic electronic properties of SWCNTs may be understood from the application of a simple band-folding technique of the graphene bandstructure. The graphene π electron bandstructure, derived within tight-binding (with the overlap matrix set to zero) [117], is:

$$E(k_x, k_y) = \pm t\{1 + 4\cos\left(\frac{3^{1/2}k_x a}{2}\right)\cos\left(\frac{k_y a}{2}\right) + 4\cos^2\left(\frac{k_y a}{2}\right)\}^{1/2}, \qquad (4.1)$$

where t is a hopping parameter and a is the lattice constant. This band structure is shown in Fig. 4.1, and is characterized by two Dirac points in the Brillouin zone at the Fermi energy where the bands take a conical structure. When a graphene sheet is rolled into a tube forming a SWCNT, additional periodic boundary conditions arise in the circumferential direction. Imposing these boundary conditions restricts the k_x and k_y values to parallel lines within the graphene phase space, as is shown in Fig. 4.1. Thus, the band structure of a SWCNT can be thought of, in this simplified tight-binding picture, as derived from various cross sections of the 2D graphene bandstructure that are consistent with the periodic boundary conditions. Additionally, depending on the rolling angle (quantified by chiral indices (m,n) [117]) and diameter, the Dirac point may or may not be included in one of the cross-sections. Thus, both semiconducting and metallic nanotubes are possible. In this chapter, we will consider the electronic and optical properties of semiconducting tubes. We discuss the properties of metallic tubes in the next chapter.

The bandstructure of the (14,0) semiconducting nanotube is shown in Fig. 4.2.

4.1 Ab Initio Results and Two-Photon Experiments

The first *ab initio* GW-Bethe-Salpeter equation calculations on SWCNTs were performed in 2004 by Spataru et. al. [127, 128]. They found remarkable results: that the electron-hole interaction qualitatively changed the nature of the optically excited states in SWCNTs. When the electron-hole interaction effects are included, the spectrum is dominated by bound and resonant exciton states. With the electron-hole interactions included, each of the optically allowed subband transitions (derived from a van Hove singularity in the non-interacting joint density of states) gives rise to a series of exciton states, labeled $1A_2$, $2A_1$, $3A_2$, $4A_1$ etc.. everywhere in this chapter for simplicity - see Fig. 4.4. The exciton states are labeled as $n\Gamma$, where $n-1$ is the number of nodes in the envelope function and Γ labels its irreducible representation in the "group of the wave vector" formalism [18]. Therefore, the $1A_2$ exciton refers to the lowest one-photon-bright exciton with a nodeless envelope function (sometimes also labeled 1u, 1s or ${}_0A_0^-$) and the $2A_1$ exciton refers to the two-photon-bright exciton with a one-node envelope function (sometimes also labeled 2g, 2p or ${}_0A_0^+$ in the literature). This notation is formally used only for chiral tubes. For zigzag tubes, the proper notation would be $1A_{2u}$ and $2A_{1g}$. For simplicity, everywhere in this chapter we use the chiral notation. Unless otherwise noted, we are always referring to excitons associated with the lowest optically allowed, E_{11}, interband transition. Due to optical selection rules only the $1A_2$, $3A_2$ ... states (i.e. states with even electron-hole envelope functions) are bright under single-photon spectra. The even n states are bright under two-photon spectroscopy. For the (8,0) tube, the lowest-energy bound exciton has a binding energy of nearly 1 eV. [127] Note that the exciton binding energy for bulk semiconductors of similar size bandgap is in general only of the order of tens of meVs, illustrating the dominance of many-electron Coulomb interaction effects in reduced dimensional systems.

More recently, we have repeated the *ab initio* GW-BSE calculations on many nanotubes between (7,0)-(20,0). We use DFT within the LDA and a plane wave basis set as our mean-field starting point. We use a 60 Ry wavefunction planewave cutoff and a 9 Ry dielectric function plane wave cutoff. The Coulomb interaction is truncated using the wire

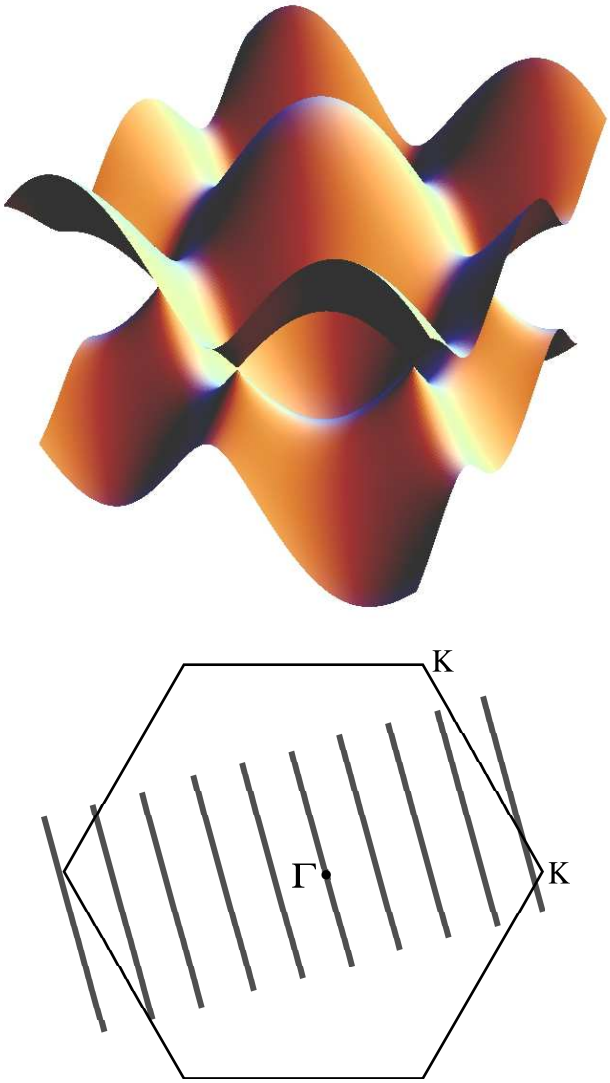

Figure 4.1: (Top) Graphene tight-binding bandstructure in the 2D plane, $E(k_x, k_y)$ plotted in arbitrary units. The points at which the top band touch the bottom band are called the Dirac points owing to the conical dispersion relation present near these points. (Bottom) First Brillouin zone in graphene and schematic of nanotube cutting lines - corresponding to cross-sections of the graphene bandstructure consistent with the nanotube periodic boundary conditions. The tube axis points along the direction of the cutting lines.

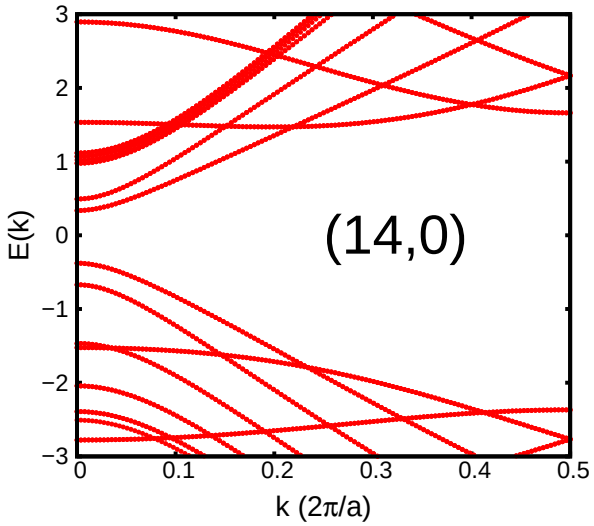

Figure 4.2: LDA bandstructure of the (14,0) SWCNT.

geometry method of Ismail-Beigi. [66] Fig. 4.3 shows the optical absorption spectra for the (14,0) nanotube. Notice the dramatic difference in both position and line-shape between the calculations, including and neglecting exctonic effects. The many-body effects are crucial for even a qualitative understanding of the excited state properties. This can be seen as a consequence of two general principles: first, the Coulomb interaction is more effective in lower dimensions, and, second, screening has a unique nature in two or less dimensional systems.

The prediction of bound exciton states that qualitatively change the absorption spectra of SWCNTs was first met with a great deal of skepticism because bound exciton states in most bulk semiconductors do not significantly affect the optical response of large frequency ranges. However, experimentalists using two-photon absorption techniques [146, 90] were able to verify the predicted excitonic picture in SWCNTs.

In the two-photon optical experiments of Wang *et al.* [146], the binding energy of the lowest exciton in a given interband transition is not directly measured. What is measured is the energy difference between the excitation energy of the lowest even envelope function exciton state, E_{1A_2}, and of the first odd envelope function exciton state, E_{2A_1}. See. Fig. 4.4. The fact that there is an energy difference of 100's of meV between these states confirms the excitonic picture shown in Fig. 4.4. Without bound exciton states, no energy difference between the 1-photon and 2-photon absorption onset would be predicted; a large energy difference only occurs when the spectra is characterized by discrete states of specific symmetry with well separated energy levels.

Using the energy difference between E_{1A_2} and E_{2A_1}, the binding energy of the lowest energy state, $E^{bind}_{1A_2}$, was then extrapolated from the measured value of $(E_{2A_1} - E_{1A_2})$. This was done by fitting this energy difference to the energy difference between the two lowest quantum states obtained from the following 1D hydrogenic-like potential by varying

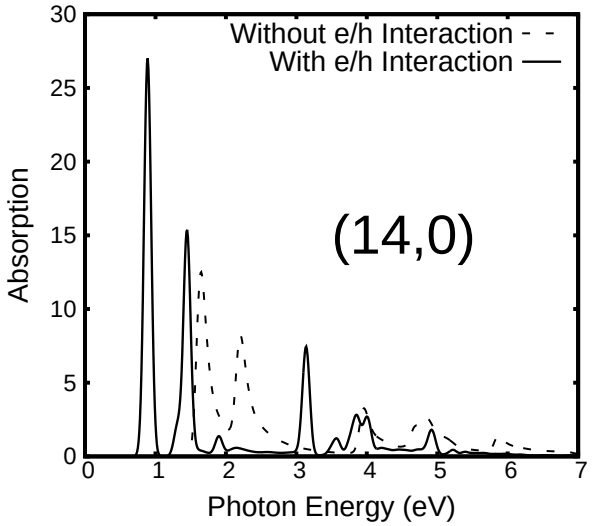

Figure 4.3: The calculated optical absorption spectra for the (14,0) SWCNT with (solid) and without (dashed) the electron-hole interaction included.

the parameter ϵ:

$$V_h(z) = \frac{-e^2}{(|z| + z_0) \cdot \epsilon} \tag{4.2}$$

where, $z_0 = 0.3d$, is a parameter approximating the diameter, d, dependence of the bare Coulomb interaction. Using this approach, one obtains a relationship for the binding energy of $E_{1A_2}^{bind} \approx 1.4 \cdot (E_{2A_1} - E_{1A_2})$. A mystery arose from this result. The value of the binding energy for the tubes measured in experiment ((7,5),(6,5) and (8,3)) were significantly lower than those predicted by calculations using the *ab initio* GW-BSE methodology [28, 127] (see Table 4.1). The *ab initio* GW-BSE predictions of the exciton binding energy in the (8,0), (10,0) and (11,0) nanotubes, for example, obey the following relationship $E_{1A_2}^{bind} \approx 2.5 \cdot (E_{2A_1} - E_{1A_2})$. The discrepancy between this relationship and the one derived from the hydrogenic model suggests an inadequacy of the hyrdogenic model interaction given in Eq. 4.2. See Fig. 4.5 for an illustration of this discrepancy.

In order to resolve the discrepancy between the binding energies obtained from the hydrogenic interaction and those obtained from the full GW-BSE methodology, we evaluate the validity and form of the model interaction. The biggest approximation in the model electron-hole interaction, Eq. 4.2, is the use of a spatially independent dielectric constant ϵ. Such an approximation may be appropriate for the study of excitons in bulk semiconductors, but is untested in one-dimensional systems. In order to evaluate the validity of this model, we relax the spatially independent screening approximation and seek a more physical electron-hole interaction.

Starting with the Bethe-Salpeter equation [115], we show in the next sections that, within an effective mass approximation, the exciton binding energy E_{ex} and envelope

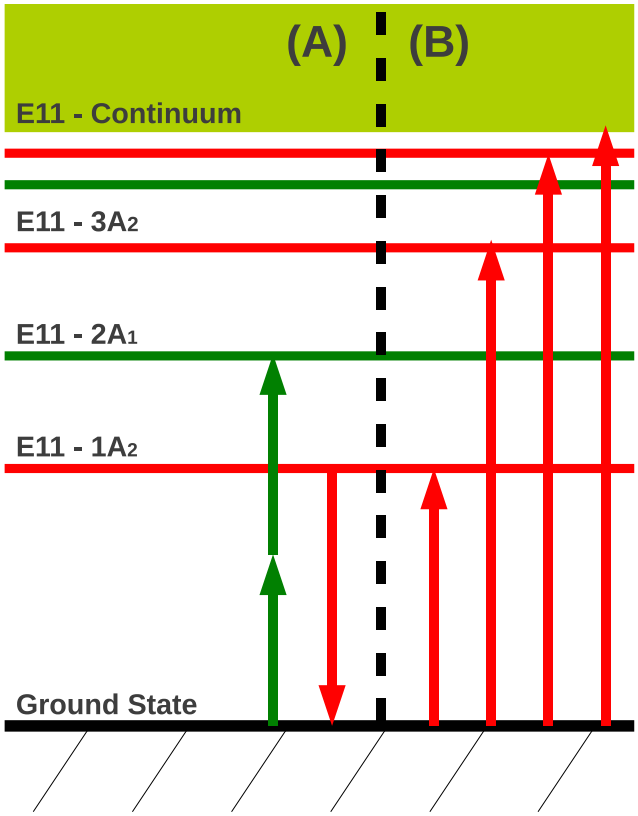

Figure 4.4: Diagram of the optically excited states of a SWCNT. (A) The two-photon luminescence spectra. The system is excited into the $2A_1$ excited states by two photon absorption and emits a single photon from the $1A_2$ state after losing energy due to scattering events. (B) The optically allowed single-photon transitions. Here E_{11} refers to the transition between the first valence and conduction subband pair that is optical allowed.

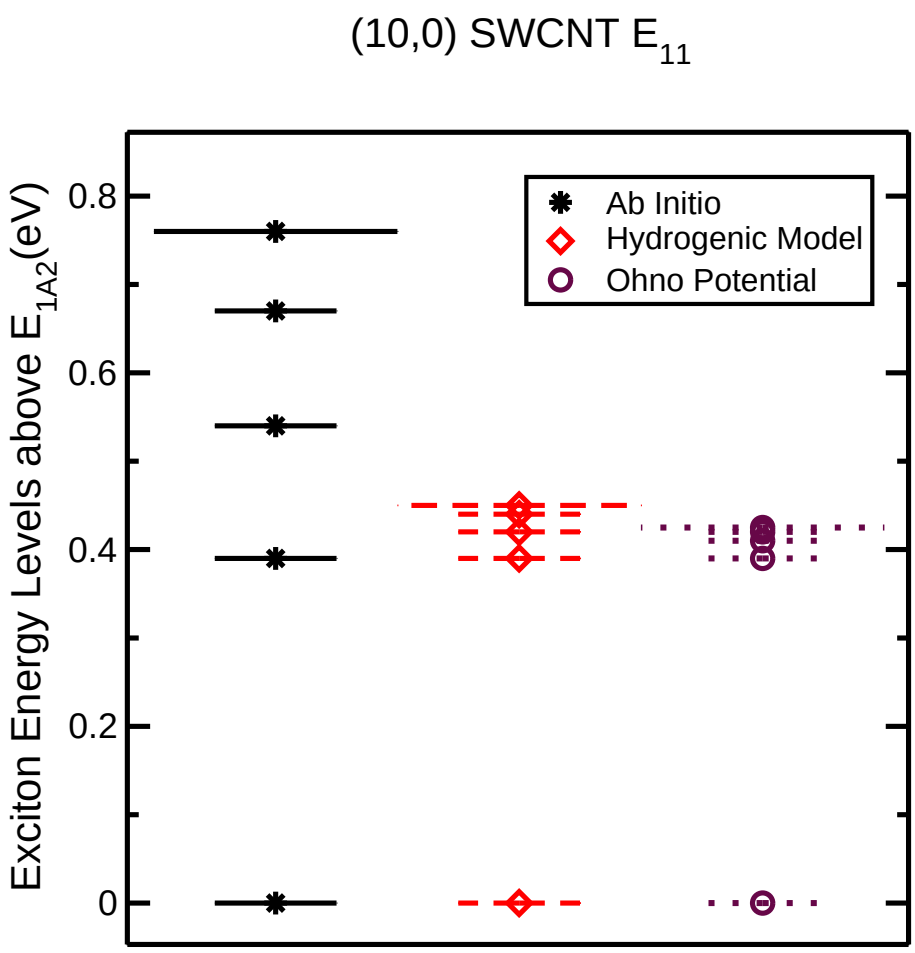

Figure 4.5: Excitation spectra predicted by various model electron-hole interactions for the (10,0) SWCNT. The $E_{2A1} - E_{1A2}$ energy has been fit in each case. The black stars represent the ab initio result with spatial dependent screening, whereas the hydrogenic and Ohno potentials have a constant dielectric screening.

function $F(z)$, for a given state, satisfy

$$\left[-\frac{\hbar^2}{2m^*}\frac{\partial^2}{\partial z^2} - V_{dir}(z) + J\delta(z)\right] F(z) = E_{ex}F(z) \tag{4.3}$$

where, J is the exchange between the valence and conduction bands and

$$V_{dir}(z) = \int dx'dy'd\mathbf{r}_2 W(\mathbf{r}' + \mathbf{r}_2, \mathbf{r}_2)\rho_c(\mathbf{r}' + \mathbf{r}_2)\rho_v(\mathbf{r}_2) \tag{4.4}$$

is the screened direct interaction, where, $\rho_c = |\psi_c(r)|^2$ and $\rho_v = |\psi_v(r)|^2$, and W is the screened Coulomb interaction. This effective interaction, V_{dir}, physically corresponds to a density weighted average of W over the electron-hole individual positions perpendicular to the tube axis (x', y') and the relative position of the center mass throughout one unit cell. We derive this interaction in the following section.

4.2 From the Bethe-Salpeter Equation to an Effective Mass Equation

We first show how to arrive at an effective mass equation for the electron-hole interaction from the Bethe-Salpeter Equation by making certain approximations. This is useful because it extracts the essential physics from the BSE necessary for the correct effective 1D electron-hole interaction potential. The full Bethe-Salpeter equation is:

$$\left[E_{cvk} - E_S\right] A^S_{cvk} + \sum_{c'v'k'} \langle\phi_{cvk}|K|\phi_{c'v'k'}\rangle A^S_{c'v'k'} = 0, \tag{4.5}$$

where capital K is the electron-hole kernel including both the direct term and the exchange term, A^S is the exciton eigenvector such that, in real space:

$$\Psi_S(\mathbf{r}_e, \mathbf{r}_h) = \sum_{cvk} A^S_{cvk}\phi_{ck}(\mathbf{r}_e)\phi_{vk}(\mathbf{r}_h) \tag{4.6}$$

and $E_{cvk} = E_c(k) - E_v(k)$ is the quasiparticle energy difference between the conduction and valence bands. Lets first discuss only the direct interaction part of K. This term can be expressed as:

$$< dir >= \int\int d\mathbf{r}_1 d\mathbf{r}_2 W(\mathbf{r}_1, \mathbf{r}_2) e^{-i(\mathbf{k}-\mathbf{k}')\cdot(\mathbf{r}_1-\mathbf{r}_2)} u^*_{ck}(\mathbf{r}_1)u_{ck'}(\mathbf{r}_1)u^*_{vk}(\mathbf{r}_2)u_{vk'}(\mathbf{r}_2). \tag{4.7}$$

where u is the periodic part of the corresponding Bloch function and W is the screened Coulomb interaction. We now make a couple of approximations. First, we assume we are dealing only with two bands (one conduction and one valence band). We also assume that the excitons are reasonably weakly bound - i.e. that A_k is sharply peaked in in k-space. We assume because of this that the u's do not change appreciably in this region. So, the term involving the four u's can be replaced by:

$$\rho_c(\mathbf{r}_1)\rho_v(\mathbf{r}_2) = |u_{ck_0}(\mathbf{r}_1)|^2|u_{vk_0}(\mathbf{r}_2)|^2 \tag{4.8}$$

where k_0 is the location of the band minimum. We can now write the direct term as:

$$< dir > \approx \int\int d\mathbf{r}_1 d\mathbf{r}_2 \rho_c(\mathbf{r}_1)\rho_v(\mathbf{r}_2)W(\mathbf{r}_1,\mathbf{r}_2)e^{-i(\mathbf{k}-\mathbf{k}')\cdot(\mathbf{r}_1-\mathbf{r}_2)}. \tag{4.9}$$

Now, we define $F(z) = \sum_k e^{ikz}A(k)$, noting that all k's are along the z direction for a nanotube. Thus, we have the following result:

$$\sum_{k'} < dir > A(k') = \int\int d\mathbf{r}_1 d\mathbf{r}_2 \rho_c(\mathbf{r}_1)\rho_v(\mathbf{r}_2)F(z_1-z_2)W(\mathbf{r}_1,\mathbf{r}_2)e^{-ik\cdot(z_1-z_2)}. \tag{4.10}$$

Now, we define $\mathbf{r}' = \mathbf{r}_1 - \mathbf{r}_2$. Noting that the Jacobian of this tranformation is 1, we can rewrite Eq. 4.10 as:

$$\sum_{k'} < dir > A(k') = \int dz' \left[\int dx'dy'd\mathbf{r}_2 \rho_c(\mathbf{r}'+\mathbf{r}_2)\rho_v(\mathbf{r}_2)W(\mathbf{r}'+\mathbf{r}_2,\mathbf{r}_2)\right] F(z')e^{-ikz'}. \tag{4.11}$$

This term is the direct part of the second term on the left hand side of Eq. 4.5; we will include the exchange below. We now multiply the entirety of Eq. 4.5 by e^{ikz} and sum over k to obtain the following effective mass equation:

$$\left[E_{cv}(-i\frac{\partial}{\partial z}) - V_{dir}(z)\right] F(z) = E_{ex}F(z) \tag{4.12}$$

where,

$$V_{dir}(z) = \int dx'dy'd\mathbf{r}_2 W(\mathbf{r}'+\mathbf{r}_2,\mathbf{r}_2)\rho_c(\mathbf{r}'+\mathbf{r}_2)\rho_v(\mathbf{r}_2). \tag{4.13}$$

and $E_{cv}(-i\frac{\partial}{\partial z})[F(z)] \approx \frac{1}{2m^*}\frac{\partial^2 F(z)}{\partial z^2}$ where m^* is the effective reduced mass of the valence and conduction bands. This is the equation and potential that we were searching for. We see that the effective 1D potential is an averaged potential of the electron-hole screened interaction over the 5 remaining free coordinates weighted by the electron and hole charge density. This function is plotted in Fig. 4.6 using the *ab initio* W and ρ from a GW-BSE calculation. In the next section, we discuss approximate models for this potential based on approximations for both W and ρ.

To include the exchange term in the Kernel, we start again with the full interaction:

$$< exchange > = \int\int d\mathbf{r}_1 d\mathbf{r}_2 \frac{-1}{|\mathbf{r}_1-\mathbf{r}_2|} u^*_{ck}(\mathbf{r}_1)u^*_{vk'}(\mathbf{r}_2)u_{vk}(\mathbf{r}_1)u_{ck'}(\mathbf{r}_2). \tag{4.14}$$

We again assume two bands and weak binding so that the u's can be replaced by u_{k_0}. Following the same procedure as above for the direct term yields an exchange potential of the form:

$$V_{ex}(z) = J\delta(z) \tag{4.15}$$

with

$$J = \int\int d\mathbf{r}_1 d\mathbf{r}_2 \frac{-1}{|\mathbf{r}_1-\mathbf{r}_2|} u^*_{ck_0}(\mathbf{r}_1)u^*_{vk_0}(\mathbf{r}_2)u_{vk_0}(\mathbf{r}_1)u_{ck_0}(\mathbf{r}_2). \tag{4.16}$$

Thus, the complete effective mass equation is:

$$\left[E_{cv}(-i\frac{\partial}{\partial z}) - V_{dir}(z) + J\delta(z)\right] F(z) = E_{ex}F(z) \tag{4.17}$$

where V_{dir} is given by Eq. 4.13.

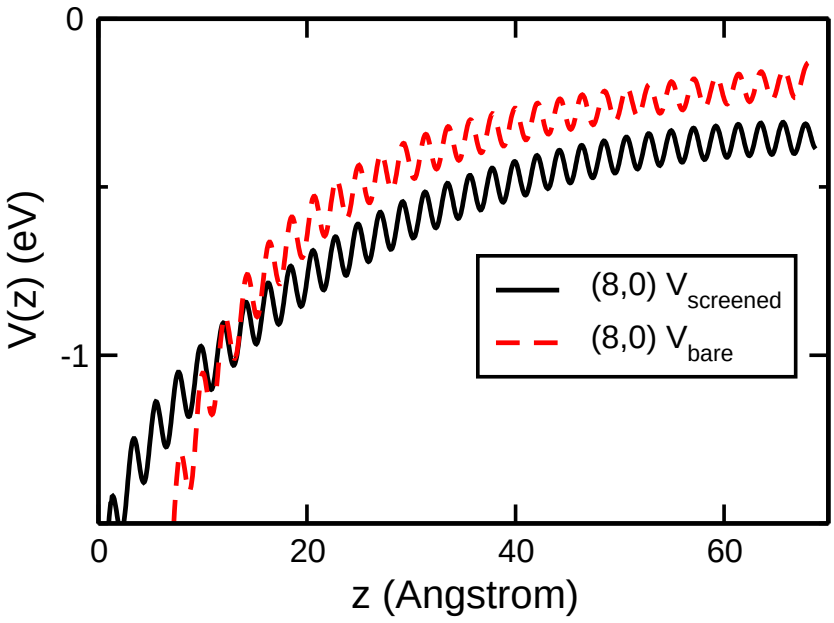

Figure 4.6: The direct interaction of Eq. 4.4 computed for the (8,0) SWCNT using the *ab initio* charge density and dielectric function from a GW-BSE calculation. The bare interaction is the same quantity with the dielectric matrix everywhere set to unity.

4.3 Model Interaction

What one would like in general is an analytic model for the 1D interaction between an excited electron and hole in nanotubes for various chiralities and diameters. Or, more gernerally, one would like a prescription for creating quickly such a model given the parameters of the nanotube. There are many physical approximations that could be made to simplify W and ρ in Eq. 4.4. We start by approximating the interaction between the two particles as the interaction between two rings of charge of plus and minus unity, noting that the charge density in Eq. 4.4 is localized very near the tube diameter (within one π-orbital) for real nanotube systems. Thus, we start by deriving the effective Coulomb interaction between two rings of charge confined to the surface of a cylinder.

4.3.1 Bare Interaction

The problem amounts to solving the following Poisson equation in the presence of rings of charge:

$$-\nabla^2\phi\left(\mathbf{r}\right) = \frac{4\pi\rho(z)}{2\pi R}\delta(s-R) \tag{4.18}$$

where $\rho(z)$ is the charge density per unit length along the z direction of the cylinder (i.e. the ring charge density) and R is the radius of the tube. In this equation, we will make use of cylindrical coordinates: (z, s, θ). We transform the equation to Fourier space in the

z direction with the following definitions: $\phi(\mathbf{r}) = \sum_q e^{iqz} F_q(s)$ and $\rho(z) = \sum_q e^{iqz}\rho(q)$. The equation for the radial function, F, now becomes:

$$\frac{1}{s}F_q' + F_q'' - q^2 F_q = \frac{2\rho(q)}{s}\delta(s-R) \tag{4.19}$$

where the derivative is with respect the radial coordinate, s. We are interested mainly in the solution to this equation at $s = R$ which corresponds to a point on another ring separated by some distance z. This solution, which can be found in the textbook of Arfken and Weber [11], is:

$$F_q(s=R) = 2\rho(q) I_0(qR) K_0(qR) \tag{4.20}$$

where I_0 and K_0 are the two zeroth modified Bessel functions. Thus, we can define the effective interaction between two rings in this geometry by:

$$v(q) \equiv 2I_0(qR)K_0(qR). \tag{4.21}$$

We can now write the 1D potential for a distribution of ring charges as $V(q) = v(q)\rho(q)$. The function v has the following real space form:

$$v(z) = \frac{-\frac{2}{\pi}K\left(\frac{d^2}{z^2}\right)}{|z|} \tag{4.22}$$

where d is the tube diameter and K is an elliptic function of the first kind. This represents the bare electrostatic interaction between two rings of charge unity with equal diameters, d. The interaction diverges as $z \to 0$ due to the fact, as we will see later, that the rings have zero thickness. However, the integral of the interaction, $\int_0^\epsilon v(z)$ remains finite, unlike the function $\frac{1}{z}$.

4.3.2 1D Dielectric Screening

Simply applying the bare interaction to the effective mass theory of excitons yields poor results for binding energies and wavefunctions. As in bulk systems, it is important to describe the dielectric screening of the medium. However, we will see that in 1D, the dielectric screening has very interesting properties.

First, note that in the previous section, we introduced the well-defined one dimensional quantities $v(q)$ and $\rho(q)$. Similarly we define the total potential $V(q)$ as the potential felt by a test *ring*-charge of charge unity in the system and the external potential $V_{ext}(q)$ as the potential function a test *ring*-charge would feel from charges external to the system alone. We next define a 1D dielectric function, $\epsilon(q)$ through the relation:

$$V(q) = \frac{V_{ext}(q)}{\epsilon(q)}. \tag{4.23}$$

Similarly, the induced *ring*-charge density is defined to be related to the total potential by:

$$\rho_{ind}(q) = \chi(q)V(q). \tag{4.24}$$

Using these definitions, we have that:

$$\begin{aligned} V(q) - V_{ext}(q) &= v(q)\left(\rho(q) - \rho_{ext}(q)\right) \\ &= v(q)\chi(q)V(q) \\ \Rightarrow \epsilon(q) &= 1 - \chi(q)v(q). \end{aligned} \tag{4.25}$$

So, putting all the pieces together, we have, similar to 3D, that the total potential from a single ring of charge unity in a dielectric cylinder is:

$$V(q) = \frac{v(q)}{1 - \chi(q)v(q)} \tag{4.26}$$

What is left now is to determine $\chi(q)$ for our cylindrical system.

4.4 Models for $\chi(q)$

4.4.1 Semiconducting Tubes

For all nanotubes, $\chi(q)$ can be expressed in the usual perturbation theory form:

$$\chi(q) = -\sum_{k,c,v} \frac{|\left\langle v,k|e^{iqz}|c,k+q\right\rangle|^2}{E_{c,k+q} - E_{v,k}} \left(f(v,k) - f(c,k+q)\right). \tag{4.27}$$

Our strategy will be to find a functional form for $\chi(q)$ based on physical grounds with adjustable parameters to obtain the electron-hole screened potential for any SWNT with a given diameter and chirality. One possible approach is to generalize the Penn model for dielectric screening [106], developed for 3D semiconductors, and use it for 1D systems [34]. In accordance with the Penn model, we expect that $\chi(q) \sim (q/E_g)^2$, where E_g is the energy gap, when $q \to 0$ and $\chi(q) \sim \textit{constant}$ when $q \to \infty$. One way of interpolation between these two limits is

$$\chi(q) = -C_2 \frac{\alpha q^2}{E_g^2 + \alpha q^2} \tag{4.28}$$

where C_2 and α are adjustable parameters. The gap energy E_G can be obtained by summing the first optical transition energy extracted from Ref. [14] and the ground state exciton binding energy calculated in Ref. [28]. In this way, $\chi(q)$ will contain diameter and chirality dependences.

In expression (4.28), the electric susceptibility goes like q^2 for small q, i.e., the electron-hole screening becomes negligible for large distances, which is consistent with the 1D geometry of carbon nanotubes. This quadratic dependence in susceptibility was also obtained by Leonard and Tersoff [82] using a simple tight-binding model.

One can derive the form of Eq. 4.28 using a 1D generalized Penn model. In the typical Penn approximations, the polarizabilty can be written as $|M(q)|^2 E^{Sum}(q) = ((\frac{2qE_w}{\sqrt{3}E_g k_f})^2/(1 + (\frac{2qE_w}{\sqrt{3}E_g k_f})^2)(\sum_{k_{occ},k+q_{unocc}} \frac{1}{E_k - E_{k+q}} + \sum_{k_{occ},k'_{unocc}} \frac{1}{E_k - E_{k'}})$. Where $k' = k + q - 2k_f$, E_w is the band width, $M(q)$ is the matrix element between state k and $k+q$ or k and k' and is roughly independent of k. In ref. [34], the explict gap dependence of

the polarizability was removed in favor of a simpler diameter dependence leaving the usual dominant d^2 scaling of χ, whereas, one might include the explicit gap dependence in order to capture accurate family patterns of exciton binding energies.

4.4.2 Metallic Tubes

In addition to the interband contribution to screening discussed in the above subsection, metallic nanotubes have a large contribution of screening from the intraband transitions (i.e. free electrons). To approximate the contribution of these electrons to screening, we will apply the Thomas-Fermi approximation. In this approximation, the local density of screening electrons is determined by occupying all states up to a local Fermi energy, $\mu - V(z)$ where μ is the chemical potential. That is to say, $\rho_{ind}(z) = [\rho_{TF}(\mu + V(z)) - \rho_{TF}(\mu)]$. So, to first order:

$$\rho_{ind}(z) = \frac{\partial \rho}{\partial \mu} V(z). \tag{4.29}$$

This implies that, for a metal, there is an additional contribution to χ given by $-\frac{\partial \rho}{\partial \mu} = -D(E_F)$, i.e. the density of states per unit length at the Fermi Energy. So, the total response for a metal is:

$$\chi_M(q) \approx -\frac{\alpha q^2}{E_g^{*2} + \alpha q^2} \cdot C_2 - \beta D(E_F) \tag{4.30}$$

where E_g^* is the lowest gap of the system apart from the two metallic like bands. Thus, we have now created a model for both semiconducting and metallic tubes that depends on their diameters, chirality (through E_g) and the parameters C_2 and α.

4.5 Properties of the Dielectric Function

Combining Eq. 4.28, Eq. 4.25 and Eq. 4.21, we see that the dielectric function for rings in semiconducting SWCNTs can be written as:

$$\epsilon(q) = 1 + 2C_2 \frac{\alpha q^2}{E_g^2 + \alpha q^2} \rho(q) I_0(qR) K_0(qR). \tag{4.31}$$

Notice that this function approaches 1 at both small and large q because the small q limit of $v(q)$ is Log(q) like all 1D potentials. In three dimensions, the q^2 term in the numerator of χ is canceled by the $\frac{1}{q^2}$ dependence of $v_{Coul}(q)$ which causes the $q = 0$ component of ϵ to equal a non-unity constant.

We compare $\epsilon(q) = 1 - \chi(q) V_{bare}(q)$ using Eq. 4.31 with the full *ab initio* dielectric function for the (8,0) tube in Fig. 4.7. In the figure we fit the two constants α and C_2 in order to best fit the *ab initio* curve. The dielectric function plotted has interesting characteristics that are not found in bulk semiconductors where a constant ϵ model is appropriate. As just mentioned, it is a unique property of reduced dimensional systems that the dielectric function approaches one in both the limit of large and small q. As was pointed out by Leonard *et al.* [82], this implies that there is no screening at both large and short distances in such systems.

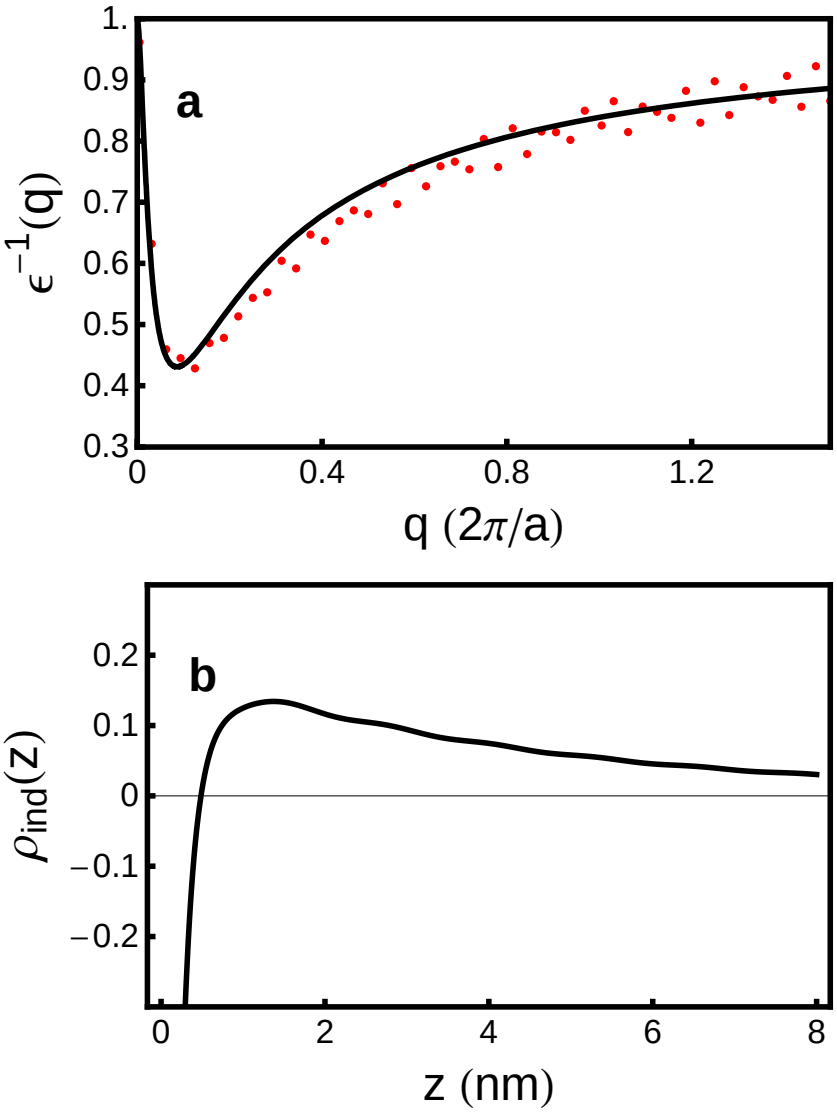

Figure 4.7: Spatially dependent dielectric screening in semiconducting SWCNTs. (a) Comparison between the q_z dependent *ab initio* inverse dielectric function, $\epsilon^{-1}_{G_{xy}=G'_{xy}=0}(q+G_z)$ (points) and the result of the 1D ring Penn model (solid line) of the (8,0) SWCNT derived in the text. The parameters of the model were fit to give the best agreement. (b) The induced ring charge distribution from the Penn model polarizability plotted around an added positive ring charge (at $z = 0$), plotted as a function of z along the tube axis. The total induced charge integrates to zero.

The spatially dependent screening can be demonstrated by the induced charge distributions in 1D vs 3D systems under a perturbation of a single additional charge or *ring*-charge. In 3D, the approximate induced charge is given by:

$$\rho_{ind}^{3D}(q) = \chi(q)V(q) \propto \frac{\frac{q^2}{1+\alpha q^2}\frac{1}{q^2}}{1+\frac{q^2}{1+\alpha q^2}\frac{1}{q^2}} \approx C, \tag{4.32}$$

for some constant C. In other words, for small q (or large distances), the induced charge is a delta function which integrates to give a finite total induced charge. In 1D however, the induced charge is:

$$\rho_{ind}^{1D}(q) = \chi(q)V(q) \propto \frac{\frac{q^2}{1+\alpha q^2}v(q)}{1+\frac{q^2}{1+\alpha q^2}v(q)} \approx q^2 v(q). \tag{4.33}$$

Unlike the 3D case, this function goes to zero as q tends to zero. Hence, the integrated induced charge is zero, which is consistent with having no screening at large distances.

Given the induced charge distribution above, we see that in bulk semiconductors, $\epsilon(q=0)$ is a non trivial constant, meaning that, at large distances, there appears to be a finite induced charge of opposite sign surrounding the external charge. In one-dimensional materials, on the other hand, the induced charge that surrounds the external charged particle actually integrates to zero. $\rho_{ind}(q) = \chi(q)V(q)$, is plotted for the (8,0) tube in real space in Fig. 4.7. Nearby the external charge, screening charges of opposite sign are induced; whereas, further away from the external charge, charges of the same sign are induced. This leads to a counter-intuitive prediction that for some electron-hole separations along the tube axis the electron-hole interaction is enhanced (see Fig. 4.8).

Fig. 4.8 shows the effective interaction in real space for the (8,0) semiconducting SWCNT. In this figure, an interesting phenomena is evident: the screened interaction drops below the bare interaction, $V_{bare}(z)$, in the region where screening charges of the same sign as the added external charge are induced (we call this the anti-screening region). Such "anti-screening" behavior was first deduced by van den Brink and Sawatzky for molecular nanostructures [143, 142] using a simple dipole interaction model. Because the screened Coulomb interaction is very different in 1D, a very distinct excitation spectrum is created for the higher energy exciton states formed from a given interband transition. The actual spectra is qualitatively and quantitatively different from those of excitons in 3D or in 1D models that neglect the anti-screening effect. The higher states in the excitonic series, with large amplitudes in this region, are considerably more bound due to the presence of the anti-screening region; evidences consistent with this prediction have been observed in recent measurements on semiconducting SWCNT. [80, 81] We have now shown that this effect comes out explicitly from the first principles GW/BSE calculations and causes the higher states in the series ($2A_1$, $3A_2$, $4A_1$, . . .) to have binding energies that are a relatively higher fraction of the $1A_2$ binding energy then is the case in a hydrogenic like electron-hole model. Therefore, the physical origin of the failure of Eq. 4.2 in describing the excitonic spectrum of isolated SWCNTs is the lack of the spatially dependent dielectric screening.

A qualitative explanation for the phenomenon is shown in Fig. 4.9 where the polarizable charge distribution of a semiconducting system is modeled by a simple ball-and-spring dielectric medium [34]. In three-dimensions, because the surface area of a spherical

	$E_{2A_1} - E_{1A_2}$		$E^{bind}_{1A_2}$		
	Expt (Measured)	Present Work	Expt. (Extrap.)	Present Work	*Ab Initio* (Extrap.)
(6,4)	0.33[b]	0.41	0.42[b]	0.98	0.91[c]
(6,5)	0.31[a], 0.28[b]	0.38	0.43[a], 0.37[b]	0.85	0.81[c]
(7,5)	0.28[a], 0.23[b]	0.34	0.39[a], 0.31[b]	0.77	0.77[c]
(7,6)	0.20[a]	0.31	0.35[a]	0.70	0.70[c]
(8,3)	0.30[a], 0.29[b]	0.37	0.42[a], 0.38[b]	0.84	0.84[c]
(8,6)	0.25[a]	0.28	0.35[a]	0.64	0.66[c]
(8,7)	0.20[a]	0.26	0.29[a]	0.60	0.61[c]
(9,1)	0.32[b]	0.38	0.42[b]	0.88	0.87[c]
(9,4)	0.24[a], 0.27[b]	0.30	0.34[a], 0.38[b]	0.69	0.71[c]
(9,5)	0.23[a]	0.28	0.33[a]	0.62	0.62[c]
(9,7)	0.22[a]	0.24	0.30[a]	0.55	0.58[c]
(10,2)	0.24[a]	0.31	0.34[a]	0.73	0.75[c]
(11,3)	0.22[a]	0.27	0.31[a]	0.62	0.65[c]
(11,6)	0.19[a]	0.21	0.27[a]	0.51	0.55[c]
(12,4)	0.20[a]	0.21	0.27[a]	0.53	0.58[c]

[a] From [146, 39] [b] From [90] [c] From [28]

Table 4.1: Comparison of experimentally measured and theoretically predicted values for the E_{11} exciton excitation energy difference, $E_{2A_1} - E_{1A_2}$, and the lowest exciton binding energy $E^{bind}_{1A_2}$ (in eV).

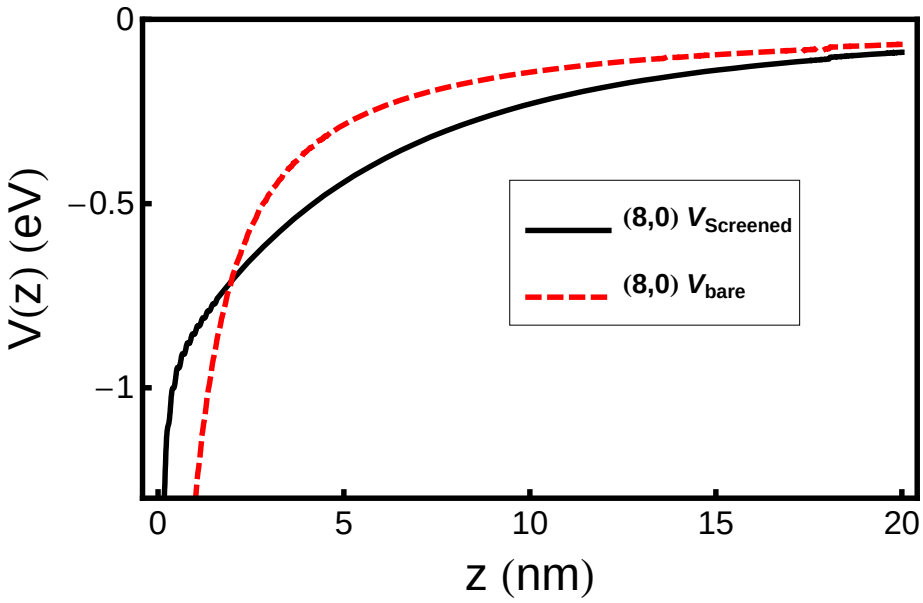

Figure 4.8: Model 1D electron-hole interaction potentials. Comparison of the Penn model screened interaction for the (8,0) zigzag tube with the bare interaction between two ring charges. There is a region in which the screened interaction becomes stronger than the bare interaction.

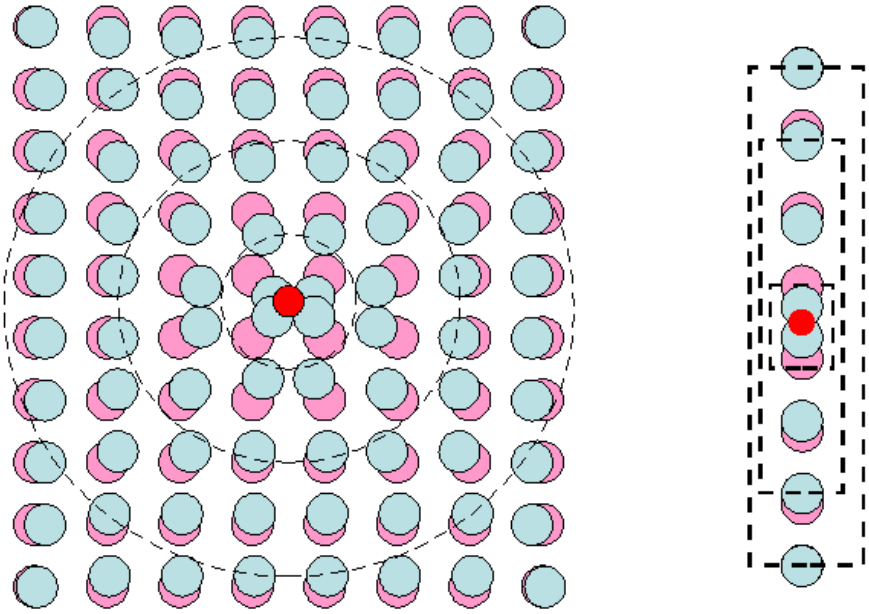

Figure 4.9: Schematic of the different screening behaviors in 3D/2D vs 1D. A positive charge (red circle) is added to the system at the origin. The screening electrons are bound to the nuclei via a spring. In 3D, the amount of charge that has crossed the surface of a spherical shell of radius r is constant with respect to r. In 1D, for pillboxes of length z, the amount of charge to cross into the pillbox is large at small z but goes to zero as $z \to \infty$.

shell is proportional to the radius squared and the force generated by a charge at the origin on a spring at distance r is proportional to $1/r^2$, the total induced charge in a shell of radius r is constant with respect to r. So, at large distances from the external charge, there is a net induced charge of the opposite sign observed surrounding the external charge. In quasi one dimension, however, the surface area perpendicular to the tube axis of a box does not change as the box length changes in the z direction. The total induced charge in larger and larger size boxes drops to zero; so at large length scales, there is effectively no screening. [34]

4.5.1 Model Results

We obtain the parameters α and C_2 in our model by fitting the exciton binding energies to several *ab initio* GW-BSE calculations on the exciton binding energies of first (E_{11}) and second (E_{22}) optically allowed interband transition. In particular, we use GW-BSE calculations on the (8,0), (10,0) and (11,0) tubes. The effective masses used in solving the BSE were taken from an interpolation formula by Jorio *et al.* [70]. Figure 4.10 compares the binding energies from the *ab initio* GW-BSE calculations on these SWCNTs and those predicted in the present model.

Table 4.1 shows the predicted exciton binding energies of the present effective

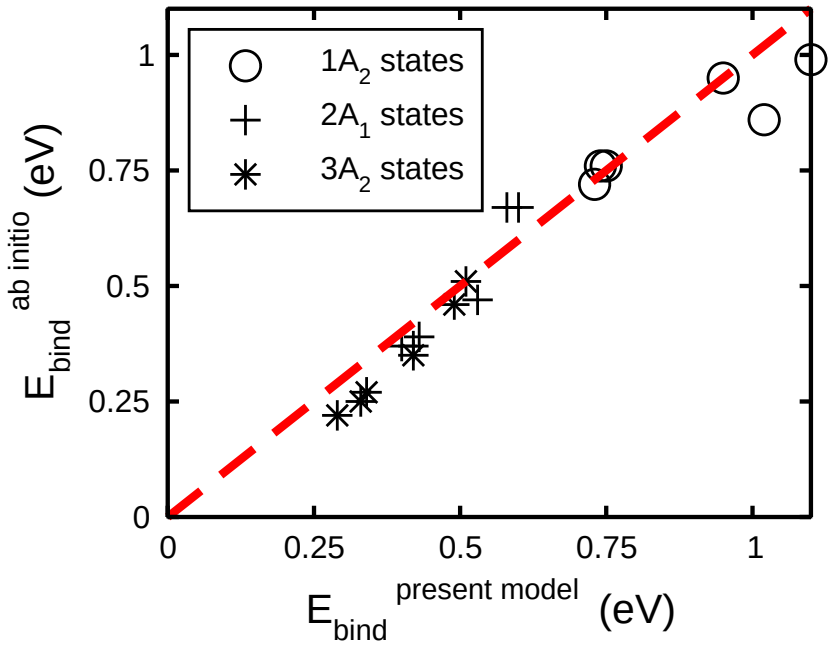

Figure 4.10: Comparison between *ab initio* GW-BSE and the present effective mass model binding energies for the $1A_2$, $2A_1$ and $3A_2$ states associated with the E_{11} and E_{22} interband transitions in the (8,0), (10,0) and (11,0) SWCNTs.

mass model for the tubes measured by Wang *et al.* [146] and Maultzsch *et al.* [90]. The new model results agree to within 0.1 eV with the directly measured $E_{2A_1} - E_{1A_2}$ values. This small differences shows that dielectric environment and the anti-screening effect does not affect this quantity because electron-hole amplitude for the $2A_1$ and $1A_2$ states are concentrated near the origin of Fig. 2 where there is no anti-screening effect. We illustrate this insensitivity directly in the following paragraphs. However, the spatially screened effective mass model disagrees with the extrapolated binding energies reported based on the hydrogenic model. In particular, the exciton binding energies from the spatially screened model does not obey the $E_{1A_2}^{bind} \approx 1.4 \cdot (E_{2A_1} - E_{1A_2})$ rule. With the inclusion of the spatially-dependent dielectric function, we find, in the present model, that $E_{1A_2}^{bind} \approx 2.3 \cdot (E_{2A_1} - E_{1A_2})$ for this range of tubes, showing that the hydrogenic model systematically underestimates the exciton binding energy of isolated SWCNTs.

Recent experiments on isolated SWCNTs have been able to confirm the novel spectra we predict. Shown in Fig. 4.11 are the excitations features assigned to the $3A_1$ and $5A_1$ excitons for tubes measured in the work of Lefebvre *et al.* [80]. The observed features above the E_{1A_2} energy in the experimental spectra lie in the energy range our predicted bright $3A_2$, $5A_2$ and continuum exciton states of the E_{11} transition. Thus, our current model supports the assignment of these features to a higher exciton states. Also shown in Table 4.1 is a comparison of the binding energies predicted in the current model for the $1A_2$ exciton with binding energies extrapolated directly from *ab initio* GW-BSE calculations [28]. Thus the experiments of Lefebvre *et al.* reproduce the main qualitative difference between the spectra obtained from our "anti-screening" model and the spectra obtained from the hydrogenic model: that the present spectra contains discrete energy levels that are much more spread-out in energy.

In a more recent work, Lefebvre et al. [81] have further confirmed our assignment of the higher energy peaks in their spectra to the $3A_1$, $5A_1$ and continuum states. The most convincing piece of evidence is a comparison of the measured excitation spectra before and after a "cleaning process" - achieved by heating the staging area to 450C. After heating, the isolated nanotube is expected to be free of adsorbents present when the tube is exposed to ambient air. The experiments show a blue shift in the excitation peak energies after the cleaning process, where the peaks they assign to the continuum and $5A_2$ transitions blue shift the most and the peak they assign to the $1A_2$ blue shifts the least.

This behavior is strong evidence in support of the assignment of these peaks in the excitation energy as the $1A_2$, $3A_2$, $5A_2$ and continuum peaks. The behavior can be explained in the following way. The affect of the cleaning process is to reduce the external dielectric screening environment the tube sees. The reduction of the dielectric screening has two effects: 1) The quasiparticle band gap is increased. And 2) the exciton binding energy is decreased. These two effects shift the optical-transition energies in opposites directions. For the $1A_2$ state, the exciton binding energy is also large, and, therefore, the change in the exciton binding energy due to the different dielectric environment almost perfectly cancels the change in the quasiparticle gap. For the $5A_2$ and continuum states however, the binding energy is small to begin with; thus, when changing the dielectric environment, the transition energy is increased by approximately the same amount as the quasiparticle gap is increased.

To demonstrate this, we have done calculations using our effective mass model on

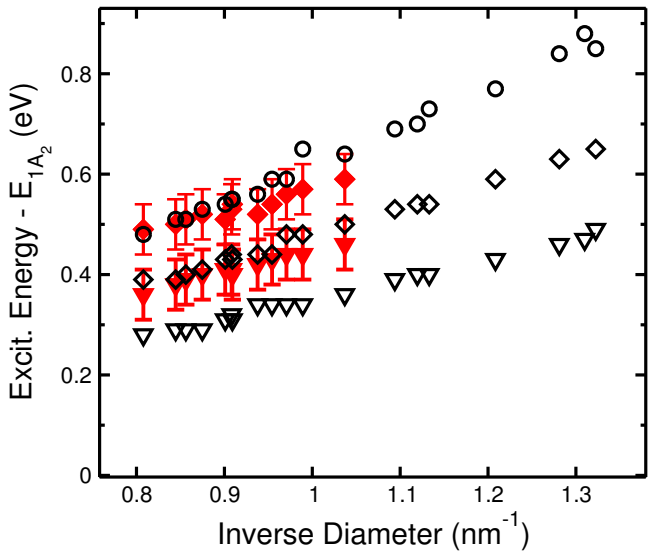

Figure 4.11: Comparison of the energy positions of the absorption spectra features for different diameter semiconducting tubes in the work of Lefebvre et. al. [80] to the $3A_2$ and $5A_2$ excitonic state energies in the E_{11} interband transition exciton series. The black circles represent the calculated continuum level in the present model while the black diamonds and triangles represent the $5A_2$ and $3A_2$ states respectively. The red diamonds and triangles are the L1* and L1 features in the work by Lefebvre et. al. [80]

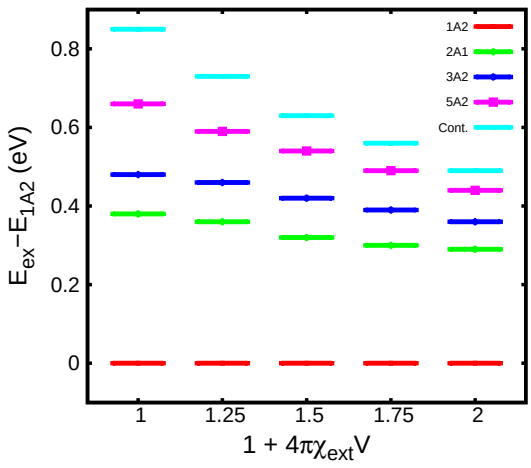

Figure 4.12: Excitation energies calculated within the present effective mass model for the $2A_1$, $3A_2$, $5A_2$ and continuum levels relative to the $1A_2$ excitation energy for the (10,0) SWCNT subject to different levels of external screening - $4\pi\chi_{ext}(q)V(q)$.

the (10,0) SWCNT. We add an external screening source to mimic the cleaning/dirtying process:

$$\epsilon(q) = 1 + 4\pi\chi_{NT}(q)V(q) + 4\pi\chi_{ext}(q)V(q) \tag{4.34}$$

where χ_{NT} is the polarizability intrinsic to the nanotube and χ_{ext} is a parameter representing the screening from external sources, such as adsorbents. Fig 4.12 shows our calculated excitation spectra with values of $4\pi\chi_{ext}(q)V(q)$ between 0 and 1. The results show the exact same trend as discovered in the experiment [81]. Notice also that the anti-screening effect is almost completely wiped out with external screening that amounts to $4\pi\chi_{ext}(q)V(q) \approx 1$. Thus, only isolated and relatively clean nanotubes will demonstrate the anti-screening effect. The anti-screening effect is thus not expected to be apparent for tubes in solution, for example. A full experimental and theoretical treatment of the environmental effects would be a fruitful avenue for future research.

Chapter 5

Graphene and Metallic Nanotubes

The quasi-electron and quasi-hole states near the Fermi energy in graphene and metallic armchair single-walled carbon nanotubes (SWNTs) obey a linear energy-wave-vector dispersion relation characteristic of two- and one- dimensional Dirac fermions. The existence of Dirac-fermions in graphene has been verified by recent experiments demonstrating a unique quantum Hall effect in single graphene sheets [101, 153]. These experiments were able to infer a Fermi velocity between 1 and $1.1 \cdot 10^6 m/s$ for the quasielectrons in graphene.

In bulk metals, excitonic effects play only a small role in the optical absorption spectra, and bound exciton states are non-existent due to the almost perfect screening of the interaction between the electron and hole. As we discussed in Chapter 2, there is perfect screening at large distances because the inverse dielectric function in reciprocal space, $\epsilon^{-1}(q) \propto q^2$, cancels the long wavelength divergence of the Coulomb interaction. Within the Thomas-Fermi model the screened interaction is effectively:

$$W(\mathbf{r}) = \frac{1}{r} e^{-k_0 r} \tag{5.1}$$

where k_0 is the inverse Thomas-Fermi screening length related to the density of states at the Fermi energy. The interaction is particularly weak in three-dimensions because the phase space near the singularity (the origin) vanishes as r^2. Thus, the singularity does not greatly affect the potential energy of the electron-hole pair and one sees no bound exciton states in bulk metals. This is illustrated schematically in Fig. 5.1.

However, as we discussed in Chapter 4, the effective Coulomb interaction is stronger in lower dimensions due to the confinement of the electrons and holes. In two dimensions the phase space near the singularity goes to zero only as r, the same rate at which the interaction diverges. In one dimension, the phase space near the origin is constant and the singularity is unavoidable. We therefore expect the interactions in two- and one-dimensions to be stronger. We therefore investigate the many-body interaction effects on the quasi-particle and optical properties in two dimensions on graphene (a semi-metal) and in one dimension on the metallic single-walled carbon nanotubes.

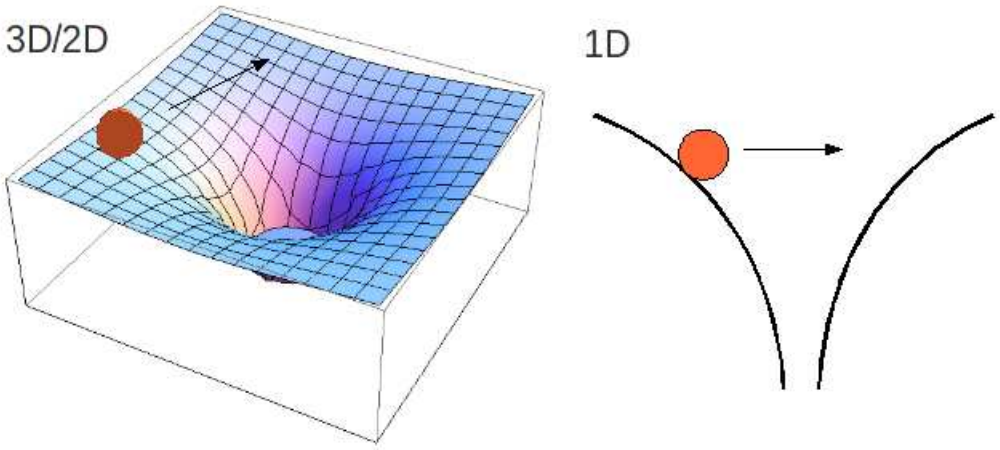

Figure 5.1: Schematic of the 1D vs 2D/3D Coulomb interaction as described in the text.

5.1 Graphene

Excitonic (electron-hole pair) effects are noticeable in the optical response of semiconductors, while they are typically unimportant in that of conventional bulk metals because of the strong screening effect from free carriers. As we mentioned above, it is of considerable interest to see if there are significant excitonic effects in two-dimensional metallic or semi-metallic systems.

The electronic and optical properties of graphene have been a subject of tremendous research effort in the last several years [102, 153, 47]. The optical properties in particular display interesting characteristics. For example, the low frequency absorbance per sheet is a constant with respect to the frequency of light/ [52, 95, 86, 147] More recently the discovery of a Fano line shape in the main absorption peak has been observed [87, 77]. It is therefore of great interest to study the quasiparticle and optical properties of graphene within a first-principles approach to discover whether novel many-body effects may be in play. In this section, we present the result of such a study on both graphene and bilayer graphene. We employ the basic GW-BSE methodology [115] laid out in Chapter 2 within the BerkeleyGW package.

As we will discuss more in the following sections of this chapter, we find that there is a significant renormalization of the graphene Fermi velocity near the Dirac point due to electron-electron interaction effects. Only with the inclusion of this self-energy renormalization can agreement with experiment be achieved for the velocity of the Dirac quasiparticles. Secondly, predict a major shift in the energy of the predominant absorption peak in the spectra. We show at the end of the section, that our predicted position and lineshape agree well with recent experimental measurements. Again, it is only with the inclusion of many-body effects that agreement can be reached. Despite the large many-body effects on the main absorption peak, we are able to confirm that the infrared spectral absorbance is relatively insensitive to many body effects and remains approximately a constant 2.4%, in

agreement with experiment [95, 86]

We use a relaxed atomic structure within the LDA for the Kohn-Sham exchange-correlation potential. However, in bilayer graphene we choose the experimental sheet separation (0.334 nm). All calculations are done in a supercell geometry [32] with a 1.2nm inter sheet distance. We studied graphene using both a Coulomb truncation scheme and without Coulomb truncation. Due to the semi-metal nature of graphene, we find only small differences between the two techniques. We use a plane-wave basis and the BerkeleyGW code as laid out in Chapter 2. The biggest computational problem in describing graphene within the GW-BSE technique is the large number of k-points required for the various levels of computation. We find that a 32x32x1 k-point grid is required to accurately describe the charge-density with DFT. A 64x64x1 k-point grid is required to accurately compute the electron self-energy. For computing the optical absorption spectra we use a 64x64x1 k-point grid in order to compute the electron-hole interaction kernel. The diagonalization of the kernel and computation of the absorption spectra requires a 200x200x1 k-point grid. For the self-energy computation, we include 96 empty orbitals. For the computation of the absorption spectra, we include 2 valence and 2 conduction bands (tested with 4 valence and 4 conduction bands). This is found to be sufficient for describing the absorption spectra below 10 eV. The Bethe-Salpeter equation is solved within the static approximation. All spectra is broadened with a 0.5 eV Lorentzian broadening.

We show that a large renormalization of band-velocity in graphene near the Dirac point due to self-energy effects in Fig. 5.2 (a). As we discuss more in the next sections, only when self-energy effects are included, do we have agreement with experiment. In particular, the LDA Fermi velocity of graphene is 0.82×10^6 m/s, while we find the GW value is 1.05×10^6 m/s. The GW number is agrees very well with the measurement of the velocity using the quantum Hall effect [153]. It also agrees favorably with previous GW calculations [138, 13]. We show the GW bandstructure for bilayer graphene in Fig. 5.2 (b). Again we see that there is a large renormalization effect on the band dispersion near the Dirac point.

The optical spectra of graphene is shown in Figure 5.3 (a). In order to obtain a quantity that is not dependent on the inter-sheet separation, we show the quantity $\alpha_2(\omega)$, the imaginary part of the polarizability per unit area, which is defined as the product of the polarizability of the supercell geomerty, $\chi = (\varepsilon - 1)/4\pi$, and the distance between adjacent graphene layers. In order to compare to a measurable polarizability, one should multiply by the graphene or bilayer graphene area. When electron-hole interaction effects are neglected, we see a major peak in the optical absorption spectra at around 5.15 eV. This corresponds to the interband transition near the M point in the graphene Brillouin zone. In this region the bands form a saddlepoint Van Hove singularity. When the electron-hole interaction is turned on (through the solution of the Bethe-Salpeter equation) however, the peak is significantly red-shifted. The peak position with excitonic effects included occurs at 4.55 eV - a 600 meV redshift from the position in the non-interacting spectra. Additionally, the peak lineshape is significantly more asymmetric with excitonic effects included. As we discuss at the end of this section, both of these predictions have been confirmed by recent experiment.

What causes the apparent shift of 600 meV in the absorption peak? It is tempting to attribute the shift to the formation of strongly bound exciton states as in the case of the

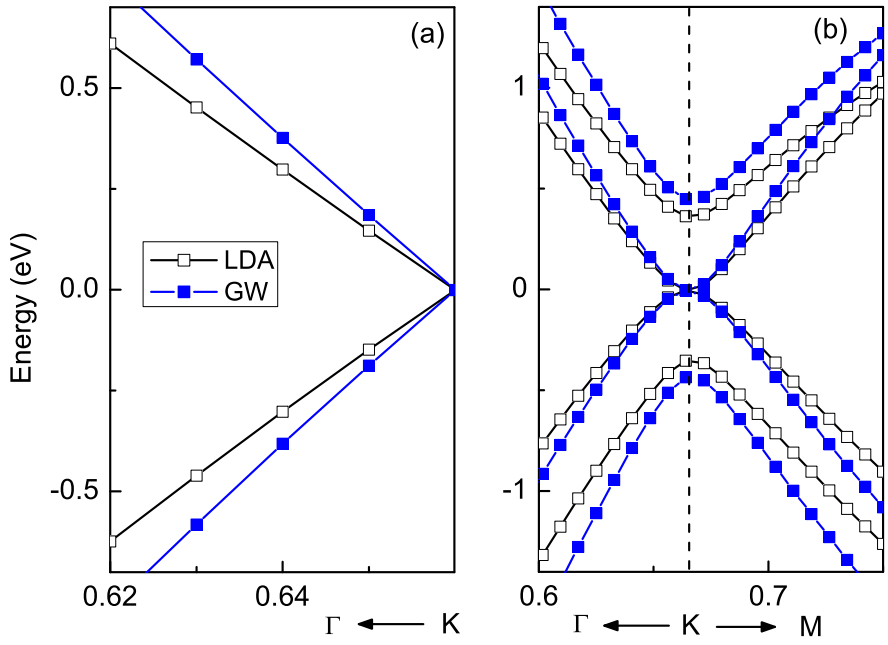

Figure 5.2: A comparison between the LDA (open squares) and the GW (solid squares) band structure for graphene (a) and bilayer graphene (b). k is in units of $\frac{2\pi}{a}$, where a is the in-plane lattice constant.

semiconducting tubes of the previous chapter. However, it should be remembered that bulk metals and semi-metals typically do not allow bound exciton states (due to the nearly perfect screening between the electron and hole). Additionally, as we will see in the next section, even one dimensional metals have bound exciton states with binding energies of only 100 meV. Indeed, upon further analysis, we find that bound-exciton states are not responsible for the change in peak position. This is illustrated in Fig. 5.3 (b), where we see that the joint density of states (JDOS) of quasiparticles from the GW calculation and the density of excitonic states are identical. This implies that the change in peak position results from a redistribution of oscillator strength from high energy to low energy transitions, and not from the creation of bound exciton states. This is similar to the redistribution of oscillator strength that occurs in bulk silicon for example [115].

In order to determine the mechanism behind this redistribution of oscillator strength, we consider the optical transition probability of a transition from the ground state to an exciton state resolved in terms of the interband electron-hole transitions at a given energy, ω, that make up the exciton state:

$$\langle 0|\mathbf{v}|i\rangle = \sum_{v}\sum_{c}\sum_{k} A^{i}_{vck}\langle vk|\vec{v}|ck\rangle = \int S_i(\omega)d\omega, \tag{5.2}$$

where

$$S_i(\omega) = \sum_{v,c,k} A^{i}_{vck}\langle vk|\vec{v}|ck\rangle\delta[\omega - (E_{ck} - E_{vk})]. \tag{5.3}$$

Note that for graphene, $S_i(\omega)$ is as a real function because the system has inversion symmetry. This quantity is plotted in Fig. 5.4.

Figure 5.4 (a) shows $S_i(\omega)$ for an exciton state with excitation energy 1.6 eV. The state is composed of electron-hole pairs from only a very narrow energy window. This implies that the electron-hole interaction is unimportant in this area of the spectra since it does not significantly mix the non-interacting states. In Fig. 5.4 (b) the exciton state considered has an excitation energy around 4.5 eV. The energy distribution of states that contribute to this exciton is much more broad - this is characteristic of a region in the spectra where electron-hole interaction effects are important.

Notice that the function $S_i(\omega)$ in Fig. 5.4 changes sign as ω, the energy contributing electron-hole pairs, crosses the excitation energy of the correlated exciton state in consideration. In Fig. 5.4 (b), for the exciton state around 4.5 eV, there is a long tail of contributions to the exciton state from higher energy electron hole pairs. Therefore, this exciton state effectively "steals" oscillator strength from higher energy excitons. In other words, much of the absorption spectra that was in the peak at 5.1 eV is shifted into exciton states near 4.5 eV.

On the other hand, Fig. 5.4 (c) shows $S_i(\omega)$ for a typical exciton with excitation energy near 5.1 eV. Electron-hole pairs from a large energy window again contribute to the state, however, the exciton has a greater contribution of oscillator strength deriving from states to the left of the excitation energy. Additionally, the contribution of states to the left and right of the excitation energy effectively cancel, giving an overall reduction to the absorption peak near 5.1 eV. Therefore, we see that the oscillator strength from the peak originating at 5.1 eV in the non-interacting spectra has moved to 4.5 eV. The fact that

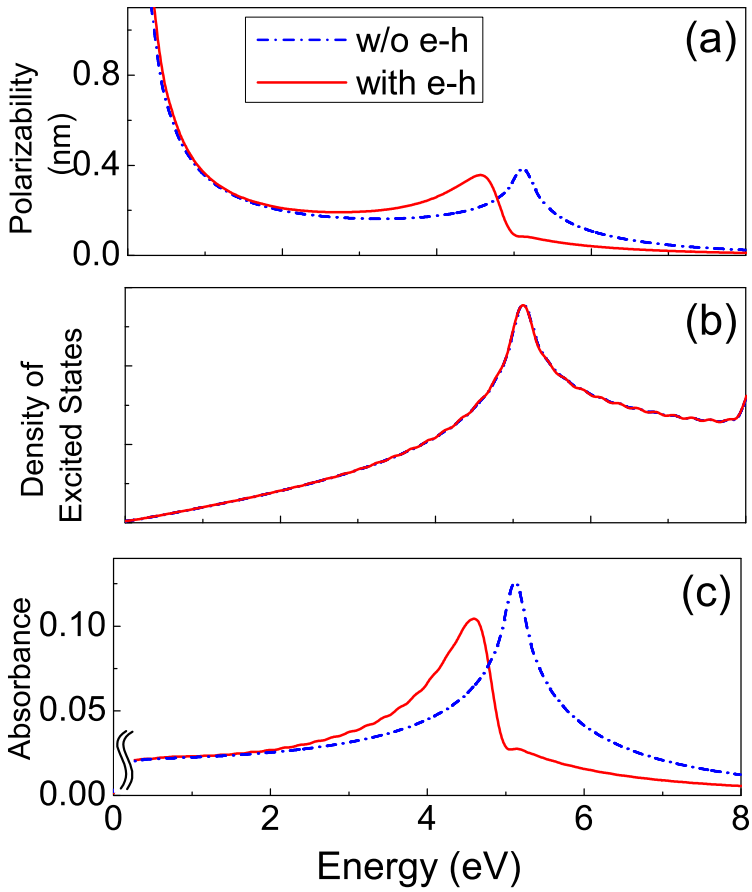

Figure 5.3: (Color online) (a) The GW-BSE predicted optical absorption spectra, (b) joint density of excited states, and (c) absorbance of a single layer of graphene with and without electron-hole interaction effects included.

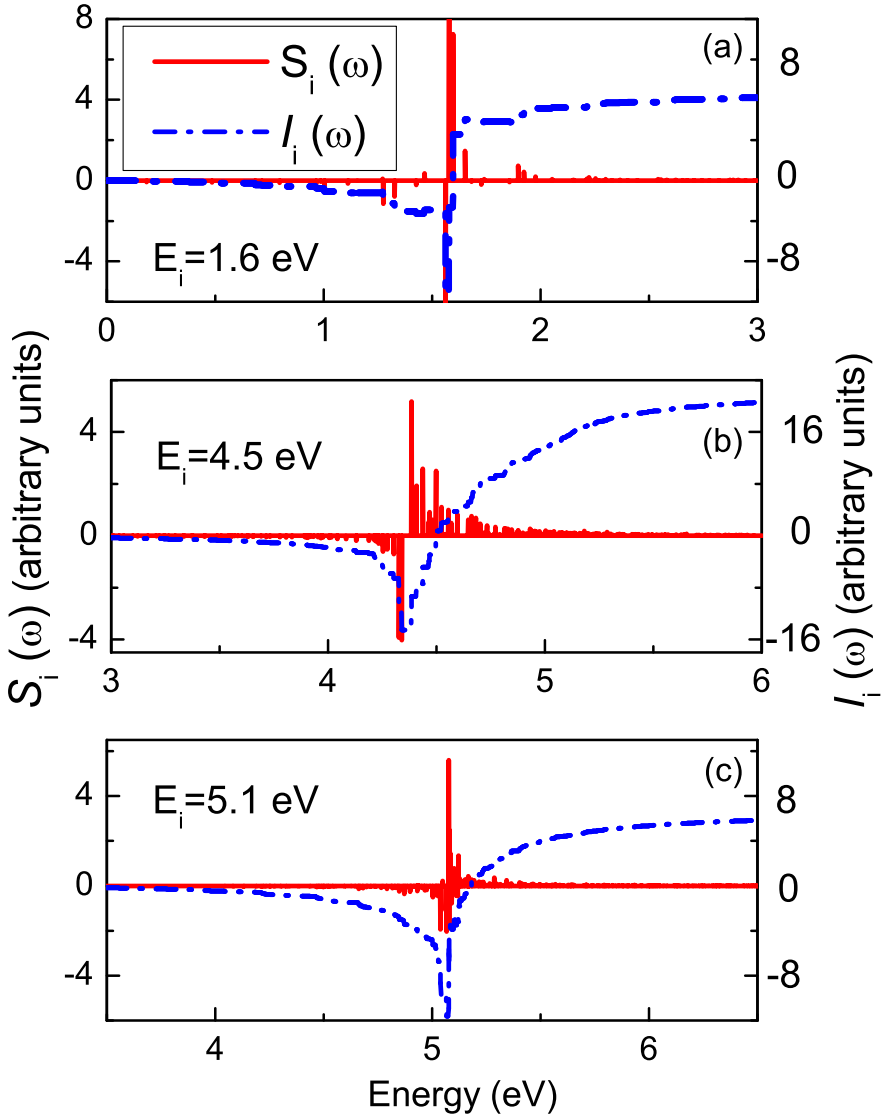

Figure 5.4: $S_i(\omega)$, defined in text, and the partial integration $I(\omega) = \int_0^\omega S_i(\omega')d\omega'$ for exciton states with excitation energies of 1.6 eV, 4.5 eV and 5.1 eV.

	graphene	bilayer graphene	graphite
E_{peak} (Expt.)	4.6 [87, 77]		4.55 [133, 38]
E_{peak} (GW+BSE)	4.55	4.52	4.50
$\delta\Sigma$	+1.10	+0.91	+0.69
$\delta_{exciton}$	-0.60	-0.45	-0.27

Table 5.1: Position (in eV) of the main absorption peak in graphene, bilayer graphene and graphite, and change in peak position from the inclusion of self-energy effects ($\delta\Sigma$) and from electron-hole interaction effects ($\delta_{exciton}$).

an enhancement in the absorption spectra in one region is accompanied by a reduction in another energy region is required by the f-sum rules discussed in Chapter 2.

From simple models for the graphene bandstructure, the low-frequency optical absorption per graphene sheet is expected to be a constant (2.29%) [52, 95, 130, 150, 86, 10, 97, 109]. Figure 5.3 (c) shows the calculated absorbance of graphene in this region, $A(\omega) = \frac{4\pi\omega}{c}\alpha_2(\omega)$. Notice that the low-frequency value does not change significantly between the calculations including and excluding excitonic effects. In both cases, we find that the absorbance is around 2.4%. This result agrees well with experiment and previous calculations. [73, 86, 95].

We have additionally computed the optical absorption spectra of bilayer graphene and graphite within the GW-BSE methodology. Table 5.1 shows a comparison between the absorption peak positions predicted by our GW-BSE calculations with experiement. As multiple layers are added, we find that the change in the quasiparticle self-energy correction and excitonic effects nearly cancel, and that the main absorption peak position is nearly unchanged between graphene, bilayer graphene and graphite. This cancellation effect is discussed more in the following sections in regards to metallic nanotubes. [127, 148, 110].

Our predicted absorption peak position for single-layer graphene has been recently confirmed by experiments [87, 77] where significant asymmetry due to excitonic affects is also confirmed. Figure 5.5 shows a comparison between our calculated absorption spectra, broadened by 350 meV and the two experimental results.

To conclude, we have utilized the GW-BSE methodology implemented in the BerkeleyGW code to study the electronic and optical properties of graphene. We found that many-body effects play an important role. We see a large renormalization of the Fermi velocity near the Dirac point due to self-energy effects. Additionally, we see a 600 meV red-shift of the main absorption peak due to excitonic effects. These large effects are surprising for a semi-metallic system. But, as we will see in the next section, they are part of a larger trend - that many-body effects are important in reduced-dimensional metals and semi-metals, and failing to include them can lead to a lack of even qualitative predictability.

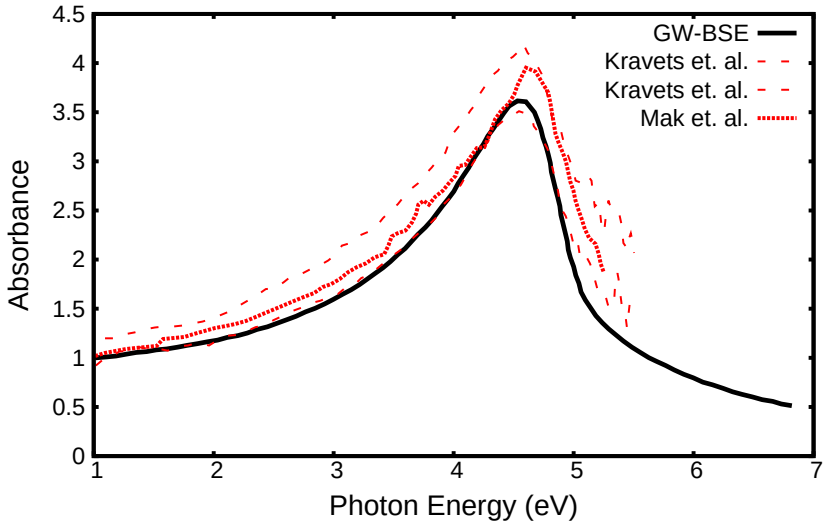

Figure 5.5: Comparison of the calculated absorbance, in units of $\pi\alpha$, (broadened 350 meV) with recent experiments: After Mak et. al. [87] after Kravets et. al. [77]

5.2 Metallic Nanotubes

As we discussed in Chapter 4, single-walled carbon nanotubes (SWCNTs) are quasi-1D structures made by rolling up graphene strips. The electronic properties of SWCNTs are determined uniquely by the tube diameter and chiral angle; SWCNTs can be either metallic or semiconducting depending on these two features. [118, 54, 91] As discussed in the previous chapter, *ab initio* calculations [127, 128, 34] have predicted that excitonic effects qualitatively change the optical properties of single-walled carbon nanotubes. The effects are stronger in one-dimension due to the enhanced electron-electron interaction in quasi-1D materials. [9, 34] In the previous chapter, we showed *ab initio* GW-BSE calculations predict the existence of discrete strongly bound exciton states below the electron-hole continuum. [115, 63, 62] It is therefore of great interest to determine whether such excitonic states exist and affect the optical spectrum of metallic SWCNTs with large enough diameters to be measured in experiment. [127, 128] Previous *ab initio* calculations of metallic tubes were limited to tubes with small diameter, (3,3) and (5,0), and the analysis was only done for excitons associated with the first peak in the spectrum. We present in this section *ab initio* calculations showing the existence of excitons in larger diameter metallic SWCNTs. In particulary we consider the (10,10), (12,0) and (21,21) tubes. We will show that bound excitons do exist in such systems, a prediction that has been experimentally verified.

As mentioned in the previous chapter, the excitonic picture for the optical spectra of semiconducting tubes has been confirmed both by experiments using two-photon photoluminescence spectroscopy techniques and more recently by single-photon luminescence experiments on isolated tubes. [146, 90, 80, 81] However, because metallic tubes do not have efficient luminescence, these experimental technique are not directly applicable to metallic

SWCNTs. However, as we show below, new experiments (along with the predictions of the GW-BSE technique) that measure directly the absorption spectra and lineshape of isolated single-walled nanotubes can determine whether bound exciton states exist in metallic tubes.

Using the GW-BSE methodology of Chapter 2, we find that the optical spectra of the lowest allowed interband transition, E_{11}, in the (10,10), (12,0) and (21,21) SWCNTs are characterized by bound exciton states with binding energies ranging between 40 and 60 meV. Additionally we find that the 2nd and 3rd optical peaks (corresponding to E_{22}, E_{33} in the notation of the previous chapter) are also characterized by narrow resonant exciton states with similar exciton binding energies as E_{11}. In the previous chapter, we saw that for semiconducting SWCNTs, each interband transition (E_{11}, E_{22} etc...) gave rise to a series of bound exciton states similar to a hydrogenic series. In metallic nanotubes, however, we find only a single noticeably bright bound or resonant excitonic state per optically allowed interband transition because of the strong metallic screening. It is possible higher bound states just below the continuum do exist but are not optically distinguishable from the continuum. The single bound or resonant state from each optically active interband transition, acounts for more than 50% of the absorption probability associated with the interband transition. Thus, the peak in the optical spectrum is derived predominatly from a single Lorentzian peak and acquires a highly symmetric lineshape. This is distinguishable from that of an independent electron-hole picture where the lineshape follows that of a one-dimensional Van Hove singularity $1/\sqrt{E - E_b}$, where E_b is the optical band gap. These theoretical predictions of the GW-BSE methodology have been recently verified by experiment as we discuss below.

5.2.1 Calculation Details and Results

We use the GW-BSE method implemented in the BerkeleyGW package laid out in Chapter 2. In particular, we compute the ground-state atomic coordinates and charge density of the (10,10), (12,0) and (21,21) tubes within *ab initio* pseudopotential density functional theory (DFT) using a planewave basis and the local density approximation, LDA, for the exchange correlation potential as discussed in Chapter 1. The quasiparticle energy are obtained by solving the Dyson's equation using first order perturbation theory within the diagonal approximation for Σ in the LDA basis. [59, 63, 62] We then calculate the two-particle electron-hole excitation energies, exciton wavefunctions and optical response functions by solving the two-particle Bethe-Salpeter equation (BSE). [115] The DFT eigenstates and eigenvalues are obtained using a 60 Ry cutoff for the plane wave basis and using *ab initio* Troullier-Martins pseudopotentials in the Kleinman-Bylander form with a cutoff of $r_c = 1.4$ a.u.). [139, 74] In order to simulate isolated tubes, we use a hexagonal supercell geometry with an intertube distance of at least 7.6 Å in all cases.

Figure 5.6 shows the Kohn-Sham bandstructure for an isolated (10,10) tube calculated within the LDA. We find that only light with an electric field polarized along the tube axis gives a strong optical response. [127, 128, 6] For the rest of this section, this polarization is taken for all the optical properties presented below. In this case, the optical perturbation Hamiltonian, $\propto \vec{A} \cdot \vec{p}$, is unchanged under application of operators from the symmetry group of the k-vector. Therefore, it is a unique property of nanotubes, that the optical transitions obey well defined selection rules that apply to the entire Brillouin zone

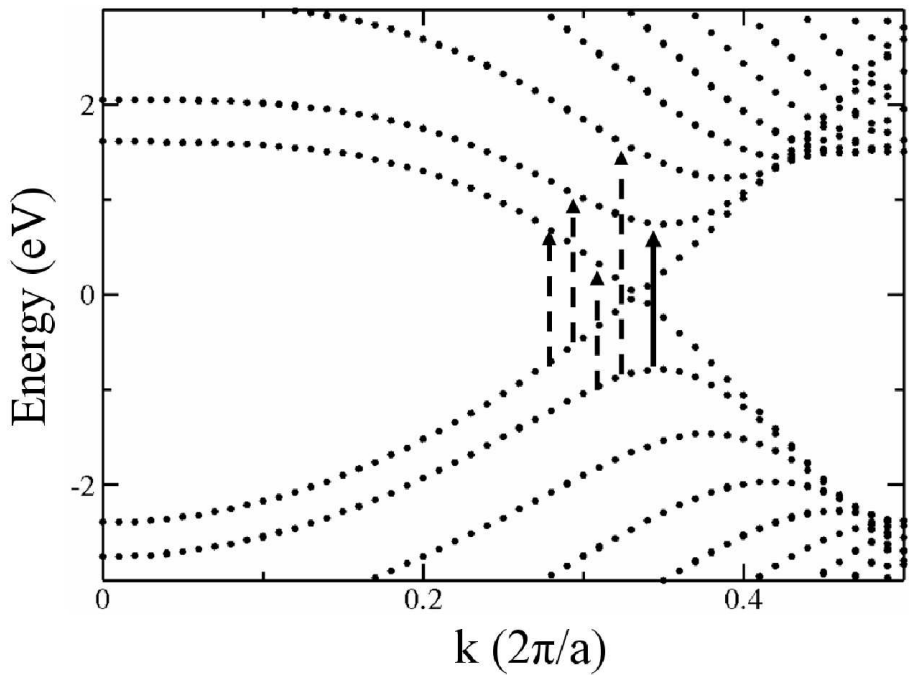

Figure 5.6: The (10,10) SWCNT LDA bandstructure with the zero of energy set at the Fermi energy. Dashed arrows indicate optically forbidden transitions for light polarized along the tube axis. The solid arrow indicates the lowest allowed optical transition.

in k-space. [18, 33] - i.e. transitions obey well-defined optical selection rules across entire bands. In SWCNTs, every k-point lies along a high symmetry direction! In the (10,10) tube, this first allowed optical transition, E_{11}, occurs between the second and third valence subbands (where each band is doubly degenerate) and the second and third conduction subbands. This allowed transition is shown by the solid black transition line drawn in Figure 5.6). The dashed lines correspond to optically forbidden interband transitions under the polarized light. The E_{22} and E_{33} transitions occur between higher sets of valence and conduction subbands. Similar symmetry rules hold for the allowed transitions in the (12,0) and (21,21) SWCNTs.

Figure 5.7 shows the GW quasiparticle corrections to the Kohn-Sham, within the LDA, eigenvalues for the (10,10) tube. For the (10,10) and (12,0) tubes, we find the slope in the quasiparticle dispersion relation of the metallic bands increases by approximately 24% over the LDA result. We find that the (21,21) tube has an 29% increase in the slope of the quasiparticle energy dispersion relation. However, the (5,5) tube sees a 19% increase, while the (3,3) tube sees a 15% increase [127]. We discuss in the next subsection the origin of the diameter dependence of the quasiparticle shifts. The corrections are somewhat larger for larger diameter tubes because of the reduced local screening as the average distance of neighboring atoms increases with tube diameter.

Figure 5.8 shows the optical absorption spectra for the first allowed transition

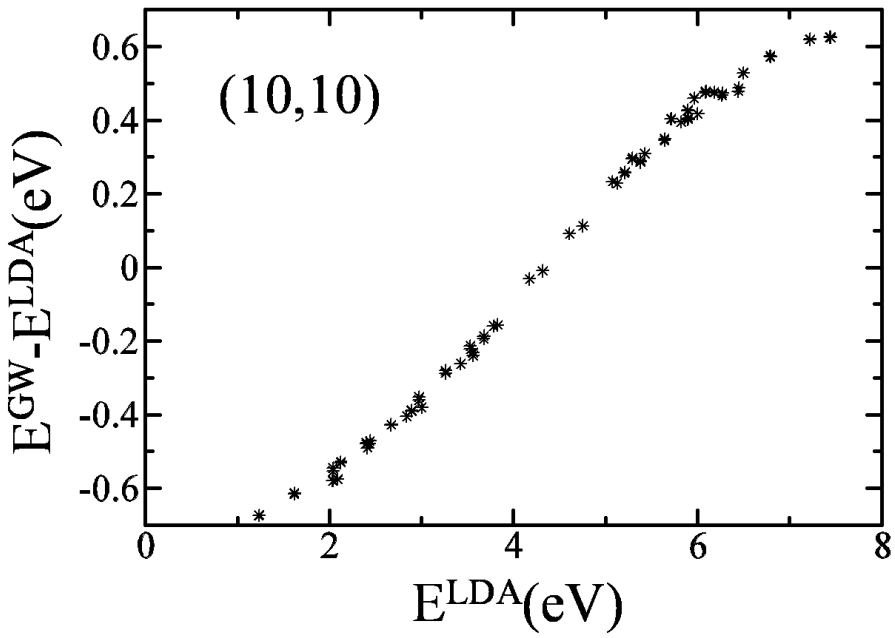

Figure 5.7: The quasiparticle energy corrections versus E_{LDA} for the (10,10) SWCNT. The linear regression slope is approximately 0.24. This slope represents a scaling of LDA energy eigenvalues by 25 percent for the (10,10) tube due to self-energy effects.

(denoted E_{11}) in the (10,10) tube. If one were to increase the supercell size by including more vacuum in the calculation, ϵ_2 would systematically approach 0. In order to show a quantity that independent of supercell size, we use $\alpha = \mathrm{Im}\{A(\epsilon - 1)/4\pi\}$, where ϵ is the calculated supercell dielectric function, and A is the cross-sectional area of the supercell in the calculation. This quantity is the imaginary part of the polarizability per tube and can be directly related to the absorption cross-section per tube length. To get a experimentally measured dielectric response, one should multiply this quantity by the density of tubes per unit area in a matrix or in a bundle.

For the (10,10), (12,0) and (21,21) nanotubes studied, the existence of bound excitons qualitatively changes the absorption spectrum. In the (10,10) tube, the first bound exciton state has a binding energy of 50 meV. In Fig. 5.8a, it is seen that the optical transition probability for the interband transition comes predominantly from the single exciton state (this is also obtained directly from the exciton wavefunctions and oscillator strengths resulting from the BSE). The absorption lineshape with excitonic effects included is significantly different from the interband transition case. As a guide to the eye, Fig. 5.8c shows both the interacting and non-interacting spectra with the non-interacting specturm scaled and shifted to match the peak height in the interacting case. It is clear that the non-interacting spectra is qualitatively different, being significantly more asymmetric than the interacting spectrum. As we will discuss below, this line-shape analysis has been used by experiments on the isolated metallic (21,21) SWCNT to confirm the excitonic picture.

For the (12,0) and (21,21) tubes, the first transition is due to a similar bound exciton state with a binding energy of approximately 50 meV and 40 meV respectively. These bindings are significantly smaller than the binding energies of semiconducting tubes that we discussed in the previous chapter due to the enhanced metallic screening and the absence of the anti-screening effect. However, these binding energies are still large compared even to bulk semiconductors such as Si and GaAs. In bulk metals, no bound excitons are found due to the nearly perfect screening between the electron and hole. The fact that optically active bound excitons can exist in lower dimensions, even in the presence of metallic screening, is due to quantum confinement effects and of the existence of optical symmetry gaps described above. In one dimensional quantum systems, it has been shown that any external potential (other than $V = 0$) that satisfies $\int V(x)dx \leq 0$ (i.e. is negative on average) is guaranteed to have at least one negative energy eigenstate. [75, 124] However, if the binding energy were low and the absorption from the metallic bands strong enough, we would still be unable to see bound exciton states manifested in the absorption spectra. Luckily, in metallic SWCNTs, the valence and conduction bands that make up the metallic-like bands that cross at the Fermi level are from different representations of the group of the k-vector [18, 17], and the probability of optical transitions between those states is zero. Thus, as long as the repulsive exchange term in the electron-hole kernel is weak, a bound state from the E_{11} interband transition, as predicted with the GW-BSE methodology, is expected to be present in measured absorption spectra. Unlike the case in similar diameter semiconducting tubes discussed in the previous chapter, however, GW-BSE guarantee only one bound state per van Hove singularity and not a series of measurable excitonic states. This is due to the absence of the anti-screening effect in metallic tubes.

One comparison that can be made to experiment is in the absolute energy of

absorption peak. For the (10,10) tube, the GW-BSE absorption peak position is calculated to be 1.84 eV. The experimental peak position has been measured to 1.89 eV. [42] For the (12,0) tube the calculated absorption peak position is 2.25 eV, whereas experimental measurement gives a value of 2.16 eV for the position. [42, 135] In both cases, we see agreement to better than 5 percent, demonstrating the accuracy of the combined GW-BSE technique.

Figure 5.9 shows the optical absorption spectra for the E_{22} and E_{33} optically allowed transitions in the (10,10) SWCNT. The absorption spectra look qualitatively similar to the E_{11} spectra. However, in this case, we find resonant exciton states because the electron-hole states from the E_{11} continuum, being of the same symmetry, mix with the E_{22} and E_{33} electon-hole states forming the exciton states. However, due to the relatively lower JDOS of the E_{11} continuum when compared to the van Hove singularity of the E_{22}, for example, we find that the resonant states contain only a 15% contribution from the E_{11} continuum. Thus, the E_{22} and E_{33} transitions are still associated with excitons with binding energy of 60 meV and 50 meV respectively. Thus, the GW-BSE calculations predicts that the second and third van Hove singularities give rise to resonant excitonic states with binding energies similar to the exciton arising from E_{11} transition.

From analyzing the exciton wave functions, A^S_{vck}, computed by diagonalizing the BSE as discussed in Chapters 1 and 2, we find that the bright exciton states that dominate the absorption spectra are composed from a single valence and conduction band pair and mixed over a range of k-points making up approximately $\frac{1}{50}$ of the Brillouin zone centered at the band minima. The exciton wavefunction in real space are defined from the electron and hole wavefunction as:

$$\Psi\left(\vec{r}_e, \vec{r}_h\right) = \sum_{kvc} A_{vck} \phi_{c,k}(\vec{r}_e) \phi_{v,k}(\vec{r}_h). \tag{5.4}$$

Figure 5.10 shows the squared magnitude, $|\Psi\left(\vec{r}_e, \vec{r}_h = const\right)|^2$, of this function for the lowest optically bright bound exciton state for the (10,10) tube with the hole position fixed in the center of a orbital. In Fig. 5.10 (a), the electron's amplitude is averaged over the radial and angular components and plotted along the tube axis, with the hole positioned at the origin (z=0). For the (10,10) tube, the axial width (analogous to the exciton radius) is approximately 30 Å. Figure 5.10 (b) shows $|\Psi|^2$ in a cross-sectional cut across the tube axis where again the hole is represented by the green dot. The wavefunction is delocalized over the entire diameter is a result of the fact that exciton derives from only one interband transition.

In this section, we showed that bound excitons exist in metallic SWCNTs and that excitonic effects are important in achieving even a qualitative description of the optical properties of metallic SWCNTs. [127, 37] These predictions were tested by experiment on an absorption study of an isolated (21,21) SWCNT. [145] The experimental absorption lineshape can be compared to the theoretical lineshapes calculated with and without the excitonic effects included to both qualitatively confirm whether excitonic effects are present in the experimental spectra and to quantitatively determine the binding energy.

Our initial comparison to experiments utilized the *ab initio* calculation of (10,10) SWCNT E_{11} described above. This approach is justified because the E_{22} transition in the

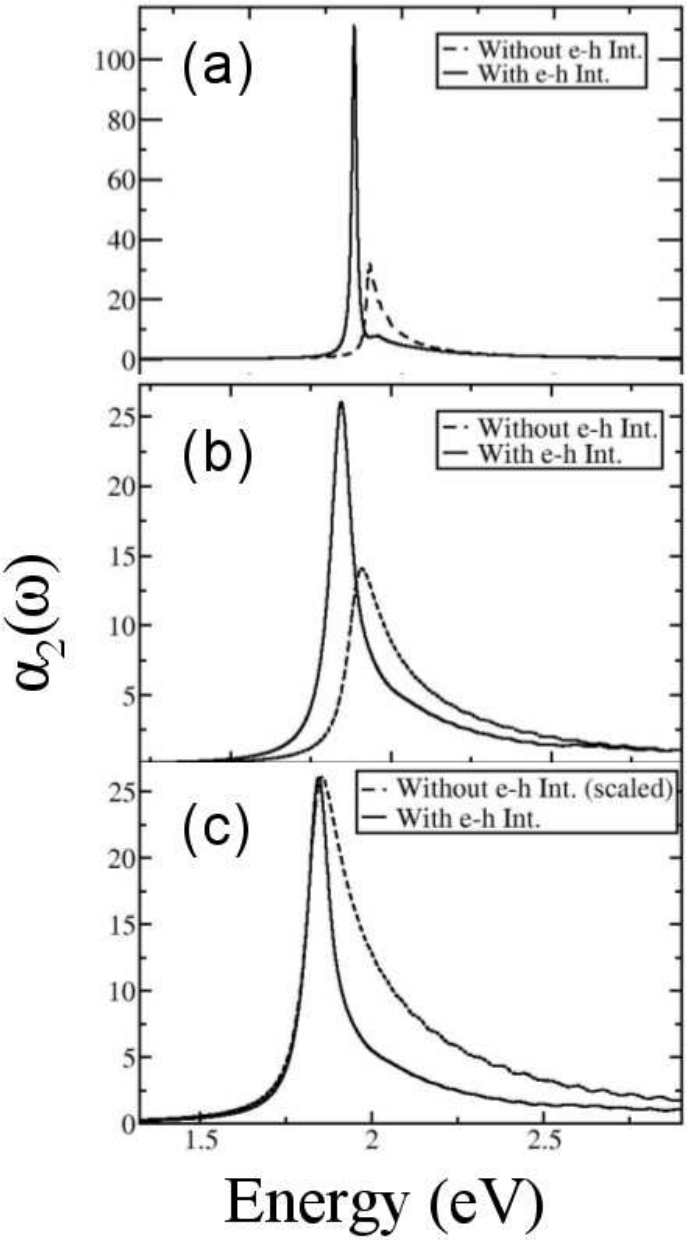

Figure 5.8: The calculated E_{11} absorption lineshape in the (10,10) SWCNT with (a) 20 meV and (b) 80 meV Lorentzian broadening. The solid curves include excitonic effects and the dashed curves were calculated without the electron-hole interaction. Panel (c) compares the two spectra with 80 meV broadening where the noninteracting spectrum has been scaled and shifted to match the peak in the interacting case.

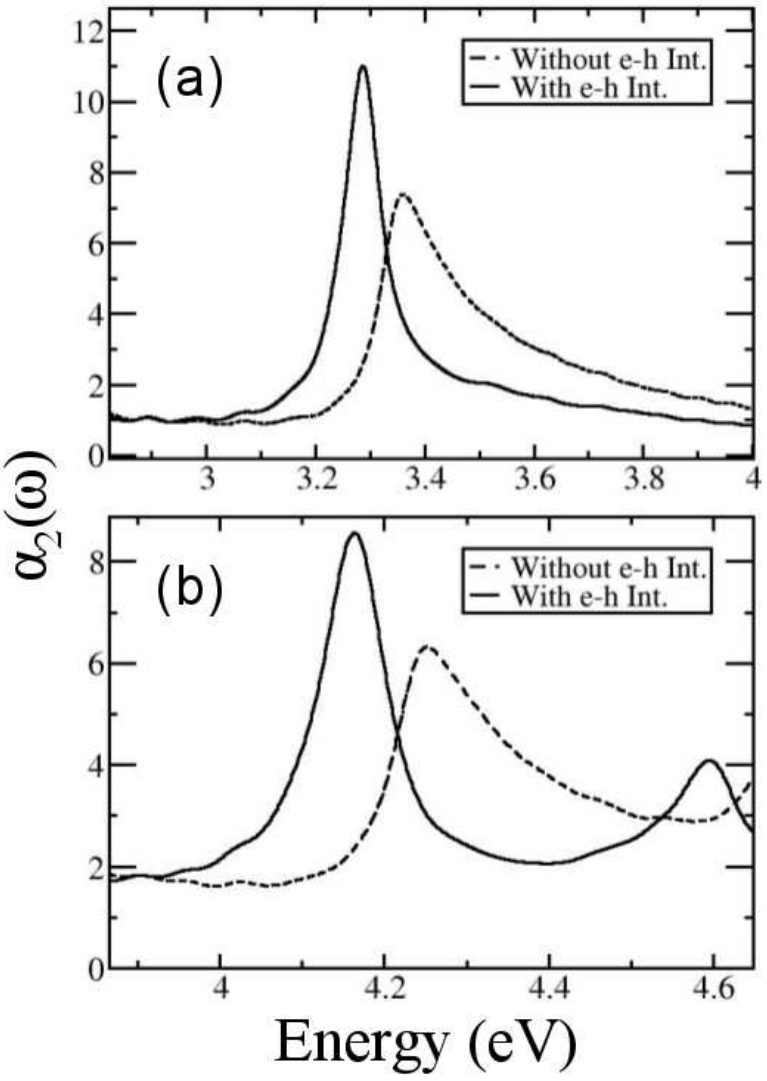

Figure 5.9: Calculated optical absorption peaks the (a) E_{22} and (b) E_{33} interband transitions in the (10,10) SWCNT.

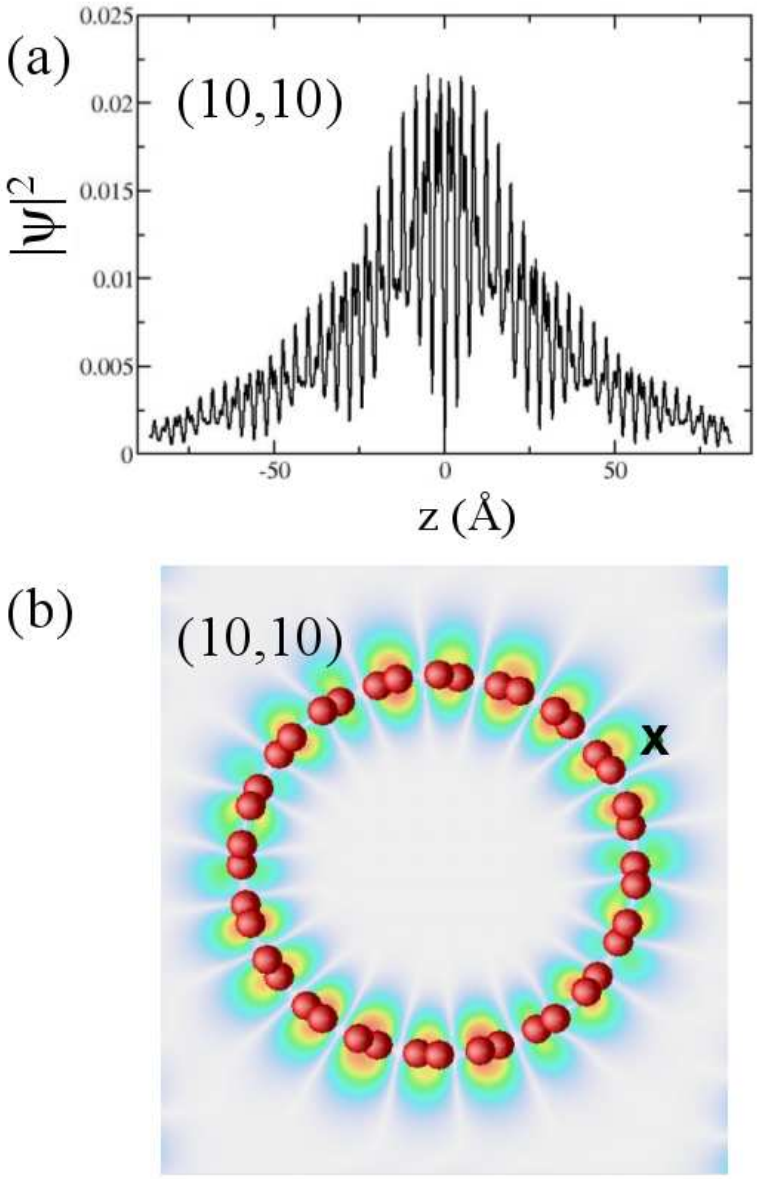

Figure 5.10: The exciton wave function in real-space: the electron amplitude squared in real space with the hole position fixed (a) plotted along the tube axis with the hole located at the origin and radial and angular degrees of freedom integrated out and (b) plotted on a cross section cut across the tube axis. The hole is located at the X in the figure.

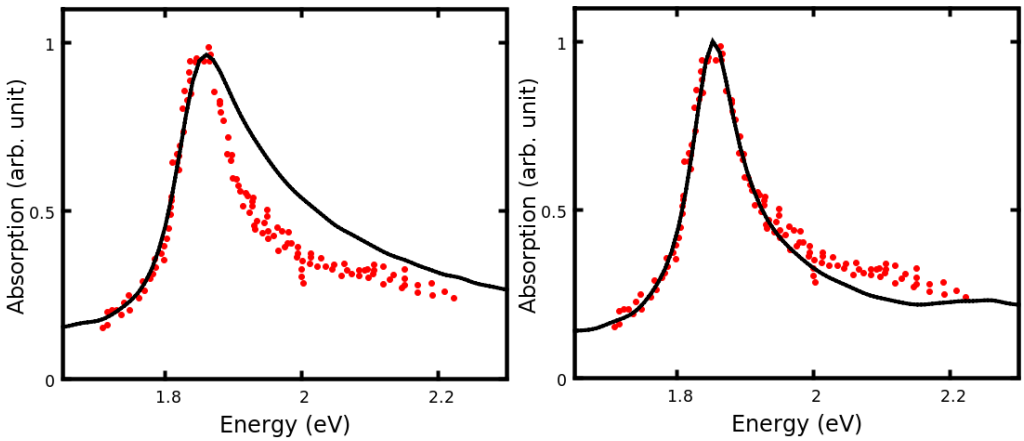

Figure 5.11: Comparison of the *ab intio* absorption spectrum to experiment. The (left) theoretical spectra without excitonic effects and (right) theoretical spectra with excitonic effects are compared to the experiment from Wang. et. al. [145]

(21,21) tube comes from nearly the same k-points in the graphene bandstructure as the E_{11} transition. To further match the (21,21) tube spectra we scaled the electron-hole kernel in order to tune the theoretical binding energy to reach maximum agreement with experiment. Surprisingly, the best agreement was achieved when the Kernel was left unchanged (100% of its original value). We discuss in the next sub-section the origin of this insensitivity to the tube diameter.

With the improvements to the BerkeleyGW package as described in Chapter 2, it is now possible to study directly the optical spectrum of the (21,21) SWCNT. Figure 5.11 shows a comparison of the ab initio optical absorption spectrum of the (21,21) nanotube to experiment. The line shape when excitonic effects (yielding an exciton with binding energy of 40 meV) have been turned on in the calculation agrees remarkably well with the experimental lineshape [145]. Without excitonic effects included however, the spectra disagrees even qualitatively - the non-interacting spectra having a significantly more asymmetrical peak.

5.2.2 Diameter Dependence

Density functional theory calculations using the local density approximation (LDA) described above yield a Fermi velocity for graphene of $\approx 0.82 \cdot 10^6 m/s$ [149]. Comparison of this estimate with experiment ($1.1 \cdot 10^6 m/s$ [101, 153]) suggests that there must be a strong positive renormalization of the graphene Fermi velocity due to electron-electon interactions. Indeed, *ab initio* calculations of the electron self-energy of undoped graphene show an electron-electron positive energy renormalization of 30% and an electron-phonon contribution to the electron self-energy that contributes a reduction in the Fermi Velocity of ≈ 4 percent [105, 149]. This large quasiparticle energy shift is also anticipated in metallic nanotubes of large diameter. Additionally, as we just saw, excitonic effects have been shown to

be strong in metallic SWCNTs [37] and the exciton binding seemingly insensitive to changes in the tube diameter in the small number of tubes studied from first-principles [34]. For these reasons, it is interesting to investigate the diameter dependence of the many-body effects in metallic nanotubes.

In order to study the diameter dependence of the quasiparticle renormalization and exciton binding energies, we use the GW approximation [62] to the electron self-energy to accurately estimate the electron-electron contribution to the Fermi velocity renormalization of graphene. We further carry out similar computations of the band and excitonic renormalization in the armchair (3,3) [128], (5,5), (10,10) and (21,21) single walled carbon nanotubes, whose bands at the Fermi level are derived from cross sections cutting through the center of the graphene Dirac cone.

We calculate the ground state electronic properties within *ab initio* Kohn-Sham density functional theory [76, 83] with a planewave basis using the local density approximation (LDA) to the exchange-correlation potential and using *ab initio* Troullier-Martins pseudopotentials in the Kleinman-Bylander form [139, 74] with a cutoff of $r_c = 1.4$ a.u.. The calculations for graphene are performed in a supercell geometry with a planar separation of 10 Å for graphene and at least 8 Å tube-tube separation in the case of metallic nanotubes. In graphene, this separation was tested by extending the separation to 20 Å and noting the small change in E_{nk}^{QP} for tested k on the Dirac cone. The graphene calculations were performed on a 64x64 k-point grid, while the nanotube calculations were done using a 32 k-point one dimensional grid. For the optical spectroscopy on SWCNT, fine grids of up to 512 k-points are used.

Figure 5.12 shows the LDA bandstructures for the (5,5) and (20,20) metallic nanotubes along the $\Gamma - K$ direction. The LDA values for the Fermi velocities near the Dirac points in each case can be found in Table 5.2. While the (10,10) LDA Fermi velocity agrees well with the graphene Fermi velocity, the smaller diameter tubes have an increasingly smaller Fermi velocity. [24, 112]

An important implementation detail worth discussing is the convergence of the GW corrections for graphene. The bandstructure near the Dirac point converges extremely slowly, showing a small, unphysical energy gap for calculations with too few k-points. This gap is systematically reduced when one converges with respect to k-points, however away from the Dirac point, the spectra remains nearly the same. Convergence is only reached when a k-point sampling of the Brillouin zone greater than a 64x64 grid is used.

The calculated renormalization of the LDA energy dispersion using the GW method is shown in Fig. 5.14 for the (5,5) and (10,10) nanotubes as well as for graphene. Having included the electron-electron contribution to the quasiparticle self-energy, we obtain a linear quasiparticle dispersion near the Dirac point, but with a rescaled Fermi velocity. This energy scaling is largest in the case of graphene which has a calculated renormalization of 30 percent as illustrated in Fig. 5.15. Thus, the calculated Fermi velocity using the GW method is $1.05 \cdot 10^6 m/s$. The effective change in the graphene bandstructure is shown in Figure 3. Shown in Table 5.2 are the corresponding LDA and GW calculated Fermi velocities for the (3,3) [127], (5,5) and (10,10) armchair SWNTs. Similarly to the LDA Fermi Velocity, the GW renormalization factor is reduced as the tube diameter is reduced. In the case of the (3,3) SWCNT, the quasiparticle Fermi Velocity is predicted to be, $0.65 \cdot 10^6 m/s$,

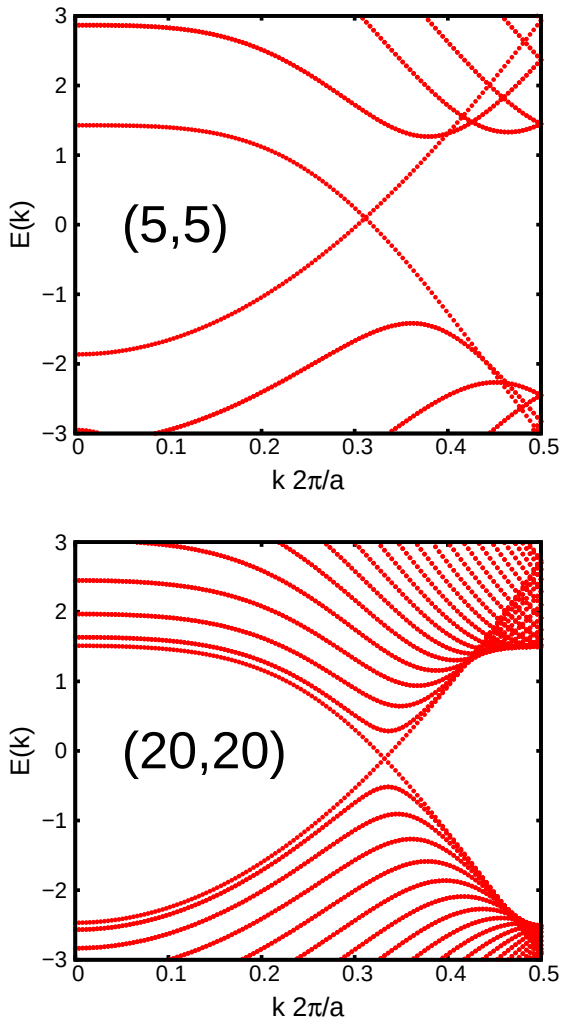

Figure 5.12: LDA bandstructure for the (5,5) (Top) and (20,20) (Bottom) SWCNTs

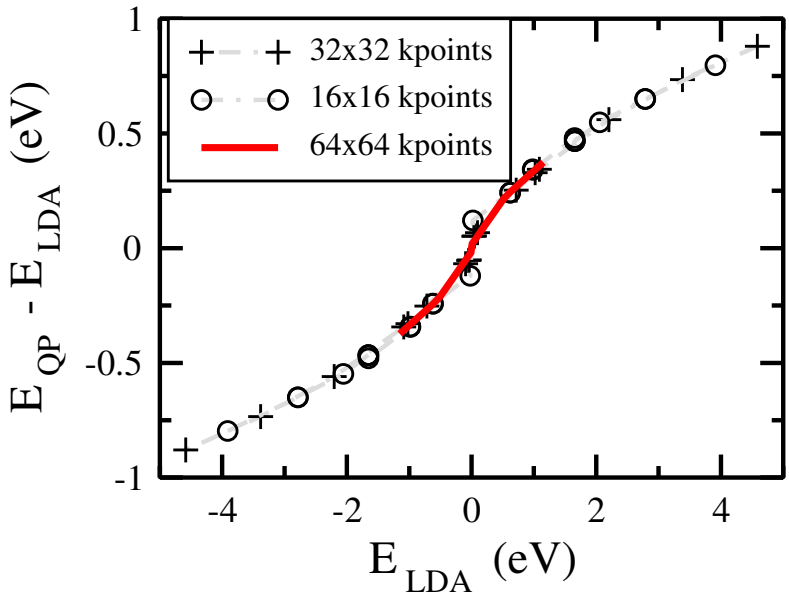

Figure 5.13: Convergence of the GW energy renormalization in graphene with respect to k-point sampling.

which is only two thirds of the measured graphene Fermi velocity. It is important to note that we do not consider the effects of the Luttinger liquid behavior of graphene near the Dirac point within our theory.

The trend in the value of the renormalization with changing tube diameter can be understood in terms of the decrease in effective screening as the tube diameter increases. As effective screening is decreased, the screened-exchange term in the self-energy operator reduces to the bare or Hartree-Fock exchange term which is well known to increase band gaps and band widths, which in turn cause an increase in band velocity. In a Thomas-Fermi approximation for screening, the dielectric function of a quasi-one-dimensional nanotube is:

$$\epsilon(q) = \epsilon_\infty - \chi_M(q) V_{bare}(q) \text{ where } \chi_M(q) \propto -D(E_f) \tag{5.5}$$

System	LDA ($10^6 m/s$)	GW Shift	QP ($10^6 m/s$)
(3,3)	0.56	15%	0.65
(5,5)	0.72	19%	0.85
(10,10)	0.81	24%	1.00
Graphene	0.82	30%	1.05

Table 5.2: Quasiparticle Fermi Velocities

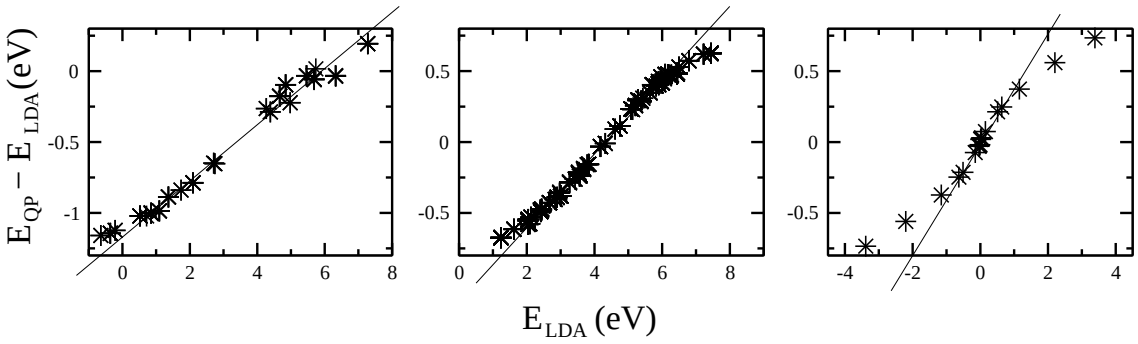

Figure 5.14: Quasiparticle corrections to LDA energies for the (a) (5,5), (b) (10,10) SWCNTs and (c) Graphene.

[145]. Here we have approximated the electron-electron interaction by the interaction between charged rings along the tube diameter, as in Chapter 4, and ϵ_∞ represents the residual screening at distances smaller than the tube diameter where this approximation breaks down. $V_{bare}(q)$ is the bare coulomb interaction from a ring of charge, $V_{bare}(q) = 2I_0\left(\frac{qd}{2}\right)\left(K_0\frac{qd}{2}\right)$, where I_0 and K_0 are the zeroth order modified Bessel functions of the first and second kind. Here $D(E_f)$ is the density of states at the Fermi level. This bare interaction is appropriate for describing the electron-hole interaction since the electron-hole wavefunctions from a single set of bands are delocalized around the tube diameter. As we mention below, for the exchange interaction of two electrons, this approximation is not appropriate, but the qualitative features of the dielectric function are the same for the description of the screening. This density of states at the Fermi energy is approximately constant with changing diameter, decreasing slightly with increased diameter due to the larger LDA Fermi velocities. Thus, the number of free screening electrons per distance around the tube circumference decreases with increasing tube diameter and the screening becomes less effective. The velocity renormalization continues to rise in armchair tubes with increasing nanotube diameter until it reaches the value in graphene, where the normalized $D(E_f)$ is zero

The discrepancy between the computed band velocity of graphene within Density Functional Theory using LDA and experiment can therefore be explained, as anticipated, by the large quasiparticle energy corrections obtained by the *ab initio* GW technique. The GW value for the band velocity (away from the exact Dirac point), $1.05 \cdot 10^6 m/s$ agrees very well with the values obtained in experiments on the quantum Hall effect in graphene [101, 153] even after the inclusion of the smaller than 4 percent negatve band renormalization due to the electron-phonon interaction. The Fermi velocity renormalization of 30 percent is larger than that of $\approx 10 - 15$ percent typically found in metallic systems due to the vanishing density of states at the Fermi energy in graphene. The renormalization in the case of small diameter nanotubes becomes more typical of a metallic system due to the relative increase in magnitude of the density of states at the Fermi energy.

This diameter dependence of the dielectric screening has important consequences

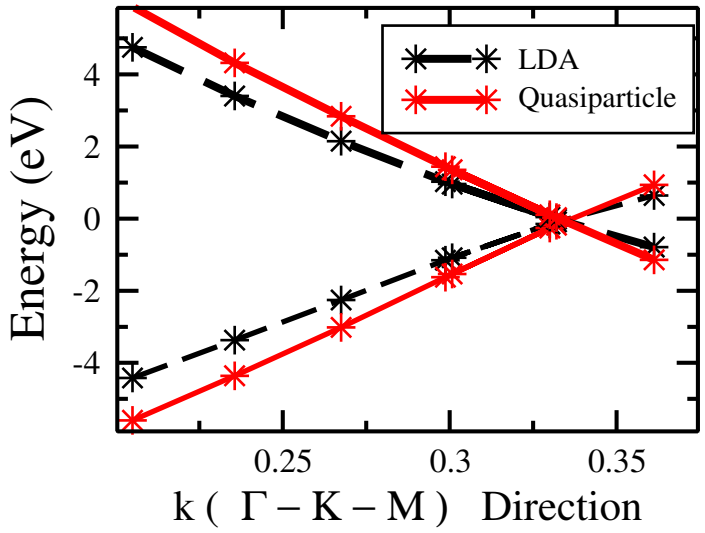

Figure 5.15: Quasiparticle energy renormalization of the graphene bandstructure along the $\Gamma - K - M$ directions.

Tube	$E^{bind}(eV)$
(5,5)	0.07
(10,10)	0.05
(21,21)	0.04

Table 5.3: Exciton binding energy of various metallic nanotubes calculated within the GW-BSE formalism.

in the optical spectra of metallic SWCNTs as well. Table 5.3 shows the diameter dependence of the exciton binding energy in metallic SWCNTs. The exciton binding energy is relatively insensitive to the tube diameter due to a cancellation effect. The bare Coulomb interaction between ring charges (appropriate since the exciton interaction involves only a single set of degenerate bands) is deeper for small diameter tubes, but the interaction is shorter in range due to increased screening.

The net result of the of the GW quasiparticle energy correction and the exciton binding energy dependence on diameter is that optical transitions in larger diameter tubes have a blue-shift compared to the corresponding transitions in smaller diameters tubes. For example, the E_{22} transition of the (20,20) tube has a higher transition energy than the E_{11} transition energy in the (10,10) tube, though they come from the same k-points in the graphene Brillouin zone. The reason why the GW shifts are diameter dependent whereas the exciton-binding energies are insensitive to the diameter is explained when we

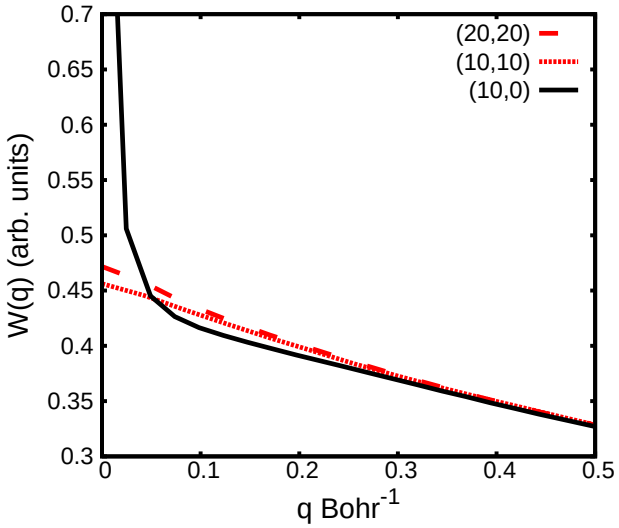

Figure 5.16: Effective screened interaction for two-point particles on the surface of a nanotube for the (20,20), (10,10) and (10,0) SWCNTs.

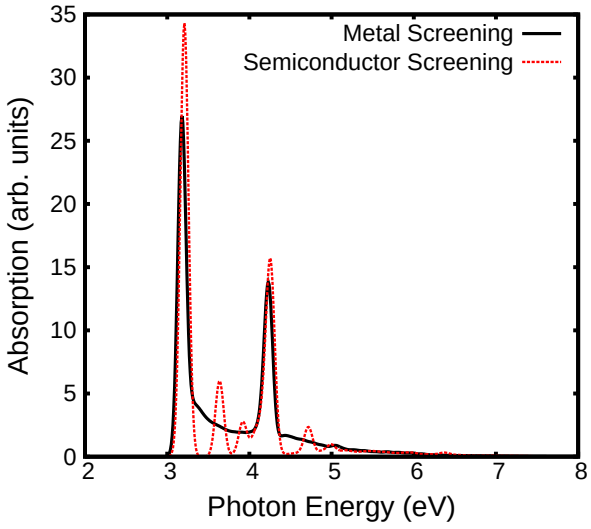

Figure 5.17: The absorption spectra vs. photon energy for the (5,5) tube using the default metallic dielectric matrix $\epsilon(q)$ (black-solid) and using an altered, semiconductor like, dielectric matrix where $\epsilon(q=0)$ is set to 1 (red-dashed).

look at the exchange contribution to the valence band self-energy (the term that contributes most to the bandgap opening). We see that, unlike the excitonic interaction, where only a single pair of bands is involved, the exchange interaction involves many valence bands. This means that the exchange-hole, unlike the excitonic electron-hole amplitude, is localized in the circumferential direction as well as the axial direction. Therefore, in the exchange interaction, treating the electrons as interacting rings, as we did in the previous chapter, is inappropriate. We should instead treat the electrons and holes as point particles on the surface of the tube. In this case, the bare interaction is the same for tubes of different diameters, but the screening remains more efficient for tubes of smaller diameters (there is more screening charge nearby). Thus we get a stronger exchange interaction for larger diameter tubes and hence a larger quasiparticle band-gap opening. This is illustrated in Fig. 5.16 where we compare the effective screened Coulomb interactions for three nanotube systems. Notice that the semiconducting (10,0) interaction only differs from the metallic interaction substantially near $q = 0$, and, outside of this region, the strongest interaction corresponds to the largest diameter tube.

Interestingly, this net blue shift in the optical transition energies of larger diameter metallic tubes also exists for semiconducting tubes. This is because, the reduction in screening in semi-conducting nanotubes leads to both an increase in the quasiparticle gap and an increased exciton binding energy. When going from a metallic tube to a semiconducting with similar diameter, the increased quasiparticle gap and increased exciton binding energy nearly perfectly cancel. To illustrate this, we calculated the absorption spectra for the (5,5) tube normally (i.e., with metallic screening) and with artificial semiconducting screening, achieved by setting $\epsilon(q = 0) = 1$. The results are shown in Figure 5.17. Notice that, though the lineshape changes, the peak position remains relatively unchanged between the two approaches.

The unusual diameter dependence mentioned above is yet one more example of the trend we have seen in this chapter, that, unlike bulk metals, the quasiparticle and optical properties of reduced dimensional metals and semi-metals are greatly affected by many-body interactions.

Bibliography

[1] BerkeleyGW - http://berkeleygw.org/.

[2] BuildBot - http://buildbot.net/.

[3] PARATEC - http://www.nersc.gov/projects/paratec/.

[4] Subversion - http://subversion.tigris.org/.

[5] Trac - http://trac.edgewall.org/.

[6] H. Ajiki and T. Ando. Aharonov-Bohm effect in carbon nanotubes. *Physica B*, 201:349, 1994.

[7] Stefan Albrecht, Lucia Reining, Rodolfo Del Sole, and Giovanni Onida. Ab initio calculation of excitonic effects in the optical spectra of semiconductors. *Phys. Rev. Lett.*, 80(20):4510–4513, May 1998.

[8] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999.

[9] Tsuneya Ando. Excitons in carbon nanotubes. *Journal of the Physical Society of Japan*, 66(4):1066–1073, 1997.

[10] Tsuneya Ando, Yisong Zheng, and Hidekatsu Suzuura. Dynamical conductivity and zero-mode anomaly in honeycomb lattices. *Journal of the Physical Society of Japan*, 71(5):1318–1324, 2002.

[11] G. Arfken and H. Weber. *Mathematical Methods for Physicists*. Harcourt Academic Press, New York, 2001.

[12] F. Aryasetiawan, L. Hedin, and K. Karlsson. Multiple plasmon satellites in Na and Al spectral functions from ab initio cumulant expansion. *Phys. Rev. Lett.*, 77(11):2268–2271, 1996.

[13] C. Attaccalite, A. Grüeneis, T. Pichler, and A. Rubio. Ab-initio band structure of doped graphene. *ArXiv e-prints*, August 2008.

[14] Sergei M. Bachilo, Michael S. Strano, Carter Kittrell, Robert H. Hauge, Richard E. Smalley, and R. Bruce Weisman. Structure-assigned optical spectra of single-walled carbon nanotubes. *Science*, 298(5602):2361–2366, 2002.

[15] A. Baldereschi and E. Tosatti. Mean-value point and dielectric properties of semiconductors and insulators. *Phys. Rev. B*, 17(12):4710–4717, Jun 1978.

[16] V. Barone, J. E. Peralta, and G. E. Scuseria. Optical transitions in metallic single-walled carbon nanotubes. *Nano Letters*, 9:1830–3, 2005.

[17] E. B. Barros, R. B. Capaz, A. Jorio, G. G. Samsonidze, A. G. Souza Filho, S. Ismail-Beigi, C. D. Spataru, S. G. Louie, G. Dresselhaus, and M. S. Dresselhaus. Selection rules for one- and two-photon absorption by excitons in carbon nanotubes. *Phys. Rev. B*, 73:241406, 2006.

[18] Eduardo B. Barros, Ado Jorio, Georgii G. Samsonidze, Rodrigo B. Capaz, Antonio G. Souza Filho, Josue Mendes Filho, Gene Dresselhaus, and Mildred S. Dresselhaus. Review on the symmetry-related properties of carbon nanotubes. *Physics Reports*, 431:261, 2006.

[19] Lorin X. Benedict and Eric L. Shirley. Ab initio calculation of $\epsilon_2(\omega)$ including the electron-hole interaction: Application to GaN and CaF_2. *Phys. Rev. B*, 59(8):5441–5451, 1999.

[20] Lorin X. Benedict, Eric L. Shirley, and Robert B. Bohn. Optical absorption of insulators and the electron-hole interaction: An ab initio calculation. *Phys. Rev. Lett.*, 80(20):4514–4517, May 1998.

[21] Lorin X. Benedict, Catalin D. Spataru, and Steven G. Louie. Quasiparticle properties of a simple metal at high electron temperatures. *Phys. Rev. B*, 66(8):085116, Aug 2002.

[22] J. A. Berger, Lucia Reining, and Francesco Sottile. Ab initio calculations of electronic excitations: Collapsing spectral sums. *Phys. Rev. B*, 82:041103(R), 2010.

[23] L. S. Blackford, J. Choi, A. Cleary, E. D'Azevedo, J. Demmel, I. Dhillon, J. Dongarra, S. Hammarling, G. Henry, A. Petitet, K. Stanley, D. Walker, and R. C. Whaley. *ScaLAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997.

[24] X. Blase, Lorin X. Benedict, Eric L. Shirley, and Steven G. Louie. Hybridization effects and metallicity in small radius carbon nanotubes. *Phys. Rev. B*, 72 no 12:1878, 1994.

[25] Fabien Bruneval and Xavier Gonze. Accurate GW self-energies in a plane-wave basis using only a few empty states: Towards large systems. *Phys. Rev. B*, 78:085125, 2008.

[26] Fabien Bruneval, Nathalie Vast, and Lucia Reining. Effect of self-consistency on quasiparticles in solids. *Phys. Rev. B*, 74:045102, 2006.

[27] K. A. Bulashevich, R. A. Suris, and S. V. Rotkin. Excitons in single-wall carbon nanotubes. *Int. Journ. Nanoscience*, 2 issue 6:561, 2003.

[28] Rodrigo B. Capaz, Catalin D. Spataru, Sohrab Ismail-Beigi, and Steven G. Louie. Diameter and chirality dependence of exciton properties in carbon nanotubes. *Phys. Rev. B*, 74:121401, 2006.

[29] Alberto Castro, Heiko Appel, Micael Oliveira, Carlo A. Rozzi, Xavier Andrade, Florian Lorenzen, M. A. L. Marques, E. K. U. Gross, and Angel Rubio. octopus: a tool for the application of time-dependent density functional theory. *Phys. Status Solidi B*, 243(11):2465–2488, 2006.

[30] D. M. Ceperley and B. J. Alder. Ground state of the electron gas by a stochastic method. *Phys. Rev. Lett.*, 45(7):566–569, Aug 1980.

[31] M. L. Cohen and T. K. Bergstresser. Band structures and pseudopotential form factors for fourteen semiconductors of the diamond and zinc-blende structures. *Phys. Rev.*, 141:789, 1966.

[32] Marvin L. Cohen, M. Schlüter, James R. Chelikowsky, and Steven G. Louie. Self-consistent pseudopotential method for localized configurations: Molecules. *Phys. Rev. B*, 12:5575, 1975.

[33] P. Delaney, H. J. Choi, J. Ihm, S. G. Louie, and M. L. Cohen. Broken symmetry and pseudogaps in ropes of carbon nanotubes. *Phys. Rev. B*, 60:7899–7904, 1999.

[34] Jack Deslippe, Mario Dipoppa, David Prendergast, Marcus V. O. Moutinho, Rodrigo B. Capaz, and Steven G. Louie. Electronhole interaction in carbon nanotubes: Novel screening and exciton excitation spectra. *Nano Letters*, 9(4):1330–1334, 2009.

[35] Jack Deslippe, Georgy Samsonidze, Manish Jain, Marvin L. Cohen, and Steven G. Louie. Converging coulomb-hole summations and absolute energies in traditional GW calculations with limited numbers of conduction bands. unpublished.

[36] Jack Deslippe, Georgy Samsonidze, David Strubbe, Manish Jain, Marvin L. Cohen, and Steven G. Louie. BerkeleyGW: A massively parallel computer package for the calculation of the quasiparticle and optical properties of materials. unpublished.

[37] Jack Deslippe, Catalin D. Spataru, David Prendergast, and Steven G. Louie. Bound excitons in metallic single-walled carbon nanotubes. *Nano Lett.*, 7:1626, 2007.

[38] Aleksandra B. Djurisic and E. Herbert Li. Optical properties of graphite. 85(10):7404–7410, 1999.

[39] Gordana Dukovic, Feng Wang, Daohua Song, Matthew Y. Sfeir, Tony F. Heinz, and Louis E. Brus. Structural dependence of excitonic optical transitions and band-gap energies in carbon nanotubes. *Nano Lett.*, 5 no 11:2314, 2005.

[40] R. J. Elliott, J. A. Krumhansl, and P. L. Leath. The theory and properties of randomly disordered crystals and related physical systems. *Rev. Mod. Phys.*, 46:465–543, 1974.

[41] Sergey V. Faleev, Mark van Schilfgaarde, and Takao Kotani. All-electron self-consistent GW approximation: Application to Si, MnO, and NiO. *Phys. Rev. Lett.*, 93(12):126406, 2004.

[42] C. Fantini, A. Jorio, M. Souza, M. S. Strano, M. S. Dresselhaus, and M. A. Pimenta. Optical transition energies for carbon nanotubes from resonant raman spectroscopy: Environment and temperature effects. *Phys. Rev. Lett.*, 93:147406, 2004.

[43] A. L. Fetter and J. D. Walecka. *Quantum Theory of Many-Body Systems.* McGraw Hill, San Francisco, 1971.

[44] A. Fleszar. PhD thesis, University of Trieste, 1985.

[45] Matteo Frigo and Steven G. Johnson. The design and implementation of FFTW3. *Proceedings of the IEEE*, 93(2):216–231, 2005. Special issue on "Program Generation, Optimization, and Platform Adaptation".

[46] Jr. G. E. Jellison, M. F. Chisholm, and S. M. Gorbatkin. Optical functions of chemical vapor deposited thin-film silicon determined by spectroscopic ellipsometry. *Appl. Phys. Lett.*, 62(25):3348–3350, 1993.

[47] A. K. Geim and K. S. Novoselov. The rise of graphene. *Nature Materials*, 6:183, 2007.

[48] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, Andrea Dal Corso, Stefano de Gironcoli, Stefano Fabris, Guido Fratesi, Ralph Gebauer, Uwe Gerstmann, Christos Gougoussis, Anton Kokalj, Michele Lazzeri, Layla Martin-Samos, Nicola Marzari, Francesco Mauri, Riccardo Mazzarello, Stefano Paolini, Alfredo Pasquarello, Lorenzo Paulatto, Carlo Sbraccia, Sandro Scandolo, Gabriele Sclauzero, Ari P Seitsonen, Alexander Smogunov, Paolo Umari, and Renata M Wentzcovitch. Quantum espresso: a modular and open-source software project for quantum simulations of materials. *Journal of Physics: Condensed Matter*, 21(39):395502, 2009.

[49] F. Giustino, M. L. Cohen, and S. G. Louie. GW method with the self-consistent sternheimer equation. *Phys. Rev. B*, 81:115105, 2010.

[50] R. W. Godby, M. Schlüter, and L. J. Sham. Self-energy operators and exchange-correlation potentials in semiconductors. *Phys. Rev. B*, 37:10159, 1988.

[51] Jeffrey C. Grossman, Michael Rohlfing, Lubos Mitas, Steven G. Louie, and Marvin L. Cohen. High accuracy many-body calculational approaches for excitations in molecules. *Phys. Rev. Lett.*, 86(3):472–475, 2001.

[52] V. P. Gusynin, S. G. Sharapov, and J. P. Carbotte. Unusual microwave response of Dirac quasiparticles in graphene. *Phys. Rev. Lett.*, 96(25):256802, Jun 2006.

[53] P. H. Hahn, W. G. Schmidt, and F. Bechstedt. Molecular electronic excitations calculated from a solid-state approach: Methodology and numerics. *Phys. Rev. B*, 72(24):245425, 2005.

[54] Noriaki Hamada, Shin-ichi Sawada, and Atsushi Oshiyama. New one-dimensional conductors: Graphitic microtubules. *Phys. Rev. Lett.*, 68(10):1579–1581, Mar 1992.

[55] W. Hanke. Dielectric theory of elementary excitations in crystals. *Adv. Phys.*, 27(2):287–341, 1978.

[56] W. Hanke and L. J. Sham. Many-particle effects in the optical spectrum of a semiconductor. *Phys. Rev. B*, 21(10):4656–4673, May 1980.

[57] Roger Haydock. The recursive solution of the Schrodinger equation. *Comput. Phys. Commun.*, 20(1):11 – 16, 1980.

[58] Lars Hedin. New method for calculating the one-particle Green's function with application to the electron-gas problem. *Phys. Rev.*, 139(3A):A796–A823, Aug 1965.

[59] Lars Hedin and Stig Lundqvist. Effects of electron-electron and electron-phonon interactions on the one-electron states of solids. In Frederick Seiz, David Turnbull, and Henry Ehrenreich, editors, *Advances in Research and Applications*, volume 23 of *Solid State Physics*, pages 1 – 181. Academic Press, 1970.

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

[61] B. Holm and U. von Barth. Fully self-consistent GW self-energy of the electron gas. *Phys. Rev. B*, 57(4):2108–2117, 1998.

[62] M. S. Hybertsen and S. G. Louie. Electron correlation in semiconductors and insulators: Band gaps and quasiparticle energies. *Phys. Rev. B*, 34:5390, 1986.

[63] Mark S. Hybertsen and Steven G. Louie. First-principles theory of quasiparticles: Calculation of band gaps in semiconductors and insulators. *Phys. Rev. Lett.*, 55(13):1418–1421, 1985.

[64] Mark S. Hybertsen and Steven G. Louie. Ab initio static dielectric matrices from the density-functional approach. i. formulation and application to semiconductors and insulators. *Phys. Rev. B*, 35(11):5585–5601, Apr 1987.

[65] S. IIjima. Helical microtubules of graphitic carbon. *Nature*, 354(7):56, 1991.

[66] Sohrab Ismail-Beigi. Truncation of periodic image interactions for confined systems. *Phys. Rev. B*, 73:233103, 2006.

[67] Uichi Itoh, Yasutake Toyoshima, Hideo Onuki, Nobuaki Washida, and Toshio Ibuki. Vacuum ultraviolet absorption cross sections of sih[sub 4], geh[sub 4], si[sub 2]h[sub 6], and si[sub 3]h[sub 8]. *J. Chem. Phys.*, 85(9):4867–4872, 1986.

[68] Manish Jain, Jack Deslippe, Georgy Samsonidze, Marvin L. Cohen, and Steven G. Louie. G_0W_0 diagonalization using the static COHSEX approximation. unpublished.

[69] J. F. Janak. Proof that $\frac{\partial e}{\partial n_i} = \epsilon$ in density-functional theory. *Phys. Rev. B*, 18(12):7165–7168, Dec 1978.

[70] A. Jorio, C. Fantini, M. A. Pimenta, R. B. Capaz, G. G. Samsonidze, G. Dresselhaus, J. Jiang M. S. Dresselhaus, N. Kobayashi, A. Grneis, and R. Saito. Resonance raman spectroscopy (n,m)-dependent effects in small-diameter single-wall carbon nanotubes. *Phys. Rev. B*, 71:075401, 2005.

[71] C. L. Kane and E. J. Mele. Ratio problem in single carbon nanotube fluorescence spectroscopy. *Phys. Rev. Lett.*, 90:207401, 2003.

[72] Wei Kang and Mark S. Hybertsen. Enhanced static approximation to the electron self-energy operator for efficient calculation of quasiparticle energies. *arXiv:1008.4320v1*, 2010.

[73] M. I. Katsnelson. Optical properties of graphene: The Fermi-liquid approach. *EPL (Europhysics Letters)*, 84(3):37001, 2008.

[74] Leonard Kleinman and D. M. Bylander. Efficacious form for model pseudopotentials. *Phys. Rev. Lett.*, 48:1425, 1982.

[75] C. A. Kocher. Criteria for bound-state solutions in quantum mechanics. *Am. J. Phys.*, 45:71–74, 1977.

[76] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133, 1965.

[77] V. G. Kravets, A. N. Grigorenko, R. R. Nair, P. Blake, S. Anissimova, K. S. Novoselov, and A. K. Geim. Spectroscopic ellipsometry of graphene and an exciton-shifted van hove peak in absorption. *Phys. Rev. B*, 81(15):155413, Apr 2010.

[78] Landolt-Börnstein. *Numerical Data and Functional Relationships in Science and Technology*, volume 17, Pt A. of *New Series Group III*. Springer, New-York, 1982.

[79] S. Lebègue, B. Arnaud, M. Alouani, and P. E. Bloechl. Implementation of an all-electron GW approximation based on the projector augmented wave method without plasmon pole approximation: Application to Si, SiC, AlAs, InAs, NaH, and KH. *Phys. Rev. B*, 67:155208, 2003.

[80] J. Lefebvre and P. Finnie. Polarized photoluminescence excitation spectroscopy of single-walled carbon nanotubes. *Phys. Rev. Lett.*, 98:167406, 2007.

[81] J. Lefebvre and P. Finnie. Excited excitonic states in single-walled carbon nanotubes. *Nano Lett.*, 8 no 7:1890, 2008.

[82] Francois Leonard and J. Tersoff. Dielectric response of semiconducting carbon nanotubes. *Appl. Phys. Lett.*, 81 no 25:4835, 2002.

[83] S. G. Louie. *Conceptual Foundations of Materials: A standard model for ground- and excited-state properties*. Contemporary Concepts of Condensed Matter Science. Elsevier, 2006.

[84] S.G. Louie. In C. Y. Fong, editor, *Topics in Computational Materials Science*. World Scientific, Singapore, 1997.

[85] G. D. Mahan. *Many-Particle Physics*. Plenum, New York, 1981.

[86] Kin Fai Mak, Matthew Y. Sfeir, Yang Wu, Chun Hung Lui, James A. Misewich, and Tony F. Heinz. Measurement of the optical conductivity of graphene. *Phys. Rev. Lett.*, 101(19):196405, Nov 2008.

[87] Kin Fai Mak, Jie Shan, and Tony F. Heinz. Seeing many-body effects in single- and few-layer graphene: Observation of two-dimensional saddle-point excitons. *Phys. Rev. Lett.*, 106(4):046401, Jan 2011.

[88] Miguel A. L. Marques, Alberto Castro, George F. Bertsch, and Angel Rubio. Octopus: a first-principles tool for excited electron-ion dynamics. *Comput. Phys. Commun.*, 151(1):60 – 78, 2003.

[89] Richard M. Martin. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, 2004.

[90] J. Maultzsch, R. Pomraenke, S. Reich, E. Chang, D. Prezzi, A. Ruini, E. Molinari, M. S. Strano, C. Thomsen, and C. Lienau. Exciton binding energies in carbon nanotubes from two-photon photoluminescence. *Phys. Rev. B*, 72:241402, 2005.

[91] J. W. Mintmire, B. I. Dunlap, and C. T. White. Are fullerene tubules metallic? *Phys. Rev. Lett.*, 68(5):631–634, Feb 1992.

[92] Zeeya Mirali. ...error...why scientific programming does not compute. *Nature*, 467:775 – 777, 2010.

[93] T. Miyake and F. Aryasetiawan. Efficient algorithm for calculating noninteracting frequency-dependent linear response functions. *Phys. Rev. B*, 61:7172, 2000.

[94] Hendrik J. Monkhorst and James D. Pack. Special points for Brillouin-zone integrations. *Phys. Rev. B*, 13:5188, 1976.

[95] R. R. Nair, P. Blake, A. N. Grigorenko, K. S. Novoselov, T. J. Booth, T. Stauber, N. M. R. Peres, and A. K. Geim. Fine structure constant defines visual transparency of graphene. *Science*, 320(5881):1308, 2008.

[96] J. B. Neaton, Mark S. Hybertsen, and Steven G. Louie. Renormalization of molecular electronic levels at metal-molecule interfaces. *Phys. Rev. Lett.*, 97(21):216405, 2006.

[97] E. J. Nicol and J. P. Carbotte. Optical conductivity of bilayer graphene with and without an asymmetry gap. *Phys. Rev. B*, 77(15):155409, Apr 2008.

[98] Risto Nieminen. Supercell methods for defect calculations. In David Drabold and Stefan Estreicher, editors, *Theory of Defects in Semiconductors*, volume 104 of *Topics in Applied Physics*, pages 29–68. Springer Berlin / Heidelberg, 2007.

[99] John E. Northrup, Mark S. Hybertsen, and Steven G. Louie. Theory of quasiparticle energies in alkali metals. *Phys. Rev. Lett.*, 59(7):819–822, 1987.

[100] John E. Northrup, Mark S. Hybertsen, and Steven G. Louie. Quasiparticle excitation spectrum for nearly-free-electron metals. *Phys. Rev. B*, 39(12):8198–8208, 1989.

[101] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, M. I. Katsnelson, I. V. Grigorieva, S. V. Dubonos, and A. A. Firsov. Two-dimensional gas of massless Dirac fermions in graphene. *Nature*, 438 no 7065:197, 2005.

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

[103] Giovanni Onida, Lucia Reining, and Angel Rubio. Electronic excitations: density-functional versus many-body Green's-function approaches. *Rev. Mod. Phys.*, 74:601–659, 2002.

[104] Pablo Ordejón, Emilio Artacho, and José M. Soler. Self-consistent order-N density-functional calculations for very large systems. *Phys. Rev. B*, 53:R10441, 1996.

[105] Cheol-Hwan Park, Feliciano Giustino, Marvin L. Cohen, and Steven G. Louie. Velocity renormalization and carrier lifetime in graphene from the electron-phonon interaction. *Phys. Rev. Lett.*, 99(8):086804, Aug 2007.

[106] David R. Penn. Wave-number-dependent dielectric function of semiconductors. *Phys. Rev.*, 128 no 5:2093, 1962.

[107] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865, 1996.

[108] John P. Perdew and Alex Zunger. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B*, 23:5048, 1981.

[109] N. M. R. Peres, F. Guinea, and A. H. Castro Neto. Electronic properties of disordered two-dimensional carbon. *Phys. Rev. B*, 73(12):125411, Mar 2006.

[110] Deborah Prezzi, Daniele Varsano, Alice Ruini, Andrea Marini, and Elisa Molinari. Optical properties of graphene nanoribbons: The role of many-body effects. *Phys. Rev. B*, 77(4):041404, Jan 2008.

[111] Su Ying Quek, David A. Strubbe, Hyoung Joon Choi, Steven G. Louie, and J. B. Neaton. First-principles approach to charge transport in single-molecule junctions with self-energy corrections. unpublished.

[112] S. Reich, C. Thomsen, and P. Ordejon. Electronic band structure of isolated and bundled carbon nanotubes. *Phys. Rev. B*, 65:155411, 2002.

[113] L. Reining, G. Onida, and R. W. Godby. Elimination of unoccupied-state summations in ab initio self-energy calculations for large supercells. *Phys. Rev. B*, 56:R4301, 1997.

[114] Patrick Rinke, Abdallah Qteish, Jörg Neugebauer, Christoph Freysoldt, and Matthias Scheffler. Combining GW calculations with exact-exchange density-functional theory: an analysis of valence-band photoemission for compound semiconductors. *New J. Phys.*, 7:126, 2005.

[115] Michael Rohlfing and Steven G. Louie. Electron-hole excitations and optical spectra from first principles. *Phys. Rev. B*, 62:4927, 2000.

[116] H. N. Rojas, R. W. Godby, and R. J. Needs. Space-time method for ab initio calculations of self-energies and dielectric response functions of solids. *Phys. Rev. Lett.*, 74:1827, 1995.

[117] R. Saito, G. Dresselhaus, and M. S. Dresselhaus. *Physical Properties of Carbon Nanotubes*. Imperial College Press, London, 1999.

[118] R. Saito, M. Fujita, G. Dresselhaus, and M. S Dresselhaus. Electronic structure of chiral graphene tubules. *Applied Physics Letters*, 60(18):2204 –2206, may 1992.

[119] Georgy Samsonidze, Manish Jain, Jack Deslippe, Marvin L. Cohen, and Steven G. Louie. Simple approximate physical orbitals for GW quasiparticle calculations. unpublished.

[120] M. Y. Sfeir, F. Wang, L. Huang, C. C. Chuang, J. Hone, S. P. O'Brien, T. F. Heinz, and L. E. Brus. Probing electronic transitions in individual carbon nanotubes by Rayleigh scattering. *Science*, 306 no. 5701:1540–1543, 2004.

[121] L. J. Sham and M. Schlüter. Density-functional theory of the energy gap. *Phys. Rev. Lett.*, 51(20):1888–1891, Nov 1983.

[122] Bi-Ching Shih, Yu Xue, Peihong Zhang, Marvin L. Cohen, and Steven G. Louie. Quasiparticle band gap of zno: High accuracy from the conventional G0W0 approach. *Phys. Rev. Lett.*, 105:146401, 2010.

[123] M. Shishkin and G. Kresse. Implementation and performance of the frequency-dependent GW method within the PAW framework. *Phys. Rev. B*, 74:035101, 2006.

[124] B. Simon. The bound state of weakly coupled Schrodinger operators in one and two dimensions. *Ann. Phys.*, 97:279–288, 1976.

[125] J. C. Slater. A simplification of the Hartree-Fock method. *Phys. Rev.*, 81(3):385–390, Feb 1951.

[126] J. M. Soler, E. Artacho, J. D. Gale, A. Garcia, J. Junquera, P. Ordejon, and D. Sanchez-Portal. The SIESTA method for ab initio order-N materials simulation. *J. Phys.: Condens. Matt.*, 14:2745, 2002.

[127] C. D. Spataru, S. Ismail-Beigi, L. X. Benedict, and S. G. Louie. Excitonic effects and optical spectra of single-walled carbon nanotubes. *Phys. Rev. Lett.*, 92:077402, 2004.

[128] C. D. Spataru, S. Ismail-Beigi, L .X. Benedict, and S .G. Louie. Quasiparticle energies, excitonic effects and optical absorption spectra of small-diameter single-walled carbon nanotubes. *Appl. Phys. A*, 78:1129, 2004.

[129] Catalin-Dan Spataru. *Electron excitations in solids and novel materials*. PhD thesis, University of California, Berkeley, 2004.

[130] T. Stauber, N. M. R. Peres, and A. K. Geim. Optical conductivity of graphene in the visible region of the spectrum. *Phys. Rev. B*, 78(8):085432, Aug 2008.

[131] L. Steinbeck, A. Rubio, L. Reining, M. Torrent, I. D. White, and R. W. Godby. Enhancements to the GW space-time method. *Comput. Phys. Commun.*, 125:105, 2000.

[132] G. Strinati. Application of the Green's functions method to the study of the optical properties of semiconductors. *Riv. Nuovo Cimento*, 11:1, 1988.

[133] E. A. Taft and H. R. Philipp. Optical properties of graphite. *Phys. Rev.*, 138(1A):A197–A202, Apr 1965.

[134] Chenggang Tao, Jibin Sun, Xiaowei Zhang, Ryan Yamachika, Daniel Wegner, Yasaman Bahri, Georgy Samsonidze, Marvin L. Cohen, Steven G. Louie, T. Don Tilley, Rachel A. Segalman, and Michael F. Crommie. Spatial resolution of a type ii heterojunction in a single bipolar molecule. *Nano Lett.*, 9:3963, 2009.

[135] H. Telg, J. Maultzsch, S. Reich, F. Hennrich, and C. Thomsen. Chirality distribution and transition energies of carbon nanotubes. *Phys. Rev. Lett.*, 93:177401, 2004.

[136] Murilo L. Tiago and James R. Chelikowsky. Optical excitations in organic molecules, clusters, and defects studied by first-principles Green's function methods. *Phys. Rev. B*, 73:205334, 2006.

[137] John S. Toll. Causality and the dispersion relation: Logical foundations. *Phys. Rev.*, 104:1760, 1956.

[138] Paolo E. Trevisanutto, Christine Giorgetti, Lucia Reining, Massimo Ladisa, and Valerio Olevano. Ab initio GW many-body effects in graphene. *Phys. Rev. Lett.*, 101(22):226405, Nov 2008.

[139] N. Troullier and José Luriaas Martins. Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B*, 43:1993, 1991.

[140] P. Umari, G. Stenuit, and S. Baroni. Optimal representation of the polarization propagator for large-scale GW calculations. *Phys. Rev. B*, 79:201104(R), 2009.

[141] P. Umari, G. Stenuit, and S. Baroni. GW quasiparticle spectra from occupied states only. *Phys. Rev. B*, 81:115104, 2010.

[142] J. van den Brink and G. A. Sawatzky. Non-conventional screening of the coulomb interaction in low-dimensional and finite-size systems. *Electronic Properties of Novel Materials: Progress in Molecular Nanostructures*, 1998.

[143] J. van den Brink and G. A. Sawatzky. Non-conventional screening of the coulomb interaction in low-dimensional and finite-size systems. *Europhys. Journal*, 50 issue 4:447, 2000.

[144] M. van Schilfgaarde, Takao Kotani, and S. Faleev. Quasiparticle self-consistent GW theory. *Phys. Rev. Lett.*, 96:226402, 2006.

[145] Feng Wang, David Cho, Brian Kessler, Jack Deslippe, P. James Schuck, Steven G. Louie, Alex Zettl, Tony F. Heinz, and Y. Ron Shen. Observation of excitons in one-dimensional metals. *Phys. Rev. Lett.*, 2007.

[146] Feng Wang, Gordana Dukovic, Louis E. Brus, and Tony F. Heinz. The optical resonances in carbon nanotubes arise from excitons. *Science*, 308 no 5723:838, 2005.

[147] Feng Wang, Yuanbo Zhang, Chuanshan Tian, Caglar Girit, Alex Zettl, Michael Crommie, and Y. Ron Shen. Gate-variable optical transitions in graphene. *Science*, 320(5873):206–209, 2008.

[148] Li Yang, Marvin L. Cohen, and Steven G. Louie. Excitonic effects in the optical spectra of graphene nanoribbons. *Nano Letters*, 7(10):3112–3115, 2007.

[149] Li Yang, Jack Deslippe, Cheol-Hwan Park, Marvin L. Cohen, and Steven G. Louie. Excitonic effects on the optical response of graphene and bilayer graphene. *Phys. Rev. Lett.*, 103(18):186802, Oct 2009.

[150] Chao Zhang, Lei Chen, and Zhongshui Ma. Orientation dependence of the optical spectra in graphene at high frequencies. *Phys. Rev. B*, 77(24):241402, Jun 2008.

[151] G. P. Zhang, David A. Strubbe, Steven G. Louie, and Thomas F. George. First-principles prediction of optical second-harmonic generation in the endohedral N@$C_6$0 compound. *Phys. Rev. A*, 00:in press, 2011.

[152] S. B. Zhang, C.-Y. Yeh, and A. Zunger. Electronic structure of semiconductor quantum films. *Phys. Rev. B*, 48:11204, 1993.

[153] Yuanbo Zhang, Yan-Wen Tan, Horst Stormer, and Philip Kim. Experimental observation of the quantum hall effect and berry's phase in graphene. *Nature*, 438 no 7065:201, 2005.

[154] Xuejun Zhu and Steven G. Louie. Quasiparticle band structure of thirteen semiconductors and insulators. *Phys. Rev. B*, 43(17):14142–14156, 1991.

Appendix A

BerkeleyGW Additional Details

A.1 Parallelization and Performance

A.1.1 `epsilon`

The parallelization of `epsilon` is characterized by two distinct schemes for the two main sections of the code: 1) the computation of matrix elements (Eq. 2.9), and 2) the matrix multiplication (Eq. 2.15) and inversion.

For the computation of the matrix elements in Eq. 2.9, the code is parallelized with nearly linear scaling up to $N_v \cdot N_c$ processors, where N_v and N_c are the number of valence and conduction bands respectively used in the sum of Eq. 2.8. Each processor owns an approximately equal fraction of the total number of (v, c) pairs for all $\mathbf{k}$ – and performs serial FFTs to compute the matrix elements, Eq. 2.9, for all $\mathbf{G}$ and $\mathbf{k}$ associated with the pair. Note that for large systems, N_v is on the order of 100s and N_c is on the order of 1000s or more; so that this section of the code scales well up to 100,000 CPUs.

All wavefunctions are stored in memory unless the optional `comm_disk` flag is given. Each processor holds in memory the wavefunctions for all the pairs it owns. If `comm_disk` is specified (as opposed to the default `comm_mpi` option), the distribution of pairs is the same, but each processor saves the conduction wavefunctions it needs on disk and reads the wavefunctions back into memory one pair at time for the purposes of computation. Using `comm_disk` can therefore reduce the amount of memory required for the computation, but comes with a substantial performance reduction.

The processors are distributed out into valence and conduction band pools in order to minimize the memory required using a complete search algorithm. For example, if one calculates a material with few valence and many conduction bands, all the processors will be in one or two valence pools (holding all or half the valence bands in memory each) but spread over a large number conduction pools because the relative cost of holding all the valence bands in memory is much smaller than holding all the conduction bands in memory. In such as scheme, the amount of memory required per processor drops linearly with small numbers of processors and decreases as $1/\sqrt{N_{\text{proc}}}$ for large number of processors (Fig. A.1).

In the second section of the `epsilon` code, we switch from a parallelization over bands to parallelization scheme over $\mathbf{G}\mathbf{G}'$ for the polarizability (Eq. 2.8) and dielectric matrices (Eq. 2.10). We use the ScaLAPACK block-cyclic layout [23] in anticipation of

utilizing the ScaLAPACK libraries for the inversion of the dielectric matrix. The transition between the band distribution of the matrix elements in Eq. 2.9 and the block-cyclic layout of the polarizability matrix is achieved naturally in the process of doing the parallel matrix-matrix multiplication involved in Eq. 2.15. There is, however, a significant amount of communication involved at this step. To minimize this communication we have two options for the parallel multiplication.

In the first scheme, corresponding to the `gcomm_matrix` flag in `epsilon.inp`, we loop over processors (for simplicity, we label the loop index i) who own a piece of the polarizability matrix $\chi(\mathbf{G},\mathbf{G}')$. Each processor does the fraction of the matrix multiplication in Eq. 2.15 relating to the (n,n') pairs it owns and for submatrix $\chi(\mathbf{G}_i,\mathbf{G}'_i)$ that the ith processor stores. The processors then MPI-reduce their contribution to the ith processor. Thus the total communication in this scheme is an eventual reduction of the entire $\chi(\mathbf{G},\mathbf{G}')$ to the processors that store it. It is important to note that we must do this one processor at time (or at most in chunks of processors – chosen often as the number of CPUs per node) because no single processor can hold in memory the entire $\chi(\mathbf{G},\mathbf{G}')$ matrix for large systems.

In the second scheme, corresponding to the `gcomm_elements` flag in `epsilon.inp`, we again loop over processors, but this time, we have the ith processor MPI-broadcast to all processors that hold a piece of the polarizability matrix the set of matrix elements for all the (v,c) pairs it owns. Each processor then uses these matrix elements to compute the contribution of the matrix-matrix product, Eq. 2.15, for the submatrix of $\chi(\mathbf{G},\mathbf{G}')$ it stores. In this scheme, all the matrix elements (Eq. 2.6) are eventually broadcast.

Whether the use of the `gcomm_elements` flag or the `gcomm_matrix` flag is optimal depends on whether it is faster to reduce the $\chi(\mathbf{G},\mathbf{G}';\omega)$ or broadcast all the matrix elements, $M_{nn'}(\mathbf{k},\mathbf{q},\{\mathbf{G}\})$. In particular if $N_{\mathbf{G}}\cdot N_{\text{freq}} < N_v\cdot N_c\cdot N_k$, where N_{freq} is the number of frequencies in a full frequency calculation, then it is cheaper to use `gcomm_matrix`. If no flag is specified, the `epsilon` code will make this choice for the user based on the above criteria.

Because we use the block-cyclic layout, the memory required to store the $\chi(\mathbf{G},\mathbf{G}')$ decreases linearly with the number of CPUs. However, the cost of the inversion utilizing ScaLAPACK can saturate at 100s of CPUs and the cost of the summation can saturate with a few thousand CPUs, see Figure A.2. In general, the number of CPUs used for the block-cyclic distribution of χ can be tuned.

Beyond the more sophisticated level of parallelization described above, there is a more trivial level of parallelization available to small systems requiring large numbers of $\mathbf{k}$-points: Eq. 2.8 is completely separable as a function of $\mathbf{q}$. One may run a separate `epsilon` calculation for each $\mathbf{q}$ required and merge the dielectric matrices – in such a way, a user can obtain perfectly linear artificial scaling with CPUs to N_k times the number CPUs mentioned above.

The scaling of memory and computation time with respect to the number of CPUs used per $\mathbf{q}$-point in `epsilon` for the example (20,20) SWCNT calculation is shown in Fig. A.1 and Fig A.2. We find nearly linear scaling up to 3200 CPUs per $\mathbf{q}$-point. Since there are 32 $\mathbf{q}$-points in this calculation that are trivially parallelized, we find nearly linear scaling of the `epsilon` computation up to 100,000 CPUs.

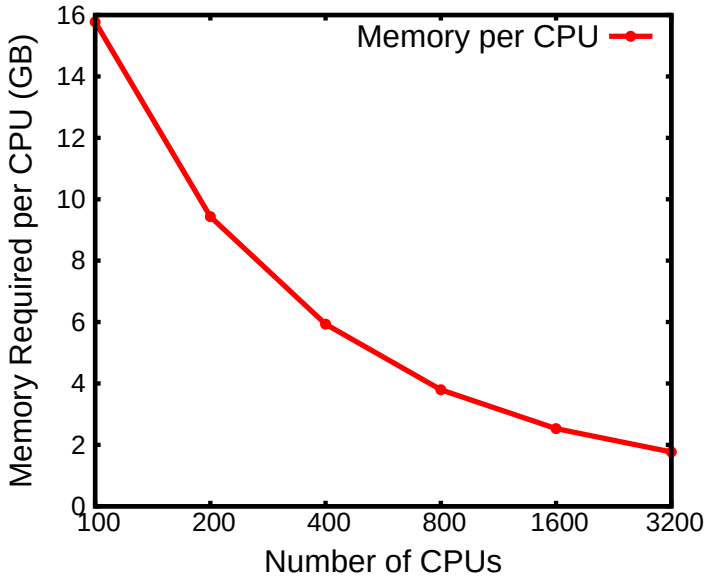

Figure A.1: The memory required per CPU vs. the number of CPUs used for a `epsilon` calculation on the (20,20) nanotube. See text for parameters used.

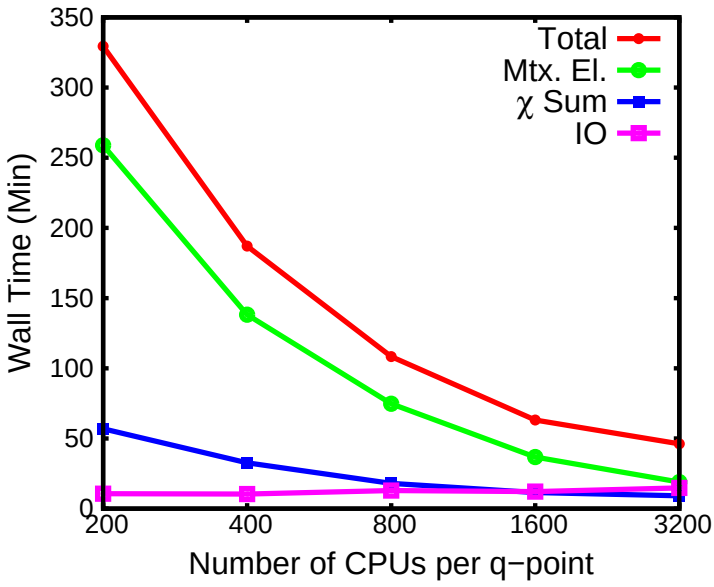

Figure A.2: The wall-time required vs. the number of CPUs per **q**-point used for a `epsilon` calculation on the (20,20) single-walled carbon nanotube. There is near linear scaling up to 1,600 CPUs. Since there is an additional layer of trivial parallelization over the 32 **q**-points required, the `epsilon` calculation scales to over 50,000 CPUs. See text for parameters used.

A.1.2 `sigma`

Within a `sigma` calculation, one computes a requested number of diagonal, Eq. 2.19, or off-diagonal, Eq. 2.18, Σ matrix elements. For each matrix element there are two computationally intensive steps. The first is to calculate all the plane-wave matrix elements $M_{nn''}$ and $M_{n'n''}$, Eq. 2.6, for the outer states of interest, n and n', and for all occupied and empty states, n''. Secondly, we compute the sum over states, n'', as well as $\mathbf{G}$, $\mathbf{G}'$ and $\mathbf{q}$ in the expressions in Eqs. 2.20 – 2.29.

The `sigma` execution is parallelized over both outer bands, n and n', and inner bands, n''. As in the case of `epsilon` this is done by defining pools and distributing the n, n' pairs evenly among the pools. We then distribute the n'' bands evenly within the pools. As was the case for `epsilon`, we define the number of pools using a complete search algorithm to minimize the amount of memory per CPU required to store the inner and outer wavefunctions.

As described above, the CPU time required for the computation of all plane-wave matrix elements, $M_{nn''}$ and $M_{n'n''}$, scales as $N^2 \log N$, where N is the number of atoms, for each Σ matrix element of interest. As described above, the outer-state pairs are parallelized over pools and the inner states are parallelized over the CPUs within each pool. The wall-time for the computation of all the plane-wave matrix elements required for every Σ matrix element scales as $N \log N$ with unlimited CPU resources. As was the case in the `epsilon` executable, each CPU computes the plane-wave matrix elements between all the (n, n'') and (n', n'') pairs it owns for all $\mathbf{G}$ through serial FFTs using FFTW [45].

The summations required in Eqs. 2.20 – 2.29 are parallelized by again distributing the outer-state pairs over the pools and then distributing the inner states over the CPUs within each pool. The wall-time for the summations, therefore, scales as N^2 (for the sums over $\mathbf{G}$ and $\mathbf{G}'$) regardless of the number of diagonal or off-diagonal elements requested, given unlimited CPU resources.

As was the case for `epsilon`, the wavefunctions are distributed in memory, with each CPU owning only the n, n' and n'' wavefunctions that it needs for the computations described above. The dielectric matrix, $\epsilon^{-1}_{\mathbf{G},\mathbf{G}'}(\mathbf{q}; E)$ for each $\mathbf{q}$ and E, is distributed globally over the matrix rows, $\mathbf{G}$.

The scaling of memory and computation time with respect to the number of CPUs used per $\mathbf{k}$-point in `sigma` for the example (20,20) SWCNT calculation is shown in Fig. A.3 and Fig A.4. We find nearly linear scaling up to 1600 CPUs per $\mathbf{k}$-point. Since there are 16 irreducible $\mathbf{k}$-points in this calculation that are trivially parallelized, we find nearly linear scaling of the `sigma` computation up to 25,000 CPUs.

A.1.3 `BSE`

As mentioned in the above sections, in the `kernel` executable, for each $\mathbf{k}$ and $\mathbf{k}'$, we must calculate all the matrix elements $M_{vv'}$, $M_{cc'}$, and M_{vc} and then perform the summations involved Eq. 2.35 and Eq. 2.36 for each $(vc\mathbf{k}, v'c'\mathbf{k}')$ pair. BerkeleyGW automatically parallelizes this in different schemes depending on the system and number of CPUs provided.

If the number of CPUs is less than N_k^2, the square of the number of coarse $\mathbf{k}$-points,

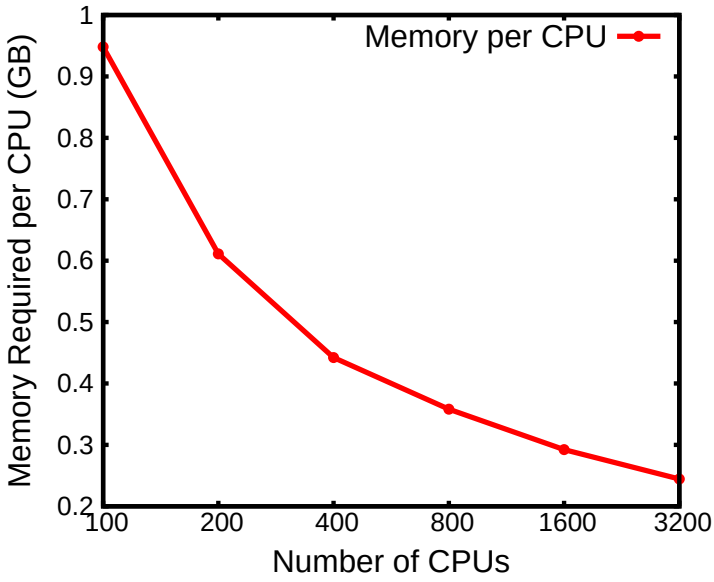

Figure A.3: The memory required per CPU vs. the number of CPUs used for a `sigma` calculation on the (20,20) nanotube. See text for parameters used.

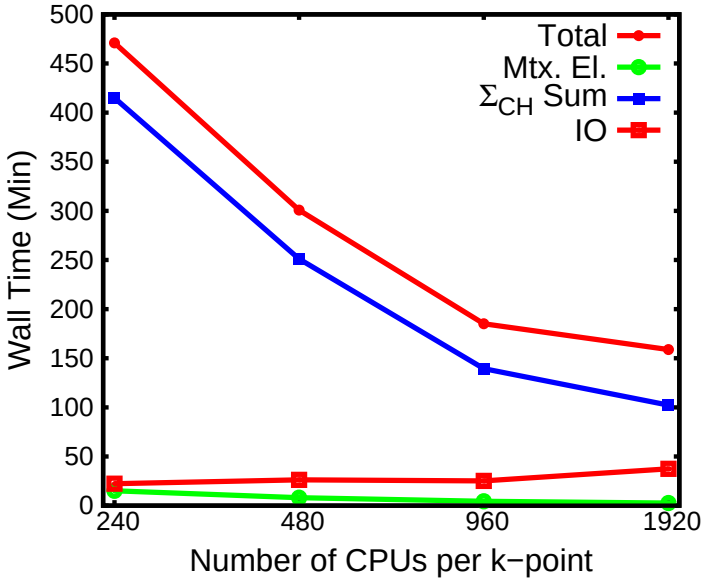

Figure A.4: The wall-time required vs. the number of CPUs per **k**-point used for a `sigma` calculation on the (20,20) single-walled carbon nanotube. There is near linear scaling up to 1,920 CPUs. Since there is an additional layer of trivial parallelization over the 16 **k**-points required, the `epsilon` calculation scales to over 30,000 CPUs. See text for parameters used.

we distribute out the $(\mathbf{k}, \mathbf{k}')$ pairs evenly over the CPUs and each CPU calculates all the matrix elements, $M_{vv'}$, $M_{cc'}$, and M_{vc}, required for the **k**-point pairs it owns through serial FFTs. It then computes the sums in Eq. 2.35 and Eq. 2.36 for all of its pairs.

If the number of CPUs is greater than N_k^2, as is often the case for large systems and molecules, but less than $N_k^2 \cdot N_c^2$, we distribute the $(c\mathbf{k}, c'\mathbf{k}')$ pairs out evenly among the processors; first distributing the processors evenly over **k**-point pairs and then creating pools to distribute the (c, c') evenly among the pools. In this scheme, each CPU computes $M_{cc'}$ for only the $(c\mathbf{k}, c'\mathbf{k}')$ it owns but computes $M_{vv'}$ and M_{vc} for all v and v' at each $(c\mathbf{k}, c'\mathbf{k}')$ pairs it owns. Each CPU does the summations in Eq. 2.35 and Eq. 2.36 for all v, v' for the $(c\mathbf{k}, c'\mathbf{k}')$ it owns.

If the number of CPUs is greater than $N_k^2 \cdot N_c^2$, we distribute the entire set of $(vc\mathbf{k}, v'c'\mathbf{k}')$ pairs out evenly among the processors; first distributing the processors evenly over $(c\mathbf{k}, c'\mathbf{k}')$ pairs and then creating pools to distribute the (v, v') evenly among the pools. In this scheme, each CPU computes only the $M_{vv'}$, $M_{cc'}$, and M_{vc} for the $(vc\mathbf{k}, v'c'\mathbf{k}')$ pairs it owns. Additionally, each CPU does the summations in Eq. 2.35 and Eq. 2.36 for only the pairs it owns.

In the last scheme, the calculation of the matrix elements has a parallel wall-time scaling of N and the summations scale as N^2 (accounting for the sum over $\mathbf{GG}'$) if a limitless number of CPU resources is assumed.

The large arrays that must be stored in memory are the dielectric matrix, the wavefunctions and the computed kernel itself. The computed kernel is distributed evenly in memory among the processors and computed directly by the CPUs who own the various $(vc\mathbf{k}, v'c'\mathbf{k}')$ pairs. The dielectric matrix is distributed, as in the `sigma` executable, over its rows **G** and must be broadcast during each calculation of the sums in Eq. 2.35 and Eq. 2.36. It is for the purposes of minimizing the communication of the dielectric matrix that we use the three different parallelization schemes above – *i.e.*, so that we might work on the biggest blocks of (v, v') and (c, c') at once. For example, in the first scheme, where the number of CPUs is less than N_k^2, we do the sums in Eq. 2.35 and Eq. 2.36 for all $(vc, v'c')$ at once; so that we need only broadcast the dielectric matrix one time.

The wall-time scaling for the example `kernel` (20,20) SWCNT calculation is shown in Fig. A.5. We see nearly linear scaling up to 1024 CPUs – which is the square of the number of **k**-points, 32.

In the `absorption` executable, as described above, the first computational challenge is the interpolation of the kernel from the coarse **k**-grid onto the fine **k**-grid. The computation of the interpolation coefficients, Eq. 2.38, is done by distributing the fine-grid **k**-points evenly among the processors. For molecules or other large systems, this does not represent a problem because the computation of the coefficients is very quick in these cases regardless of the lack of parallelization.

The parallelization of the kernel interpolation is different from the parallelization scheme in the `kernel` executable. However, the parallelization is also described by three different schemes:

First, if the number of CPUs is less than N_k, where N_k is the number of fine-grid **k**-points, we distribute the N_k **k**-vectors on the fine grid evenly among the CPUs. Each processor owns all the $(vc\mathbf{k}, v'c'\mathbf{k}')$ pairs consistent with the **k**-vectors it was assigned. It

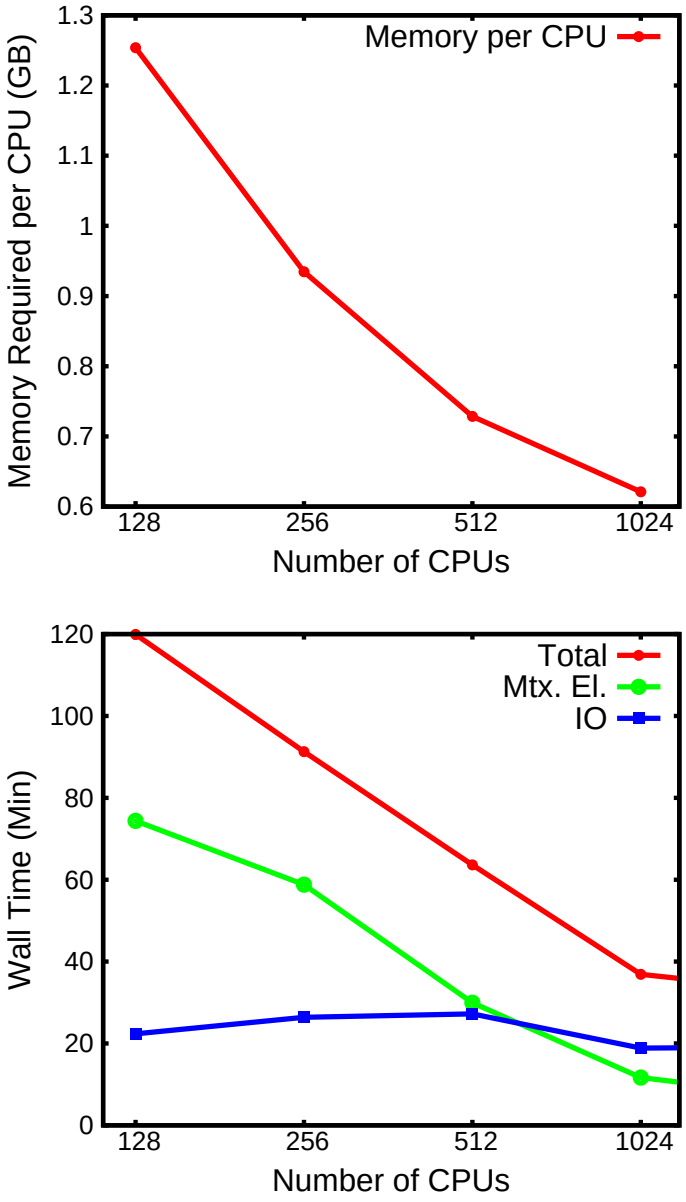

Figure A.5: (left) Memory per CPU required vs the number of CPUs for a `kernel` calculation on the (20,20) SWCNT. (right) The wall-time required vs. the number of CPUs used for a `kernel` calculation on the (20,20) SWCNT. The parameters used are described in the text.

then serially performs the interpolation in Eq. 2.39 by replacing the simple loops that represent the sums with four matrix-matrix multiplications. For example, for each n_1 and n_3, one can write the sum over n_2 as a matrix-matrix product between the coarse kernel matrix (whose outer dimension is n_4) and $C^{*\mathbf{k}_{co}}_{v,n_2}$ matrix (whose outer dimension is N_v on the fine grid). We write the remaining sums as a similar matrix-matrix product.

If the number of CPUs is greater than N_k but less than $N_k \cdot N_c$, we distribute the $N_k \cdot N_c$ **k**-point and conduction-band pairs evenly among the processors. Again, each processor owns all the $(vc\mathbf{k}, v'c'\mathbf{k}')$ pairs consistent with the $(\mathbf{k}, c)$ it was assigned. In the previous scheme, we utilized matrix-matrix products to interpolate many kernel elements at once. This required the allocation of an array of size $N_v^2 \cdot N_c^2$ as described above. In the present scheme, we avoid storing large intermediate arrays by doing more of the sums in Eq. 2.39 as simple loops rather than matrix products. By default we do two simple loops of two matrix products. However, if the user selects the `low_memory` option, we do all four summations as four nested loops without the aid of matrix-matrix multiplications, obtaining one fine-grid matrix element at a time without the need for any intermediate matrices.

If the number of CPUs is greater than $N_k \cdot N_c$, we distribute the $N_k \cdot N_c \cdot N_v$ **k**-point and conduction- and valence-band pairs evenly among the processors. Each processor owns all the $(vc\mathbf{k}, v'c'\mathbf{k}')$ pairs consistent with the $(\mathbf{k}, c, v)$ it was assigned. The interpolation is done exactly as in the previous case.

Once the fine-grid kernel has been constructed, in the `absorption` executable, the matrix is diagonalized using ScaLAPACK with a block-cyclic layout [23]. This diagonalization scales well to $O(1000)$ CPUs but quickly saturates beyond this point.

In order to calculate the absorption spectrum as per Eq. 2.5, we compute all the necessary matrix elements, Eq. 2.6, by distributing the fine-grid **k**-points evenly over processors.

In order to diagonalize large matrices (*e.g*, graphene on a 256×256 **k**-point grid) [149], we turn to iterative diagonalization methods, in particularly the Haydock Recursion Iteration – since it requires only matrix-vector products, this method scales well to larger number of processors. It should be pointed out that the kernel matrix is in general not sparse, so methods designed for the diagonalization of sparse matrices are not appropriate here.

A.2 Symmetry and degeneracy

A.2.1 Mean field

As was mentioned in the Chapter 2, the largest cost when performing a GW calculation with the `BerkeleyGW` package is the generation of the input mean-field states. In order to reduce this cost, all the codes allow the user to input the wavefunctions in only the reduced Brillouin zone and construct the wavefunctions in the full zone by the following relation:

$$\phi_{\mathbf{R}(\mathbf{k})}(\mathbf{G}) = \phi_{\mathbf{k}}(\mathbf{R}^{-1}(\mathbf{G}))e^{-i\mathbf{k}\cdot\tau}. \tag{A.1}$$

where the symmetry operation is defined by a rotation matrix $\mathbf{R}$ and a fractional translation τ.

A.2.2 Dielectric matrix

The wavefunctions in the full zone are then used for the sum over $\mathbf{k}$ in Eq. 2.8 in `epsilon`. `sigma`, `kernel` and `absorption` which require not only the wavefunctions in the full zone but also the dielectric matrix. While in principle one could construct the dielectric matrices at all the $\mathbf{q}$-points required in Eqs. 2.20 – 2.29 and 2.35, in practice one can use symmetry to reduce the required $\mathbf{q}$-points. The `epsilon` code requires the user to calculate the dielectric matrices on a reduced set of $\mathbf{q}$-points and the other codes generate the dielectric matrices in the full zone. If one defines $\mathbf{q}_1 = \mathbf{R}(\mathbf{q}) + \mathbf{G}_R$, where $\mathbf{G}_R$ is a $\mathbf{G}$-vector chosen to ensure that $\mathbf{q}$ and $\mathbf{q}_1$ are in the first Brillouin zone,that can be required to ensure that, and $\mathbf{R}, \tau$ are rotation matrix and translation respectively, then one can use the relation [62, 64] :

$$\epsilon^{-1}_{\mathbf{G}\mathbf{G}'}(\mathbf{q}_1; E) = e^{-i(\mathbf{G}-\mathbf{G}')\cdot\tau}\epsilon^{-1}_{\mathbf{G}_1\mathbf{G}'_1}(\mathbf{q}; E) \tag{A.2}$$

where $\mathbf{G}_1 = \mathbf{R}^{-1}(\mathbf{G} + \mathbf{G}_R)$. Given ϵ^{-1} in the reduced zone, this allows one to construct it in the full zone.

A.2.3 Truncation of sums

Both the dielectric matrix and the self-energy operator involve infinite sums over unoccupied states, which must in practice be truncated in the `epsilon` and `sigma` codes. The dielectric matrix and self-energy operator only retain the full symmetry of the system if the truncation does not cut through any degenerate subspaces. Consider a subspace of states belonging to a degenerate representation of a symmetry operation. Only the whole subspace is invariant under that operation, while just a part of it is not necessarily. As a result, a calculation using only part of the subspace will produce self-energies that break degeneracies due to that operation. Moreover, the actual values obtained are not well defined because the states used are arbitrary linear combinations in the subspace, which could even differ from run to run of a DFT code depending on how the calculation is initialized. These considerations are particularly acute for sums over a small number of states, since the contribution of the last few bands may be significant. Therefore, the `epsilon` and `sigma` codes check that the highest band requested for the sum is not degenerate with the next one, and block calculations that will break degeneracy. However, it is also possible to override this behavior with the flag `degeneracy_check_override`, for testing purposes and because in some cases there may be overlapping degenerate subspaces on different $\mathbf{k}$-points that make it difficult to find acceptable numbers of bands; for large numbers of bands, the effect of truncation in a degenerate subspace will be small.

A.2.4 Self-energy operator

Degeneracy is also important from the point of view of the states on which self energies are calculated, as opposed to those appearing in the sum. Since the self-energy operator has the full symmetry of the system, the matrix elements between states belonging

to different representations are zero by symmetry. In the presence of high symmetry, this consideration can make the matrix quite sparse. To take advantage fully of symmetry here would require a careful analysis of each wavefunction's behavior under various symmetry operations and comparison to character tables of space groups. Users can certainly do this in deciding which off-diagonal self-energy matrix elements to calculate. The `Sigma` takes a very simple approach to identify some of the elements which are zero by symmetry, based on degeneracy. The multiplicity of the degenerate subspace to which each states belongs is counted (1, 2, or 3 for the standard space groups), and clearly two states in subspaces of different multiplicity must belong to different representations, and their matrix element can be set to zero without calculation. This saves time and enforces symmetry.

Application of symmetry in a degenerate subspace can also speed up calculation of diagonal elements of the self-energy operator. The expressions for the exchange, screened exchange, and Coulomb-hole parts contain a sum over $\mathbf{q}$. In general, this must be done over the whole Brillouin zone, but to calculate the sum of the self energies within a degenerate subspace it is sufficient to use the irreducible part of the Brillouin zone. Each part of Σ, in the various approximations, has the generic form

$$\langle n\mathbf{k}|\,\Sigma\,|n'\mathbf{k}\rangle = -\sum_{n''}\sum_{\mathbf{qGG}'}\left\langle n\mathbf{k}\left|e^{i(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}}\right|n''\mathbf{k}-\mathbf{q}\right\rangle\left\langle n''\mathbf{k}-\mathbf{q}\left|e^{-i(\mathbf{q}+\mathbf{G}')\cdot\mathbf{r}}\right|n'\mathbf{k}\right\rangle F\left(\mathbf{q},\mathbf{G},\mathbf{G}'\right) \tag{A.3}$$

The summand is invariant under application of a symmetry operation O in the subgroup of $\mathbf{k}$ provided that $n = n'$ and n and n'' are non-degenerate, since in that case the action of the operation simply introduces a phase: $O\,|mk\rangle = e^{i\theta}\,|mk\rangle$ (degenerate states may instead transform into linear combinations in the degenerate subspace). These phases are cancelled by the fact that each state appears also with its complex conjugate. If the states n'' in the sum are degenerate, the summand is not invariant but the sum is, if the whole degenerate subspace is summed over, since then we are taking the trace of the projector matrix $|n''\mathbf{k}\rangle\langle n''\mathbf{k}|$ in that subspace, which is invariant [64]. If n is degenerate, then $\langle n\mathbf{k}|\,\Sigma\,|n\mathbf{k}\rangle$ is not invariant, but the trace of the self-energy in the degenerate subspace, $\sum_n \langle n\mathbf{k}|\,\Sigma\,|n\mathbf{k}\rangle$, is invariant. Therefore, to calculate diagonal elements for a whole degenerate subspace, for each state we sum only over $\mathbf{q}$ in the irreducible zone, with weight $W_\mathbf{q}$ from the number of $\mathbf{q}$-vectors related to $\mathbf{q}$ by symmetry. We then symmetrize by assigning the average to each:

$$\begin{aligned}\langle m\mathbf{k}|\,\Sigma\,|m\mathbf{k}\rangle &= \frac{1}{N_{\mathrm{deg}}}\sum_{n}^{\mathrm{deg}}\langle n\mathbf{k}|\,\Sigma\,|n\mathbf{k}\rangle \\ &= -\sum_{n}^{\mathrm{deg}}\sum_{n''}\sum_{\mathbf{GG}'}\sum_{\mathbf{q}}^{\mathrm{irr}} W_\mathbf{q}\left\langle n\mathbf{k}\left|e^{i(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}}\right|n''\mathbf{k}-\mathbf{q}\right\rangle\left\langle n''\mathbf{k}-\mathbf{q}\left|e^{-i(\mathbf{q}+\mathbf{G}')\cdot\mathbf{r}}\right|n\mathbf{k}\right\rangle \\ &\quad\times F\left(\mathbf{q},\mathbf{G},\mathbf{G}'\right)\end{aligned} \tag{A.4}$$

If we are calculating only part of a degenerate subspace, this trick does not work, and we must perform the complete sum. For diagonal elements, the code by default uses the irreducible **q**-sum and will write an error if the calculation requires the full sum because of degeneracy, directing the user to enable via the flag `no_symmetries_q_grid`, or include all states in the degenerate subspace. For off-diagonal elements ($n \neq n''$), even if both are non-degenerate, application of the symmetry operation introduces in general different phases from the two states, which are not canceled. Thus the contributions from different **q**-points related by symmetry differ, so that the full sum must always be used.

A.2.5 Bethe-Salpeter equation

Degeneracy must be considered in BSE calculations as well, when choosing the subspace in which to work. If the set of occupied or unoccupied states includes only part of a degenerate subspace, then the solutions found by `absorption` will break symmetry and can give qualitatively incorrect results. For example, an excitation that should have zero oscillator strength by symmetry, due to interference between transitions to two degenerate states, may not be dark if only one of those transitions is included. This issue is quite general and applies to the choice of active spaces in other theories as well, such as configuration interaction [151].

A.2.6 Degeneracy utility

We provide a utility called `degeneracy_check.x` which reads wavefunction files and writes out a list of acceptable numbers of bands. Multiple wavefunction files can be checked at once, for example the shifted and unshifted grids in `epsilon` or shifted, unshifted, coarse, and fine grids for Bethe-Salpeter equation calculations, in which case the utility will identify numbers of bands which are consistent with degeneracy for every file.

A.2.7 Real and complex flavors

The component executables come in two "flavors," real and complex, specified at compile time and denoted by the suffix `.real.x` or `.cplx.x`. When the system has inversion and time-reversal symmetry, we can choose the wavefunctions to be real in reciprocal space. The plane-wave expansions are:

$$u(\mathbf{r}) = \sum_{\mathbf{G}} u_{\mathbf{G}} e^{i\mathbf{G}\cdot\mathbf{r}} \tag{A.5}$$

$$u(-\mathbf{r}) = \sum_{\mathbf{G}} u_{\mathbf{G}} e^{-i\mathbf{G}\cdot\mathbf{r}} \tag{A.6}$$

$$u^*(\mathbf{r}) = \sum_{\mathbf{G}} u^*_{\mathbf{G}} e^{-i\mathbf{G}\cdot\mathbf{r}} \tag{A.7}$$

The symmetry conditions mean that wavefunctions can be chosen to satisfy $u(-\mathbf{r}) = au(\mathbf{r})$ (inversion symmetry) and $u^*(\mathbf{r}) = bu(\mathbf{r})$ (time-reversal, equivalent to taking the complex conjugate of the Schrödinger equation), with a, b each equal to ± 1 depending on whether the

wavefunction belongs to an odd or even representation. Thus we can choose $u(-\mathbf{r}) = cu^*(\mathbf{r})$ with $c = ab$ also equal to ± 1. Combining this with the plane-wave expansions,

$$\sum_{\mathbf{G}} u_{\mathbf{G}} e^{-i\mathbf{G}\cdot\mathbf{r}} = c \sum_{\mathbf{G}} u^*_{\mathbf{G}} e^{-i\mathbf{G}\cdot\mathbf{r}} \tag{A.8}$$

$$u_{\mathbf{G}} = cu^*_{\mathbf{G}} \tag{A.9}$$

The choice $c = 1$ corresponds to real coefficients; $c = -1$ corresponds to pure imaginary coefficients. Most plane-wave electronic-structure codes always use complex coefficients, and so the coefficients will in general not be real, even in the presence of inversion and time-reversal symmetry. For a non-degenerate state, the coefficients will be real times an arbitrary global phase, determined by the initialization of the solution procedure. We must divide out this global phase to make the coefficients real. In a degenerate subspace, the states need not be eigenstates of inversion, and so in general they may not just be real times a global phase. Instead, in each subspace of degeneracy n we take the $2n$ vectors given by the real and imaginary parts of each wavefunction, and then use a Gram-Schmidt process to find n real orthonormal wavefunctions spanning the subspace. The density and exchange-correlation potential are real already in the presence of inversion symmetry and there is no arbitrary phase possible.

The real-space density is always real: $\rho(\mathbf{r}) = \rho^*(\mathbf{r})$. With inversion symmetry, we also have $\rho(\mathbf{r}) = \rho(-\mathbf{r})$. In reciprocal space,

$$\rho(\mathbf{r}) = \sum_{\mathbf{G}} \rho_{\mathbf{G}} e^{i\mathbf{G}\cdot\mathbf{r}} \tag{A.10}$$

$$\rho^*(\mathbf{r}) = \sum_{\mathbf{G}} \rho^*_{\mathbf{G}} e^{-i\mathbf{G}\cdot\mathbf{r}} \tag{A.11}$$

$$\rho(-\mathbf{r}) = \sum_{\mathbf{G}} \rho_{\mathbf{G}} e^{-i\mathbf{G}\cdot\mathbf{r}} \tag{A.12}$$

Together, these relations imply $\rho_{\mathbf{G}} = \rho^*_{\mathbf{G}}$, *i.e.* the reciprocal-space coefficients are real. Precisely the same equations apply for the exchange-correlation potential.

The wavefunction, density, and exchange-correlation potential are then all stored as real coefficients, saving disk space (for the files), memory, and operations compared to the complex representation.

A.3 Computational Issues

A.3.1 Memory estimation

In the beginning of each run, all the major code components print the amount of memory available per CPU and an estimate of memory required per CPU to perform the calculation. If the latter exceeds the former the job is likely to fail with memory allocation error. The amount of memory required is estimated by determining the sizes of the largest arrays after reading in the parameters of the system from the input files. A straightforward approach to estimating the amount of available memory is to allocate memory by

incremental amounts until the allocation call returns with an error. Unfortunately, in many implementations the allocation call returns without an error even if the requested amount of memory is not physically available, but the system fails when trying to access this "allocated" memory. We implement another approach based on the Linux `/proc` file system. First, each CPU opens file `/proc/meminfo` and reads in the values of `MemFree` and `Cached`. The sum of these two values gives the amount of memory available per node. Second, each CPU calls `hostnm` routine (`hostnm_` for XLF and `hostnam` for Intel compilers) that returns the host name which is unique for each node. By comparing host names reported by different CPUs we identify the number of CPUs per node. The amount of memory available per CPU is then given by the ratio of the amount of memory available per node to the number of CPUs per node. This approach works on almost all modern high-performance computing systems where the Linux `/proc` File System is accessible.

A.3.2 Makefiles

The main codes are in the `Epsilon`, `Sigma`, `BSE`, `PlotXct`, and `MeanField` directories. Routines used by all parts are in the `Common` directory, and routines common to some of the `MeanField` codes are in the `Symmetry` directory. The `Makefile`s are designed for GNU Make, and enable targets in a directory to be built from any level of the directory hierarchy. They contain a full set of dependencies, including those between directories, to ensure that the build is correct after any changes to source, for ease in development and modification. This also enables use of parallel make on large numbers of processors for rapid builds – any omissions in the dependencies generally cause a failure for a parallel `make`. The special `make` target `all-j` (i.e. `make -j all-j`) begins by using all processes to build the `Common` and `Symmetry` directories, which contain files required by files in a large number of directories; otherwise, the build would fail due to attempts by multiple processes to read and write the same files in the `Common` and `Symmetry` directories. Commonly, Fortran Makefiles are set up with object files depending on other object files. However, the real situation is that object files depend on module files (`.mod`) for the modules they use, and only executables depend on object files. Therefore we have dependencies directly on the module files to ensure the required files are present for compilation, particularly for parallel builds.

A.3.3 Installation instructions

The code can be installed via the following steps:

```
cp [flavor_real.mk/flavor_cplx.mk] flavor.mk
ln -s config/[mysystem].mk arch.mk
make all
make check
```

First a flavor is selected by copying the appropriate file to `flavor.mk`. Then a configuration file must be put as `arch.mk`. Configurations appropriate for various supercomputers as well as for using standard Ubuntu packages are provided in the `config` directory. Appropriate paths, libraries, and compiler flags can be selected here for other systems.

A.3.4 Validation and verification

The importance of verification and validation of complicated scientific software packages is receiving increasing attention. We use standard open-source tools for code development, following accepted best practices. [92] Development is done with the subversion (SVN) version-control system [4] and Trac, an issue-tracking system and interface to SVN [5]. All code runs identify the version and revision number used in the output for traceability of results, implemented via a special source file called `svninfo.f90` which all SVN revisions must modify (enforced via a pre-commit hook). Debug mode can be enabled via `-DDEBUG` in the `arch.mk` file, which produces more verbose output and also performs extra checking of dynamic memory allocation and deallocation. A macro enables a check of the status returned by the system after an allocation attempt, and reports failures, identifying the array name, size, source file, and line number, as well as which processor failed to allocate the array. Additionally, it keeps track of the amount of memory dynamically allocated and deallocated, so the code can report at the end of each run how much memory remains allocated, and the maximum and minimum memory 'high-water-mark' among the processors. In debug mode, a stack trace can also be enabled, either on just the root processor, or on all processors (causing the code to run much slower), which can be used to locate where problems such as segmentation faults are occurring (possibly on only one processor).

The package contains a comprehensive testsuite to test the various executables, run modes, and options, in the `testsuite` directory. Calculations of several different physical systems, with mean-field, `epsilon`, `sigma`, and BSE calculations, are carried out (including use of `PlotXct` and some utilities), detecting any run-time errors and showing any warnings generated. Then selected results are extracted from the output and compared to reference information within a specified tolerance. The actual calculated values, as well as timing for each step, are displayed. Each match is shown as either `OK` or `FAIL`, and a final summary is written of failures. The calculations are small and generally underconverged, to make them quick enough for routine testing and rapid feedback. The mean-field steps are either EPM (quick serial calculations) or stored compressed output from DFT calculations. The `Epsilon`, `Sigma`, and BSE calculations are run either in serial or on 4 processors (for parallel builds). The testsuite has numerous uses. It is useful for users to verify the success of a new build of the code on their platform (failures could be due to library problems, excessive optimizations, etc.). It is used for developers to verify that the code is giving reproducible answers, ensure consistency between serial/parallel runs, as well as real/complex and spin-polarized/unpolarized runs, and check that the code works with new compilers or libraries. On a routine basis, the testsuite is also useful for developers to check that changes to the code do not introduce problems. The driver scripts (`run_testsuite.sh` and `run_regression_test.pl`) and specifications for the files defining the test steps are originally based on, and developed in conjunction with, those of the Octopus code. [29, 88] This framework is quite general and can easily be used for constructing a testsuite for another code. It can be run in serial with the command `make check` (or `make check-save` to retain the working directories from the runs), or in parallel with `make check-jobscript` (or `make check-jobscript-save`). The system configuration file `arch.mk` can specify how to submit an appropriate jobscript for parallel execution on a supercomputer using a scheduler. Scripts are provided in the `testsuite` directory for some supercomputers. The testsuite

is used with a continuous-integration system, the open-source tool BuildBot, [2] to ensure the integrity of the code during development. Each commit to the SVN repository triggers a build of the code on each of 10 "buildslaves," which have different configurations with respect to serial/parallel, compilers, and libraries. After the build, the testsuite is run. BuildBot will report to the developers if the either the build or test runs failed, so the problem can be quickly remedied. Use of the various different buildslave configurations helps ensure that the code remains portable across different platforms and in accordance with the language standards. Two of the buildslaves are on a supercomputer with a scheduler, a situation for which standard BuildBot usage is problematic. We provide a Perl script `buildbot_mpi.pl` that can submit jobs, monitor their status, capture their output for BuildBot, and determine success or failure. This script is general for any PBS scheduler and can be used for other codes too.