Title
Spin-Orbit Coupling, Broken Time-Reversal Symmetry, and Polarizability Self-Consistency in GW and GW-BSE Theory with Applications to Two-Dimensional Materials

Author
Wu, Meng

Publication Date
2020

Peer reviewed|Thesis/dissertation

Spin-Orbit Coupling, Broken Time-Reversal Symmetry, and Polarizability Self-Consistency in GW and GW-BSE Theory with Applications to Two-Dimensional Materials

by

Meng Wu

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:

Professor Steven G. Louie, Chair
Professor Graham R. Fleming
Professor Feng Wang

Fall 2020

Spin-Orbit Coupling, Broken Time-Reversal Symmetry, and Polarizability Self-Consistency in GW and GW-BSE Theory with Applications to Two-Dimensional Materials

Abstract

Spin-Orbit Coupling, Broken Time-Reversal Symmetry, and Polarizability Self-Consistency in GW and GW-BSE Theory with Applications to Two-Dimensional Materials

by

Meng Wu

Doctor of Philosophy in Physics

University of California, Berkeley

Professor Steven G. Louie, Chair

Interactions (e.g., spin-orbit coupling (SOC), electron-hole, magnetic ordering, etc.) often give rise to dramatic new features in the excited-state physics of two-dimensional (2D) materials. Advanced first-principles methods greatly deepen our understanding of these interactions, and enable us to predict novel phenomena with high accuracy. In this dissertation, I discuss the formalism of several new features – i.e., full-spinor wavefunctions, magneto-optical (MO) effects, and self-consistency with vertex corrections in screening – in the framework of the GW and GW plus Bethe-Salpeter equation (GW-BSE) methods, the state-of-the-art many-body theoretical tools to explore condensed matter physics (Chapters 1–4). These techniques are then applied to 2D materials of recent interest (Chapters 5–8). This dissertation not only aims to understand and predict the excited-state physics of 2D materials with theory and first-principles calculations but also to elucidate relevant experimental data when available. The contents of this dissertation are organized as follows:

- In Chapter 1, I briefly review some important concepts used throughout the dissertation: density-functional theory (DFT), many-body perturbation theory (MBPT), dielectric responses, and 2D materials. In particular, I review the basics of DFT and MBPT, from which the first-principles GW and GW-BSE methods are derived. Dielectric responses of materials are introduced as an application of the linear response theory to a many-electron system under external electromagnetic perturbations. Relevant physical quantities measured in experiments are explained and connected to first-principles calculations.
- In Chapter 2, I introduce the SOC effect in solids and the formalism of full-spinor GW and GW-BSE methods. I focus on the total dielectric function, matrix elements involving spinor wavefunctions, the macroscopic transverse dielectric function tensor

calculated at the GW-BSE level, and matrix elements of the current operator. Benchmark results of the full-spinor GW and GW-BSE methods are also presented.

- In Chapter 3, I discuss the formalism of first-principles modeling of MO effects. The basics of magneto-optics are introduced, emphasizing the magneto-optical Kerr effect (MOKE) and Faraday effect (FE). MO signals are connected to the macroscopic transverse dielectric function tensor that can be calculated from first principles. Since this formalism will be applied to 2D magnetic insulators in Chapter 8, I also discuss the definition of dielectric function in 2D materials.

- In Chapter 4, I present a new first-principles method – self-consistent with appropriate polarizability GW (swapGW). With swapGW, we can perform self-consistent GW calculations and incorporate the effects of vertex corrections in the polarizability through a BSE. Different self-consistent GW methods and the effect of vertex corrections are reviewed in detail. Our implementation of the swapGW method is benchmarked using bulk silicon.

- In Chapter 5, I demonstrate a new set of optical selection rules dictated by the winding number of interband optical matrix elements, which is in fact due to a topological effect on optical transitions in 2D materials [1]. These selection rules are later verified by GW and GW-BSE calculations of gapped graphene systems.

- In Chapter 6, I present a work in collaboration with experimentalists to study the strain engineering of the band gap in 2D InSe flakes [2]. We discover the ultrasensitive tunability of the direct band gap in few-layer InSe flakes by photoluminescence spectroscopy. We also develop a theoretical understanding of the strain-induced band gap change through first-principles DFT and GW calculations.

- In Chapter 7, I discuss the important roles of the excitonic exchange interaction and SOC in reshaping the exciton states and modifying the optical properties of monolayer transition metal dichalcogenides [3]. Full-spinor GW and GW-BSE methods are employed to demonstrate the exchange-driven mixing of exciton states in monolayer MoS_2. Our experimental collaborators use the 2D electronic spectroscopy (an ultrafast four-wave mixing spectroscopy technique) to demonstrate the intravalley exchange interaction unambiguously in both time and frequency domains.

- In Chapter 8, I investigate the physical origin of giant excitonic and MO responses in 2D ferromagnetic insulators [4]. We show, with the full-spinor GW and GW-BSE methods, that excitonic effects dominate the optical and MO responses in the prototypical 2D ferromagnetic insulator, monolayer CrI_3. In this work, we also predict the sensitive frequency- and substrate-dependence of MO responses by simulating the MOKE and FE signals in realistic experimental setups.

In addition to the projects mentioned above, I am also involved in three others during my Ph.D. program, including two studies of heterojunctions made from one-dimensional (1D) bottom-up-synthesized graphene nanoribbons [5, 6], as well as one study of orbital texture in a 2D Mott insulator, monolayer 1T-$TaSe_2$ [7]. The discoveries made in these works are also interesting but do not seem to fit in the scope of this dissertation, and hence they are not included.

To Fanglin.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I want to express the most sincere gratitude to my Ph.D. advisor, Prof. Steven G. Louie. It is because of you that I can touch the frontier of modern condensed matter physics and materials science, become proficient in using the state-of-the-art first-principles tools, and dive into the uncharted territories of the method development and new physics. Thank you for always being so patient, insightful, and inspiring. Whenever I discuss with you in the office, I could feel your sharp insights into the underlying physics, broad scope of knowledge in both theory and experiment, and endless passion for the truth. From you, I learn to think independently and critically in the research, to collaborate synergistically with experimentalists, to present my work in front of a broad audience, to build up clear physical pictures from complicated phenomena, and to become a responsible person both inside and outside academia. All in all, thank you so much for your support to my family and me, and I am so proud and grateful to have you as my academic advisor. Moreover, I would like to thank Prof. Louie and Mrs. Jane Louie for the invitations to their Thanksgiving Eve dinner all these years. I really miss those tasty foods and warm-hearted conversations.

Thank you, my wonderful colleagues and lovely friends in the Louie group. Ting, it is hard to imagine that we have known each other for over 8 years ever since our undergraduate study at Peking University. You have always been a role model for me in both science and daily life. May we continue our friendship and collaboration in the future. Zhenglu, I am deeply impressed by your enthusiasm for physics, as well as your basketball and skiing skills. Fangzhou and Jingwei, it has been a great experience doing research in the Cave and enjoying casual time around Berkeley with you two. Diana, thanks a lot for teaching me how to use BerkeleyGW and how to prepare allocation proposals. Felipe, thank you for a lot of valuable discussions in the Corner Office. Chin Shen, my officemate, thank you for being so kind and supportive. I will miss those days when we worked in the Corner Office until midnight. I would also like to thank Sinisa, Marco, Andrea, Gabriel, Yea-Lee, Yang-Hao, Mit, Jiawei, Jamal, Derek, Brad, and Omar. Your presence makes my Ph.D. days so bright and pleasant. I will miss your friendly or grilling questions at group meetings, anecdotes and jokes shared during happy hours, and all the enjoyable moments we spent together. I also want to thank Anne, Donna, Katherine, Arica, and Maria for their help and professional support.

Thank you, my experimental collaborators: Giang D. Nguyen, Hsin-Zon Tsai, Daniel J. Rizzo, Liang Guo, Daniele M. Monahan, Yi Chen, Wei Ruan, Shujie Tang, Yang Li, Tianmeng Wang, Dr. Sung-Kwan Mo, Prof. Su-Fei Shi, Prof. Cheng-Yan Xu, Prof. Felix R. Fischer, Prof. Michael F. Crommie, Prof. Zhi-Xun Shen, and Prof. Graham R. Fleming. All these exciting discoveries out of our collaboration have enriched my Ph.D. life and deepened my understanding of the fantastic scientific world.

During this special time of the global pandemic when I was writing this dissertation, those great moments kept flashing in my mind: heated discussion of physics in different meeting rooms, group hiking in the Pinnacles National Park, annual Thanksgiving Eve dinner at the Louie's, annual BerkeleyGW workshops, Louie Group reunions at APS March Meetings, Prof. Louie's birthday symposium, red egg and ginger parties, many farewell parties, so on

and so forth. These moments appear to be so precious when everything goes virtual at this time. I will always cherish these invaluable memories and remember these brilliant people.

I would like to thank my dissertation and qualifying exam committee members: Prof. Alex Zettl, Prof. Feng Wang, and Prof. Graham R. Fleming. Thank you for your kindness and patience in qualifying me and reading this dissertation. And many thanks to my undergraduate advisor, Prof. Ji Feng, for introducing me to this exciting field of condensed matter physics.

I also want to thank my parents for their everlasting love. Through thick and thin, your support and encouragement are always there for me.

Lastly, I feel so lucky to meet my wife, Fanglin, at Berkeley. It feels like a gift from heaven to marry you. When I feel down and helpless, you encourage and enlighten me. When I make some progress, you get excited and curious about the details. The world has turned brighter since you entered my life. The year 2020 has been very hard for many of us, but I know that being together with you will make me invincible.

Chapter 1

Introduction

1.1 Density-functional theory

1.1.1 Kohn-Sham equation

The density-functional theory (DFT) is based on the *Hohenberg-Kohn theorem* [13]. It states that: the ground-state wavefunction and the ground-state expectation value of any observable physical quantity is a unique functional of the ground-state electron density. Furthermore, the total energy as a functional of the density takes the minimum value at the true ground-state density. This theorem, however, only proves the existence of such an energy functional without giving any clue of its actual form. It turns out that, by mapping the many-body problem to an effective one-body problem with an appropriately chosen single-particle potential $V_{\mathsf{KS}}(\boldsymbol{r})$ such that they have the same ground-state electron density, we can construct a self-consistent procedure to calculate the ground-state density and energy, as proposed by Kohn and Sham [14].

Such a noninteracting auxiliary problem can be described by a single-particle Schrödinger equation [8, 15]:

$$\left[-\frac{\hbar^2}{2m_e}\nabla^2 + V_{\mathsf{KS}}(\boldsymbol{r})\right]\phi_i(\boldsymbol{r}) = \epsilon_i\phi_i(\boldsymbol{r}), \tag{1.1}$$

where m_e is the electron rest mass, along with an electron density,

$$n_e(\boldsymbol{r}) = \sum_i^{\mathsf{occ}} |\phi_i(\boldsymbol{r})|^2, \tag{1.2}$$

where the summation goes over all the occupied states. The many-body wavefunction of such a noninteracting system is nothing but a Slater determinant formed by all the single-particle wavefunctions ϕ_i. We then introduce the exchange-correlation energy $E_{\mathsf{xc}}[n_e(\boldsymbol{r})]$ as a functional of n_e, which by its definition contains all the corrections beyond the Hartree approximation to the many-body problem. By defining the exchange-correlation potential as an derivative of E_{xc} with respect to the variation of electron density, $V_{\mathsf{xc}}[n_e(\boldsymbol{r})] = \delta E_{\mathsf{xc}}/\delta n_e$

and adding other ingredients, we can write down the famous *Kohn-Sham equation* as,

$$\left[-\frac{\hbar^2}{2m_e}\nabla^2 + V_{\mathsf{ext}}(\boldsymbol{r}) + V_{\mathsf{H}}[n_e(\boldsymbol{r})] + V_{\mathsf{xc}}[n_e(\boldsymbol{r})]\right]\phi_i(\boldsymbol{r}) = \epsilon_i\phi_i(\boldsymbol{r}), \tag{1.3}$$

where V_{H} is the Hartree potential, and V_{ext} the external potential [8]. If we know the form of $V_{\mathsf{ext}}(\boldsymbol{r})$ (i.e., ionic potentials, external electromagnetic fields, etc.) and the exchange-correlation energy functional $E_{\mathsf{xc}}[n_e(\boldsymbol{r})]$, the ground-state density and energy can then be calculated self-consistently.

The above-mentioned procedure is in principle exact, but we still need an approximate $V_{\mathsf{xc}}[n_e]$ to start with. In the *local-density approximation* (LDA), we assume that for a system with slowly varying $n_e(\boldsymbol{r})$, the exchange-correlation potential $V_{\mathsf{xc}}^{\mathsf{LDA}}$ can be approximated as a function of the local electron density, such that,

$$\begin{aligned} E_{\mathsf{xc}}^{\mathsf{LDA}}[n_e] &= \int \mathrm{d}\boldsymbol{r}\, n_e(\boldsymbol{r})\epsilon_{\mathsf{xc}}^{\mathsf{hom}}[n_e(\boldsymbol{r})], \\ V_{\mathsf{xc}}^{\mathsf{LDA}}(\boldsymbol{r}) &\equiv \frac{\delta E_{\mathsf{xc}}^{\mathsf{LDA}}}{\delta n_e} = \epsilon_{\mathsf{xc}}^{\mathsf{hom}}[n_e(\boldsymbol{r})] + n_e(\boldsymbol{r})\left.\frac{\mathrm{d}\epsilon_{\mathsf{xc}}^{\mathsf{hom}}(n_e)}{\mathrm{d}n_e}\right|_{n_e(\boldsymbol{r})}, \end{aligned} \tag{1.4}$$

where $\epsilon_{\mathsf{xc}}^{\mathsf{hom}}(n_e)$ is the exchange-correlation energy of a homogeneous electron gas with density n_e. The exchange energy of a homogeneous electron gas is given by [16] (with $\alpha = 2/3$ in the Slater's $X\alpha$ method),

$$\epsilon_{\mathsf{x}}^{\mathsf{hom}} = -\frac{3q_e^2}{4}\left(\frac{3}{\pi}\right)^{1/3} n_e^{1/3}, \tag{1.5}$$

which leads to the LDA exchange potential,

$$V_{\mathsf{x}}^{\mathsf{LDA}}(\boldsymbol{r}) = -q_e^2\left(\frac{3}{\pi}\right)^{1/3} n_e^{1/3}(\boldsymbol{r}). \tag{1.6}$$

Note that throughout this dissertation, q_e denotes the elementary charge of electron $q_e = -e = -1.602176 \times 10^{-19}$ C. As for the correlation part, various parameterized analytical forms have been used by the community, e.g., the Ceperley-Alder data [17] with Perdew-Zunger [18] parametrization, etc.

In the *generalized gradient approximation* (GGA), the exchange-correlation energy is written as a function of both the density and the gradient of the density,

$$E_{\mathsf{xc}}^{\mathsf{GGA}} = \int \mathrm{d}\boldsymbol{r}\, f(n_e(\boldsymbol{r}), \nabla n_e(\boldsymbol{r})). \tag{1.7}$$

Some popular GGA exchange functionals used in solids include the Perdew-Wang [19], Perdew-Burke-Ernzerhof (PBE) [20], and Perdew-Burke-Ernzerhof revised for solids (PBEsol) [21] parametrizations, etc.

Until now, we have assumed that the system is non-magnetic: that is, the density of spin-up and spin-down electrons are equal. These non-magnetic systems can be characterized

by the electron density alone. In magnetic systems, however, we need both spin-up and spin-down densities. It can be shown that the Kohn-Sham equation can be generalized to spin-polarized systems as follows [22, 23],

$$\sum_{\sigma'}\left\{\left[-\frac{\hbar^2}{2m_e}\nabla^2+V_{\mathsf{KS}}(\boldsymbol{r})\right]\delta_{\sigma\sigma'}-\frac{1}{2}g_e\mu_B\boldsymbol{\sigma}_{\sigma\sigma'}\cdot\boldsymbol{B}_{\mathsf{KS}}(\boldsymbol{r})\right\}\phi_{i\sigma'}(\boldsymbol{r})=\epsilon_{i\sigma}\phi_{i\sigma}(\boldsymbol{r}), \tag{1.8}$$

where $\boldsymbol{B}_{\mathsf{KS}}(\boldsymbol{r})$ is an effective magnetic induction. Consider a simpler case with a uniform magnetic field or collinear spin polarization, the spin-up and spin-down densities can be used instead of the total electron density and magnetization density. In this case, the Kohn-Sham equation is given by,

$$\left[-\frac{\hbar^2}{2m_e}\nabla^2+V_{\mathsf{ext}}(\boldsymbol{r})+V_{\mathsf{H}}[n_e(\boldsymbol{r})]+V_{\mathsf{xc};\sigma}[n_\uparrow(\boldsymbol{r}),n_\downarrow(\boldsymbol{r})]-\frac{1}{2}g_e\mu_B\sigma B\right]\phi_{i\sigma}(\boldsymbol{r})=\epsilon_{i\sigma}\phi_{i\sigma}(\boldsymbol{r}), \tag{1.9}$$

where the spin-resolved exchange-correlation potential $V_{\mathsf{xc};\sigma}$ is defined by the following derivative of E_{xc},

$$V_{\mathsf{xc};\sigma}[n_\uparrow(\boldsymbol{r}),n_\downarrow(\boldsymbol{r})]\equiv\frac{\delta E_{\mathsf{xc}}[n_\uparrow(\boldsymbol{r}),n_\downarrow(\boldsymbol{r})]}{\delta n_\sigma(\boldsymbol{r})},\quad \sigma=\uparrow,\downarrow. \tag{1.10}$$

LDA can still be used here to get an approximate expression of the spin-dependent E_{xc}, as named by the *local-spin-density approximation* (LSDA). In this approximation, the exchange-correlation energy can be written as,

$$E_{\mathsf{xc}}^{\mathsf{LSDA}}[n_\uparrow,n_\downarrow]=\int \mathrm{d}\boldsymbol{r}\,\epsilon_{\mathsf{xc}}[n_\uparrow(\boldsymbol{r}),n_\downarrow(\boldsymbol{r})]n_e(\boldsymbol{r}) \tag{1.11}$$

The exchange contribution is simply the summation of contributions from each spin, because there is no exchange between electrons with different spin polarization,

$$E_{\mathsf{x}}^{\mathsf{LSDA}}[n_\uparrow,n_\downarrow]=\frac{1}{2}\left(E_{\mathsf{x}}^{\mathsf{LDA}}[2n_\uparrow]+E_{\mathsf{x}}^{\mathsf{LDA}}[2n_\downarrow]\right). \tag{1.12}$$

Parametrization of the correlation energy can be performed in different forms, such as the Perdew-Zunger, von Barth-Hedin [23], and Vosko-Wilk-Nusair [24] parameterizations, etc.

1.1.2 Pseudopotentials

Another equally important ingredient in first-principles modeling of materials is the electron-ion interaction described with pseudopotentials. The theory of pseudopotentials is based on the *Phillips-Kleinman cancellation theorem* [25]: the wavefunctions of the valence and higher-energy electronic states are expected to be smooth away from the atomic sites and oscillatory with atomic character in the core region, and it is possible to derive a pseudopotential that guarantees the orthogonality between the valence electronic states and the implicitly included core electronic states [8]. This formalism justifies the intuition that

only valence electrons participate in the chemical bonding, while core electrons are mostly inactive. Pseudopotentials eliminate the core electrons from the Kohn-Sham equation, remove the original strong electron-ion interaction, and therefore allow the use of simple basis sets (e.g., plane waves) for a numerical solution with nice convergence behavior. In general, the construction of first-principles pseudopotentials involves first solving the all-electron (AE) Kohn-Sham equation (including both core and valence states) for a given atom in a particular configuration. Given the AE wavefunctions, a set of pseudowavefunctions can be built by some standardized pseudization procedures (e.g., connection to AE wavefunctions beyond a radius, smoothness up to high-order derivatives, norm-conserving or not, parameterization, etc.). Pseudopotentials can then be constructed by inverting the Schrödinger equation using the AE eigenvalues of valence states and corresponding pseudowavefunctions. To ensure the transferability of pseudopotentials, it is important to "unscreen" them by removing the electronic contributions arising from V_{H} and V_{xc}, because these components are specific to a particular configuration. The resulting potentials are called ionic pseudopotentials and they are widely used in first-principles calculations (DFT and beyond) for different physical properties in different chemical environments [26, 27].

The combination of DFT and the first-principles pseudopotential method prove to be very effective and versatile in investigating a wide range of materials (e.g., molecules, solids, nanostructures, interfaces, etc.) and various ground-state properties (e.g., electronic, structural, vibrational, etc.) [8, 27, 28, 29]. As long as a physical quantity (e.g., Born effective charge, piezoelectric tensor, dynamical matrix, etc.) can be formulated as changes in the total energy in response to some perturbations, it can be reasonably modeled by DFT. With DFT, it is possible to make direct and quantitative comparison with experiments or to predict certain phenomena before any measurements.

1.1.3 The band gap problem

Even though the charge density and the total energy of the ground state can be accurately calculated by DFT, it is important to point out that the Kohn-Sham eigenvalues and eigenstates in Eq. (1.1) cannot be interpreted as the actual quasiparticle energies and wavefunctions in real systems. It is because Kohn-Sham eigenvalues are just Lagrange multipliers used in the variational problem of minimizing the energy functional. In fact, this leads to the famous band gap problem of DFT as shown in Fig. 1.1. The band gap is an important quantity of semiconductors and insulators, and it is defined by,

$$E_{\mathsf{g}} \equiv (E_{N-1} - E_N) - (E_N - E_{N+1}), \tag{1.13}$$

where E_N is the ground-state total energy of the neutral system with N electrons, and $E_{N\pm1}$ is the ground-state total energy of the charged system with $N \pm 1$ electrons. The size of band gap will have a huge influence on the dielectric screening and optical properties of materials. However, DFT suffers from the notorious band gap problem: the Kohn-Sham band gaps $E_{\mathsf{g}}^{\mathsf{KS}}$ of semiconductors and insulators are usually much smaller (by about 40%) than experimental

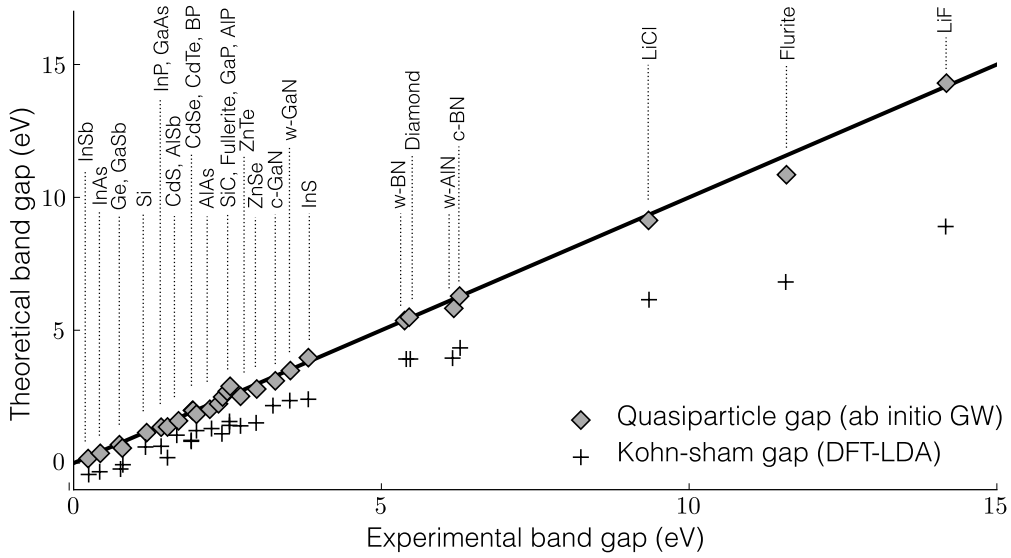

Figure 1.1: Comparison of calculated band gaps from DFT and from first-principles *GW* approach with experimental values for conventional semiconductors and insulators. This figure is adapted from Ref. [8].

band gaps (also called quasiparticle band gap $E_{\mathsf{g}}^{\mathsf{QP}}$). For some narrow-gap semiconductors, such as Ge and InSb, DFT could even give a metallic ground state [30, 31, 32]. In solids, the change in the electron density (Δn_e) upon adding or removing an electron is on the order of 10^{-20} of the existing density n_e. It has been shown in Ref. [33] that $E_{\mathsf{g}}^{\mathsf{QP}}$ and $E_{\mathsf{g}}^{\mathsf{KS}}$ differ by a constant Δ_{xc},

$$\lim_{\Delta n_e \to 0} E_{\mathsf{g}}^{\mathsf{QP}} = E_{\mathsf{g}}^{\mathsf{KS}} + \Delta_{\mathsf{xc}}, \tag{1.14}$$

where Δ_{xc} is given by,

$$\Delta_{\mathsf{xc}} \equiv \lim_{\Delta n_e \to 0} \left(V_{\mathsf{xc}}[n_e + \Delta n_e] - V_{\mathsf{xc}}[n_e - \Delta n_e] \right). \tag{1.15}$$

That is, if the exact E_{xc} is used, we would expect a discontinuity in V_{xc} with respect to the ground-state n_e in the neutral system. However, in LDA or similar approximations of V_{xc}, we assume a continuous functional with respect to n_e, which leads to $\Delta_{\mathsf{xc}} = 0$. In this way, the band gap problem is deeply rooted in the formalism of DFT and some higher-level theories are needed to solve this problem. The good news is that Kohn-Sham eigenvalues and eigenstates could serves as a reasonable starting point or an "educated guess" for more sophisticated and higher-level calculations of excited-state properties, such as quasiparticle bandstructure [9, 30] and optical properties [34, 35]. One such solution is the *GW* method, first proposed by Hedin [36] and later formulated from *ab initio* and applied to real materials by Hybertsen and Louie [9, 30]. In the *GW* method, we replace the local or semilocal static V_{xc} by an nonlocal frequency-dependent non-Hermitian self-energy operator Σ which well

captures the discontinuity of Δ_{xc} and yields good agreement with experiments [9, 30, 37]. Figure 1.1 compares the calculated band gaps from DFT and those from the first-principles GW method as well as with experimental band gaps for conventional semiconductors and insulators [8]. A brief review of the GW method is presented in the following sections.

1.2 Many-body perturbation theory

Electronic structure in solids is a typical many-body problem, involving a huge number ($\sim$ the Avogadro's constant) of electrons, ions, as well as under various external fields. The most systematic and elegant approach to this problem is based on the quantum field theory and the many-body perturbation theory (MBPT) (also called the Green's function method) [36, 38, 39, 40], from which the GW and GW-BSE methods are derived. MBPT is probably one of the most powerful, predictive and versatile formalisms in physics. This section is just a sketch of important concepts and conclusions most relevant to the development of the GW and GW-BSE methods, and we only focus on the Green's function formalism for electrons in the following. More details and other topics about MBPT can be found in several excellent reviews and textbooks [8, 40, 41, 42, 43, 44].

We start from a general Hamiltonian $\hat{H}$ for a system of interacting electrons under an external potential $V_{\mathsf{ext}}(\boldsymbol{x})$ [8, 41, 42, 45]:

$$\hat{H} = \int \mathrm{d}\boldsymbol{x}\, \hat{\psi}^\dagger(\boldsymbol{x}) \left[-\frac{\hbar^2}{2m_e}\nabla^2 + V_{\mathsf{ext}}(\boldsymbol{x}) \right] \hat{\psi}(\boldsymbol{x}) + \frac{1}{2}\int \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}'\, v(\boldsymbol{r},\boldsymbol{r}')\hat{\psi}^\dagger(\boldsymbol{x})\hat{\psi}^\dagger(\boldsymbol{x}')\hat{\psi}(\boldsymbol{x}')\hat{\psi}(\boldsymbol{x}), \tag{1.16}$$

where $\boldsymbol{x} = \{\boldsymbol{r}\sigma\}$ and $\int \mathrm{d}\boldsymbol{x} = \sum_\sigma \int \mathrm{d}\boldsymbol{r}$. The spin-resolved electron density operator is defined as, $\hat{n}_e(\boldsymbol{x}) \equiv \hat{\psi}^\dagger(\boldsymbol{x})\hat{\psi}(\boldsymbol{x})$. The Coulomb interaction is given by, $v(\boldsymbol{r},\boldsymbol{r}') = \frac{q_e^2}{4\pi\varepsilon_0|\boldsymbol{r}-\boldsymbol{r}'|}$, where ε_0 is the vacuum permittivity.

The *time-ordered single-particle Green's function* G in the zero-temperature formalism is defined as:

$$\begin{aligned} G(12) &\equiv (-\frac{i}{\hbar})\frac{\langle\Psi_0|T[e^{i\hat{H}t_1/\hbar}\hat{\psi}(\boldsymbol{x}_1)e^{-i\hat{H}t_1/\hbar}e^{i\hat{H}t_2/\hbar}\hat{\psi}^\dagger(\boldsymbol{x}_2)e^{-i\hat{H}t_2/\hbar}]|\Psi_0\rangle}{\langle\Psi_0|\Psi_0\rangle} \\ &\equiv (-\frac{i}{\hbar})\frac{\langle\Psi_0|T[\hat{\psi}(1)\hat{\psi}^\dagger(2)]|\Psi_0\rangle}{\langle\Psi_0|\Psi_0\rangle}, \end{aligned} \tag{1.17}$$

where $1 \equiv \{\boldsymbol{x}_1 t_1\}$, T is the time-ordering operator, and $|\Psi_0\rangle$ the ground state of an interacting system with the Hamiltonian in Eq. (1.16). Throughout the following derivation, the time-dependence of the field operator $\hat{\psi}(1)$ comes from the Hamiltonian in Eq. (1.16),

$$\hat{\psi}(1) \equiv \hat{\psi}(\boldsymbol{x}_1, t_1) = e^{i\hat{H}t_1/\hbar}\hat{\psi}(\boldsymbol{x}_1)e^{-i\hat{H}t_1/\hbar}. \tag{1.18}$$

Note that the following discussion also applies to the finite-temperature G, as shown in Ref. [39]. After a time Fourier transform with respect to $t_1 - t_2$ (Sec. A), we get $G(\boldsymbol{x}_1, \boldsymbol{x}_2; \omega)$, defined in the frequency domain.

The *retarded single-particle Green's function* G^{R} is defined as,

$$G^{\mathsf{R}}(12) \equiv (-\frac{i}{\hbar})\Theta(t_1 - t_2)\frac{\langle\Psi_0|\{\hat{\psi}(1), \hat{\psi}^\dagger(2)\}|\Psi_0\rangle}{\langle\Psi_0|\Psi_0\rangle}, \tag{1.19}$$

where the commutator $\{A, B\} \equiv AB + BA$ is for fermionic operators. In the following, quantities without an explicit R superscript are all time-ordered. Applications of retarded correlation functions, including G^{R}, will be discussed in the next section of dielectric responses. The single-particle Green's function is at the center of MBPT because (i) the Feynman rules are simpler for G than for other operators, and (ii) the expectation value of any single-particle operator in the ground state of the system can be calculated using G [41]. In this dissertation, we will adopt the functional derivative approach [39, 40, 45] to reduce the many-body problem to the solution of a coupled set of nonlinear integral equations. This approach avoids the cumbersome conventional diagrammatic expansion of relevant quantities.

We first add an external time-dependent local potential φ to the Hamiltonian in Eq. (1.16) in order to perturb the system,

$$\hat{H}'(t) \equiv \int \mathrm{d}\boldsymbol{x}\, \hat{\psi}^\dagger(\boldsymbol{x})\varphi(\boldsymbol{x}, t)\hat{\psi}(\boldsymbol{x}). \tag{1.20}$$

This perturbation potential will go to zero at the end of our derivation. Since there is an extra term in the Hamiltonian,

$$\hat{H}''(t) \equiv \hat{H} + \hat{H}'(t), \tag{1.21}$$

we adopt the interaction picture from now on,

$$\hat{H}'_{\mathsf{I}}(t) \equiv e^{i\hat{H}t/\hbar}\hat{H}'(t)e^{-i\hat{H}t/\hbar}, \tag{1.22}$$

and introduce the S matrix,

$$\hat{S} \equiv \exp\left\{-\frac{i}{\hbar}\int_{-\infty}^{\infty}\mathrm{d}t\, \hat{H}'_{\mathsf{I}}(t)\right\}. \tag{1.23}$$

The generalized single-particle Green's functions with respect to $\hat{H}''(t)$ is then,

$$G(12) = \left(-\frac{i}{\hbar}\right)\frac{\langle\Psi_0|T[\hat{S}\hat{\psi}(1)\hat{\psi}^\dagger(2)]|\Psi_0\rangle}{\langle\Psi_0|T[\hat{S}]|\Psi_0\rangle}. \tag{1.24}$$

When we take the limit of $\hat{H}' \to 0$, the S matrix will be reduced to an identity operator. We now introduce a total potential V_{tot} averaged over the many-body ground state $|\Psi_0\rangle$ as,

$$V_{\mathsf{tot}}(1) \equiv \varphi(1) + \int \mathrm{d}2\, v(12)\langle\hat{n}_e(2)\rangle = \varphi(1) - i\hbar\int \mathrm{d}2\, v(12)G(22^+), \tag{1.25}$$

where $v(12) = v(\boldsymbol{r}_1, \boldsymbol{r}_2)\delta(t_1 - t_2)$, $\langle \hat{n}_e(2) \rangle \equiv \langle \Psi_0 | \hat{n}_e(2) | \Psi_0 \rangle = -i\hbar G(22^+)$, $2 = \{\boldsymbol{x}_2 t_2^+\}$, and $t_2^+ = t_2 + \eta$ with $\eta \to 0^+$. We relate V_{tot} to the external potential φ by introducing the *reducible polarizability* χ and the *inverse dielectric function* ε^{-1} as functional derivatives:

$$\chi(12) \equiv \frac{\delta \langle \hat{n}_e(1) \rangle}{\delta \varphi(2)}, \tag{1.26}$$

$$\varepsilon^{-1}(12) \equiv \frac{\delta V_{\mathsf{tot}}(1)}{\delta \varphi(2)} = \delta(12) + \int \mathrm{d}3\, v(13)\chi(32). \tag{1.27}$$

Eq. (1.27) can be inverted to get,

$$\varepsilon(12) = \delta(12) - \int \mathrm{d}3\, v(13)\chi^\star(32), \tag{1.28}$$

where we have defined the *irreducible polarizability* $\chi^\star$ as,

$$\chi^\star(12) \equiv \frac{\delta \langle \hat{n}_e(1) \rangle}{\delta V_{\mathsf{tot}}(2)}. \tag{1.29}$$

The *screened Coulomb interaction* W is defined intuitively as,

$$W(12) \equiv \int \mathrm{d}3\, \varepsilon^{-1}(13) v(32). \tag{1.30}$$

Combine Eqs. (1.27), (1.29) and (1.30), and we get the following expressions of W,

$$\begin{aligned} W(12) &= v(12) + \int \mathrm{d}(34)\, v(13)\chi(34)v(42) \\ &= v(12) + \int \mathrm{d}(34)\, v(13)\chi^\star(34)W(42) \end{aligned} \tag{1.31}$$

The *electron self-energy* Σ, the Hartree term Σ_{H} and the *mass operator* $M(12)$ are defined as follows,

$$\Sigma(12) \equiv \Sigma_{\mathsf{H}}(12) + M(12), \tag{1.32}$$

$$\Sigma_{\mathsf{H}}(12) \equiv \delta(12) \int \mathrm{d}2\, v(12) \langle \hat{n}_e(2) \rangle, \tag{1.33}$$

$$M(12) \equiv i\hbar \int \mathrm{d}(34) v(13) \frac{\delta G(14)}{\delta \varphi(3)} G^{-1}(42). \tag{1.34}$$

Σ drives the equation of motion of G,

$$\left[i\hbar \frac{\partial}{\partial t_1} + \frac{\hbar^2}{2m_e} \nabla_1^2 \right] G(12) - \int \mathrm{d}3\, \Sigma(13) G(32) = \delta(12). \tag{1.35}$$

By defining a noninteracting single-particle Green's function $G_0(12)$ at the absence of the Coulomb interaction, Eq. (1.35) can be reformulated into the *Dyson's equation*,

$$G(12) = G_0(12) + \int \mathrm{d}(34)\, G_0(13)\Sigma(34)G(42). \tag{1.36}$$

The *irreducible three-point vertex function* $\Gamma^\star$ is defined as,

$$\Gamma^\star(123) \equiv -\frac{\delta G^{-1}(12)}{\delta V_{\mathsf{tot}}(3)} = \delta(12)\delta(13) + \int \mathrm{d}(4567)\, \frac{\delta M(12)}{\delta G(45)} G(46)G(75)\Gamma^\star(673), \tag{1.37}$$

M can then be expressed as an integral involving G, $\Gamma^\star$, and W:

$$M(12) = i\hbar \int \mathrm{d}(34)\, G(13)W(41)\Gamma^\star(324). \tag{1.38}$$

$\chi^\star$ is also related to $\Gamma^\star$ according to,

$$\chi^\star(12) = -i\hbar \int \mathrm{d}(34)\, G(13)G(41)\Gamma^\star(342). \tag{1.39}$$

Until now, we get the well-known *Hedin's equations* [36, 40] for a system of interacting electrons:

$$\begin{aligned}
G(12) &= G_0(12) + \int \mathrm{d}(34)\, G_0(13)\Sigma(34)G(42), && (1.40)\\
\Gamma^\star(123) &= \delta(13)\delta(23) + \int \mathrm{d}(4567) \frac{\delta M(12)}{\delta G(45)} G(46)G(75)\Gamma^\star(673), && (1.41)\\
\chi^\star(12) &= -i\hbar \int \mathrm{d}(34)\, G(13)G(41)\Gamma^\star(342), && (1.42)\\
W(12) &= v(12) + \int \mathrm{d}(34)\, v(13)\chi^\star(34)W(42), && (1.43)\\
M(12) &= i\hbar \int \mathrm{d}(34)\, G(13)W(41)\Gamma^\star(324) = \Sigma(12) - \Sigma_{\mathsf{H}}(12). && (1.44)
\end{aligned}$$

This set of equations build up a large self-consistent loop involving the single-particle Green's function G, irreducible vertex function $\Gamma^\star$, irreducible polarizability $\chi^\star$, screened Coulomb interaction W and mass operator M.

1.3 The GW method

The most important approximation of the GW method is to take the zeroth order vertex (i.e., no vertex corrections) in Eq. (1.41),

$$\Gamma^\star(123) = \delta(13)\delta(23), \tag{1.45}$$

which leads to the irreducible polarizability within the *random-phase approximation* (RPA),

$$\chi^\star(12) = -i\hbar G(12)G(21), \tag{1.46}$$

and the iconic GW mass operator,

$$M(12) = i\hbar G(12)W(21). \tag{1.47}$$

The number of self-consistent Hedin's equations are now reduced to four. Underneath the simple look of the GW method, it is actually one of the conserving approximations, as discussed by Kadanoff and Baym [46, 47], which means it is expected to satisfy the general (number, momentum, and energy) conservation laws.

As discussed before, the band gap problem of DFT was successfully resolved by the development of the first-principles GW method for quasiparticle excitations by Hybertsen and Louie [9, 30]. In real materials, W is usually much weaker than the bare Coulomb interaction, which leads to the laudable accuracy and versatility of the GW method [8]. Even though the GW method is much simpler than the original Hedin's equations, we still need further approximations for a realistic calculation. In practice, DFT results are usually used as a starting point for the GW method by replacing the many-body wavefunctions by a Slater determinant of Kohn-Sham eigenstates. This allows us to evaluate the GW self-energy (mass operator) as a first-order perturbation with respect to the Kohn-Sham eigenvalues,

$$\epsilon^{\mathsf{QP}}_{n\boldsymbol{k}} = \epsilon^{\mathsf{KS}}_{n\boldsymbol{k}} + \langle n\boldsymbol{k}|M(\epsilon^{\mathsf{QP}}_{n\boldsymbol{k}}) - V_{\mathsf{xc}}|n\boldsymbol{k}\rangle. \tag{1.48}$$

It has been shown that the overlap between DFT-LDA and quasiparticle wavefunctions is greater than 99.9% in some conventional semiconductors and insulators [9]. Moreover, the full self-consistency is often out of reach for real materials and therefore further approximations have to be adopted, such as the one-shot GW method (also called the G_0W_0 method). In the G_0W_0 method, both G_0 and W_0 are constructed using Kohn-Sham eigenvalues and eigenstates in the quasiparticle approximation. G_0 is given by,

$$G_0(\boldsymbol{x}_1, \boldsymbol{x}_2; \omega) = \sum_{n\boldsymbol{k}} \frac{\phi^{\mathsf{KS}}_{n\boldsymbol{k}}(\boldsymbol{x}_1)(\phi^{\mathsf{KS}}_{n\boldsymbol{k}}(\boldsymbol{x}_2))^*}{\hbar\omega - \epsilon^{\mathsf{KS}}_{n\boldsymbol{k}} + i\eta\,\mathsf{sgn}(\epsilon^{\mathsf{KS}}_{n\boldsymbol{k}} - \epsilon_{\mathsf{F}})}, \quad \eta \to 0^+, \tag{1.49}$$

where ϵ_{F} denotes the Fermi level and sgn is the sign function. After space and time Fourier transforms (Sec. A), W_0 can be constructed from the RPA (denoted by the superscript of 0) irreducible polarizability $\chi^{\star;0}$ with Kohn-Sham eigenvalues and eigenstates,

$$\begin{aligned}\chi^{\star;0}_{\boldsymbol{G}_1\boldsymbol{G}_2}(\boldsymbol{q};\omega) = \frac{1}{N_k\Omega}\sum_{cv\boldsymbol{k}} \Bigg\{ &\frac{\langle v(\boldsymbol{k}-\boldsymbol{q})|e^{-i(\boldsymbol{q}+\boldsymbol{G}_1)\cdot\boldsymbol{r}_1}|c\boldsymbol{k}\rangle\langle c\boldsymbol{k}|e^{i(\boldsymbol{q}+\boldsymbol{G}_2)\cdot\boldsymbol{r}_2}|v(\boldsymbol{k}-\boldsymbol{q})\rangle}{\hbar\omega - (\epsilon^{\mathsf{KS}}_{c\boldsymbol{k}} - \epsilon^{\mathsf{KS}}_{v(\boldsymbol{k}-\boldsymbol{q})}) + i\eta} \\ &- \frac{\langle c\boldsymbol{k}|e^{-i(\boldsymbol{q}+\boldsymbol{G}_1)\cdot\boldsymbol{r}_1}|v(\boldsymbol{k}+\boldsymbol{q})\rangle\langle v(\boldsymbol{k}+\boldsymbol{q})|e^{i(\boldsymbol{q}+\boldsymbol{G}_2)\cdot\boldsymbol{r}_2}|c\boldsymbol{k}\rangle}{\hbar\omega + (\epsilon^{\mathsf{KS}}_{c\boldsymbol{k}} - \epsilon^{\mathsf{KS}}_{v(\boldsymbol{k}+\boldsymbol{q})}) - i\eta} \Bigg\}, \quad \eta \to 0^+,\end{aligned} \tag{1.50}$$

where N_k is the number of k-points in the Brillouin zone and it is also equal to the number of unit cells used in the Born–von Kármán boundary condition [8], and Ω is the volume of a unit cell. Equation (1.50) is the famous Adler-Wiser expression of the RPA polarizability [48, 49]. The evaluation of the GW self-energy correction can be performed either by incorporating the full frequency dependence of $\chi^\star$, and hence ε^{-1}, through a numerical integral [11] or by employing the generalized plasmon-pole model of ε^{-1} with a sum-rule constraint [9].

1.4 Dielectric responses

In this section, we discuss the microscopic theory of dielectric responses in solids using the linear response theory. When an external electromagnetic perturbation is switched on, several physical quantities of the material may change accordingly. One external field couples directly to an observable (e.g., the scalar potential couples to the charge density and the vector potential couples to the charge current density) such that their product serves as a perturbation Hamiltonian. When the external field is relatively weak, the change in the expectation values of those directly related observables is proportional to the field amplitude and to the corresponding susceptibility (as a special correlation function). These susceptibilities can be calculated as retarded correlation functions of the equilibrium system [41, 42, 50, 51, 52]. Electromagnetic perturbations represent many state-of-the-art experimental characterization techniques in solid state physics: electron energy-loss spectroscopy using the scalar potential of external electrons; linear optics using the vector potential of external photons; transport measurements using the external electric field; magnetic neutron scattering using the magnetic field of neutrons, etc.

Here, we use the linear response theory to study a system of interacting electrons with mass m_e and charge q_e moving under the influence of a total vector potential $\boldsymbol{A}(\boldsymbol{r},t)$ and an external scalar potential $\varphi^{\mathsf{ext}}(\boldsymbol{r},t)$ [42, 45]. The perturbation Hamiltonian in the Schrödinger picture is given by,

$$\begin{aligned}\hat{H}'(t) &= -q_e \int \mathrm{d}\boldsymbol{r}\, \left[\hat{\boldsymbol{j}}_{\mathsf{p}}(\boldsymbol{r}) + \hat{\boldsymbol{j}}_{\mathsf{d}}(\boldsymbol{r},t)\right] \cdot \boldsymbol{A}(\boldsymbol{r},t) + q_e \int \mathrm{d}\boldsymbol{r}\, \hat{n}_e(\boldsymbol{r}) \varphi^{\mathsf{ext}}(\boldsymbol{r},t) \\ &= -\int \mathrm{d}\boldsymbol{x}\, \left[\hat{\boldsymbol{J}}_{\mathsf{p}}(\boldsymbol{r}) + \hat{\boldsymbol{J}}_{\mathsf{d}}(\boldsymbol{r},t)\right] \cdot \boldsymbol{A}(\boldsymbol{r},t) + \int \mathrm{d}\boldsymbol{r}\, \hat{\rho}(\boldsymbol{r}) \varphi^{\mathsf{ext}}(\boldsymbol{r},t),\end{aligned} \tag{1.51}$$

where $\hat{\boldsymbol{j}}_{\mathsf{p}}$ ($\hat{\boldsymbol{J}}_{\mathsf{p}} = q_e \hat{\boldsymbol{j}}_{\mathsf{p}}$) refers to the paramagnetic particle (charge) current density and $\hat{\boldsymbol{j}}_{\mathsf{d}}$ ($\hat{\boldsymbol{J}}_{\mathsf{d}} = q_e \hat{\boldsymbol{j}}_{\mathsf{d}}$) refers to the diamagnetic particle (charge) current density. $\hat{\rho} = q_e \hat{n}_e$ refers to the electron charge density. Detailed discussion of current operators can be found in Sec. B.

In the following, we introduce four different susceptibilities that will be used in the linear responses of induced charge density and induced charge current density. First, let's define the deviation operators for the charge density and paramagnetic charge current density,

$$\hat{\rho}'(1) \equiv \hat{\rho}(1) - \langle\Psi_0|\hat{\rho}(1)|\Psi_0\rangle, \tag{1.52}$$

$$\hat{\boldsymbol{J}}'_{\mathsf{p}}(1) \equiv \hat{\boldsymbol{J}}_{\mathsf{p}}(1) - \langle\Psi_0|\hat{\boldsymbol{J}}_{\mathsf{p}}(1)|\Psi_0\rangle, \tag{1.53}$$

with the following identities,

$$[\hat{\rho}'(1), \hat{\rho}'(2)] = [\hat{\rho}(1), \hat{\rho}(2)]\,, \tag{1.54}$$

$$\left[\hat{\rho}'(1), \hat{\boldsymbol{J}}'_{\mathsf{p}}(2)\right] = \left[\hat{\rho}(1), \hat{\boldsymbol{J}}_{\mathsf{p}}(2)\right], \tag{1.55}$$

$$\left[\hat{\boldsymbol{J}}'_{\mathsf{p}}(1), \hat{\boldsymbol{J}}'_{\mathsf{p}}(2)\right] = \left[\hat{\boldsymbol{J}}_{\mathsf{p}}(1), \hat{\boldsymbol{J}}_{\mathsf{p}}(2)\right]. \tag{1.56}$$

The *retarded reducible susceptibilities* (*density-density*, *current-density*, *density-current*, *paramagnetic current-current*) can be defined as follows,

$$\boldsymbol{\chi}^{\mathsf{R}}(12) \equiv \left(-\frac{i}{\hbar}\right)\Theta(t_1-t_2)\frac{\langle\Psi_0|\left[\hat{\rho}'(1),\hat{\rho}'(2)\right]|\Psi_0\rangle}{\langle\Psi_0|\Psi_0\rangle}, \tag{1.57}$$

$$\overleftarrow{\boldsymbol{\chi}}^{\mathsf{R}}(12) \equiv \left(-\frac{i}{\hbar}\right)\Theta(t_1-t_2)\frac{\langle\Psi_0|\left[\hat{\boldsymbol{J}}'_{\mathsf{p}}(1),\hat{\rho}'(2)\right]|\Psi_0\rangle}{\langle\Psi_0|\Psi_0\rangle}, \tag{1.58}$$

$$\overrightarrow{\boldsymbol{\chi}}^{\mathsf{R}}(12) \equiv \left(-\frac{i}{\hbar}\right)\Theta(t_1-t_2)\frac{\langle\Psi_0|\left[\hat{\rho}'(1),\hat{\boldsymbol{J}}'_{\mathsf{p}}(2)\right]|\Psi_0\rangle}{\langle\Psi_0|\Psi_0\rangle}, \tag{1.59}$$

$$\overleftrightarrow{\boldsymbol{\chi}}^{\mathsf{p;R}}(12) \equiv \left(-\frac{i}{\hbar}\right)\Theta(t_1-t_2)\frac{\langle\Psi_0|\left[\hat{\boldsymbol{J}}'_{\mathsf{p}}(1),\hat{\boldsymbol{J}}'_{\mathsf{p}}(2)\right]|\Psi_0\rangle}{\langle\Psi_0|\Psi_0\rangle}, \tag{1.60}$$

where $[A,B]=AB-BA$ is the commutator for bosonic operators. Their time-ordered counterparts are defined as,

$$\boldsymbol{\chi}(12) \equiv \left(-\frac{i}{\hbar}\right)\frac{\langle\Psi_0|T\left[\hat{\rho}'(1)\hat{\rho}'(2)\right]|\Psi_0\rangle}{\langle\Psi_0|\Psi_0\rangle}, \tag{1.61}$$

$$\overleftarrow{\boldsymbol{\chi}}(12) \equiv \left(-\frac{i}{\hbar}\right)\frac{\langle\Psi_0|T\left[\hat{\boldsymbol{J}}'_{\mathsf{p}}(1)\hat{\rho}'(2)\right]|\Psi_0\rangle}{\langle\Psi_0|\Psi_0\rangle}, \tag{1.62}$$

$$\overrightarrow{\boldsymbol{\chi}}(12) \equiv \left(-\frac{i}{\hbar}\right)\frac{\langle\Psi_0|T\left[\hat{\rho}'(1)\hat{\boldsymbol{J}}'_{\mathsf{p}}(2)\right]|\Psi_0\rangle}{\langle\Psi_0|\Psi_0\rangle}, \tag{1.63}$$

$$\overleftrightarrow{\boldsymbol{\chi}}^{\mathsf{p}}(12) \equiv \left(-\frac{i}{\hbar}\right)\frac{\langle\Psi_0|T\left[\hat{\boldsymbol{J}}'_{\mathsf{p}}(1)\hat{\boldsymbol{J}}'_{\mathsf{p}}(2)\right]|\Psi_0\rangle}{\langle\Psi_0|\Psi_0\rangle}. \tag{1.64}$$

Retarded and time-ordered susceptibilities are related in many ways, as discussed in Sec. C. Note that all the susceptibilities in Eqs. (1.61)–(1.64) refer to the charge density or charge current density instead of particle density or particle current density, and they are marked in bold to emphasize the difference. With the Kubo-Greenwood formula, the induced charge density and the induced charge current density due to electromagnetic fields can then be expressed with these susceptibilities,

$$\rho^{\mathsf{ind}}(1)=\int \mathrm{d}2\,\boldsymbol{\chi}^{\mathsf{R}}(12)\varphi^{\mathsf{ext}}(2)-\int \mathrm{d}2\,\overrightarrow{\boldsymbol{\chi}}^{\mathsf{p;R}}(12)\cdot\boldsymbol{A}(2), \tag{1.65}$$

and

$$\begin{aligned}\boldsymbol{J}^{\mathsf{ind}}(1)&=\int \mathrm{d}2\,\overleftarrow{\boldsymbol{\chi}}^{\mathsf{R}}(12)\varphi^{\mathsf{ext}}(2)-\frac{q_e^2}{m_e}n_e(1)\boldsymbol{A}(1)-\int \mathrm{d}2\,\overleftrightarrow{\boldsymbol{\chi}}^{\mathsf{p;R}}(12)\cdot\boldsymbol{A}(2)\\&=\int \mathrm{d}2\,\overleftarrow{\boldsymbol{\chi}}^{\mathsf{R}}(12)\varphi^{\mathsf{ext}}(2)-\int \mathrm{d}2\,\overleftrightarrow{\boldsymbol{\chi}}^{\mathsf{R}}(12)\cdot\boldsymbol{A}(2),\end{aligned} \tag{1.66}$$

with the definition of the *total current-current susceptibility*,

$$\overleftrightarrow{\boldsymbol{\chi}}^{\mathsf{R}}(12)\equiv\frac{q_e^2}{m_e}n_e(1)\delta(12)+\overleftrightarrow{\boldsymbol{\chi}}^{\mathsf{p;R}}(12). \tag{1.67}$$

1.4.1 Longitudinal dielectric function and electron energy-loss spectroscopy

In electron energy-loss spectroscopy (EELS), a material is hit by a beam of electrons with a narrow range of kinetic energies. Some electrons will encounter inelastic scattering, in which they lose energy and get deflected. The electron energy-loss spectrum can be roughly split into three different regions: zero loss, low energy loss (up until about 50 eV in energy loss) and high energy loss. Different scattering mechanisms dominate each region:

- **Zero-loss peak**: the elastic scattering is responsible for the zero-loss peak in EELS. Elastic scattering is a process where an incident electron is scattered by an atomic nucleus. Due to the great mass difference between an electron and a nucleus, the energy exchange between the incident electron and the nucleus is small.
- **Low-loss spectrum**: inelastic scatterings from valence electrons and other quasiparticles (e.g., plasmons, excitons, etc.) contribute to the low-loss spectrum, which contains information about the bandstructure and dielectric properties of the sample.
- **High-loss spectrum**: the energy loss from the core shell excitation process can be much higher, which is due to the fact that the binding energies of core shell electrons are around hundreds or even thousands of eV. If sufficient energy from an incident electron is transferred through an inelastic scattering process, a core shell electron can be excited to an unoccupied state above the Fermi level and shows up as an ionization edge.

We are interested in the low-loss spectrum where quasiparticle excitations play an important role. During an inelastic scattering, the momentum transfer is unnecessarily zero and it is possible to detect the change in both the energy and momentum in a carefully done EELS measurement. To model EELS in the low-loss region, we need to consider the scattering of an incident electron by the electrons in the solid. A general assumption is that the interaction between an incident electron and the existing electrons can be written as a summation of pairwise potentials [15], $\sum_i \varphi(\boldsymbol{r}-\boldsymbol{r}_i)$, where i denotes all the existing electrons in the solid. Incident electrons within an infinitesimal patch of cross-sectional area $\mathrm{d}\sigma$ will be scattered into a corresponding infinitesimal solid angle $\mathrm{d}\Omega$. The differential inelastic scattering cross section $\frac{\mathrm{d}^2\sigma}{\mathrm{d}\omega\mathrm{d}\Omega}$ characterizing both the momentum transfer and the energy loss is then proportional to the dynamical structure factor $S(\boldsymbol{q};\omega)$, which is further related to the imaginary part of the macroscopic component ($\boldsymbol{G}_1 = \boldsymbol{G}_2 = 0$) of the density-density susceptibility $\boldsymbol{\chi}^{\mathsf{R}}$ through the *fluctuation-dissipation theorem* [53],

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}\omega\mathrm{d}\Omega} \propto S(\boldsymbol{q};\omega) \propto -\mathsf{Im}\,\boldsymbol{\chi}^{\mathsf{R}}_{00}(\boldsymbol{q};\omega). \tag{1.68}$$

Recall the definition of the time-ordered inverse dielectric function in Eq. (1.27), we define its retarded counterpart as the *longitudinal dielectric function* [48, 49, 54, 55, 56, 57, 58],

$$\varepsilon^{\mathsf{R};-1}(12) = \delta(12) + \int \mathrm{d}3\, v(13)\chi^{\mathsf{R}}(32) = \delta(12) + \frac{1}{q_e^2}\int \mathrm{d}3\, v(13)\boldsymbol{\chi}^{\mathsf{R}}(32), \tag{1.69}$$

where $\boldsymbol{\chi}^{\mathsf{R}} \equiv q_e^2\chi^{\mathsf{R}}$ according to Eq. (1.57). In this way, we define the dimensionless *energy-loss function* $\mathcal{L}$ as,

$$\mathcal{L}(\boldsymbol{q};\omega) \equiv -\mathsf{Im}\,\varepsilon_{00}^{\mathsf{R};-1}(\boldsymbol{q};\omega). \tag{1.70}$$

$\varepsilon_{00}^{\mathsf{R};-1}(\boldsymbol{q};\omega)$ is calculated by taking the time and space Fourier transforms (Sec. A) of Eq. (1.69) and only keeping the macroscopic component.

1.4.2 Transverse dielectric function and optical spectra

In linear optics, a beam of low-energy photons (usually in the optical spectrum) is shined on the materials. These photons can be either absorbed, transmitted or reflected, which leads to different optical spectra such as absorption and reflection spectra. The light intensity is kept low enough such that dielectric responses stay in the linear regime and therefore relevant signals are proportional to the amplitude of the electric field of light. Once a photon is absorbed by an insulating sample, an electron-hole pair can be created with Coulomb interaction. These interacting electron-hole pairs are called excitons, and they could strongly modify the optical spectra going beyond the noninteracting picture. In particular, bound excitons show up below the fundamental band gap in the absorption spectrum and determine the so-called optical band gap, as shown in Fig. 1.2. Typical exciton binding energies are around 50 meV in conventional bulk semiconductors [8]. Unlike impurity-trapped bound states, excitons are usually delocalized and be able to propagate in the material. Traditionally, excitons are classified as Wannier excitons where the electron and the hole are weakly bound (exciton radius is larger than the unit cell), and Frenkel excitons where the electron-hole interaction is strong (exciton radius is smaller than the unit cell).

In this section, we use the linear response theory and the microscopic Maxwell's equations to model the response of a crystal to external electromagnetic fields of light. To avoid the discussion of surface charge and current, we introduce the polarization field $\boldsymbol{P}$ which is related to the induced charge density,

$$\nabla\cdot\boldsymbol{P} = -\rho^{\mathsf{ind}}. \tag{1.71}$$

The electric displacement vector $\boldsymbol{D}$ is defined as,

$$\boldsymbol{D} \equiv \varepsilon_0\boldsymbol{E} + \boldsymbol{P} \Rightarrow \nabla\cdot\boldsymbol{D} = \rho^{\mathsf{ext}}. \tag{1.72}$$

The induced current in the most rigorous form is given by,

$$\boldsymbol{J}^{\mathsf{ind}} = \nabla\times\boldsymbol{M} + \frac{\partial\boldsymbol{P}}{\partial t}, \tag{1.73}$$

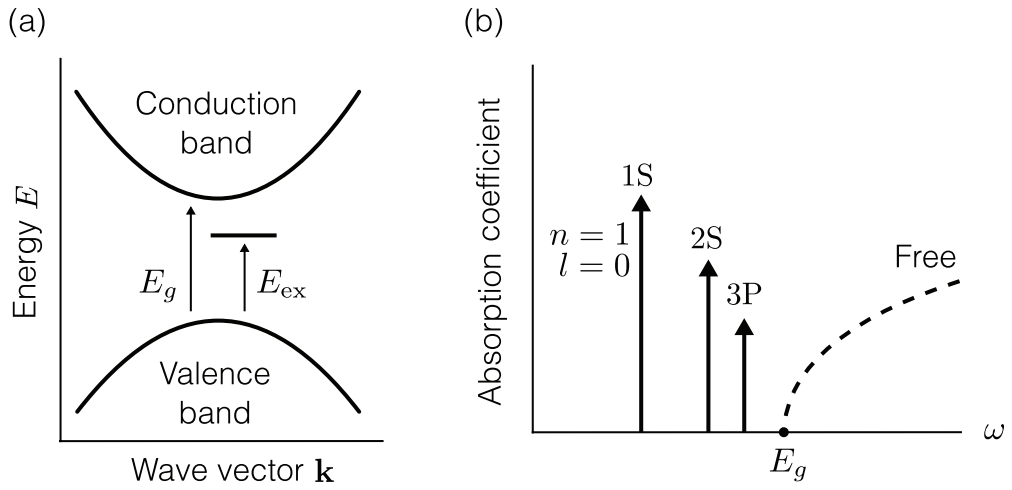

Figure 1.2: (a) Schematic of exciton energy levels. Quasiparticle bandstructure with band gap E_{g} and an exciton state at energy E_{ex} are indicated. (b) Sketch of the absorption spectrum of a dipole-allowed direct-gap semiconductor with the lowest three excitonic transitions shown. This figure is adapted from Ref. [8].

where $\boldsymbol{M}$ is the magnetization field. Within the optical frequency range, it is a good approximation to assume that the magnetic permeability is just equal to that of the vacuum, i.e., $\mu = \mu_0$, and therefore, $\boldsymbol{M} \approx 0$. In this way, we can only focus on the electric field and ignore the magnetic field. We decompose these physical quantities into their Fourier components (e.g., $\boldsymbol{D}(\boldsymbol{r}, t) \Rightarrow \boldsymbol{D}_{\boldsymbol{G}}(\boldsymbol{q}; \omega)$), and compute how the system responds. To be specific, we are interested in the macroscopic dielectric function tensor, $\boldsymbol{\varepsilon}_{\mathsf{M}}(\boldsymbol{q}; \omega) \equiv \varepsilon^{\alpha\beta}_{\mathsf{M}}(\boldsymbol{q}; \omega) = \varepsilon^{\alpha\beta}_{00}(\boldsymbol{q}; \omega)$, as given in the following expression,

$$D^{\alpha}_{0}(\boldsymbol{q}; \omega) = \sum_{\beta=x,y,z} \varepsilon^{\alpha\beta}_{\mathsf{M}}(\boldsymbol{q}; \omega) E^{\beta}_{0}(\boldsymbol{q}; \omega), \quad \forall \alpha = x, y, z. \tag{1.74}$$

We start from the definition of $\boldsymbol{D}$ in terms of $\boldsymbol{E}$ and $\boldsymbol{P}$,

$$\boldsymbol{D} = \varepsilon_0 \boldsymbol{E} + \boldsymbol{P} \Leftrightarrow \boldsymbol{D}_{\boldsymbol{G}} = \varepsilon_0 \boldsymbol{E}_{\boldsymbol{G}} + \boldsymbol{P}_{\boldsymbol{G}}, \tag{1.75}$$

and the induced current from the variation of $\boldsymbol{P}$,

$$\boldsymbol{J}^{\mathsf{ind}} = \frac{\partial \boldsymbol{P}}{\partial t} \Leftrightarrow \boldsymbol{J}^{\mathsf{ind}}_{\boldsymbol{G}} = -i\omega \boldsymbol{P}_{\boldsymbol{G}}. \tag{1.76}$$

And then we define a perturbing electric field $\boldsymbol{E}^{\mathsf{P}}$ following Ref. [45],

$$\boldsymbol{E}^{\mathsf{P}} \equiv -\nabla \varphi^{\mathsf{ext}} - \frac{\partial \boldsymbol{A}}{\partial t}. \tag{1.77}$$

This perturbing electric field is useful to simplify Eqs. (1.65) and (1.66),

$$\begin{aligned}\rho^{\mathsf{ind}}_{\boldsymbol{G}} &= \sum_{\boldsymbol{G}_1} \boldsymbol{\chi}^{\mathsf{R}}_{\boldsymbol{G}\boldsymbol{G}_1}\varphi^{\mathsf{ext}}_{\boldsymbol{G}_1} - \frac{1}{i\omega}\sum_{\boldsymbol{G}_1}\overrightarrow{\boldsymbol{\chi}}^{\mathsf{R}}_{\boldsymbol{G}\boldsymbol{G}_1}\cdot\left[\boldsymbol{E}^{\mathsf{P}}_{\boldsymbol{G}_1} + i(\boldsymbol{q}+\boldsymbol{G}_1)\varphi^{\mathsf{ext}}_{\boldsymbol{G}_1}\right] \\ &= -\frac{1}{i\omega}\sum_{\boldsymbol{G}_1}\overrightarrow{\boldsymbol{\chi}}^{\mathsf{R}}_{\boldsymbol{G}\boldsymbol{G}_1}\cdot\boldsymbol{E}^{\mathsf{P}}_{\boldsymbol{G}_1},\end{aligned} \tag{1.78}$$

and

$$\begin{aligned}\boldsymbol{J}^{\mathsf{ind}}_{\boldsymbol{G}} &= \sum_{\boldsymbol{G}_1} \overleftrightarrow{\boldsymbol{\chi}}^{\mathsf{R}}_{\boldsymbol{G}\boldsymbol{G}_1}\varphi^{\mathsf{ext}}_{\boldsymbol{G}_1} - \frac{1}{i\omega}\sum_{\boldsymbol{G}_1}\overleftrightarrow{\boldsymbol{\chi}}^{\mathsf{R}}_{\boldsymbol{G}\boldsymbol{G}_1}\cdot\left[\boldsymbol{E}^{\mathsf{P}}_{\boldsymbol{G}_1} + i(\boldsymbol{q}+\boldsymbol{G}_1)\varphi^{\mathsf{ext}}_{\boldsymbol{G}_1}\right] \\ &= -\frac{1}{i\omega}\sum_{\boldsymbol{G}_1}\overleftrightarrow{\boldsymbol{\chi}}^{\mathsf{R}}_{\boldsymbol{G}\boldsymbol{G}_1}\cdot\boldsymbol{E}^{\mathsf{P}}_{\boldsymbol{G}_1}.\end{aligned} \tag{1.79}$$

where we have used the Ward identities in Sec. D. Plug Eqs. (1.76) and (1.79) into Eq. (1.75), and we will get,

$$\boldsymbol{D}_{\boldsymbol{G}} = \varepsilon_0\boldsymbol{E}_{\boldsymbol{G}} + \frac{1}{-i\omega}\boldsymbol{J}^{\mathsf{ind}}_{\boldsymbol{G}} = \varepsilon_0\boldsymbol{E}_{\boldsymbol{G}} - \frac{1}{\omega^2}\sum_{\boldsymbol{G}_1}\overleftrightarrow{\boldsymbol{\chi}}_{\boldsymbol{G}\boldsymbol{G}_1}\cdot\boldsymbol{E}^{\mathsf{P}}_{\boldsymbol{G}_1}. \tag{1.80}$$

On the other hand, $\boldsymbol{E}^{\mathsf{P}}$ is related to the total electric field $\boldsymbol{E}$ by,

$$\begin{aligned}&\boldsymbol{E}^{\mathsf{p}} = \boldsymbol{E} + \nabla\varphi^{\mathsf{ind}}, \quad \nabla^2\varphi^{\mathsf{ind}} = -\frac{\rho^{\mathsf{ind}}}{\varepsilon_0} \\ \Leftrightarrow &\boldsymbol{E}_{\boldsymbol{G}} = \boldsymbol{E}^{P}_{\boldsymbol{G}} + \frac{(\boldsymbol{q}+\boldsymbol{G})}{\varepsilon_0\omega(\boldsymbol{q}+\boldsymbol{G})^2}\sum_{\boldsymbol{G}_1}\overrightarrow{\boldsymbol{\chi}}^{\mathsf{R}}_{\boldsymbol{G}\boldsymbol{G}_1}\cdot\boldsymbol{E}^{\mathsf{P}}_{\boldsymbol{G}_1} = \boldsymbol{E}^{\mathsf{P}}_{\boldsymbol{G}} + \frac{(\boldsymbol{q}+\boldsymbol{G})(\boldsymbol{q}+\boldsymbol{G})}{\varepsilon_0\omega^2(\boldsymbol{q}+\boldsymbol{G})^2}\cdot\sum_{\boldsymbol{G}_1}\overleftrightarrow{\boldsymbol{\chi}}^{\mathsf{R}}_{\boldsymbol{G}\boldsymbol{G}_1}\cdot\boldsymbol{E}^{\mathsf{P}}_{\boldsymbol{G}_1},\end{aligned} \tag{1.81}$$

where the Ward identities are used again. Employ the *macroscopic approximation* of $\boldsymbol{E}^{\mathsf{P}}$, $\boldsymbol{E}^{\mathsf{P}}_{\boldsymbol{G}} = \delta_{\boldsymbol{G}0}\boldsymbol{E}^{\mathsf{P}}_0$, and we will get,

$$\boldsymbol{E}_{\boldsymbol{0}} = \boldsymbol{E}^{\mathsf{P}}_{\boldsymbol{0}} + \frac{\boldsymbol{q}\boldsymbol{q}}{\varepsilon_0\omega^2 q^2}\cdot\overleftrightarrow{\boldsymbol{\chi}}^{\mathsf{R}}_{00}\cdot\boldsymbol{E}^{\mathsf{P}}_0 = \mathcal{M}\cdot\boldsymbol{E}^{\mathsf{P}}_0, \tag{1.82}$$

where $\mathcal{M} \equiv \mathcal{M}^{\alpha\beta}(\boldsymbol{q};\omega)$ is given by,

$$\mathcal{M}^{\alpha\beta}(\boldsymbol{q};\omega) = \delta_{\alpha\beta} + \sum_{\gamma}\frac{q_\alpha q_\gamma}{\varepsilon_0\omega^2 q^2}\overleftrightarrow{\boldsymbol{\chi}}^{\mathsf{R};\gamma\beta}_{00}(\boldsymbol{q};\omega), \quad \forall\alpha,\beta = x,y,z. \tag{1.83}$$

Luckily, the inverse matrix $\mathcal{M}^{-1}$ has a closed form [45],

$$\left(\mathcal{M}^{-1}\right)^{\alpha\beta} = \delta_{\alpha\beta} - \frac{1}{\omega}\frac{q_\alpha\overrightarrow{\boldsymbol{\chi}}^{\mathsf{R};\beta}_{00}}{\varepsilon_0 q^2 + \boldsymbol{\chi}^{\mathsf{R}}_{00}}, \tag{1.84}$$

which leads to an explicit relation between $\boldsymbol{D}_0$ and $\boldsymbol{E}_0$,

$$\boldsymbol{D}_0 = \left[\varepsilon_0\mathbb{1} - \frac{1}{\omega^2}\overleftrightarrow{\boldsymbol{\chi}}^{\mathsf{R}}_{00} + \frac{1}{\omega^2}\frac{\overleftrightarrow{\boldsymbol{\chi}}^{\mathsf{R}}_{00}\overrightarrow{\boldsymbol{\chi}}^{\mathsf{R}}_{00}}{\varepsilon_0 q^2 + \boldsymbol{\chi}^{\mathsf{R}}_{00}}\right]\cdot\boldsymbol{E}_0, \tag{1.85}$$

where $\mathbb{1} \equiv \delta_{\alpha\beta}$ is an identity tensor.

In summary, the *macroscopic transverse dielectric function tensor* is given by,

$$\boldsymbol{\varepsilon}_{\mathsf{M}} = \varepsilon_0 \mathbb{1} - \frac{1}{\omega^2} \overleftrightarrow{\boldsymbol{\chi}}^{\mathsf{R}}_{00} + \frac{1}{\omega^2} \frac{\overleftarrow{\boldsymbol{\chi}}^{\mathsf{R}}_{00} \overrightarrow{\boldsymbol{\chi}}^{\mathsf{R}}_{00}}{\varepsilon_0 q^2 + \boldsymbol{\chi}^{\mathsf{R}}_{00}}. \tag{1.86}$$

1.5 Bethe-Salpeter equation (BSE)

The Bethe-Salpeter equation (BSE) was initially proposed in 1951 [59] to study the bound-state problem for two interacting Fermi-Dirac particles considering relativistic effects. It finds many applications in different branches of theoretical physics, and particularly it can be used to describe different kinds of bound states in condensed matter, such as excitons (bound state of an electron-hole pair) [34], polarons (phonon-dressed electrons) [60], etc. The application of BSE to studying excitons from first principles was pioneered by Rohlfing and Louie [34, 35].

Because we need to study a two-particle problem, it is intuitive to introduce the two-particle Green's function G_2,

$$G_2(1234) \equiv \left(-\frac{i}{\hbar}\right)^2 \frac{\langle \Psi_0 | T[\hat{S}\hat{\psi}(1)\hat{\psi}^\dagger(2)\hat{\psi}(3)\hat{\psi}^\dagger(4)] | \Psi_0 \rangle}{\langle \Psi_0 | T[\hat{S}] | \Psi_0 \rangle}, \tag{1.87}$$

as well as the two-particle correlation function L,

$$L(1234) \equiv -G_2(1234) + G(12)G(34). \tag{1.88}$$

By generalizing the local potential in Eq. (1.20) to a nonlocal instantaneous potential $\varphi(12) = \varphi(\boldsymbol{x}_1, \boldsymbol{x}_2; t_1)\delta(t_1 - t_2)$ such that,

$$\hat{H}'(t) = \int \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}'\, \hat{\psi}^\dagger(\boldsymbol{x})\varphi(\boldsymbol{x}, \boldsymbol{x}'; t)\hat{\psi}(\boldsymbol{x}'), \tag{1.89}$$

we can relate L with G using the following expression,

$$\begin{aligned} L(1234) &= \frac{\delta G(12)}{\delta \varphi(43)} = -\int \mathrm{d}(56)\, G(15)\frac{\delta G^{-1}(56)}{\delta \varphi(43)} G(62) \\ &= G(14)G(32) + \int \mathrm{d}(5678) G(15)G(62)\frac{\delta \Sigma(56)}{\delta G(87)}\frac{\delta G(87)}{\delta \varphi(43)} \\ &= L_0(1234) + \int \mathrm{d}(5678) L_0(1265)\Xi(5678)L(8734). \end{aligned} \tag{1.90}$$

The noninteracting two-particle correlation function L_0 and the interaction kernel Ξ are defined as follows,

$$L_0(1234) \equiv G(14)G(32), \tag{1.91}$$

$$\Xi(5678) \equiv \frac{\delta \Sigma(56)}{\delta G(87)}. \tag{1.92}$$

Eq. (1.90) is the BSE for L. Note that L is preferred over G_2 because it gets rid of those disconnected diagrams in G_2 [41] and it is directly related to the susceptibilities introduced in Eqs. (1.61) – (1.64) [45]:

$$\boldsymbol{\chi}(12) = -i\hbar q_e^2 L(11^+22^+), \tag{1.93}$$

$$\overleftarrow{\boldsymbol{\chi}}(12) = -\frac{q_e^2\hbar^2}{2m_e}\left\{(\nabla_1 - \nabla_{1'})L(11'22^+)\right\}\Big|_{1'=1^+}, \tag{1.94}$$

$$\overrightarrow{\boldsymbol{\chi}}(12) = -\frac{q_e^2\hbar^2}{2m_e}\left\{(\nabla_2 - \nabla_{2'})L(11^+22')\right\}\Big|_{2'=2^+}, \tag{1.95}$$

$$\overleftrightarrow{\boldsymbol{\chi}}^{\mathsf{P}}(12) = \frac{i\hbar^3 q_e^2}{4m_e^2}\left\{(\nabla_1 - \nabla_{1'})(\nabla_2 - \nabla_{2'})L(11'22')\right\}|_{\substack{1'=1^+\\2'=2^+}}, \tag{1.96}$$

where $1^+ = \{\boldsymbol{x}_1 t_1^+\}$ and $2^+ = \{\boldsymbol{x}_2 t_2^+\}$.

The BSE formalism is usually combined with the GW method, and therefore it is often called the GW-BSE method, which is among the most accurate first-principles methods available to calculate optical properties of real materials. It has proven very effective in dealing with bulk semiconductors, metals, low-dimensional materials and nanodevices [61]. By combining Eqs. (1.47) and (1.92), we get the interaction kernel at the GW level [34, 45, 46],

$$\begin{aligned}\Xi(1234) &= i\hbar\frac{\delta}{\delta G(43)}\left(\Sigma_{\mathsf{H}}(12) + G(12)W(21)\right)\\ &= -i\hbar\delta(12)\delta(43)v(14) + i\hbar\delta(1,4)\delta(23)W(21)\\ &+ i\hbar G(12)\left[W(24)W(31^+) + W(23)W(41^+)\right]G(34)\\ &\approx -i\hbar\delta(12)\delta(43)v(14) + i\hbar\delta(14)\delta(23)W(21),\end{aligned} \tag{1.97}$$

where we have ignored terms with quadratic order of W. Those higher order terms correspond to the change of dielectric screening upon excitation, and it is assumed to be negligible in semiconductors and insulators [34]. In Eq. (1.97), we identify two kernels: the one with the bare Coulomb interaction is called the *exchange kernel*,

$$\Xi_v(1234) \equiv -i\hbar\delta(12)\delta(43)v(14), \tag{1.98}$$

and the one with the screened Coulomb interaction is called the *direct kernel*,

$$\Xi_W(1234) \equiv i\hbar\delta(14)\delta(23)W(21). \tag{1.99}$$

The direct kernel can be reduced to an attractive Coulomb potential when the electron-hole separation is large, and it is largely responsible for the exciton binding energy in semiconductors and insulators [34, 45]. The repulsive exchange kernel, on the other hand, is usually much smaller than the direct kernel in bulk semiconductors and it is responsible for the longitudinal-transverse splitting of excitonic states and mixing of different spin-polarized states [45]. The exchange kernel is proportional to the wavefunction overlap between the

electron and the hole, and it plays an important role in low-dimensional materials where the quantum confinement forces the electron and the hole to come close to each other [8, 62]. One caveat here is that directly solving Eq. (1.90) cannot give the answer to Eq. (1.86). We need to introduce another two-particle correlation function $\tilde{L}$ in order to calculate the macroscopic transverse dielectric function. To do so, we first split the exchange kernel into a short-range part Ξ_v^{SR} and a long-range part Ξ_v^{LR},

$$\begin{aligned} \Xi_v^{\mathsf{LR}}(1234) &\equiv -i\hbar\delta(12)\delta(43)v^{\mathsf{LR}}(14), \text{ with } v_{\boldsymbol{G}}^{\mathsf{LR}}(\boldsymbol{q}) \equiv \delta_{\boldsymbol{G}0}v_0(\boldsymbol{q}), \\ \Xi_v^{\mathsf{SR}}(1234) &\equiv \Xi_v(1234) - \Xi_v^{\mathsf{LR}}(1234). \end{aligned} \tag{1.100}$$

$\tilde{L}$ satisfies a BSE with the kernel $\Xi_v^{\mathsf{SR}} + \Xi_W$ such that,

$$\tilde{L}(1234) = L_0(1234) + \int \mathrm{d}(5678) L_0(1265) \left[\Xi_v^{\mathsf{SR}}(5678) + \Xi_W(5678)\right] \tilde{L}(8734). \tag{1.101}$$

$\tilde{L}$ is related to L through the following expression,

$$L(1234) = \tilde{L}(1234) + \int \mathrm{d}(5678) \tilde{L}(1265) \Xi_v^{\mathsf{LR}}(5678) L(8734). \tag{1.102}$$

Equation (1.101) is shown diagrammatically in Fig. 1.3.

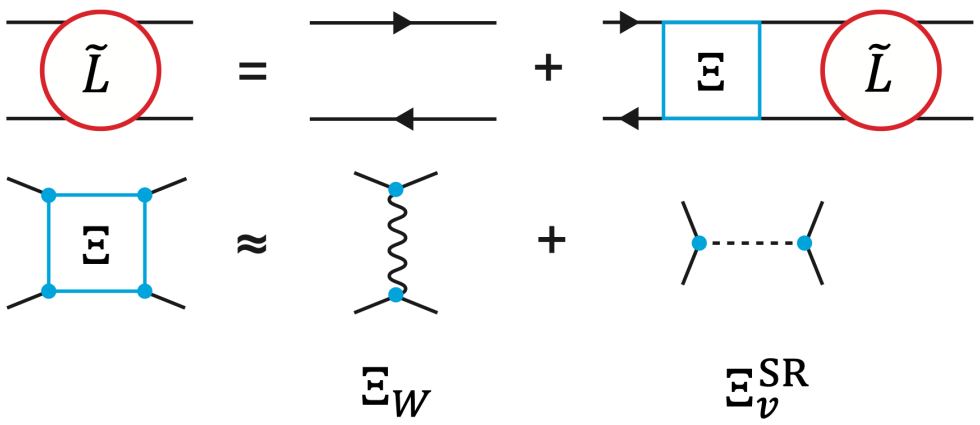

Figure 1.3: Top line: Bethe-Salpeter equation (BSE) for the two-particle correlation function $\tilde{L}$. Bottom line: approximate interaction kernels of the BSE in the top line, including the direct kernel Ξ_W and the short-range exchange kernel Ξ_v^{SR}. Arrowed lines denote single-particle Green's functions G. The wavy line denotes the screened Coulomb interaction W while the dashed line denotes the short-range part of the bare Coulomb interaction v. All the quantities are time-ordered.

Using the Ward identities, we can prove the following identities between the susceptibilities derived from $\tilde{L}$ and L,

$$\overleftrightarrow{\tilde{\boldsymbol{\chi}}}_{00} = \frac{\overleftrightarrow{\boldsymbol{\chi}}_{00}}{1 + \frac{\boldsymbol{\chi}_{00}}{\varepsilon_0 q^2}}, \tag{1.103}$$

and

$$\overleftrightarrow{\boldsymbol{\chi}}^{\mathsf{P}}_{00} - \overleftrightarrow{\tilde{\boldsymbol{\chi}}}^{\mathsf{P}}_{00} = \frac{1}{\varepsilon_0 q^2} \overleftrightarrow{\tilde{\boldsymbol{\chi}}}_{00} \overrightarrow{\boldsymbol{\chi}}_{00}. \tag{1.104}$$

Combine Eqs. (1.103) and (1.104), and set everything to be retarded, we can get a simpler expression of Eq. (1.86) [45, 58, 63],

$$\boldsymbol{\varepsilon}_{\mathsf{M}}(\boldsymbol{q};\omega) = \varepsilon_0 \mathbb{1} - \frac{1}{\omega^2} \overleftrightarrow{\tilde{\boldsymbol{\chi}}}^{\mathsf{R}}_{00}(\boldsymbol{q};\omega). \tag{1.105}$$

In practice, we solve the BSE for $\tilde{L}$ instead that for L, and then construct $\overleftrightarrow{\tilde{\boldsymbol{\chi}}}$ from $\tilde{L}$ to calculate Eq. (1.105).

To facilitate first-principles GW-BSE calculations, we ignore the dynamical effects in Ξ_W such that W is approximated by its instantaneous component [34],

$$W(1,2) \approx W(\boldsymbol{x}_1, \boldsymbol{x}_2)\delta(t_1 - t_2). \tag{1.106}$$

The approximations made in Eqs. (1.97) and (1.106) will in principle violate the conservation law, but numerous numerical results based on these approximations agree quite satisfactorily with experiments [34, 35, 61].

Since our goal is to calculate $\overleftrightarrow{\tilde{\boldsymbol{\chi}}}$ according to Eq. (1.96), it is then intuitive to study $L(11'22')$ with $1' = \boldsymbol{x}_1' t_1^+$ and $2' = \boldsymbol{x}_2' t_2^+$. In this case, both sides of Eq. (1.101) only depend on the time difference $t_1 - t_2$ and it is possible to do a time Fourier transform of Eq. (1.101) and then invert it to get a closed expression of $\tilde{L}$,

$$\tilde{L}^{-1}(\boldsymbol{x}_1, \boldsymbol{x}_1'; \boldsymbol{x}_2, \boldsymbol{x}_2'; \omega) = L_0^{-1}(\boldsymbol{x}_1, \boldsymbol{x}_1'; \boldsymbol{x}_2, \boldsymbol{x}_2'; \omega) - \left[\Xi_W(\boldsymbol{x}_1, \boldsymbol{x}_1'; \boldsymbol{x}_2, \boldsymbol{x}_2') + \Xi_v^{\mathsf{SR}}(\boldsymbol{x}_1, \boldsymbol{x}_1'; \boldsymbol{x}_2, \boldsymbol{x}_2')\right]. \tag{1.107}$$

Note that the interaction kernels are frequency-independent.

1.6 Two-dimensional (2D) materials

In 2D materials, electrons or holes are free to move parallel to the interface, but highly confined in the perpendicular direction. In recent two decades or so, the discovery of atomically thin solid materials has attracted huge interest in different fields of condensed matter physics, materials science, chemistry, electrical engineering, etc. Ever since the successful exfoliation of monolayer graphene [64], the size of the 2D materials family has grown enormously, and now we have a myriad of 2D materials, including metals [64, 65], semiconductors [66, 67, 68, 69], insulators [70, 71], magnets [72, 73, 74], topological materials [75, 76, 77], superconductors [78, 79, 80], Mott insulators [7, 81], etc. [82, 83] It is now even possible to

stack different 2D materials together and create van der Waals heterostructures with exotic properties and promising applications [84].

Recent years have also seen an outburst of first-principles studies of 2D materials, at the DFT level and beyond. With the development of first-principles GW and GW-BSE methods, researchers were able to discover and investigate a lot of fascinating excited-state physics in various 2D materials, such as nonhydrogenic Rydberg series of exciton states [85], environmental screening effects [62], exciton dispersion relations [86], dark exciton states [87], 2D plasmonics [88], etc. The reduced dimension introduces weakened overall dielectric screening, enhanced many-electron correlation and excitonic effects, as well as topological effects in the electronic structures of 2D materials. In this dissertation, I focus on the application of the first-principles GW and GW-BSE methods to the quasiparticle bandstructure, excitonic effects, exotic optical selection rules, strain engineering, as well as optical and magneto-optical properties of 2D semiconductors and insulators.

Chapter 2

Full-spinor GW and GW-BSE methods

2.1 Spinor wavefunctions and spin-orbit coupling

In materials consisting of heavy elements, relativistic effects cannot be ignored. Spin-orbit coupling (SOC), derived from the perturbation theory of the Dirac equation [89], describes the interaction between a particle's spin with its motion within the external potentials (ionic potentials, external electromagnetic fields, etc.). SOC is an important effect in condensed matter physics, exhibiting itself in various phenomena, such as the Dresselhaus effect [90], the Rashba effect [91, 92, 93], spin-valley locking [94, 95], topological phases [96, 97, 98, 99], etc. The SOC Hamiltonian,

$$\hat{H}_{\mathsf{SOC}} = \frac{\hbar^2}{4im_e^2c^2}\boldsymbol{\sigma}\cdot((\nabla V_{\mathsf{ext}})\times\nabla), \tag{2.1}$$

is spin-dependent, and therefore spin may not be a good quantum number once H_{SOC} is added to the Kohn-Sham equation. In this way, we have to introduce the two-component spinor Bloch waves,

$$\phi_{n\boldsymbol{k}}(\boldsymbol{x}) \equiv \begin{pmatrix}\phi^{\uparrow}_{n\boldsymbol{k}}(\boldsymbol{r})\\ \phi^{\downarrow}_{n\boldsymbol{k}}(\boldsymbol{r})\end{pmatrix} = \frac{1}{\sqrt{N_k}}e^{i\boldsymbol{k}\cdot\boldsymbol{r}}\begin{pmatrix}u^{\uparrow}_{n\boldsymbol{k}}(\boldsymbol{r})\\ u^{\downarrow}_{n\boldsymbol{k}}(\boldsymbol{r})\end{pmatrix} = \frac{1}{\sqrt{N_k\Omega}}\sum_{\boldsymbol{G}}e^{i(\boldsymbol{k}+\boldsymbol{G})\cdot\boldsymbol{r}}\begin{pmatrix}C^{\uparrow}_{\boldsymbol{G}n\boldsymbol{k}}\\ C^{\downarrow}_{\boldsymbol{G}n\boldsymbol{k}}\end{pmatrix}, \tag{2.2}$$

instead of the scalar one,

$$\phi_{n\boldsymbol{k}}(\boldsymbol{r}) \equiv \frac{1}{\sqrt{N_k}}e^{i\boldsymbol{k}\cdot\boldsymbol{r}}u_{n\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{N_k\Omega}}\sum_{\boldsymbol{G}}e^{i(\boldsymbol{k}+\boldsymbol{G})\cdot\boldsymbol{r}}C_{\boldsymbol{G}n\boldsymbol{k}}. \tag{2.3}$$

Note that $\hat{H}_{\mathsf{SOC}}$ stil respects the time-reversal symmetry, and therefore the Kramers' theorem (Sec. F) still holds in non-magnetic systems with SOC.

In practical DFT calculations with pseudopotentials, we have three different types of treatment of SOC: namely, non-relativistic (NR), scalar-relativistic (SR), and fully relativistic (FR). In the NR case, relativistic effects are not included in the construction of pseudopotentials. In the FR case, the relativistic AE Kohn-Sham equation is solved for atoms

and the SOC term is incorporated in the pseudopotentials. In the SR case, the relativistic AE Kohn-Sham equation is transformed into an equation for the large component only and then averaged over relevant spin-orbit components with the same orbital angular momentum quantum number l [100].

Previous GW and GW-BSE studies of materials with heavy elements usually treated SOC as a perturbation on the results without SOC [62, 85, 87]. This approach could get quantitatively good results when the systems have extra symmetries (e.g., a mirror symmetry) to guarantee that the spin is a good quantum number, such that there is a well-defined correspondence between spinor Bloch waves and scalar Bloch waves. To treat general systems involving heavy elements with arbitrary crystal symmetry, a rigorous full-spinor formalism is required. The formalism of the GW and GW-BSE methods introduced in the last section is very general and able to deal with spinor wavefunctions. In this section, we focus on possible pitfalls and several technical details about the full-spinor GW and GW-BSE methods.

2.2 Total dielectric function

Naïvely, one expects to take a 2×2 tensor form of the polarizability and the dielectric function in the spinor case, with tentative forms of $\varepsilon_{\boldsymbol{G}_1\boldsymbol{G}_2;\sigma_1\sigma_2}(\boldsymbol{q};\omega)$ and $\chi_{\boldsymbol{G}_1\boldsymbol{G}_2;\sigma_1\sigma_2}(\boldsymbol{q};\omega)$, respectively. For example, the RPA polarizability can be written as,

$$\chi^{0;\star}_{\sigma_1\sigma_2}(\boldsymbol{r}_1,\boldsymbol{r}_2;t_1-t_2) = -i\hbar G_{\sigma_1\sigma_2}(\boldsymbol{r}_1,\boldsymbol{r}_2;t_1-t_2)G_{\sigma_2\sigma_1}(\boldsymbol{r}_2,\boldsymbol{r}_1;t_2-t_1). \tag{2.4}$$

However, this practice is totally unnecessary because the bare Coulomb interaction v is blind to the spin (i.e., no spin indices in the expression):

$$v_{\boldsymbol{G}_1\boldsymbol{G}_2;\sigma_1\sigma_2}(\boldsymbol{q}) = \delta_{\boldsymbol{G}_1\boldsymbol{G}_2}\frac{e^2}{\varepsilon_0}\frac{1}{|\boldsymbol{q}+\boldsymbol{G}_1|^2} = \delta_{\boldsymbol{G}_1\boldsymbol{G}_2}v_{\boldsymbol{G}_1}(\boldsymbol{q}). \tag{2.5}$$

Recall Eq. (1.31),

$$W(12) = v(12) + \int \mathrm{d}(34)\, v(13)\chi(34)v(42). \tag{2.6}$$

If we take time and space Fourier transforms of Eq. (2.6) and plug in Eq. (2.5), we will get,

$$\begin{aligned} W_{\boldsymbol{G}_1\boldsymbol{G}_2;\sigma_1\sigma_2}(\boldsymbol{q};\omega) &= v_{\boldsymbol{G}_1\boldsymbol{G}_2;\sigma_1\sigma_2}(\boldsymbol{q};\omega) + \sum_{\substack{\boldsymbol{G}_3\boldsymbol{G}_4\\ \sigma_3\sigma_4}} v_{\boldsymbol{G}_1\boldsymbol{G}_3;\sigma_1\sigma_3}(\boldsymbol{q};\omega)\chi_{\boldsymbol{G}_3\boldsymbol{G}_4;\sigma_3\sigma_4}(\boldsymbol{q};\omega)v_{\boldsymbol{G}_4\boldsymbol{G}_2;\sigma_4\sigma_2}(\boldsymbol{q};\omega) \\ &= v_{\boldsymbol{G}_1}(\boldsymbol{q})\delta_{\boldsymbol{G}_1\boldsymbol{G}_2} + v_{\boldsymbol{G}_1}(\boldsymbol{q})\left(\sum_{\sigma_3\sigma_4}\chi_{\boldsymbol{G}_1\boldsymbol{G}_2;\sigma_3\sigma_4}(\boldsymbol{q};\omega)\right)v_{\boldsymbol{G}_2}(\boldsymbol{q}) = W_{\boldsymbol{G}_1\boldsymbol{G}_2}(\boldsymbol{q};\omega), \end{aligned} \tag{2.7}$$

which means W is also spin-blind. $\chi^\star$ and χ are therefore independent of spin indices. In this way, we only need to calculate the total polarizability by summing over all the four entries in the tensor form,

$$\chi^\star_{\boldsymbol{G}_1\boldsymbol{G}_2}(\boldsymbol{q};\omega) = \sum_{\sigma_1\sigma_2}\chi^\star_{\boldsymbol{G}_1\boldsymbol{G}_2;\sigma_1\sigma_2}(\boldsymbol{q};\omega), \tag{2.8}$$

as well as the total dielectric function,

$$\varepsilon_{\boldsymbol{G}_1\boldsymbol{G}_2}(\boldsymbol{q};\omega) = \delta_{\boldsymbol{G}_1\boldsymbol{G}_2} - v_{\boldsymbol{G}_1}(\boldsymbol{q})\chi^{\star}_{\boldsymbol{G}_1\boldsymbol{G}_2}(\boldsymbol{q};\omega). \tag{2.9}$$

2.3 Matrix elements involving spinor wavefunctions

In the *GW* and *GW*-BSE methods, we frequently encounter the following matrix elements via a Fourier transform,

$$\begin{aligned}&\langle n_1(\boldsymbol{k}-\boldsymbol{q})|e^{-i(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}}|n_2\boldsymbol{k}\rangle \\ &= \begin{cases}\sum_{\boldsymbol{G}_1\boldsymbol{G}_2} C^{*}_{\boldsymbol{G}_1 n_1(\boldsymbol{k}-\boldsymbol{q})} C_{\boldsymbol{G}_2 n_2\boldsymbol{k}} \frac{1}{\Omega}\int_\Omega \mathrm{d}\boldsymbol{\xi}\, e^{i(-\boldsymbol{G}_1-\boldsymbol{G}+\boldsymbol{G}_2)\cdot\boldsymbol{\xi}}, & \text{scalar.} \\ \sum_{\boldsymbol{G}_1\boldsymbol{G}_2;\sigma} C^{\sigma;*}_{\boldsymbol{G}_1 n_1(\boldsymbol{k}-\boldsymbol{q})} C^{\sigma}_{\boldsymbol{G}_2 n_2\boldsymbol{k}} \frac{1}{\Omega}\int_\Omega \mathrm{d}\boldsymbol{\xi}\, e^{i(-\boldsymbol{G}_1-\boldsymbol{G}+\boldsymbol{G}_2)\cdot\boldsymbol{\xi}}, & \text{spinor.}\end{cases}\end{aligned} \tag{2.10}$$

Note that the integrals are within a unit cell. Because the Fourier-transformed density operator $e^{-i(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}}$ is spin-diagonal, we only need to consider the inner product with the same spin component. The cost of each matrix element double in the spinor case. Besides, the number of bands also doubles in the spinor case compared with that in the scalar case, which leads to more matrix elements to be calculated in the spinor case. For example, the number of BSE matrix elements in the spinor case will be 16 times that in the scalar case, which means the computational cost of diagonalization of the BSE matrix in the spinor case will be 64 times that in the scalar case.

2.4 Macroscopic transverse dielectric function tensor

We use Kohn-Sham eigenstates to transform the BSE in Eq. (1.107) from the coordinate representation into the Bloch-wave representation (Sec. E),

$$\tilde{L}^{-1}_{\lambda\mu;\alpha\beta}(\omega) = L^{-1}_{0;\lambda\mu;\alpha\beta}(\omega) - \left[\Xi_{W;\lambda\mu;\alpha\beta} + \Xi^{\mathsf{SR}}_{v;\lambda\mu;\alpha\beta}\right]. \tag{2.11}$$

L_0^{-1} and the interaction kernels in the Bloch-wave representation are given by,

$$L^{-1}_{0;\lambda\mu;\alpha\beta}(\omega) = -i\hbar\delta_{\lambda\beta}\delta_{\mu\alpha}(n_\alpha - n_\beta)[\hbar\omega - (\epsilon_\beta - \epsilon_\alpha)], \tag{2.12}$$

$$\Xi_{W;\lambda\mu;\alpha\beta} = i\hbar\int \mathrm{d}\boldsymbol{x}_1\mathrm{d}\boldsymbol{x}_2\, \phi^*_\lambda(\boldsymbol{x}_1)\phi_\mu(\boldsymbol{x}_2)\phi^*_\alpha(\boldsymbol{x}_2)\phi_\beta(\boldsymbol{x}_1)W(\boldsymbol{x}_1,\boldsymbol{x}_2;\omega=0), \tag{2.13}$$

$$\Xi^{\mathsf{SR}}_{v;\lambda\mu;\alpha\beta} = -i\hbar\int \mathrm{d}\boldsymbol{x}_1\mathrm{d}\boldsymbol{x}_2\, \phi^*_\lambda(\boldsymbol{x}_1)\phi_\mu(\boldsymbol{x}_1)\phi^*_\alpha(\boldsymbol{x}_2)\phi_\beta(\boldsymbol{x}_2)v(\boldsymbol{r}_1,\boldsymbol{r}_2), \tag{2.14}$$

where n_α is the occupation number for the state ϕ_α, and the indices λ, μ, α, and β label the Bloch waves. The size of the BSE matrix in the Bloch-wave representation can be very large, which poses a formidable challenge for numerical diagonalization algorithms. To make our life easier, it is often the case that the Tamm-Dancoff approximation (TDA) is used to get rid of 3/4 of the matrix elements [34, 41]. To be specific, we only keep those band-to-band

transitions with positive transition energy in Eq. (2.12): that is, only the transitions from a valence state ($n_\alpha = 1$) to a conduction state ($n_\beta = 0$) are considered. Negative-energy solutions of the full BSE are inferred from the corresponding positive-energy solutions, and the coupling of these two sets of solutions is ignored. It is widely recognized that, for bulk semiconductors and metals, the TDA only introduces a very small error in the calculated optical properties. However, in the case of low-dimensional systems with incident light polarized along a confined direction, the TDA may incur large errors [101, 102].

Moreover, from the translational symmetry of the crystal,

$$\boldsymbol{\chi}(\boldsymbol{r}_1 + \boldsymbol{R}, \boldsymbol{r}_2 + \boldsymbol{R}; \omega) \equiv \boldsymbol{\chi}(\boldsymbol{r}_1, \boldsymbol{r}_2; \omega), \tag{2.15}$$

where $\boldsymbol{R}$ is a lattice vector, we know that a center-of-mass momentum $\boldsymbol{Q}$ can be assigned to label the eigenvalues and eigenstates from the solution of BSE [86]. Under these considerations, the BSE for one $\boldsymbol{Q}$ can be reformulated as an eigenvalue problem in the Bloch-wave representation,

$$A^{S\boldsymbol{Q}}_{cv\boldsymbol{k}}(\epsilon_{c\boldsymbol{k}} - \epsilon_{v(\boldsymbol{k}-\boldsymbol{Q})}) + \sum_{c'v'\boldsymbol{k}'} A^{S\boldsymbol{Q}}_{c'v'\boldsymbol{k}'} \langle cv, \boldsymbol{k}; \boldsymbol{Q}|K|c'v', \boldsymbol{k}'; \boldsymbol{Q}\rangle = A^{S\boldsymbol{Q}}_{cv\boldsymbol{k}} \Omega_{S\boldsymbol{Q}}, \tag{2.16}$$

where $K \equiv \frac{1}{-i\hbar}(\Xi_W + \Xi_v^{\mathsf{SR}})$ is the conventional interaction kernel in the unit of energy [34]. For an eigenstate $|S\boldsymbol{Q}\rangle$, $A^{S\boldsymbol{Q}}_{cv\boldsymbol{k}}$ denotes the amplitude of each free electron-hole pair $|cv, \boldsymbol{k}; \boldsymbol{Q}\rangle = \phi^*_{v(\boldsymbol{k}-\boldsymbol{Q})}\phi_{c\boldsymbol{k}}$ involving a valence state $\phi_{v(\boldsymbol{k}-\boldsymbol{Q})}$ and a conduction state $\phi_{c\boldsymbol{k}}$. If we are only interested in the $\boldsymbol{Q} \to 0$ limit, as in the case of optical absorption, the $\boldsymbol{Q}$ index can be dropped. In this way, we only consider direct transitions where the electron and hole states have the same $\boldsymbol{k}$. Equation (2.16) can then be simplified as,

$$A^{S}_{cv\boldsymbol{k}}(\epsilon_{c\boldsymbol{k}} - \epsilon_{v\boldsymbol{k}}) + \sum_{c'v'\boldsymbol{k}'} A^{S}_{c'v'\boldsymbol{k}'} \langle cv, \boldsymbol{k}|K|c'v', \boldsymbol{k}'\rangle = A^{S}_{cv\boldsymbol{k}} \Omega_S. \tag{2.17}$$

The exciton eigenstate $|S\rangle = \sum_{cv\boldsymbol{k}} A^S_{cv\boldsymbol{k}}|cv, \boldsymbol{k}\rangle = \sum_{cv\boldsymbol{k}} A^S_{cv\boldsymbol{k}}\phi^*_{v\boldsymbol{k}}\phi_{c\boldsymbol{k}}$ is a coherent superposition of free electron-hole pairs at different k-points, as shown in Fig. 2.1.

The eigenvalues Ω_S and eigenstates $|S\rangle$ from the diagonalization of Eq. (2.17) can be used to construct $\varepsilon^{\alpha\beta}_{\mathsf{M}}(\boldsymbol{q} \to 0; \omega)$,

$$\begin{aligned}
&\varepsilon^{\alpha\beta}_{\mathsf{M}}(\boldsymbol{q} \to 0; \omega) \\
&= \varepsilon_0 \delta_{\alpha\beta}(1 - \frac{\omega_p^2}{\omega^2}) - \frac{1}{\omega^2 N_k \Omega} \sum_S \langle 0|\hat{j}^\alpha_{\mathsf{p}}|S\rangle\langle S|\hat{j}^\beta_{\mathsf{p}}|0\rangle \left[\frac{1}{\hbar\omega - \Omega_S + i\eta} - \frac{1}{\hbar\omega + \Omega_S + i\eta}\right].
\end{aligned} \tag{2.18}$$

where $\eta \to 0^+$, $\alpha, \beta = x, y, z$. Matrix elements of the paramagnetic current operator is given by,

$$\langle 0|\hat{j}^\alpha_{\mathsf{p}}|S\rangle = \sum_{cv\boldsymbol{k}} A^S_{cv\boldsymbol{k}} \langle v\boldsymbol{k}|\hat{j}^\alpha_{\mathsf{p}}|c\boldsymbol{k}\rangle. \tag{2.19}$$

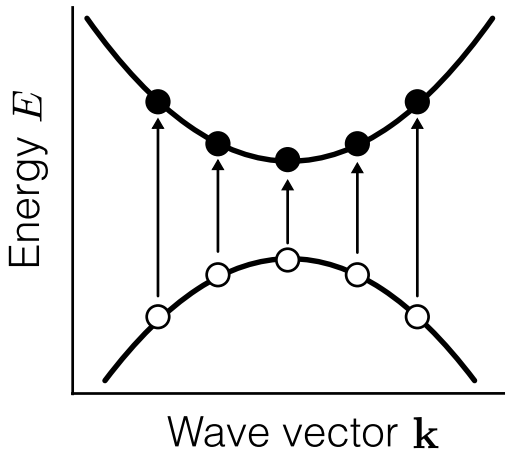

Figure 2.1: Schematic of an exciton state as a coherent superposition of free electron-hole pairs at different k-points. This figure is adapted from Ref. [8].

To avoid numerical instability around $\omega = 0$, we calculate the imaginary part first using the expression,

$$\mathsf{Im}\,\varepsilon_{\mathsf{M}}^{\alpha\beta}(\boldsymbol{q} \to 0; \omega) = \frac{\pi\hbar^2}{N_k\Omega} \sum_S \frac{1}{\Omega_S^2} \langle 0|\hat{j}_{\mathsf{p}}^{\alpha}|S\rangle\langle S|\hat{j}_{\mathsf{p}}^{\beta}|0\rangle \delta(\hbar\omega - \Omega_S). \tag{2.20}$$

And then the real part of $\varepsilon_{\mathsf{M}}^{\alpha\beta}$ can be easily evaluated using the Kramers-Krönig relation,

$$\begin{aligned} \mathsf{Re}\,\varepsilon_{\mathsf{M}}^{\alpha\beta}(\boldsymbol{q} \to; \omega) &= \mathsf{Re}\,\varepsilon_{\mathsf{M}}^{\alpha\beta}(\boldsymbol{q} \to 0; \infty) + \frac{1}{\pi}\mathcal{P}\int_{-\infty}^{\infty} \mathrm{d}\omega' \, \frac{\mathsf{Im}\,\varepsilon_{\mathsf{M}}^{\alpha\beta}(\boldsymbol{q} \to 0; \omega')}{\omega' - \omega} \\ &= \varepsilon_0\delta_{\alpha\beta} - \frac{\hbar^2}{N\Omega} \sum_S \frac{1}{\Omega_S^2} \langle 0|\hat{j}_{\mathsf{p}}^{\alpha}|S\rangle\langle S|\hat{j}_{\mathsf{p}}^{\beta}|0\rangle \frac{(\hbar\omega - \Omega_S)}{(\hbar\omega - \Omega_S)^2 + \eta^2}. \end{aligned} \tag{2.21}$$

We also follow the convention and define the relative dielectric function tensor (also called the relative permittivity),

$$\boldsymbol{\varepsilon}(\omega) \equiv \boldsymbol{\varepsilon}_{\mathsf{M}}(\boldsymbol{q} \to 0; \omega)/\varepsilon_0, \tag{2.22}$$

as a ratio relative to the vacuum permittivity ε_0.

If scalar wavefunctions are used, we can rearrange band indices such that two consecutive degenerate states have exactly the same real-space wavefunction, but with opposite spin polarization. The selection rule of direct and exchange kernels then gives us a simple form of the BSE matrix. If we perform an unitary transformation from the original band-to-band transition into spin-singlet and spin-triplet electron-hole pairs, we will have a block-diagonal form of the BSE matrix, where spin-singlet states and spin-triplet states can be calculated separately, as discussed in Ref. [34].

2.5 Matrix elements of the paramagnetic current density operator

In this section, we discuss how to evaluate matrix elements of the paramagnetic current density operator $\langle v\boldsymbol{k}|\hat{j}^{\alpha}_{\mathsf{p}}|c\boldsymbol{k}\rangle$ in Eq. (2.19). With a local potential, the matrix element can be expanded as,

$$\begin{aligned}\langle v\boldsymbol{k}|\hat{j}^{\alpha}_{\mathsf{p}}|c\boldsymbol{k}\rangle &= \langle v\boldsymbol{k}|\frac{-i\hbar q_e}{2m_e}\left(\overrightarrow{\partial_\alpha}-\overleftarrow{\partial_\alpha}\right)|c\boldsymbol{k}\rangle \\ &= \frac{-i\hbar q_e}{2m_e}\int \mathrm{d}\boldsymbol{x}_1\left[\phi^*_{v\boldsymbol{k}}(\boldsymbol{x}_1)(\partial_\alpha\phi_{c\boldsymbol{k}}(\boldsymbol{x}_1))-(\partial_\alpha\phi^*_{v\boldsymbol{k}}(\boldsymbol{x}_1))\phi_{c\boldsymbol{k}}(\boldsymbol{x}_1)\right],\end{aligned} \tag{2.23}$$

where $\overleftarrow{\partial_\alpha}$ acts on the left function and $\overrightarrow{\partial_\alpha}$ acts on the right one. It is easy to prove that Eq. (2.23) can be simplified using the momentum operator $\hat{\boldsymbol{p}} = -i\hbar\nabla$,

$$\langle v\boldsymbol{k}|\hat{j}^{\alpha}_{\mathsf{p}}|c\boldsymbol{k}\rangle = \frac{q_e}{m_e}\langle v\boldsymbol{k}|\hat{p}_\alpha|c\boldsymbol{k}\rangle. \tag{2.24}$$

In a DFT calculation with semilocal pseudopotentials [103], however, we need to add the contribution from the nonlocal part of the pseudopotential V_{NL} to the original $\hat{\boldsymbol{p}}$,

$$\hat{\boldsymbol{p}} \to -i\hbar\nabla - \frac{im_e}{\hbar}[\boldsymbol{r}, V_{\mathsf{NL}}]. \tag{2.25}$$

2.6 Benchmark

In this section, we benchmark our implementation of the full-spinor GW and GW-BSE methods on a conventional semiconductor, bulk silicon. Implementation in support of the spinor wavefunctions in association with SOC is tested against the experimental absorption spectrum. The experimental lattice constant $a = 5.43\,\text{Å}$ is used. We use ONCVPSP pseudopotentials with the PBE exchange-correlation functional [104, 105]. DFT calculations are performed using the `Quantum ESPRESSO` package [106, 107]. GW and GW-BSE calculations are performed using the `BerkeleyGW` package [108].

We first compare the DFT bandstructure using NR, SR, and FR pseudopotentials shown in Fig. 2.2. These three bandstructures look almost the same in Fig. 2.2a. The indirect band gap at the DFT-PBE level is calculated to be 0.57 eV, 0.55 eV, and 0.54 eV for NR, SR, and FR, respectively. These band gap values are far smaller than the experimental value of 1.12 eV (at 300 K) and 1.17 eV (at 0 K) [109]. The spin-orbit splitting of the valence bands at Γ is calculated to be 48.5 meV in the FR case as shown in Fig. 2.2b, in good agreement with the experimental value of $42.6 \sim 44.1$ meV [110, 111, 112, 113].

We then use the FR DFT results as a starting point to calculate the GW self-energy correction at the G_0W_0 level on an $8 \times 8 \times 8$ k-grid. The dynamical screening effect is

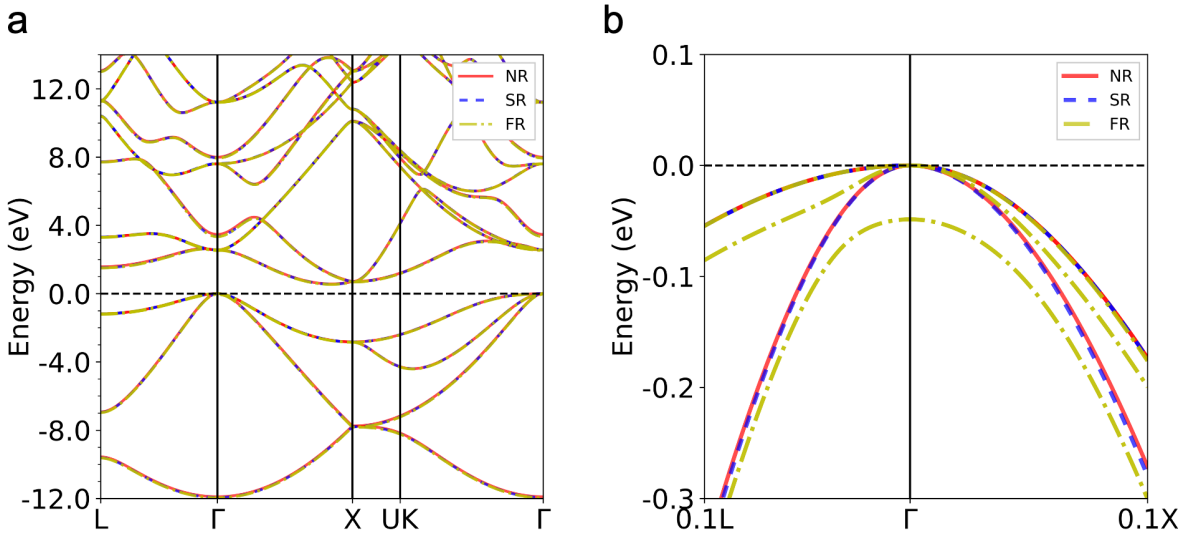

Figure 2.2: (a) DFT-PBE bandstructures of bulk Si using the non-relativistic (NR, red solid), scalar-relativistic (SR, blue dashed), and fully relativistic (FR, yellow dash-dotted) pseudopotentials, respectively. *GW* calculations are performed at the G_0W_0 level with the generalized plasmon-pole model [9]. (b) The same bandstructures in (a) around Γ.

incorporated through the generalized plasmon-pole model [9]. The dielectric function is computed with a cutoff energy of 10 Ry. A total of 400 bands are used in calculating the dielectric function and the self-energy correction. A linear interpolation method [34] is used to get the quasiparticle bandstructure shown in Figs. 2.3 and 2.4. We obtain a fundamental quasiparticle band gap of 1.17 eV, in excellent agreement with experiment. The spin-orbit splitting of the valence bands at the Γ point is 49.9 meV after the self-energy correction, barely changed from the DFT value of 48.5 meV as shown in Fig. 2.3b. These results demonstrate the great predictive power of the *GW* method, as well as the reliability of our full-spinor implementation of the *GW* method. We also present the *GW* bandstructure in the SR case in Fig. 2.4, where the indirect quasiparticle band gap is calculated to be 1.18 eV.

We further calculate the absorption spectrum of bulk Si in the FR case as shown in Fig. 2.5, considering transitions from the top six valence bands to the bottom eight conduction bands on a $20 \times 20 \times 20$ k-grid. Note that the number of bands double in the FR case compared with that in the SR or NR case. The calculated imaginary part of the *GW*-BSE

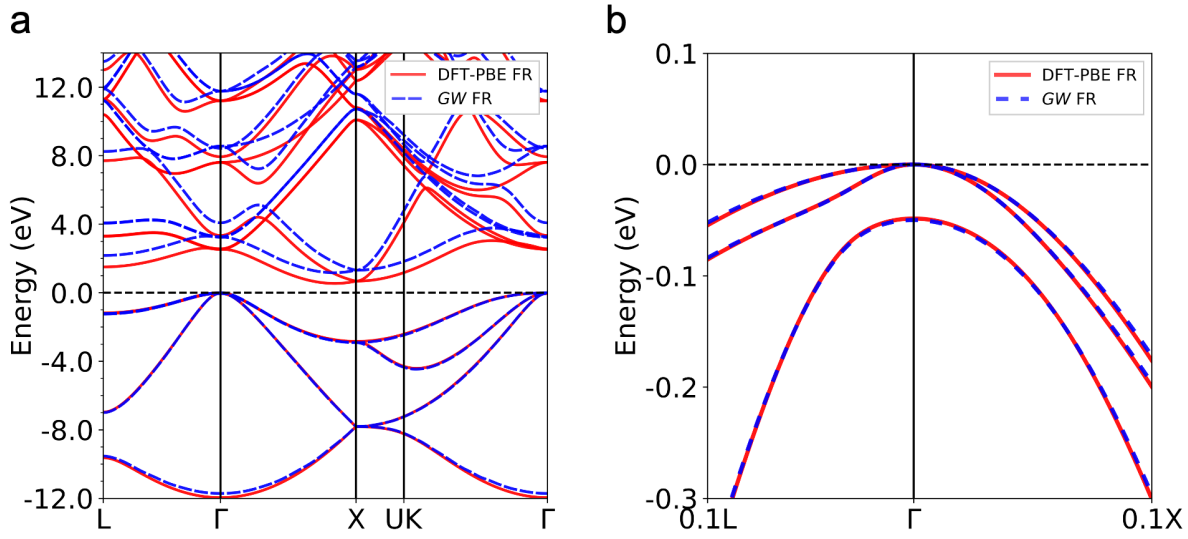

Figure 2.3: (a) DFT-PBE (red solid) and *GW* (blue dashed) bandstructures of bulk Si using the fully relativistic (FR) pseudopotential. The *GW* calculation is performed at the G_0W_0 level with the generalized plasmon-pole model. Spin-orbit coupling is included in both calculations within the full-spinor formalism. (b) The same bandstructures in (a) around Γ.

(with electron-hole interaction) dielectric function, $\varepsilon_2 \equiv \mathsf{Im}\,\varepsilon_{\alpha\alpha}(\omega)$, $\forall\alpha = x, y, z$, greatly improves agreement with the experimental data [10] compared with the *GW*-RPA (without electron-hole interaction) result, consistent with previous works using scalar wavefunctions [108]. Note that due to the cubic symmetry of bulk Si, we have $\varepsilon_{xx} = \varepsilon_{yy} = \varepsilon_{zz}$. We also compare the *GW*-BSE ε_2 in FR and SR cases, as shown in Fig. 2.5b. These two curves overlap with each other because the SOC effect is very weak in Si (also evident in Fig. 2.2a), which proves the validity of our full-spinor *GW*-BSE method.

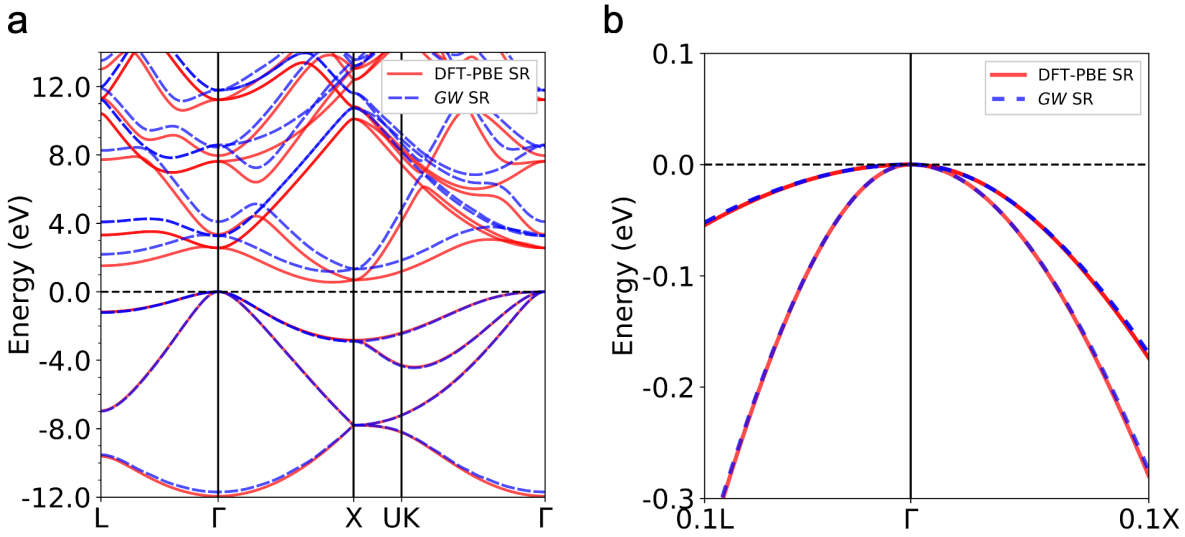

Figure 2.4: (a) DFT-PBE (red solid) and *GW* (blue dashed) bandstructures of bulk Si using the scalar-relativistic (SR) pseudopotential. The *GW* calculation is performed at the G_0W_0 level with the generalized plasmon-pole model. (b) The same bandstructures in (a) around Γ.

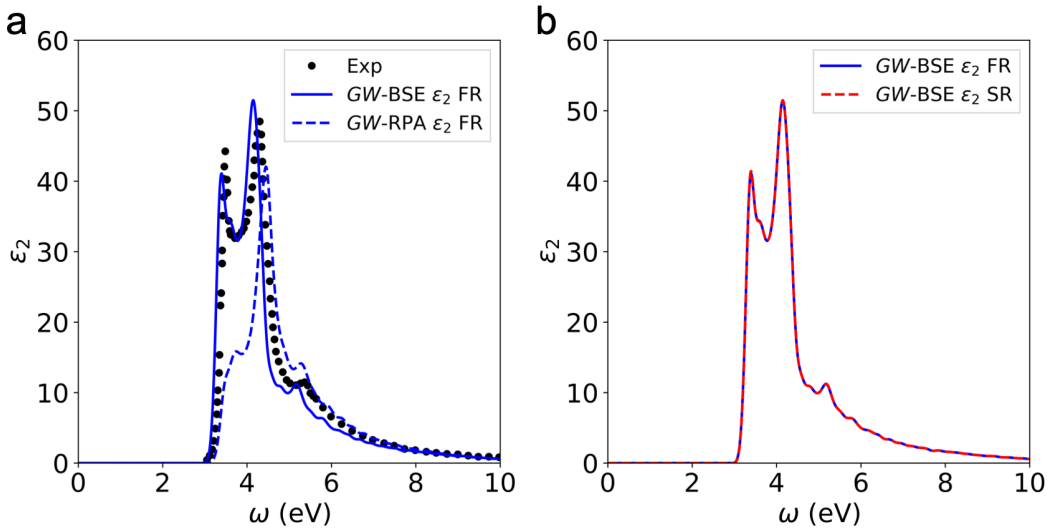

Figure 2.5: (a) Comparison of experimental (solid dots) and theoretical absorption spectra of bulk Si using the fully relativistic (FR) pseudopotential and full-spinor formalism. The imaginary part of dielectric function ε_2 at the GW-BSE level (blue solid) is calculated with the electron-hole interaction, while ε_2 at the GW-RPA level (blue dashed) is calculated without the electron-hole interaction. The experimental data is extracted from Ref. [10]. (b) Comparison of GW-BSE ε_2 between the full-spinor formalism using the FR pseudopotential (blue solid) and the scalar formalism using the scalar-relativistic (SR) pseudopotential (red dashed). A constant 80 meV Gaussian broadening is used.

Chapter 3

First-principles modeling of magneto-optics

3.1 Magneto-optics

In magneto-optics, a linearly polarized light propagating through a medium is modified by the presence of a magnetic field, where the left and right circularly polarized components propagate at different speeds (with different dielectric functions). This is because the time-reversal symmetry is broken by the magnetic field, and therefore left and right circularly polarized lights are no longer equivalent to each other. There are several important magneto-optical (MO) effects, such as the Faraday effect (FE) [114], the magneto-optical Kerr effect (MOKE) [115, 116], and the Voigt effect [117], etc. In the FE, the polarization change of the transmitted light is measured, while in the MOKE, the polarization change of the reflected light is measured.

The polarization of a linearly polarized light reflected from a magnetic material will be modified in two ways: 1) the polarization plane of the reflected light is rotated by a so-called Kerr angle θ_{K}; 2) the reflected light becomes elliptically polarized, characterized by the Kerr ellipticity χ_{K}. There are three different MOKE setups [118] depending on the propagation direction of the light and the direction of magnetization: the polar MOKE (P-MOKE), the longitudinal MOKE and the transversal MOKE, as shown in Fig. 3.1. Similar quantities and setups can be defined for the FE.

There have been a wide range of applications of MO effects, such as MO recording [119], the examination of surface magnetism [120, 121], visualization of magnetic domains and domain walls [122, 123], etc. Moreover, the recent discovery of magnetism in atomically thin layers of van der Waals crystals has created great opportunities for exploring light-matter interactions and MO phenomena in the 2D limit. Recent measurements using MOKE have led to the discovery of 2D magnets and demonstrated their rich magnetic behaviors [72, 73].

MO effects stem from the coupling between photons and the orbital motion of spin-polarized electrons, which is further interacting with the spin degree of freedom via the SOC

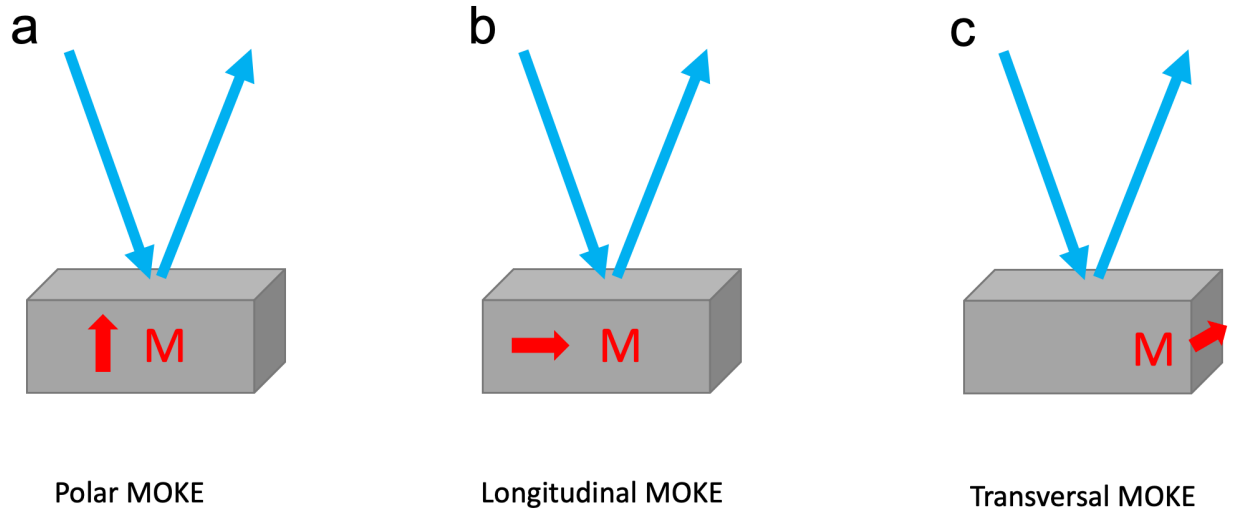

Figure 3.1: Schematic of MOKE setups: (a) the polar MOKE (P-MOKE), (b) the longitudinal MOKE (L-MOKE), and (c) the transversal MOKE (T-MOKE). Propagation directions of the incident and reflected lights are denoted by arrows. In the P-MOKE, the magnetization ($\boldsymbol{M}$) is perpendicular to the reflection surface and parallel to the incidence plane. In the L-MOKE, $\boldsymbol{M}$ is parallel to both the reflection surface and the incidence plane. In the T-MOKE, $\boldsymbol{M}$ is parallel to the reflection surface and perpendicular to the incidence plane.

effect. Both the spin splitting and SOC are required to achieve non-zero MO effects, as discussed by Argyres using the perturbation theory [124]. According to Ref. [124], MO signals are to the lowest order linear in the SOC strength. There have been many early numerical works – either model calculations [125] or first-principles calculations [126, 127, 128, 129] – investigating the dependence of MO signals on the spin splitting and SOC. It has been shown that MOKE signals are vanishing if SOC is not included in the calculation, and a linear dependence on the SOC strength has also been confirmed numerically [126, 128]. The dependence on the spin splitting through the magnetic exchange interaction is more complicated, beyond a simple linear relation. But there are no MO signals when the spin splitting is absent [118]. There have been a surge of first-principles studies on MO effects and related anomalous Hall conductivity in recent decades [130, 131, 132, 133, 134, 135], due to the popularization of first-principles modeling programs and ever-increasing computation power. However, these works are all based on the RPA dielectric function, where the critical electron-hole interaction has been missing. The essence of theoretical modeling of MO effects lies in accurately accounting for both the diagonal and off-diagonal elements of the frequency-dependent transverse dielectric function tensor, which are readily available from our *GW* and *GW*-BSE calculations.

3.2 Normal modes of light in materials

In the P-MOKE configuration with at least C_3 rotational symmetry along the spin polarization direction (z-axis), the frequency-dependent relative dielectric function tensor as a function of the magnetic field takes the following form,

$$\boldsymbol{\varepsilon}(\omega, \boldsymbol{B}) = \begin{pmatrix} \varepsilon_{xx}(\omega, \boldsymbol{B}) & \varepsilon_{xy}(\omega, \boldsymbol{B}) & 0 \\ -\varepsilon_{xy}(\omega, \boldsymbol{B}) & \varepsilon_{xx}(\omega, \boldsymbol{B}) & 0 \\ 0 & 0 & \varepsilon_{zz}(\omega, \boldsymbol{B}), \end{pmatrix}, \tag{3.1}$$

which is proven in Sec. G. The Fresnel equation for the normal modes is given by,

$$\left(n^2 \mathbb{1} - \boldsymbol{\varepsilon} - \boldsymbol{n} : \boldsymbol{n}\right) \cdot \boldsymbol{E} = 0, \tag{3.2}$$

where $\boldsymbol{n}$ is the complex refractive index, $\boldsymbol{n} = \frac{c\boldsymbol{k}}{\omega}$, with the wave vector $\boldsymbol{k}$ of light. After setting the normal incidence condition ($\boldsymbol{k} \parallel \hat{\boldsymbol{e}}_z$) and solving the Fresnel equation with the dielectric function tensor in Eq. (3.2), we get the normal modes as σ^+ and σ^- circularly polarized plane waves, with distinct refractive indices,

$$(n^{\pm}(\omega, \hat{\boldsymbol{e}}_z))^2 = \varepsilon_{xx}(\omega, B\hat{\boldsymbol{e}}_z) \pm i\varepsilon_{xy}(\omega, B\hat{\boldsymbol{e}}_z), \tag{3.3}$$

where $+(-)$ in $n^{\pm}$ denotes the circularly polarized light with the complex electric field amplitude along the direction of the spherical basis:

$$\hat{\boldsymbol{e}}_{\pm} = \frac{\mp}{\sqrt{2}}(\hat{\boldsymbol{e}}_x \pm i\hat{\boldsymbol{e}}_y). \tag{3.4}$$

Here $\hat{\boldsymbol{e}}_\alpha$, $\alpha = x, y, z$ denotes unit vectors along each direction. The distinction of refractive index between σ^+ and σ^- circularly polarized lights leads to the phenomena called magnetic circular dichroism (MCD), where the absorption of σ^+ and σ^- lights are different in a sample under a magnetic field or with intrinsic magnetization.

There are other symmetries of the dielectric function tensor regarding the direction of magnetization and the off-diagonal elements. According to the Onsager reciprocal relations [136], we have,

$$\varepsilon_{\alpha\beta}(\omega, \boldsymbol{B}) = \varepsilon_{\beta\alpha}(\omega, -\boldsymbol{B}), \quad \forall \alpha, \beta = x, y, z. \tag{3.5}$$

3.3 Kerr signals and Faraday signals

We consider a polar setup as shown in Fig. 3.2a. Suppose the incident light is linearly polarized along the x-axis. Its electric field can be described by the following expression,

$$\begin{aligned} \boldsymbol{E}^{\mathsf{i}} &= \tilde{\boldsymbol{E}}_0^{\mathsf{i}} e^{i(-kz+\omega t)}, \\ \tilde{\boldsymbol{E}}_0^{\mathsf{i}} &= \tilde{E}_{0x}^{\mathsf{i}} \hat{\boldsymbol{e}}_x = \tilde{E}_{0+}^{\mathsf{i}} \hat{\boldsymbol{e}}_+ + \tilde{E}_{0-}^{\mathsf{i}} \hat{\boldsymbol{e}}_-, \end{aligned} \tag{3.6}$$

where $\tilde{E}^{\mathsf{i}}_0$ is the complex amplitude of the electric field at $t = 0$ and $z = 0$. According to Eq. (3.6), $\tilde{E}^{\mathsf{i}}_{0+} = -\tilde{E}^{\mathsf{i}}_{0-} = -\tilde{E}^{\mathsf{i}}_{0x}/\sqrt{2}$. The electric field of the reflected light is given by,

$$\begin{aligned} \boldsymbol{E}^{\mathsf{r}} &= \tilde{\boldsymbol{E}}^{\mathsf{r}}_0 e^{i(+kz-\omega t)}, \\ \tilde{\boldsymbol{E}}^{\mathsf{r}}_0 &= \tilde{E}^{\mathsf{r}}_{0x}\hat{\boldsymbol{e}}_x + \tilde{E}^{\mathsf{r}}_{0y}\hat{\boldsymbol{e}}_y = \tilde{E}^{\mathsf{r}}_{0+}\hat{\boldsymbol{e}}_+ + \tilde{E}^{\mathsf{r}}_{0-}\hat{\boldsymbol{e}}_-. \end{aligned} \tag{3.7}$$

Similarly, the electric field of the transmitted light is given by,

$$\begin{aligned} \boldsymbol{E}^{\mathsf{t}} &= \tilde{\boldsymbol{E}}^{\mathsf{t}}_0 e^{i(-kz+\omega t)}, \\ \tilde{\boldsymbol{E}}^{\mathsf{t}}_0 &= \tilde{E}^{\mathsf{t}}_{0x}\hat{\boldsymbol{e}}_x + \tilde{E}^{\mathsf{t}}_{0y}\hat{\boldsymbol{e}}_y = \tilde{E}^{\mathsf{t}}_{0+}\hat{\boldsymbol{e}}_+ + \tilde{E}^{\mathsf{t}}_{0-}\hat{\boldsymbol{e}}_-. \end{aligned} \tag{3.8}$$

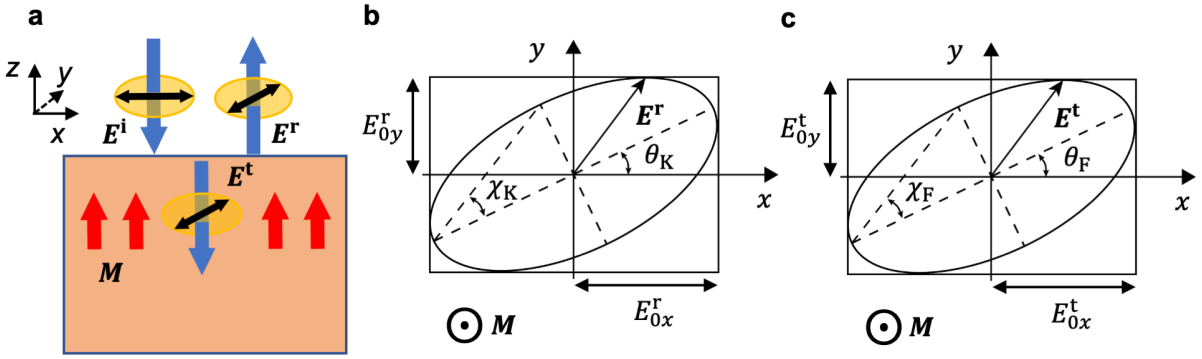

Figure 3.2: (a) Configuration of the polar MO effects. The incident light is linearly polarized along the x-axis. Red arrows pointing along the $+z$ direction denote the magnetization ($\boldsymbol{M}$). (b) The polarization plane of the reflected light. The polarization ellipse is oriented at a Kerr angle θ_{K} with respect to the x-axis. The Kerr ellipticity is defined through the ellipticity angle χ_{K}. (c) The polarization plane of the transmitted light. The polarization ellipse is oriented at a Faraday angle θ_{F} with respect to the x-axis. The ellipticity is defined through the Faraday ellipticity angle χ_{F}.

Suppose the upper space in Fig. 3.2a has a refractive index $n_1^{\pm}$ for the $\sigma^{\pm}$ polarized light. The material also has a refractive index $n_2^{\pm}$ for the $\sigma^{\pm}$ polarized light. The complex reflectivity $\tilde{r}_{\pm}$ at normal incidence is given by [137],

$$\tilde{r}_{\pm} \equiv \frac{\tilde{E}^{\mathsf{r}}_{0\pm}}{\tilde{E}^{\mathsf{i}}_{0\pm}} = \frac{n_1^{\pm} - n_2^{\pm}}{n_1^{\pm} + n_2^{\pm}}, \tag{3.9}$$

while the complex transmissivity $\tilde{t}_{\pm}$ at normal incidence is given by,

$$\tilde{t}_{\pm} \equiv \frac{\tilde{E}^{\mathsf{t}}_{0\pm}}{\tilde{E}^{\mathsf{i}}_{0\pm}} = \frac{2n_1^{\pm}}{n_1^{\pm} + n_2^{\pm}}. \tag{3.10}$$

According to the definition of $\tilde{r}_\pm$, we can rewrite Eq. (3.7) as,

$$\tilde{\boldsymbol{E}}_0^{\mathsf{r}} = \frac{1}{2}\hat{\boldsymbol{e}}_x \tilde{E}_{0x}^{\mathsf{i}}(\tilde{r}_+ + \tilde{r}_-) + \frac{1}{2}\hat{\boldsymbol{e}}_y \tilde{E}_{0x}^{\mathsf{i}}(i\tilde{r}_+ - i\tilde{r}_-), \tag{3.11}$$

which means,

$$\frac{\tilde{E}_{0x}^{\mathsf{r}}}{\tilde{E}_{0y}^{\mathsf{r}}} = \frac{1 + \tilde{r}_-/\tilde{r}_+}{i(1 - \tilde{r}_-/\tilde{r}_+)}. \tag{3.12}$$

Equation (3.12) defines an ellipse oriented away from the x-axis as shown in Fig. 3.2b. The Kerr angle θ_{K} and Kerr ellipticity χ_{K} can then be calculated as [138],

$$\tan 2\theta_{\mathsf{K}} = \frac{2E_{0x}^{\mathsf{r}} E_{0y}^{\mathsf{r}} \cos\delta}{(E_{0x}^{\mathsf{r}})^2 - (E_{0y}^{\mathsf{r}})^2}, \quad -\frac{\pi}{2} < \theta_{\mathsf{K}} \leq \frac{\pi}{2}, \tag{3.13}$$

and

$$\sin 2\chi_{\mathsf{K}} = \frac{2E_{0x}^{\mathsf{r}} E_{0y}^{\mathsf{r}} \sin\delta}{(E_{0x}^{\mathsf{r}})^2 + (E_{0y}^{\mathsf{r}})^2}, \quad -\frac{\pi}{4} < \chi_{\mathsf{K}} \leq \frac{\pi}{4}, \tag{3.14}$$

where $\delta = \mathsf{angle}(\tilde{E}_{0y}^{\mathsf{r}}/\tilde{E}_{0x}^{\mathsf{r}})$, $E_{0x}^{\mathsf{r}} = |\tilde{E}_{0x}^{\mathsf{r}}|$, $E_{0y}^{\mathsf{r}} = |\tilde{E}_{0y}^{\mathsf{r}}|$, and $\mathsf{angle}(Z)$ is a function that returns the phase angle of a complex number Z. A sign convention enters the expression of Kerr angle θ_{K}: θ_{K} is chosen to be positive if the rotation vector of the polarization plane is parallel to the magnetization vector. The Faraday angle θ_{F} and Faraday ellipticity χ_{F} are defined in a similar way for the transmitted light, as shown in Fig. 3.2c.

If the setup contains multiple interfaces, we can use the transfer matrix method to calculate $\tilde{r}_\pm$ and $\tilde{t}_\pm$ systematically [137, 139].

3.4 Dielectric function of 2D materials

As an extensive physical quantity, the dielectric function is ill-defined for 2D materials. Conventionally, we rescale the relative dielectric function which is calculated with a supercell approach by the thickness d of the 2D material (e.g., $d = c_{\mathsf{bulk}}/3 = 6.6$ Å for a monolayer CrI_3 based on the bulk structure [140]),

$$\varepsilon_{xx}(\omega) = 1 + \frac{l}{d}(\tilde{\varepsilon}_{xx}(\omega) - 1), \tag{3.15}$$

and

$$\varepsilon_{xy}(\omega) = \frac{l}{d}\tilde{\varepsilon}_{xy}(\omega), \tag{3.16}$$

where $\tilde{\varepsilon}_{xx}$ and $\tilde{\varepsilon}_{xy}$ are calculated frequency-dependent relative dielectric functions. The supercell has thickness l along the out-of-plane (z) direction.

Despite the ambiguity in defining ε for 2D materials, measurable quantities related to ε are independent of the rescaling procedure. For example, the absorbance and transimittance of light through a sheet of 2D material are governed by the quantity $A_0 = \omega d\varepsilon_2/c$ as shown in Sec. H. A_0 is obviously invariant when we change the parameter d in Eq. (3.15).

Chapter 4

swapGW

4.1 Self-consistency in the GW method

As discussed in Sec. 1.2 and Sec. 1.3, the Hedin's equations in the GW approximation should in principle be solved self-consistently, because all the physical quantities – G, W, Σ, and $\chi^\star$ – are interrelated to each other. However, such a truly self-consistent GW (scGW) calculation is very challenging in practice. In fact, the first scGW calculations of homogeneous electron gas reported less satisfactory agreement with experiments in terms of the quasiparticle bandwidth and satellite structures compared with one-shot calculations [141] because of the neglect of vertex corrections. In general, scGW calculations conserve the particle number and improve the ground-state total energy, but worsen the quasiparticle bandstructure [142, 143]. In addition to the strenuous fully self-consistent GW method, there exist different types of partially self-consistent GW methods [144], as discussed in the following with a focus on applications to solids.

We first discuss the starting-point dependence of a G_0W_0 calculation, where often the Kohn-Sham eigenvalues and eigenstates are used as inputs to construct the Green's function G and the screened Coulomb interaction W in the spirit of "best G, best W". With the development of different exchange-correlation functionals and hybrid schemes incorporating the exact exchange, it has been recognized that G_0W_0 calculations can exhibit a starting-point dependence for some materials [145, 146, 147, 148]. The reason is that the strength of dielectric screening in G_0W_0 calculations is inversely proportional to the band gap of the starting-point DFT or other mean-field results. Using the exact exchange in DFT calculations tends to overestimate the band gap and therefore underestimate the dielectric screening. Besides, spurious interactions between localized d orbitals and delocalized s and p orbitals will also cause issues of starting-point dependence, as evident in ZnO and other transition metal compounds [149, 150]. Despite less theoretical justification, the G_0W_0 method still enjoys a lot of popularity due to its feasible computational cost and good agreement with experiments.

To go beyond G_0W_0, we need to incorporate some self-consistent schemes in the Hedin's

equations. The easiest first step is the eigenvalue self-consistent GW method (evGW), where the real part of the quasiparticle energies obtained from a G_0W_0 calculation are reinserted to the expression of G and/or W. Iteration of the GW calculation is then performed until the input energies are equal to the output ones [9]. The evGW calculations where W is self-consistently updated tend to overestimate the band gap in semiconductors and insulators, because the RPA dielectric screening used in the expression of $\chi^\star$ and W will be underestimated with quasiparticle energies. This underestimation should be compensated by the missing vertex corrections. In practice, updating G alone (evGW_0) is preferred over updating W and G together (evGW) [151, 152].

In evGW calculations, the Kohn-Sham eigenstates are still used throughout the calculation, which still keeps some level of dependence on the starting point. The quasiparticle self-consistent GW method (QSGW) goes one step further and updates the wavefunctions using a static Hermitian mean-field potential derived from the self-energy operator [153, 154, 155]. This method cannot capture the satellites in the spectral function and therefore it is still in the quasiparticle approximation, hence the name. Lastly, a fully self-consistent GW approach based on the solution of the Dyson's equation with a plane wave basis set has been applied to conventional semiconductors [156], which is in principle independent of the starting point. Both latter two fully scGW methods significantly overestimate the band gap due to the underestimation of the dielectric function.

The above-mentioned self-consistent GW schemes are all within the GW approximation, i.e., no vertex corrections. The error canceling between the vertex corrections and the self-consistency has long been believed to be responsible for the great success of the G_0W_0 method [63, 157, 156]. It is therefore our motivation in this chapter to incorporate the vertex corrections in a self-consistent GW calculation, such that the removal of starting-point dependence and the good agreement of quasiparticle bandstructure with experiments can be achieved at the same time.

4.2 Vertex corrections from first principles

There have been several different strategies in the literature to add vertex corrections from first principles. The earliest approaches use the LDA exchange-correlation functional in place of the nonlocal mass operator to evaluate an analytical two-point approximate vertex function [9, 158], which can be easily incorporated into the expression of W in the so-called $GW\Gamma$ approximation. Development of time-dependent density-functional theory (TDDFT) leads to an upgraded version of the two-point approximate vertex using a TDDFT exchange-correlation kernel [63, 159, 160, 161, 162, 163], while some approximations (e.g., static screening, long-range interaction in the TDDFT kernel, etc.) are still needed. This two-point vertex approach yields also too large a band gap for ionic insulators [164].

The irreducible vertex function $\Gamma^\star$ appears explicitly in two different places inside the Hedin's equations, namely, the irreducible polarizability (Eq. (1.42)) and the mass operator (Eq. (1.44)). Full self-consistency involving $\Gamma^\star$ in all the quantities is very challenging and

expensive, and therefore updating $\Gamma^\star$ in part of the quantities seems to be the prudent way to make practical progress.

Note that $\Gamma^\star$ itself also satisfies a BSE-like equation given in Eq. (1.41). It is therefore tempting to relate the vertex function with a two-particle correlation function. To this end, we first recall Eq. (1.42) here,

$$\chi^\star(12) = -i\hbar \int \mathrm{d}(34)\, G(13)G(41)\Gamma^\star(342), \tag{4.1}$$

where the irreducible vertex function beyond the zeroth order introduces the electron-hole interaction. On the other hand, the irreducible polarizability $\chi^\star$ is related to the *irreducible two-particle correlation function* $L^\star$ with the definition,

$$L^\star(1234) \equiv \frac{\delta G(12)}{\delta V_{\mathsf{tot}}(43)}. \tag{4.2}$$

$\chi^\star$ and $L^\star$ are directly related via the following expression,

$$\chi^\star(12) = -i\hbar L^\star(11^+22^+). \tag{4.3}$$

$L^\star$ also satisfies a BSE,

$$L^\star(1234) = L_0(1234) + \int \mathrm{d}(5678) L_0(1265) \frac{\delta M(56)}{\delta G(87)} L^\star(8734). \tag{4.4}$$

Note that in Eq. (4.4) there is no exchange kernel. Given the success of the GW-BSE method in calculating optical properties [34], it is intuitive to extend the usage of BSE to $L^\star$ and to introduce vertex corrections in the calculation of $\chi^\star$ and W.

4.3 swapGW

In this section, we introduce the vertex corrections in a novel approach we named the swapGW method, which stands for *self-consistent with appropriate polarizability GW*. Within the swapGW method, we add vertex corrections to $\chi^\star$ and W via the solution of BSE for $L^\star$, while the conventional GW form of the mass operator M is kept the same (Eq. (1.47)). That is, we emphasize the importance of vertex corrections in improving the dielectric screening, and keep the GW level of complexity through the self-consistent loop of the Hedin's equations. This method has been implemented within the `BerkeleyGW` package [108].

Under similar approximations used in Sec. 1.5 and Sec. 2.4, $L^\star$ can be calculated in analogy to Eq. (2.11),

$$L^{\star -1}(\omega) = L_0^{-1}(\omega) - \Xi_W. \tag{4.5}$$

And $\chi^\star$ is assembled with Eq. (4.3). Just like $\tilde{L}$, we can choose a center-of-mass momentum $\boldsymbol{Q}$ to label the eigenvalues and eigenstates of $L^\star$ [86]. It is important to point out that we

cannot only focus on the $\boldsymbol{Q} \to 0$ limit in this case, because the integral equations of W and M need all the spatial fluctuations in $\chi^\star$ as well as in ε^{-1}. Diagrammatically, the irreducible polarizability calculated by the swapGW method contains infinite number of W ladders, as shown in Fig. 4.1.

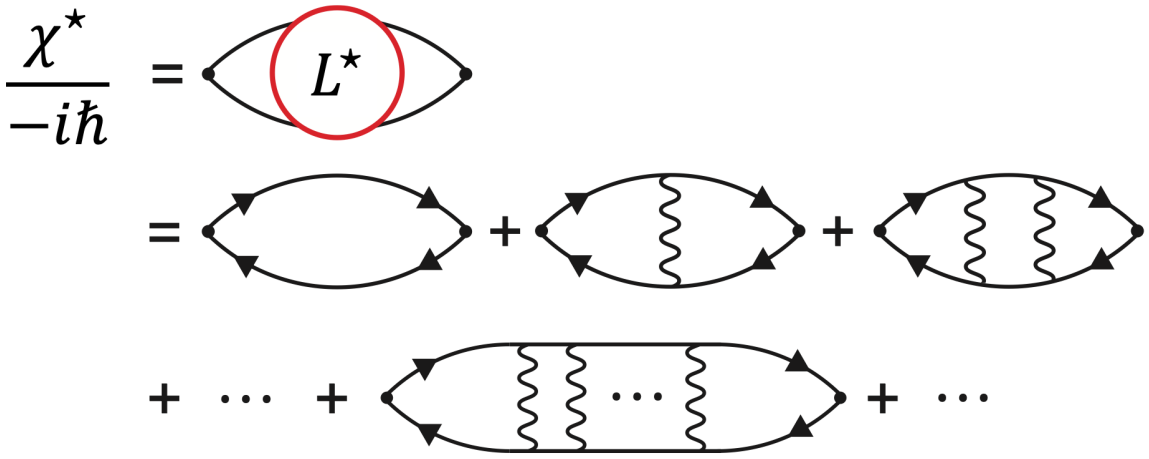

Figure 4.1: Feynman-diagrammatic representation of the expansion of the irreducible polarizability $\chi^\star$ in terms of the single-particle Green's function G (arrowed lines) and screened Coulomb interaction W (wavy lines). $\chi^\star$ is proportional to the irreducible two-particle correlation function $L^\star$. All the quantities are time-ordered.

In practice, we start from a conventional full-frequency G_0W_0 calculation (the zeroth iteration) with DFT results as a starting point. Using the quasiparticle energies and dielectric function from the G_0W_0 calculation, we solve the BSE for $L^\star$ at all the $\boldsymbol{Q}$ vectors and then construct the vertex-corrected irreducible polarizability $\chi^\star_{\boldsymbol{G}_1\boldsymbol{G}_2}(\boldsymbol{Q};\omega)$ as follows,

$$\begin{aligned}\chi^\star_{\boldsymbol{G}_1\boldsymbol{G}_2}(\boldsymbol{Q};\omega) = \frac{1}{N_k\Omega}\sum_S \langle 0|e^{-i(\boldsymbol{Q}+\boldsymbol{G}_1)\cdot\boldsymbol{r}_1}|S\boldsymbol{Q}\rangle\langle S\boldsymbol{Q}|e^{i(\boldsymbol{Q}+\boldsymbol{G}_2)\cdot\boldsymbol{r}_2}|0\rangle \\ \times\left[\frac{1}{\hbar\omega - \Omega_{S\boldsymbol{Q}} + i\eta} - \frac{1}{\hbar\omega + \Omega_{S\boldsymbol{Q}} - i\eta}\right], \quad \eta\to 0^+,\end{aligned} \tag{4.6}$$

where the matrix elements are evaluated as,

$$\langle 0|e^{-i(\boldsymbol{Q}+\boldsymbol{G}_1)\cdot\boldsymbol{r}_1}|S\boldsymbol{Q}\rangle = \sum_{cv\boldsymbol{k}} A^{S\boldsymbol{Q}}_{cv\boldsymbol{k}}\langle v(\boldsymbol{k}-\boldsymbol{Q})|e^{-i(\boldsymbol{Q}+\boldsymbol{G}_1)\cdot\boldsymbol{r}_1}|c\boldsymbol{k}\rangle. \tag{4.7}$$

The time-ordered dielectric function $\varepsilon_{\boldsymbol{G}_1\boldsymbol{G}_2}(\boldsymbol{Q};\omega)$ is calculated using Eq. (1.28). Next, W and M in the GW approximation are calculated as in a conventional GW calculation. And then we enter the next iteration and calculate a new set of quasiparticle energies. The BSE

kernel (i.e., the direct kernel Ξ_W) is constructed based on the quasiparticle energies from the current iteration and the dielectric function from the last iteration. This loop goes on until the quasiparticle energies converge within a given threshold. For now, wavefunctions are not updated throughout the iterations. The self-consistent loop of the swapGW method is summarized in Fig. 4.2.

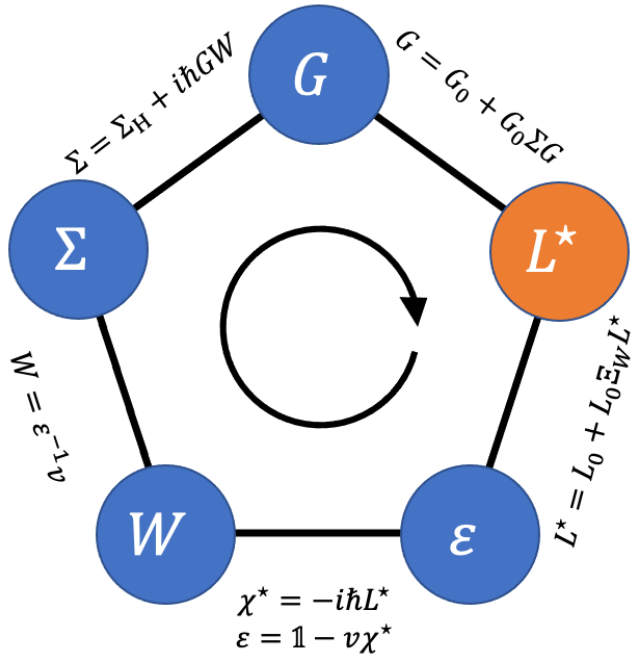

Figure 4.2: Self-consistent loop of the swapGW method. Important physical quantities in the loop include the single-particle Green's function G, the irreducible two-particle correlation function $L^\star$, the dielectric function ε, the screened Coulomb interaction W, and the self-energy Σ. G is solved with the Dyson's equation. $L^\star$ is solved with the Bethe-Salpeter equation with the direct kernel Ξ_W. ε is constructed from the irreducible polarizability $\chi^\star$, which is readily available from $L^\star$. W is straightforwardly calculated by screening the bare Coulomb interaction v with ε^{-1}. The electron self-energy Σ includes the Hartree term and the GW term, and it is fed to the Dyson's equation of G to restart the loop. The solution of $L^\star$ is crucial for adding the vertex corrections.

A large number of empty states are needed in GW calculations, which means the size of the BSE matrix will be very large. To circumvent this computational bottleneck, we treat low-energy transitions and high-energy transitions differently in realistic swapGW calculations. To be specific, low-energy transitions are treated at the BSE level and fall completely

within the loop shown in Fig. 4.2. High-energy transitions, on the other hand, are treated at the RPA level and then added to the $\chi^\star$ from the low-energy transitions to form a complete $\chi^\star$. For those high-energy bands not used in the BSE calculation, we deduce their quasiparticle energies through a scissors-shift approach with parameters derived from low-energy bands [108].

4.4 Benchmark

We benchmark our theory and implementation of the swapGW method using bulk Si. Since the SOC is rather weak in Si according to Sec. 2.6 and the computational cost of swapGW is already very high, we choose to use the SR pseudopotential and scalar GW and GW-BSE formalisms in this study. GW self-energies are solved on an $8 \times 8 \times 8$ k-grid using a contour deformation technique [11]. The energy cutoff used for the dielectric function is 10 Ry. For GW calculations, a total of 200 bands are considered in the summation, among which the lower 80 bands (4 valence bands and 76 conduction bands) are included in the BSE calculation. In this way, the transition from the 4 valence bands to the lower 76 conduction bands are treated at the BSE level while the transition from the valence bands to the higher 120 conduction bands are treated at the RPA level.

The procedure to determine the scissor-shift parameters is shown in Fig. 4.3 for the zeroth iteration (at the full-frequency G_0W_0 level). We fit a linear relation between the quasiparticle energy correction $E^{\mathsf{QP}} - E^{\mathsf{KS}}$ of conduction states and the corresponding mean-field Kohn-Sham energy E^{KS} using the following expression,

$$E^{\mathsf{QP}} - E^{\mathsf{KS}} = \Delta E + C(E^{\mathsf{KS}} - E^{\mathsf{KS}}_{\mathsf{CBM}}), \tag{4.8}$$

where CBM stands for conduction band minimum. Because all the valence bands are used in the BSE calculation, we only need to deduce the quasiparticle energy of higher conduction bands with the scissors-shift approach.

Comparison of the swapGW bandstructure in the zeroth (iter. 0), first (iter. 1) and fifth (iter. 5) iteration is shown in Fig. 4.4. We notice that the overall shape of bandstructure is well kept in our swapGW calculations, with a slight narrowing of the valence bandwidth. The detailed convergence behavior of relevant physical quantities (indirect band gap, direct band gap at the Γ point, valence bandwidth, and the head element of the static dielectric function $1/\varepsilon^{-1}_{00}(\boldsymbol{q} \to 0; \omega = 0)$) are listed in Table 4.1. In particular, the calculated indirect band gap converges within 1 meV after 3 iterations, and $1/\varepsilon^{-1}_{00}(\boldsymbol{q} \to 0; \omega = 0)$ converges within 0.01 also after 3 iterations. It is obvious that the swapGW method manages to avoid the overestimation of the band gap as well as to avoid severe underestimation of the dielectric screening during the self-consistent iterations in conventional scGW calculations [156]. On the other hand, the small difference ($\sim$ 60 meV) in the value of indirect band gap between G_0W_0 and the converged swapGW results indicates that the G_0W_0 method really benefits from the error canceling between the vertex corrections and self-consistency

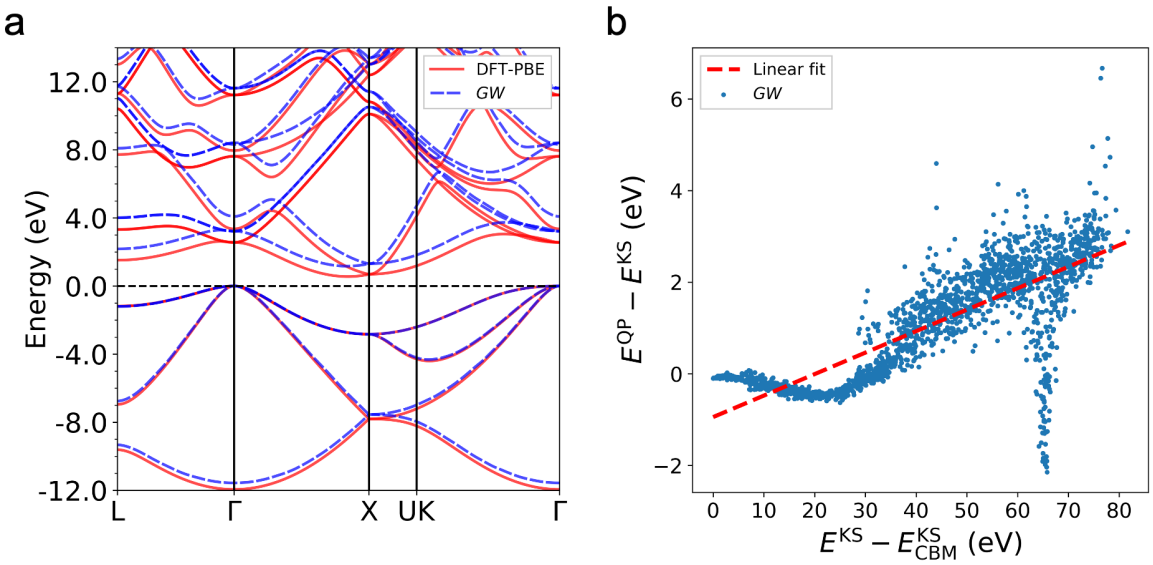

Figure 4.3: (a) DFT-PBE (red solid) and GW (blue dashed) bandstructures of bulk Si using the scalar-relativistic pseudopotential. The GW calculation is performed at the full-frequency G_0W_0 level with a contour deformation method [11]. (b) GW self-energy correction $E^{\mathsf{QP}} - E^{\mathsf{KS}}$ *versus* the Kohn-Sham energy E^{KS} for conduction states of bulk Si, with the data input from (a). The red dashed line is a linear fit to the data with an expression of $E^{\mathsf{QP}} - E^{\mathsf{KS}} = \Delta E + C(E^{\mathsf{KS}} - E^{\mathsf{KS}}_{\mathsf{CBM}})$. The fitting parameters are: $\Delta E = -0.944$ eV and $C = 0.047$ eV.

using quasiparticle energies. Moreover, we note that the RPA dielectric function using Kohn-Sham eigenvalues and eigenstates at the DFT-PBE level is already good enough for a G_0W_0 calculation, which means any further modification of the mean-field calculation with hybrid functionals or exact-exchange will likely worsen the G_0W_0 results.

The results above serve to benchmark our implementation of the swapGW method within the `BerkeleyGW` package [108]. This method will be applied to a variety of bulk and low-dimensional materials in the future. We believe that it can further improve the predictive power of the first-principles GW and GW-BSE methods.

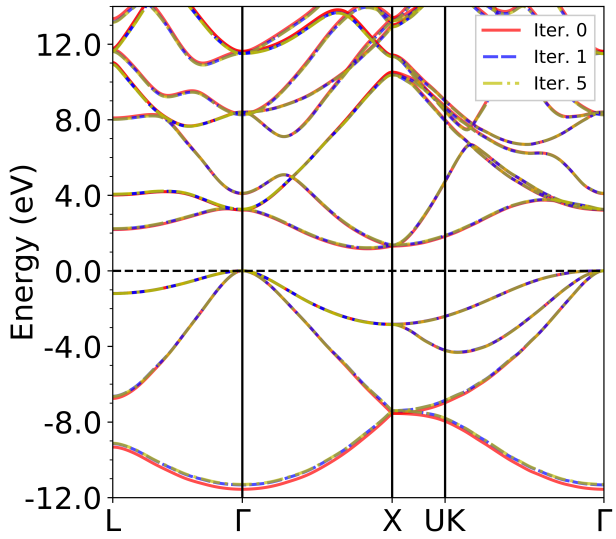

Figure 4.4: swapGW bandstructures of bulk Si in the zeroth (iter. 0, red solid), first (iter. 1, blue dashed) and fifth (iter. 5, yellow dash-dotted) iteration. The zeroth iteration corresponds to a full-frequency G_0W_0 calculation. The valence bandwidth in the fifth iteration is converged within 1 meV.

	Ind. E_g (eV)	E_g @ Γ (eV)	Val. bandwidth (eV)	$\frac{1}{\varepsilon_{00}^{-1}(\boldsymbol{q}\to 0;\omega=0)}$
DFT-PBE	0.554	2.561	11.947	N/A
Iter. 0	1.166	3.221	11.564	13.338
Iter. 1	1.215	3.251	11.332	13.241
Iter. 2	1.224	3.258	11.307	13.038
Iter. 3	1.225	3.259	11.304	13.026
Iter. 4	1.226	3.259	11.305	13.021
Iter. 5	1.226	3.260	11.305	13.020

Table 4.1: Convergence of the indirect band gap (ind. E_g), direct band gap at the Γ point (E_g @ Γ), valence bandwidth (val. bandwidth), and $1/\varepsilon_{00}^{-1}(\boldsymbol{q}\to 0;\omega=0)$ in the swapGW calculations of bulk Si with respect to the number of iterations. The same quantities at the DFT-PBE level are given in the second row. The zeroth iteration (iter. 0) denotes the full-frequency G_0W_0 results while iter. 1 through 5 denote swapGW results in each iteration.

Chapter 5

Unifying optical selection rules for excitons in two dimensions: Band topology and winding numbers

The concept of topology and relevant topological effects have revolutionized many areas of modern physics, deepening our understanding of a wide range of phenomena covering topological phase transitions [165], topological defects [166], quantum transport [167], and the emergence of symmetry protected topological phases [96, 99, 168, 169, 170], etc. In this chapter, I present a theoretical work in which we show that topology in the electronic structure can dramatically reshape the photophysics of 2D semiconductors [1]. For systems in which states near the band extrema are of multicomponent character, the spinors describing these components (pseudospins) can develop nonzero winding numbers near the extremal k-point. The oscillator strength and intrinsic polarization of an excitonic optical transition in these systems are therefore dictated by the optical matrix element winding number, a unique and heretofore unrecognized topological characteristic. We demonstrate these findings in gapped graphene systems, where the pseudospin textures are characterized by nontrivial optical matrix element winding numbers associated with the valley index and circular polarization. This winding-number physics leads to exotic exciton series and new optical selection rules, with each valley hosting multiple bright excitons coupled to the light of different circular polarization. Our work shows that the optical selection rule in two dimensions must incorporate the topological effects in the electronic structure.

5.1 Introduction

Owing to the similarity between the electron-hole binding in a semiconductor and the electron-proton binding in a hydrogen atom, the hydrogenic model and their variants (for example, including electron-hole-separation-dependent screening effects) are usually adopted in describing excitons in various dimensions, when the electron-hole correlation length of the

exciton of interest is large compared to the unit cell size [8, 171]. Within this picture, the envelope functions of the excitonic states are hydrogen-like wavefunctions with even or odd parity and are characterized by a series of quantum numbers. In linear optical spectroscopy, an exciton may be created or annihilated by absorbing or emitting a photon, respectively. Such coupling is allowed if the full many-body excitonic states have different parity from the ground state (these states are called optically active or bright excitons). For conventional semiconductors in which the electron (hole) states in the conduction (valence) band extreme forming the exciton are of single orbital character, this parity law together with the hydrogenic picture leads to the well-known optical selection rules [172], which says: in dipole-allowed materials (e.g., GaAs, monolayer transition metal dichalcogenides, etc. [8, 87, 173, 174]), s-like excitons are optically active, whereas p-like excitons are optically inactive; in dipole-forbidden materials (e.g., Cu_2O [175]), the optically active excitons are p-like states, while s-like states are optically inactive.

However, for many reduced-dimensional systems of current interest, the states near the band extrema are of multiple orbital and spin components, and the bands can have nontrivial topological characteristics. Such nontrivial topological bands may be characterized by the behavior of the amplitudes of the components that compose a band state, viewed as a multi-component spinor (the pseudospin) in k-space. The pseudospins of the electron and hole states can develop a complex texture with respect to the crystal momentum $\boldsymbol{k}$ around the band extrema [169, 168, 176, 177, 178, 179]. The pseudospin texture (viewed as a spinor field of $\boldsymbol{k}$) could in principle affect the energy levels, optical selection rules, and many other properties of the excitons. Recent studies have shown that Berry curvature flux leads to a fine energy-level splitting of the otherwise doubly degenerate hydrogenic $2p$ excitons in monolayer transition metal dichalcogenides [180, 181]. Yet, it remains unexplored whether central properties such as the optical selection rules are altered in materials with topological band characteristics.

5.2 Results

5.2.1 Winding numbers and optical selection rules in two dimensions

The eigenstate of an exciton S is a coherent superposition of free electron-hole pairs at different k-points, $|S\rangle = \Sigma_{cv\boldsymbol{k}} A^S_{cv\boldsymbol{k}} |cv, \boldsymbol{k}\rangle$. The oscillator strength that relates to the intensity for optical transition to exciton S is given by,

$$I^S_e = \frac{2|\Sigma_{cv\boldsymbol{k}} A^S_{cv\boldsymbol{k}} \hat{\boldsymbol{e}} \cdot \langle c\boldsymbol{k}|\hat{\boldsymbol{p}}|v\boldsymbol{k}\rangle|^2}{\Omega_S}, \tag{5.1}$$

where $\hat{\boldsymbol{e}}$ is the photon polarization unit vector, and $\langle c\boldsymbol{k}|\hat{\boldsymbol{p}}|v\boldsymbol{k}\rangle$ is the interband optical matrix element. In the following, we assume a two-band model, and use $A^S_{\boldsymbol{k}}$ instead of $A^S_{cv\boldsymbol{k}}$. Although the exciton energies and oscillator strengths are physical observables and thus gauge

invariant, the individual components in Eq. (5.1) (the exciton envelope functions in k-space and the interband optical matrix elements) may separately look different depending on a chosen gauge. This ambiguity arises because $|cv, \boldsymbol{k}\rangle$ could have an arbitrary phase, which would be canceled out by the complex conjugate of the same phase in $A_{\boldsymbol{k}}$. This gauge arbitrariness can be eliminated by requiring $A_{\boldsymbol{k}}$ of the lowest-energy s-like excitonic state to resemble that of a conventional hydrogen-like s orbital. Under this well-defined and intrinsically smooth gauge, we find that an analysis of Eq. (5.1) illuminates clearly the physical role of the exciton envelope function and of the topological characteristics of the interband optical matrix elements in optical transitions. In dipole-allowed conventional semiconductors, the interband optical matrix elements are nearly a constant around the extremal k-point [8, 172]. Therefore, only s-like excitons have nonzero oscillator strength, as its envelope function in k-space is isotropic in phase (i.e., no phase winding around the extremal k-point).

Having topologically nontrivial bands in two dimensions with associated pseudospin texture of nonzero winding numbers will lead to both magnitude and phase modulations of the interband optical matrix elements with k-points, represented by a 2D vector field with a certain winding pattern. To illustrate this effect, we decompose the interband optical matrix element $\langle c\boldsymbol{k}|\hat{\boldsymbol{p}}|v\boldsymbol{k}\rangle$ into $p_{\boldsymbol{k}+} = \hat{\boldsymbol{e}}_+ \cdot \langle c\boldsymbol{k}|\hat{\boldsymbol{p}}|v\boldsymbol{k}\rangle$ and $p_{\boldsymbol{k}-} = \hat{\boldsymbol{e}}_- \cdot \langle c\boldsymbol{k}|\hat{\boldsymbol{p}}|v\boldsymbol{k}\rangle$, which correspond to coupling to σ^- and σ^+ circularly polarized photon modes, respectively. For topologically nontrivial bands, as illustrated below, $p_{\boldsymbol{k}+}$ and $p_{\boldsymbol{k}-}$ are typically nonzero (except possibly at the extremal k-point), and can be viewed as two vector fields that may differ in their winding patterns. (Note: The interband optical matrix elements $p_{\boldsymbol{k}\pm}$ are complex numbers determined only by the band states and are independent of the specific excitonic states.)

We shall show that the transition strength (brightness) and the specific light polarization needed for an excitonic optical transition are dictated by the phase winding of the exciton envelope function and that of the interband optical matrix elements. For an excitonic state of which the k-space envelope function $A^S_{\boldsymbol{k}}$ is a highly localized function around an extremal k-point (Wannier excitons), $A^S_{\boldsymbol{k}}$ and $p_{\boldsymbol{k}\pm}$ in the relevant small part of the Brillouin zone are dominated by a cylindrical angular phase dependence of $\sim e^{im\theta_{\boldsymbol{k}}}$ and $e^{il_\pm\theta_{\boldsymbol{k}}}$, respectively ($\theta_{\boldsymbol{k}}$ is defined as the angle $\boldsymbol{k}$ made with respect to the k_x axis) [8, 172]. Here, and in subsequent discussion, we shall define $\boldsymbol{k}$ as the wave vector measured from the extremal k-point. Thus, m is the cylindrical angular quantum number of the exciton envelope function and $l_\pm$ are the winding numbers of $p_{\boldsymbol{k}\pm}$. From Eq. (5.1), the oscillator strength for an optical transition to an excitonic state S by $\sigma^\pm$ photon (coupling to $p_{\boldsymbol{k}\mp}$) is,

$$I^S_{\sigma^\pm} = \frac{2|\Sigma_{\boldsymbol{k}} f(|\boldsymbol{k}|)e^{i(m+l_\mp)\theta_{\boldsymbol{k}}}|^2}{\Omega_S}, \tag{5.2}$$

where $f(|\boldsymbol{k}|)$is the radial part in the summation. $I^S_{\sigma^\pm}$ is thus nonzero *only* when $m = -l_\mp$.

This set of selection rules is distinctly different from that in conventional semiconductors. For a system with discrete n-fold rotational symmetry, the general selection rule is,

$$m = -l_\mp (\mathsf{mod}\, n). \tag{5.3}$$

A generalization to systems with discrete rotational symmetries is given in the supplemental material of Ref. [1]. As a result, excitons with different angular quantum numbers (i.e., different m) would couple differently to $p_{\boldsymbol{k}+}$ and $p_{\boldsymbol{k}-}$, causing multiple bright excitons each accessible by σ^- and σ^+ photons. We note that Eq. (5.3) thus incorporates and generalizes the conventional selection rules for Wannier excitons to all 2D semiconductors with a single critical point (or no critical point) where excitonic optical transitions happen.

5.2.2 Applications: Gapped graphene systems

An ideal set of materials to illustrate the predicted novel excitonic physics is the gapped graphene systems, in which a band gap and a layer-number-dependent pseudospin texture emerge from an induced broken inversion symmetry that may be tuned. We consider three (already experimentally achieved) systems based on 1-3 layers of graphene [182, 183, 184, 185]. For monolayer graphene, inversion symmetry is broken by placing the graphene layer on top of a monolayer of hexagonal boron nitride [1]. For bilayer (in a Bernal stacking order) and trilayer graphene (in a rhombohedral stacking order), inversion symmetry is broken by applying an external electric field along the out-of-plane direction. In our first-principles GW-BSE calculations presented below, the applied electric field was set to 0.13 V/Å, an experimentally feasible value [184]. Modifying the applied electric field strength, which determines the size of the induced band gap, does not change the physics discussed here.

For the gapped graphene systems studied, DFT calculations are performed within LDA using the `Quantum ESPRESSO` package [13, 107] to determine their ground-state properties. First-principles GW and GW-BSE methods are employed to calculate the quasiparticle bandstructure and excitonic states, respectively, using the `BerkeleyGW` package [108]. In the GW and GW-BSE calculations, the dielectric matrix for the screened Coulomb interaction is constructed with a slab truncation scheme and with an energy cutoff of 8 Ry. Close scrutiny is needed for the k-point sampling in the excited-state calculations. For calculations of the quasiparticle bandstructure, a 150×150 k-grid in the Brillouin zone is necessary to converge the band gap within 3 meV. For the calculation of excitons, a patched sampling scheme is used to solve the BSE for the excitonic states in the individual K and K' valleys. The sampling density is equivalent to a uniform 450×450 k-grid in the Brillouin zone. For monolayer graphene, a 450×450 k-grid is interpolated into a 1500×1500 k-grid to converge the exciton energy levels to within 2 meV.

The gapped graphene systems of 1, 2, and 3 atomic layers studied have GW quasiparticle band gaps of 130, 159, and 185 meV, as evident in Fig. 5.1, respectively. These values are much larger than their corresponding Kohn-Sham band gaps of 62, 90, and 118 meV, respectively, owing to the self-energy effects. For biased bilayer and trilayer graphene, the top valence and bottom conduction bands at the K and K' valleys develop a Mexican-hat-like shape. The pseudospin texture of the states in bilayer graphene is schematically shown in Fig. 5.1d, where the amplitude of the carbon π orbitals develops a phase winding around

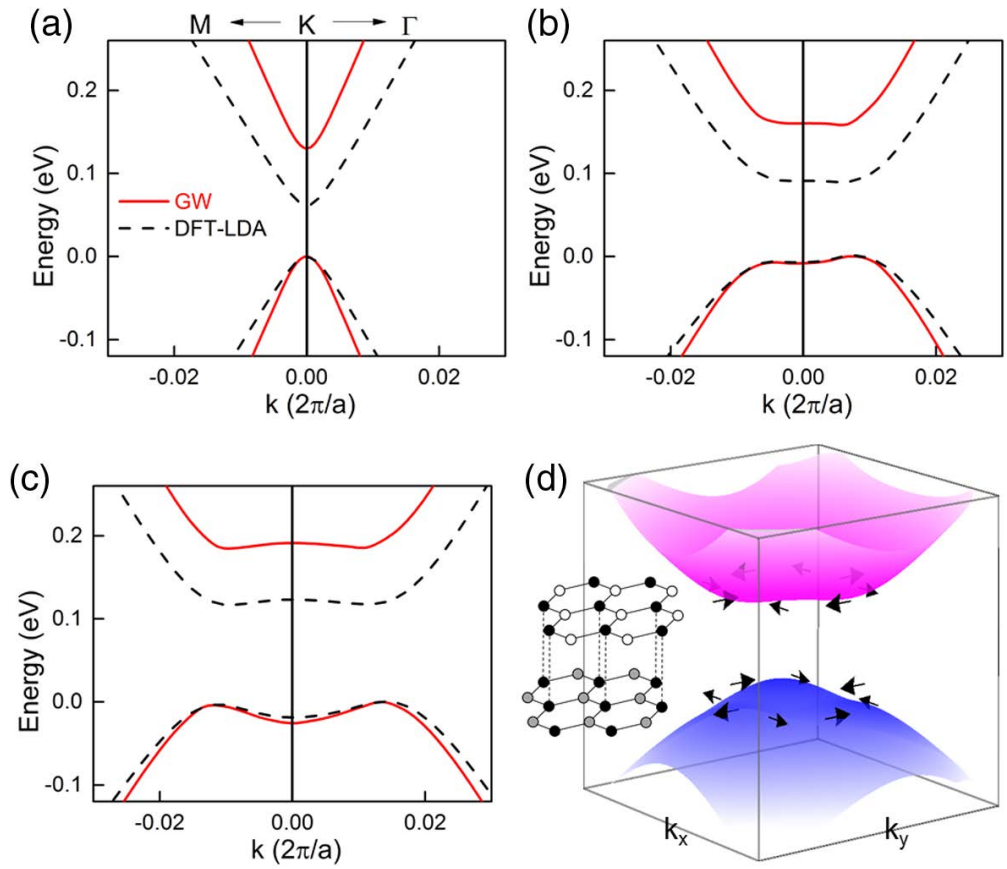

Figure 5.1: Calculated bandstructure and pseudospin (orbital components) winding of gapped graphene systems. Bottom conduction band and top valence band of (a) monolayer graphene with broken A and B sublattice symmetry, (b) biased Bernal-stacked bilayer graphene, and (c) biased rhombohedral-stacked trilayer graphene. Red solid lines and black dashed lines are GW and DFT-LDA bands, respectively. The K point is set at $k = 0$. Positive and negative k values denote the $K - \Gamma$ and $K - M$ direction, respectively. (d) Orbital pseudospin winding in biased bilayer graphene. Inset: Structure of biased bilayer graphene. The carbon atoms forming bonds with a neighboring layer are colored black.

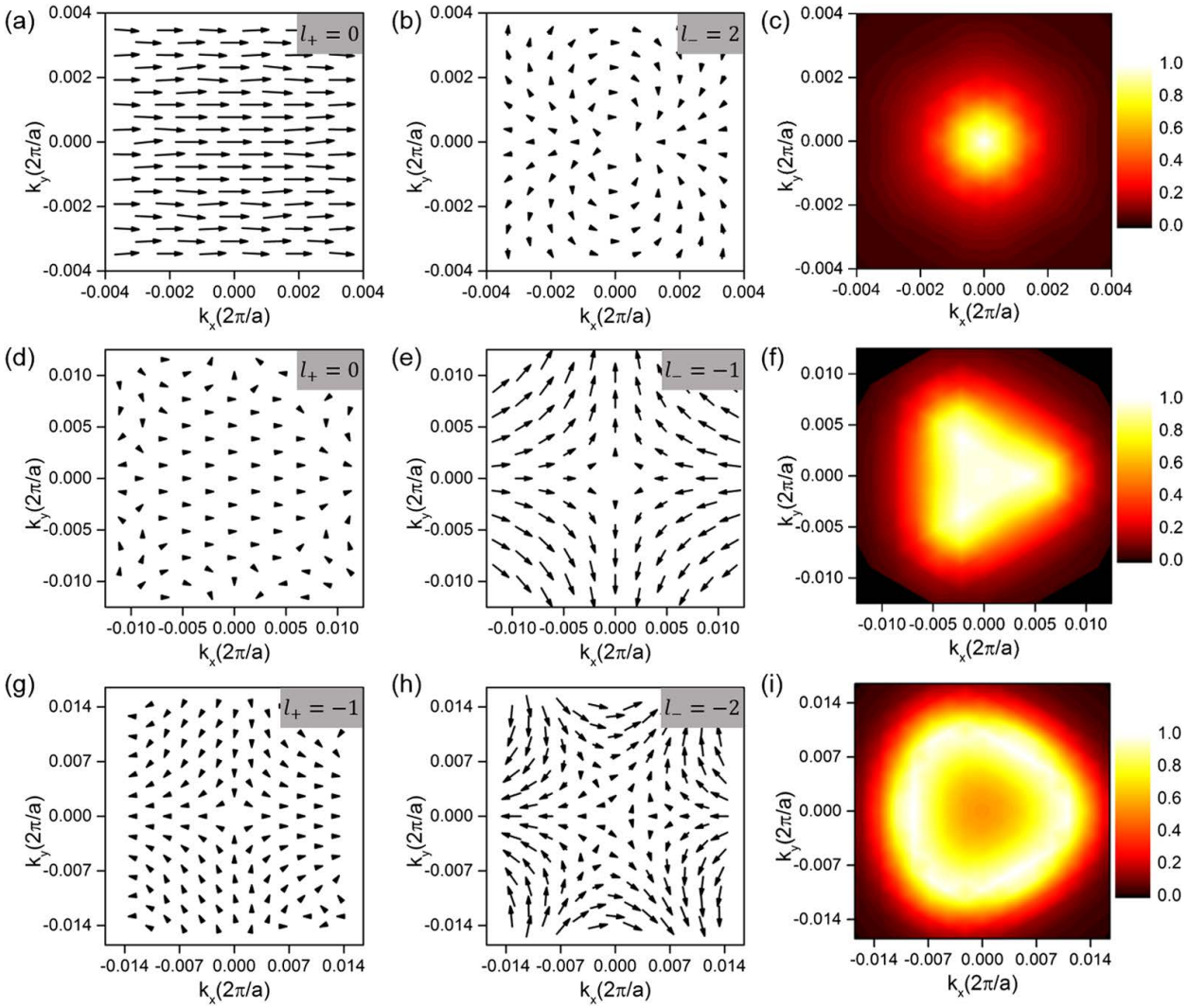

Figure 5.2: K-valley interband optical transition matrix elements and $1s$ exciton envelope function in k-space. The K point is placed at the origin. Optical interband transition matrix element and its winding number for light of (a) σ^- circular polarization $p_{\boldsymbol{k}+}$ and (b) σ^+ circular polarization $p_{\boldsymbol{k}-}$ in monolayer graphene with inequivalent sublattices. The direction and length of an arrow denote, respectively, the phase and the magnitude of the corresponding matrix element. (d) $p_{\boldsymbol{k}+}$ and (e) $p_{\boldsymbol{k}-}$ in biased bilayer graphene. (g) $p_{\boldsymbol{k}+}$ and (h) $p_{\boldsymbol{k}-}$ in biased trilayer graphene. (c, f, i) $1s$ exciton envelope function in k-space in gapped monolayer graphene, biased bilayer graphene, and biased trilayer graphene, respectively. The envelope functions show in color scale the magnitude of the free electron-hole pair excitation at each $\boldsymbol{k}$, normalized to its largest value in each plot.

the band extreme [186].

The very different pseudospin texture of the bands in the three gapped graphene systems gives a strong layer-number- and valley-index-dependent interband optical matrix element winding pattern for each. We show in Fig. 5.2 the winding pattern of $p_{\boldsymbol{k}+}$ and $p_{\boldsymbol{k}-}$ in the K valley, defined using the gauge procedure as described above. The complex quantity $p_{\boldsymbol{k}+}$ or $p_{\boldsymbol{k}-}$ (given by a magnitude and a phase $\phi_{\boldsymbol{k}}$) are represented by an arrow with its length proportional to the magnitude and its orientation pointing along the direction with angle $\phi_{\boldsymbol{k}}$ to the x-axis. In monolayer graphene with inequivalent A and B sublattices (Figs. 5.2a, b), $p_{\boldsymbol{k}+}$ is nearly constant in magnitude and phase (arrows with constant length and orientation) and has a winding number $l_+ = 0$ for any contours enclosing K, whereas $p_{\boldsymbol{k}-}$ is much smaller in magnitude and its phase (the orientation of the arrows) winds counterclockwise around the K point twice after completing any counterclockwise contour enclosing K (winding number of $l_- = 2$). This analysis, making use of the selection rules deduced above, predicts an optically active s exciton series, as well as a weakly active d exciton series ($m = -2$ in the K valley). In biased bilayer graphene, the pseudospin texture in Fig. 5.1d leads to a winding number of $l_- = -1$ for the interband optical matrix element $p_{\boldsymbol{k}-}$, as shown in Fig. 5.2e. Compared with $p_{\boldsymbol{k}-}$, $p_{\boldsymbol{k}+}$ is much smaller in magnitude, as evident in Fig. 5.2d, but remains constant in both magnitude and phase around the K point (winding number $l_+ = 0$). We, therefore, predict that (i) unlike the case of gapped monolayer graphene, the p exciton series ($m = 1$ in the K valley) are now optically very active, (ii) the s exciton series are still somewhat optically active, but has a much smaller oscillator strength than the p exciton series, and (iii) importantly, the photoexcitation of the s excitons and p excitons at a given valley (K or K') requires opposite circular polarization in biased bilayer graphene. The interband optical matrix elements in biased trilayer graphene have even more features, shown in Figs. 5.2g, h, leading to a winding number of $l_+ = -1$ and $l_- = -2$ for $p_{\boldsymbol{k}+}$ and $p_{\boldsymbol{k}-}$, respectively, at the K valley. The $1s$ exciton envelope functions of the three gapped graphene systems studied are shown in Figs. 5.2c, f, and i, respectively. Our new selection-rule predictions based on topological effects are completely borne out by our explicit GW-BSE calculations of the optical absorption spectra.

The physics of interband optical matrix element winding numbers thus leads to novel exciton series in the gapped graphene systems, with each valley hosting multiple optically active excitons whose creation requires different circular polarization. We show in Fig. 5.3 the first-principles GW-BSE calculated energy levels, required circular polarization, and oscillator strength of the first six lowest-energy excitons in the K and K' valley of each system. The calculated binding energies of the lowest-energy exciton state of the 1-, 2-, and 3-layer systems are 34, 52, and 45 meV, respectively. In gapped monolayer graphene with inequivalent sublattices (Fig. 5.3a), as expected, the s-like excitons are optically bright. The $1s$ exciton in the K and K' valleys can be selectively excited by σ^- and σ^+ light, respectively, similar to monolayer transition metal dichalcogenides [94, 95, 187, 188]. In biased bilayer graphene (Fig. 5.3b), however, the optically most active exciton becomes a $2p$ state that is located at 13 meV above the lowest-energy $1s$ state, with an oscillator strength ~ 20 times larger than that of the $1s$ exciton. Moreover, the circular polarization needed for excitation

of the $2p$ state is opposite to that of the $1s$ state, a feature that is directly predicted from the interband optical matrix element winding patterns depicted in Figs. 5.2d, e. In the biased trilayer graphene (Fig. 5.3c), the lowest-energy $1s$ exciton is optically inactive from the matrix element winding patterns in Figs. 5.2g, h. Because of a significant deviation of the band dispersion from a parabola, we are no longer able to associate the higher-energy excitonic states with a clear principal quantum number. However, a pair of nearly degenerate excitons with p-like and d-like orbital characters could still be identified, located at ~ 9 meV above the $1s$ state. They are excitable in the K valley with σ^+ polarized light, and couple strongly (optically bright) to the ground state via $p_{\boldsymbol{k}-}$ in Fig. 5.2h, either directly or through a trigonal warping effect. There is also a weakly active p-like exciton at ~ 4 meV above the $1s$ state. In all three cases, the circular polarization for excitation of every bright exciton in the K' valley is opposite to that of a degenerate-in-energy counterpart in the K valley due to time-reversal symmetry.

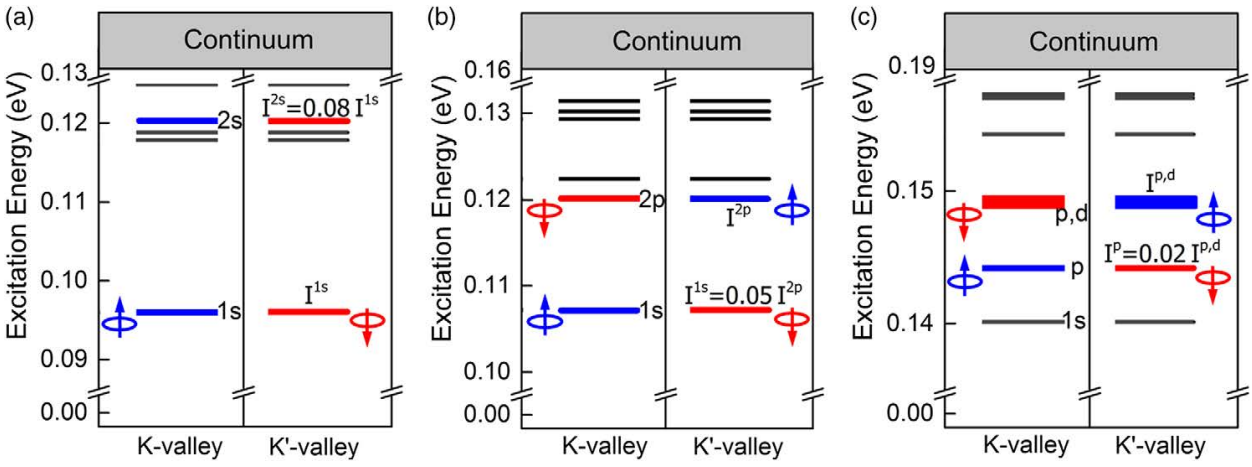

Figure 5.3: K-valley and K'-valley exciton energy levels and valley-exciton selective circular dichroism in (a) monolayer graphene with inequivalent A and B sublattices, (b) biased bilayer graphene, and (c) biased trilayer graphene. Left (right) part of each panel depicts the K-valley (K'-valley) exciton energy levels. The first six lowest-energy excitons are shown in each plot. Black lines indicate dark states (with maximum oscillator strength $< 1\%$ of the brightest exciton in each plot). The oscillator strength (I) of each bright state is expressed in terms of that of the brightest state, for unpolarized light. Blue and red lines (or circles) indicate bright states that are excitable from the ground state with σ^- and σ^+ circularly polarized light, respectively.

5.3 Conclusion

In sum, we have presented the discovery of a set of new unifying optical selection rules, as well as results of novel bright exciton series, arising from band topological effects in 2D semiconductors. This valley-exciton selective circular dichroism can be unambiguously detected using optical spectroscopy. All 2D systems, with or without nontrivial band topology, have optical selection rules given by Eq. (5.3) replacing the conventional ones for Wannier excitons. Owing to the Poincaré-Hopf theorem, which in the present context states that the sum of the winding numbers from all band extrema in the 2D Brillouin zone should equal to zero (the Euler characteristics of a 2D torus), one should look for the predicted nonconventional selection rules in the excitonic spectra of multivalley 2D materials. Our work reveals yet another important manifestation of band topology in the physical properties of materials; it also opens opportunities for use of these effects in gapped graphene systems for potential valleytronic applications.

Chapter 6

Ultrasensitive tunability of the direct band gap of 2D InSe flakes via strain engineering

One unique advantage of 2D materials is their ultrahigh flexibility [189]. It has been shown that a single-layer graphene can sustain as much as $\sim 10\%$ strain without mechanic failure [190], and similar findings have also been reported in other 2D semiconductors such as MoS_2 [190, 191, 192, 193, 194, 195]. Meanwhile, strain can effectively and reversibly change the bandstructure of 2D materials to achieve a tunable spectral response in optoelectronic devices [192, 193, 194, 196, 197, 198]. InSe, a member of the layered materials family, is a superior electronic and optical material which retains a direct band gap feature from the bulk to atomically thin few-layers and high electronic mobility down to the single-layer limit [69]. In this way, 2D InSe flakes provide an ideal platform to realize such functionality in the infrared regime.

In this chapter, I present a collaborative work with experimental groups on the discovery of ultrasensitive tunability of the electronic structure in few-layer InSe via strain engineering [2]. The experimental study was mainly performed by research groups led by Prof. Su-Fei Shi (Rensselaer Polytechnic Institute) and Prof. Cheng-Yan Xu (Harbin Institute of Technology). In this work, we exploit the strain engineering to drastically modify the band gap of 2D InSe nanoflakes. It is demonstrated that the band gap of a few-layer InSe flake can be decreased by 160 meV with an in-plane uniaxial tensile strain of 1.06%, and increased by 79 meV with an in-plane uniaxial compressive strain of 0.62%, as evidenced by photoluminescence (PL) spectroscopy. The large reversible band gap change of few-layer InSe in response to strain represents the most pronounced uniaxial strain-induced band gap strain coefficient experimentally reported in 2D materials. We also develop a theoretical understanding of the strain-induced band gap change through first-principles DFT and *GW* calculations.

6.1 Introduction

The success of graphene research has inspired tremendous interest in 2D materials beyond graphene, particularly 2D semiconductors with finite-size band gap and high carrier mobility [66, 68, 69, 199]. One particularly significant effort is on the transition metal dichalcogenides (TMDCs), which exhibit a transition from an indirect band gap to a direct-gap semiconductor when the TMDC flake is thinned down to the single-layer limit [66, 67, 199]. Single-layer TMDCs possess a unique valley-spin degree of freedom [67, 94] and strong light-matter interaction [85], which can be exploited for novel quantum optoelectronic applications. However, having the direct band gap in a limited visible spectrum range and limited carrier mobility compromise the prospect of utilizing TMDCs for light-emitting devices, solar energy harvesting or quantum electronic device applications [199, 200, 201]. In parallel, thin black phosphorous (BP) flake has been discovered as a new 2D material with carrier mobility as high as 6000 $cm^2\ V^{-1}\ s^{-1}$, which enabled the observation of quantum interference effect and quantum Hall effect (QHE) [68]. Optically, BP remains a direct band gap material with an optical band gap sensitively dependent on layer numbers, approaching the bulk value of $\sim$ 0.3 eV from the single-layer value of $\sim$ 1.5 eV quickly as the layer number increases [202, 203, 204]. However, BP is extremely unstable under ambient conditions. The challenge of device stability has to be solved before it can be implemented for optoelectronics applications [205].

As a prototypical layered material in the III-VI family, InSe is a stable nonlinear optical crystal and exhibits superior electrical and optical properties [69, 206, 207, 208, 209]. Recently, the QHE has been demonstrated in a thin flake (6 layers) with mobility exceeding 10000 $cm^2\ V^{-1}\ s^{-1}$ [69]. Moreover, in contrast to the TMDCs, InSe is a direct band gap material from the bulk to a few layers (>5 layers), which renders it a promising candidate for efficient optoelectronic devices possessing large absorbance [207, 208, 209]. In addition, the band gap of multilayer InSe is located in the infrared regime and bridges the spectrum gap between the TMDCs and BP [207]. Considering the relatively small exciton binding energy in a few-layer InSe flake [210], we expect that the optical band gap obtained from the PL spectroscopy would correspond closely to the direct band gap at the Γ point for a few-layer (>5) system in the bulk hexagonal Brillouin zone, which is folded from the Z point in the Brillouin zone of the rhombohedral primitive cell. Therefore, PL spectroscopy can be employed to investigate the strain-induced band gap modulation of few-layer InSe flakes.

6.2 Results

The atomic structure of rhombohedral γ-InSe is given in Figs. 6.1a, b, consisting of In-Se-Se-In atomic layers stacked vertically with the van der Waals interaction. Each layer is of a hexagonal structure with D_{3h} symmetry. The layer-dependent band gap of InSe has been investigated both theoretically and experimentally [207, 210]. Bulk InSe is a direct-gap semiconductor with both the conduction band minimum (CBM) and valence band maximum

(VBM) located at the Z point. InSe flakes retain the direct-gap feature until they are less than five layers thick [207, 210]. Experimentally, we exfoliated InSe flakes from the γ-InSe single crystal onto the flexible poly(ethylene terephthalate) (PET) substrate with a thickness of 120 μm. The strain was applied using a home-built instrument in which the PET film was clamped at both ends. By pushing the PET film from one clamped end through a micrometer manipulator, we can precisely control the lateral movement and determine either uniaxial tensile or compressive strain, as schematically shown in Figs. 6.1c, d. The quantitative strain value can be determined by a two-point bending method: the bent PET is assumed as a circular arc for the strain calculation with the equation: strain $= \tau/R$, where 2τ and R are the thickness of PET substrate and the radius of curvature of the bent PET, respectively (Fig. 6.1d) [191].

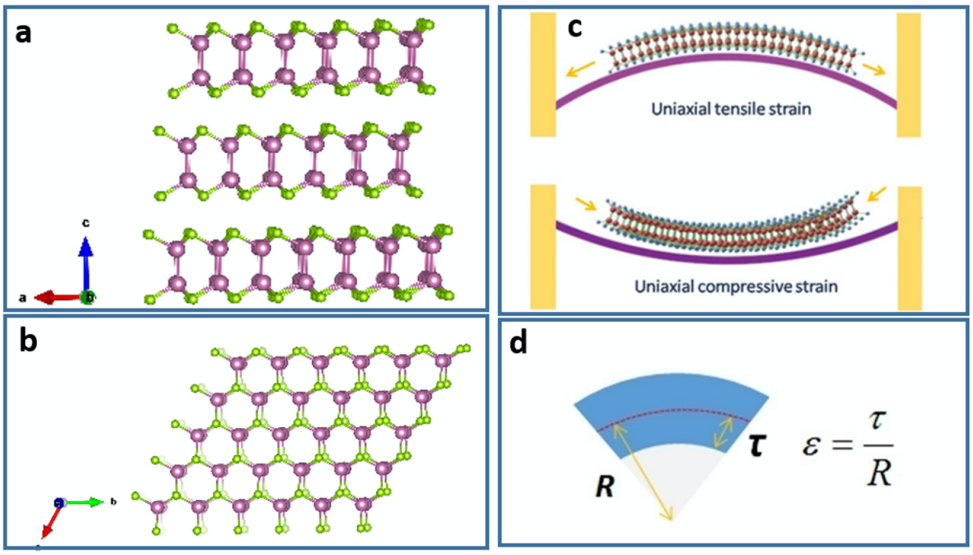

Figure 6.1: (a) A side view of the atomic structure of the InSe crystal structure. In atom: purple. Se atom: green. (b) A top view of the InSe crystal structure showing a hexagonal structure with D_{3h} symmetry. (c) A schematic of the two-point bending apparatus used for applying uniaxial tensile and compressive strain. (d) A schematic for the calculation of the strain in an InSe flake.

We investigate the evolution of the InSe bandstructure with strain using PL spectroscopy and first-principles DFT and GW calculations. Drastically different from semiconducting TMDCs, we show that the unstrained InSe with thickness more than $\sim$ 4 nm (about five layers with a single-layer thickness of 0.82 nm) can be approximately treated as a direct-gap semiconductor, with $\left|E_{\mathrm{g}}^{\Gamma} - E_{\mathrm{g}}^{\mathrm{ind}}\right| < 30$ meV based on our DFT calculations, where E_{g}^{Γ} is the direct band gap at the Γ point, and $E_{\mathrm{g}}^{\mathrm{ind}}$ the indirect band gap. Consequently, with the neglect of possible excitonic effects, the PL spectroscopy directly reads out the optical band

gap of InSe through the PL peak position. Considering the significant dielectric screening when the layer number exceeds five (exciton binding energy will be small) [203], the optical band gap of InSe is approximately the same as the quasiparticle band gap. Figure 6.2a shows the PL spectra of multilayer InSe of thickness $\sim$ 12 nm over a range of 0 – 1.06% tensile strain. The main PL peak in the unstrained InSe is located at 1.34 eV. With increasing strain, the PL spectra of InSe are significantly changed: the peak exhibits a redshift of about 118 meV with a tensile strain of $\sim$ 1.06%. We plot the PL peaks as a function of strain value for this sample in Fig. 6.2b (blue diamonds) with a linear fitting. The slope, demonstrating the sensitivity of the optical band gap to strain, is $\sim$ 101 meV/% strain for this sample. We also explored the effect of strain on the band gap of InSe flakes of different thicknesses. For all the flakes with thickness ranging from 4 nm to more than 30 nm, the PL peak exhibits a consistent redshift with increasing tensile strain, as shown in Fig. 6.2b, and the peak shift rate is determined to be 154 ± 8 meV/% for a thin flake (4 and 6 nm) and 81 ± 4 meV/% strain for a thick flake (>30 nm).

We also investigate the effect of the compressive strain on the band gap by applying a relatively small strain (from 0 to 0.62%), with the purpose of avoiding complication due to potential buckling or other mechanical deformation of the flakes. As shown in Fig. 6.2c, the PL peak of InSe of thickness $\sim$ 14 nm has a blueshift of about 56 meV when the compressive strain approaches 0.59%, corresponding to a shift rate of $\sim$ 100 meV/% strain. These results show that the band gap of InSe can be increased through a compressive strain and decreased through a tensile strain. We also measured the compressive-strain-dependent PL of samples with different thicknesses, and the PL peak position as a function of strain is shown in Fig. 6.2d. The band gap of InSe flakes obtained from the PL spectra increases with strain, and the band gap shift rate for a few-layer InSe flake ($\sim$ 5 nm) is determined to be $\sim$ 140 meV/% strain. The combination of tensile strain and compressive strain greatly enhances the band gap tunability of an InSe flake.

We compare our results with previously reported values in other 1D and 2D materials, summarized in Table 6.1. It is evident that the band gap strain coefficient with a uniaxial tensile strain in few-layer InSe is the most pronounced among what have been reported experimentally in 2D materials so far, several times larger than those of TMDCs and comparable to that of BP [192, 193, 194, 196, 211, 212, 213, 214]. This excellent tunability may find its applications in the future development of flexible electronic, optoelectronic and photonic devices with near-infrared spectral responses.

To understand the microscopic mechanism of the strain-induced band gap change, we performed first-principles DFT (implemented in the `Quantum ESPRESSO` package [106, 107]) and *GW* calculations (implemented in the `BerkeleyGW` package [108]) of InSe at the single-layer (Fig. 6.3a) and the bulk limit (Fig. 6.3b). We did not include the SOC effect in the shown calculations because it has a minimum effect on the band gap studied here, considering the CBM is mainly composed of In s orbitals and the VBM is mainly composed of Se p_z

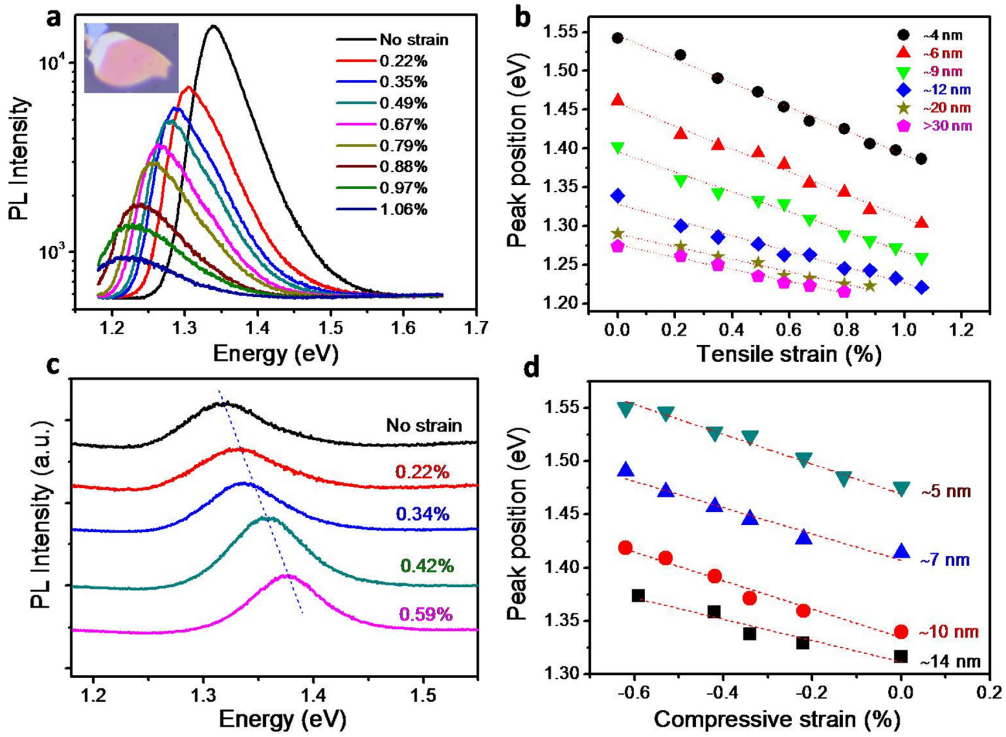

Figure 6.2: Strain-induced band gap change revealed by PL spectroscopy. (a) PL spectra of a multilayer InSe of thickness $\sim$ 12 nm. The applied uniaxial tensile strain ranges from 0 to 1.06%. Inset: the optical image of the flake on PET. The laser excitation spot is at the center of the pink flake. (b) PL peak positions *versus* tensile strain in InSe flakes of different thicknesses. (c) PL spectra of a multilayer InSe flake of thickness $\sim$ 14 nm under different uniaxial compressive strains. (d) PL peak position *versus* compressive strain in InSe flakes of different thicknesses.

orbitals [209, 210].

Unstrained single-layer InSe is found to be an indirect-gap semiconductor, with a "Mexican hat" feature near the Γ point. As shown in Fig. 6.3a, the direct band gap at Γ is $\sim$ 1.7 eV from DFT calculations, while *GW* calculations show a band gap of $\sim$ 3.1 eV. This drastic difference in the band gap between the two calculations originates from strong electron-electron interaction and weak dielectric screening in 2D materials. As the layer number increases to more than five layers, DFT calculations show that InSe quickly becomes a direct-gap semiconductor, with the band gap located at the Γ point [2]. The bandstructure of unstrained bulk InSe is shown in Fig. 6.3b. Bulk InSe is found to be a direct-gap semiconductor with the band gap located at the Z point in the Brillouin zone of the rhombohedral primitive cell. The band gap is 0.35 eV from DFT calculations and 1.25 eV from *GW* calculations. The *GW* band gap value shows an excellent agreement with the experimentally measured band

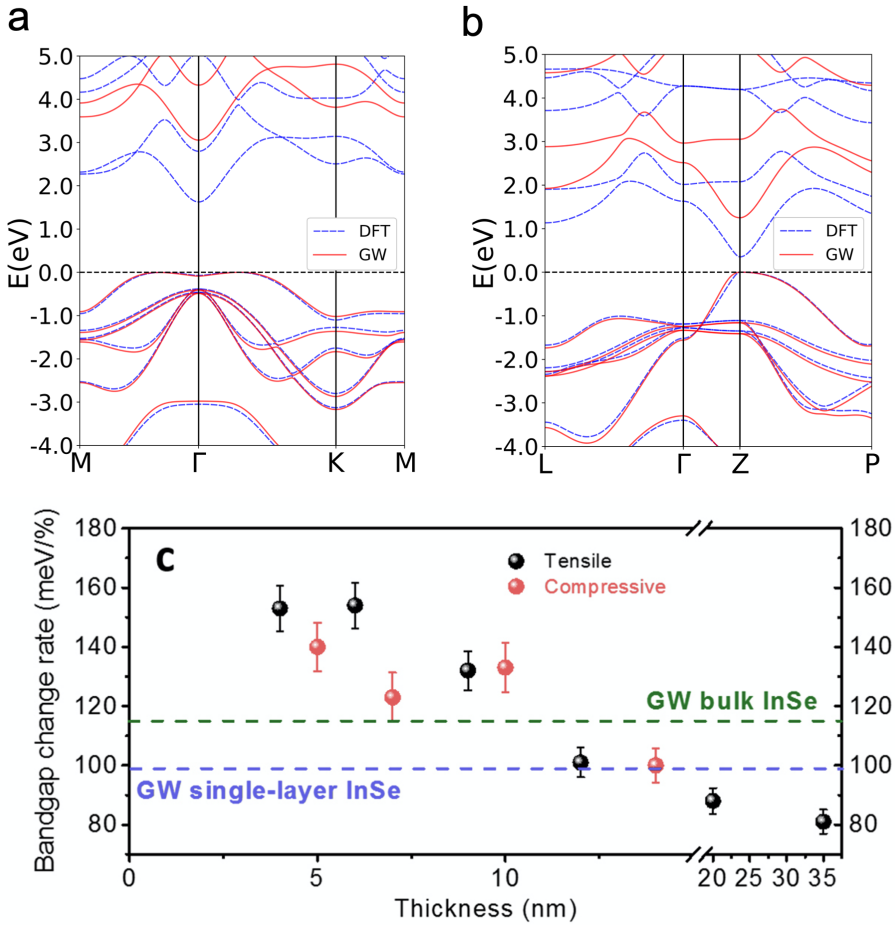

Figure 6.3: Calculation of the strain-induced band gap change in InSe. (a) DFT-PBEsol (blue dashed) and *GW* (red solid) bandstructures of single-layer InSe. (b) DFT-PBEsol (blue dashed) and *GW* (red solid) bandstructures of bulk InSe. (c) Experimentally measured band gap strain coefficients for uniaxial tensile strain (black dots) and compressive strain (red dots) as a function of the InSe flake thickness. The blue and green dashed lines label the calculated quasiparticle band gap strain coefficient of single-layer and bulk InSe, respectively.

Materials	Thickness (nm)	Strain (%)	Strain coefficient (meV/%)
InSe	4	1.06	153
	6	1.06	154
	9	1.06	132
	12	1.06	101
	20	0.88	88
	35	0.79	81
	5	−0.62	140
	7	−0.62	123
	10	−0.62	133
	14	−0.59	100
MoS_2 [192, 193]	Single-layer	2.2	45
	Single-layer	0.52	70
WSe_2 [194, 215]	Single-layer	1.04	60–70
	Single-layer	1.4	54
$MoSe_2$ [211]	Single-layer	1.1	27
WS_2 [212]	Single-layer	2.5	20–30
$ReSe_2$ [216]	Single-layer	1.64	36
BP [213, 214]	10	5	100–140
	4	0.92	117–124
MoS_2 biaxial strain [217]	Single-layer	>2	99
GaAs nanowire [218]	40	3.5	84

Table 6.1: Comparison of band gap tunability between InSe and other 1D or 2D semiconductors under uniaxial tensile, compressive and biaxial strain.

gap which is at ~ 1.27 eV for a few-layer InSe that is 30 nm thick.

Since we access the direct band gap experimentally through PL spectroscopy, in our calculations, we focus on the strain effect on the corresponding direct band gap for the single-layer and bulk InSe. We first investigate the strain effect on the single-layer InSe, and our DFT calculations show that the band gap strain coefficient is 91 meV/% for an in-plane uniaxial tensile strain [2]. Our *GW* calculation shows a coefficient of 97 meV/% (Fig. 6.3c), only 7% larger than that from the DFT calculation. We also calculated the band gap strain coefficient for bulk InSe, and the DFT calculation shows a value of 95 meV/% and the *GW* calculation results in an increased value of 114 meV/% (Fig. 6.3c).

We plot the experimental band gap strain coefficients for InSe of different thicknesses in Fig. 6.3c and compare them with the *GW* results. Since PL peaks can be determined accurately in experiments, the uncertainty in measurement is mainly from the strain calculation. Fig. 6.3c shows that for InSe flake of thickness ~ 4 nm (5 layers) and ~ 6 nm (8 layers), the band gap strain coefficient due to uniaxial tensile strain is almost the same,

within the experimental uncertainty considered. The band gap strain coefficient due to uniaxial compressive strain, which is slightly smaller, is also about the same for the InSe flake of thickness from ~ 5 nm to ~ 10 nm. The value of the nearly constant band gap strain coefficient is in better agreement with the *GW* calculation for bulk InSe (114 meV/%), and larger than what is predicted for the single-layer InSe (97 meV/%). In addition, we attribute the remaining discrepancy between experimental and calculated band gap strain coefficients to unintended out-of-plane strain in the experiment, which is confirmed by the strain-induced shift of Raman A_1 and A_1(LO) modes [2]. Calculations with additional out-of-plane tensile (compressive) strain for bulk InSe structures with fixed in-plane compressive (tensile) strain will increase the calculated band gap strain coefficient. It is interesting to note that the strain-induced effect on the band gap in few-layer InSe is closer to that of bulk InSe rather than a single-layer InSe. This can be intuitively understood by the fact that the electronic bandstructure of few-layer InSe is closer to that of the bulk InSe (direct band gap at the Γ point: ~ 1.3 eV) shown in Fig. 6.3b, much smaller than that of the single-layer InSe at the Γ point (3.14 eV) shown in Fig. 6.3a.

We also note that InSe flakes thicker than 10 nm show significantly decreased band gap strain coefficient as the thickness of the flake increases. This trend is likely due to an effective strain decay in the thick flakes. The nominal strain is calculated based on the geometry of the flexible PET film, which should be the same as that in the thin InSe flake since it exactly follows the change of the substrate [191, 192]. For thick samples, we expect inefficient strain transfer between different layers, which will result in a strain gradient in the flake. This strain gradient is confirmed by the extra broadening observed in PL width of strained thick InSe flakes [2], and it will reduce the effective strain applied to the flake and lead to reduced strain coefficient observed.

6.3 Conclusion

In summary, we have investigated the effect of uniaxial tensile and compressive strain on the electronic structure of InSe with different thicknesses. The strain-induced band gap modulation has been experimentally demonstrated, suggesting that strain engineering is an effective tool to explore the novel physics of III-VI layered compounds. Applying a relatively small strain, we have demonstrated a reversible optical band gap tunability of ~ 239 meV in the infrared regime. Such tunability originates from a sensitive response of InSe electronic structure to the external strain, which manifests itself as a substantial band gap strain coefficient as large as ~ 154 meV/% under a uniaxial tensile strain. The observed band gap strain coefficients are confirmed by our first-principles DFT and *GW* calculations. This sensitive response to strain holds great potential for a wide range of applications based on atomically thin InSe flakes, including electromechanical, piezoelectric and optoelectronic devices.

Chapter 7

Exchange-driven intravalley mixing of excitons in monolayer transition metal dichalcogenides

Monolayer transition metal dichalcogenides (TMDCs) are promising 2D semiconductors with a lot of potential applications in optoelectronics. Their optical properties are dominated by two series of photo-excited exciton states – A (XA) and B (XB) [67, 219] – that are derived from direct interband transitions near the band extrema. These exciton states have large binding energies and strong optical absorption [85, 87, 174, 220], and form an ideal system to investigate many-body effects in low dimensions. Because the SOC effect causes a large splitting between bands of opposite spins, XA and XB are usually treated as spin-polarized Ising excitons, each arising from interactions within a specific set of states induced by interband transitions between pairs of either spin-up or spin-down bands (TA or TB). Among the various types of many-body effect, the electron-hole exchange interaction is known to be an important mechanism for defining exciton landscape [221, 222] and for exciton spin relaxation in quantum wells based on III-V semiconductors [223, 224, 225]. This effect also contributes to valley depolarization [226], valley decoherence [227, 228], valley-orbit coupling for excitons [229] and the shaping of the exciton dispersion relation [86, 230] in monolayer TMDCs. Despite the extensive studies of many-body effects among excitons in monolayer TMDCs [231, 232, 233], understanding of the excitonic exchange interaction within the same valley is incomplete due to the insufficiency of existing first-principles methods, and the restricted resolution of the coupling between photo-excited states with traditional linear optical spectroscopy.

In this chapter, I present a work in collaboration with an experimental group led by Prof. Graham R. Fleming [3]. In this work, by using the full-spinor GW and GW-BSE methods and state-of-the-art 2D electronic spectroscopy, we numerically demonstrate and experimentally verify the role of the strong intravalley exchange interaction in forming XA and XB in a prototypical TMDC, monolayer MoS_2. This exchange interaction is significant because of the reduced dielectric screening and enhanced wavefunction overlap between the

electrons and the holes in a monoatomic layer and is manifested in the transient optical response.

7.1 Introduction

An exciton is an electron-hole pair, bound by Coulomb interactions. The interaction between the electron and the hole contains two terms, a direct screened Coulomb interaction and an exchange bare Coulomb interaction. The attractive direct interaction corresponds to a classical picture of Coulomb attraction between the two oppositely charged quasiparticles, producing a series of hydrogenic-like electron-hole states, of which the envelope wavefunctions have specific nodal structure and angular momentum quantum numbers ($1s$, $2s$, $2p$ and so on). The diagrams labeled TA and TB in Fig. 7.1a schematically show such electron-hole states in the K valley of monolayer TMDCs, which are bound by the direct interaction only. The states TA and TB are derived from interband transitions occurring between the spin-up and the spin-down band-pairs, respectively, and are therefore referred to as Ising excitons. The repulsive exchange interaction involving the bare Coulomb interaction, on the other hand, arises from the exchange symmetry of the many-fermion wavefunctions, and leads to exciton eigenstates consisting of electron-hole states with mixed spins, as illustrated in diagrams labeled XA and XB in Fig. 7.1a.

7.2 Results

7.2.1 Exchange-driven mixing of exciton states

We employ the full-spinor GW and GW-BSE methods implemented in the `BerkeleyGW` package [108] to fully incorporate the exchange interaction, particularly the intravalley exchange interaction between TA and TB. It is demonstrated that the true exciton eigenstates, XA and XB, are mixed states of TA and TB. Figs. 7.1d–i shows the k-space wavefunction amplitudes of the bright exciton eigenstates XA-1s, XB-1s, and XA-2s projected onto TA (Figs. 7.1d, f, h) and TB subspaces (Figs. 7.1e, g, i). TA and TB are illustrated in Figs. 7.1b and 7.1c, respectively. We label the principal quantum number of an exciton eigenstate by the nodal structure of its envelope wavefunction in its major subspace where the amplitude is the largest. The 1s state of XA (XB) primarily includes TA (TB), but also mixes in 3.6% of TB (5.3% of TA) by the intravalley exchange interaction. Moreover, the mixing also occurs between states with different principal quantum numbers. For example, the k-space wavefunction amplitude of XA-2s has components with one node in the TA subspace, but shows components without node (a 1s-like character) in the TB subspace. The mixing of TA and TB produces an important change in the ratio of the optical oscillator strengths of XA and XB in the calculations. This large asymmetry in the oscillator strengths has been observed in experiments but has not been well explained [234]. The calculated oscillator strengths of

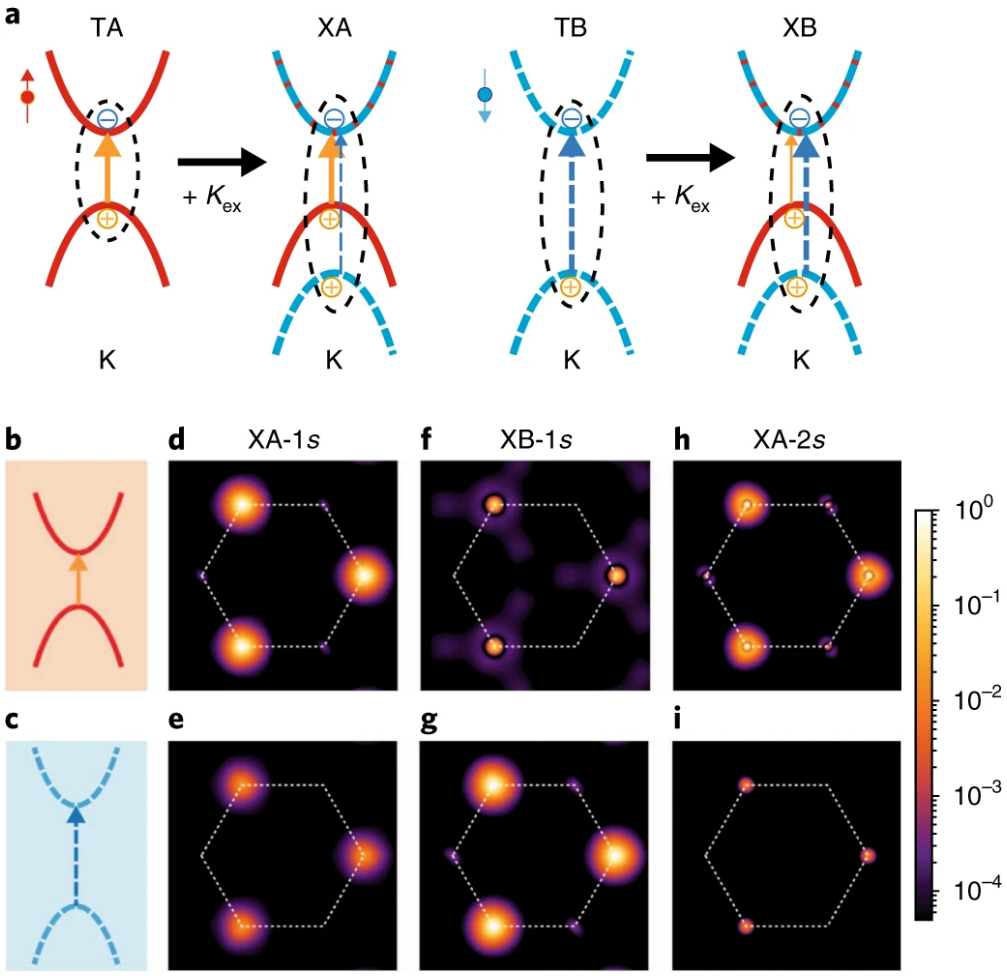

Figure 7.1: Exciton state mixing from the intravalley exchange interaction. (a) Schematic of exchange-driven mixing of the two Ising excitons, TA and TB, which produce the exciton eigenstates, XA and XB. The orange solid arrows denote the transition between spin-up states (in the TA subspace) while the blue dashed arrows denote the transition between spin-down states (in the TB subspace). The thicker arrow in the plot of XA or XB indicates the major subspace. Dashed ovals denote the exciton states composed of the corresponding interband transitions. K_{ex} represents the exchange interaction kernel. (b, c) Illustration of TA and TB subspaces, respectively. (d–i) Log-scale wavefunction amplitudes in k-space (normalized to unity at the highest amplitude for each state) for exciton eigenstates (d, e) XA-1s, (f, g) XB-1s, and (h, i) XA-2s with the exchange interaction included. The upper (d, f, h) and lower (e, g, i) panels show the projected amplitudes of the wavefunction onto the TA and TB subspace, respectively.

	XA-1s	XB-1s	XA-2s
Without TA-TB exchange interaction	1.01×10^{-5}	1.07×10^{-5}	0.34×10^{-5}
With full exchange interaction	0.67×10^{-5}	1.27×10^{-5}	0.41×10^{-5}

Table 7.1: Calculated oscillator strengths (in atomic units) of the first three bright exciton states in monolayer MoS_2.

optical transitions (in atomic units) to the different exciton states are presented in Table 7.1. If the exchange interaction between TA and TB is excluded in the calculations, the two Ising exciton 1s states have nearly equal oscillator strengths, consistent with previous theoretical works on monolayer TMDCs, which have all ignored this interaction. With inclusion of the intravalley exchange interaction, the oscillator strength of XA-1s decreases whereas that of XB-1s increases, leading to a factor of two change in their ratio. This asymmetric change in the oscillator strengths originates from the destructive and the constructive interference between TA and TB due to the intravalley spin-resolved exchange interaction according to our first-principles GW-BSE calculations.

7.2.2 2D electronic spectroscopy

An important role of the exchange interaction in exciton spin dynamics was demonstrated both theoretically and experimentally in III-V quantum wells, where the exchange interaction leads to simultaneous spin flip of the electron and the hole after an ultrafast resonant excitation of excitons [223, 224, 225]. Therefore, in monolayer MoS_2, when the excitation pulse duration is short enough, we expect an evolution from the transient Ising exciton states (which are directly launched by the pulse) towards the eigenstates XA and XB with mixed spin polarization. The evolution dynamics are mainly driven by the intravalley exchange interaction. Based on these considerations, 2D electronic spectroscopy (an ultrafast four-wave mixing spectroscopy technique) is employed to demonstrate the intravalley exchange interaction unambiguously in both time and frequency domains. This method has been utilized to map excitonic coupling and energy transport in various systems including photosynthetic complexes [235, 236] and semiconductors [228, 232, 233, 237, 238, 239, 240], with high resolution in both the time and frequency domains.

Monolayer MoS_2 with continuous size (~ 10 mm^2, polycrystalline) grown on a sapphire substrate by chemical vapour deposition is used for our experiments. The broadband laser in our experiments, covering the resonance energies of both XA-1s and XB-1s in monolayer MoS_2 as shown in Fig. 7.2a, is generated from an optical parametric amplifier driven by a Ti:sapphire femtosecond laser amplifier (Astrella, Coherent). In the experiments, the sample was cooled to 40 K by liquid helium. In this work, 2D electronic spectroscopy was performed using the boxcar geometry in the rephasing scheme with co-circularly polarized

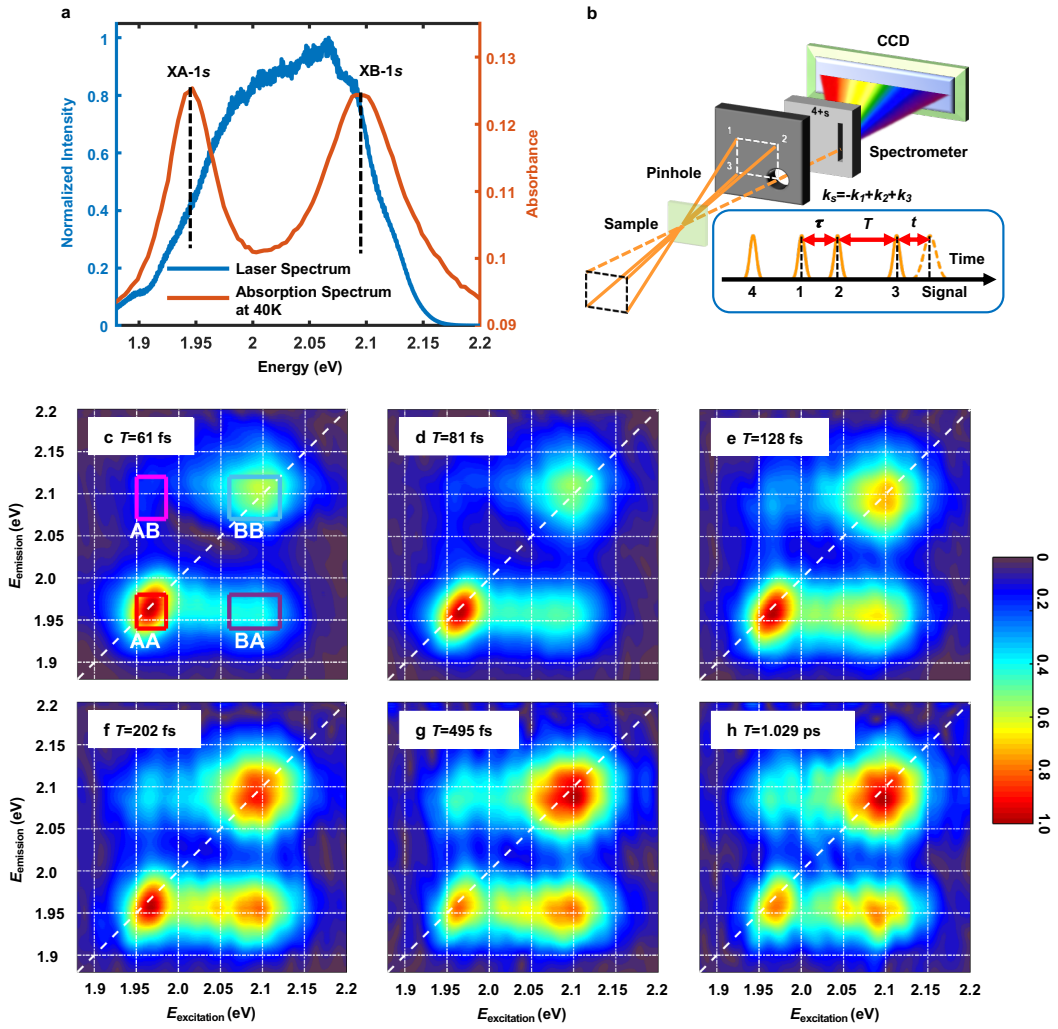

Figure 7.2: (a) Laser spectrum (left, blue, normalized) and absorption spectrum (right, red) of monolayer MoS_2 at 40 K. (b) 2D electronic spectroscopy in the boxcar geometry. Four beams, denoted 1, 2, 3 and 4, are focused into the sample. The fields with wave vectors $\boldsymbol{k}_1$, $\boldsymbol{k}_2$ and $\boldsymbol{k}_3$ interact with the sample and generate a photon echo signal with wave vector $\boldsymbol{k}_s = -\boldsymbol{k}_1 + \boldsymbol{k}_2 + \boldsymbol{k}_3$. The signal collinear with beam 4 is measured by heterodyning detection. The delay τ between pulses 1 and 2 and the delay T between the second striking pulse (pulse 2 for rephasing) and pulse 3 are adjusted. The delay time t between pulse 3 and the signal is implicit since the signal is detected in the frequency domain by a spectrometer. (c–h) Rephasing amplitude 2D spectra (normalized) at labeled waiting times. The four peaks are denoted by $\alpha\beta$ (α, β = A, B), which indicates excitation resonant with transition α and emission resonant with transition β. Peak regions are marked by colored squares. The color scale in the simulated 2D spectra denotes values between 0 and 1 due to normalization.

light, as shown in Fig. 7.2b. Three input laser fields interact with the sample and induce a third-order polarization, which emits a photon echo signal in the phase-matching direction. The signal is acquired as a function of the coherence time τ and the waiting time T. The signal field $\mathcal{E}_{\mathsf{s}}(E_{\mathsf{excitation}}, T, E_{\mathsf{emission}})$, as a 2D function of the excitation energy $E_{\mathsf{excitation}}$ and the emission energy E_{emission} at a specific T, is extracted by a Fourier transform with respect to τ. Presenting $\mathcal{E}_{\mathsf{s}}$ in 2D diagrams reveals the correlation between the excitation and the emission energies with T as the delay between the excitation and the emission events, thus tracking the energy transport landscape.

The rephasing amplitude 2D spectra (Figs. 7.2c–h) feature two diagonal peaks (AA and BB) and two cross peaks (AB and BA). The diagonal peaks indicate ground-state bleaching and stimulated emission due to excitation of each individual transient state TA or TB. The cross peaks correspond to excitation resonant with TA (TB) and photon echo emission resonant with TB (TA). In previous works on semiconductors [232, 233, 237], the presence of cross peaks in 2D spectra was explained by an excitation-induced shift of the resonance energy and excitation-induced dephasing of the electronic polarization caused by many-body effects, which offset the contributions from ground-state bleaching and excited-state absorption. In this scenario, the amount of peak shift and broadening is proportional to the excited exciton population. Therefore, the cross-peak amplitude should be maximized at small T, when the exciton population is maximal, and then decay together with the diagonal peak amplitude, which is proportional to the population of the corresponding state. However, our data deviate from the prediction of the excitation-induced shift/dephasing model. As shown in Figs. 7.2c–h, the cross peak AB is dark at $T = 61$ fs and becomes brighter relative to the diagonal peak AA as T increases. This indicates a conversion from the excited lower-energy state TA to the higher-energy state TB, which is qualitatively consistent with our theoretical analysis considering the intravalley exchange interaction. The system, which is initialized into the transient state TA by resonant excitation, evolves towards the eigenstate XA, which mixes TA and TB and therefore populates TB. Phonon-assisted population upconversion [233, 241], which was previously observed between the trion and the exciton in monolayer $MoSe_2$ and WSe_2, is excluded here since it can hardly surmount the energy gap of 150 meV between XA-1s and XB-1s. The small contribution of the Auger effect (by which two XA-1s interact and induce one XB-1s) to the cross peak AB is confirmed by the nearly constant ratio of the peak amplitude AB/AA *versus* fluence. Spin flipping of the electron component of TA by carriers and phonons could also bleach the TB transition. However, it has been shown by time-resolved Kerr rotation that the spin lifetime of electrons in TA is as long as 200 ps at 40 K [242], whereas the growth of the peak ratio AB/AA occurs within 1 ps. Spin flipping of the hole component of TA takes longer than this timeframe due to the much larger energy splitting in the valence band. Therefore, we attribute the emergence of the TB feature after the excitation of TA to the excitonic exchange interactions.

Considering the intravalley exchange interaction as the mechanism for the population transfer, we computed the 2D spectra. We model the coupling between TA and TB using a four-level diamond system constructed from a Hilbert space transformation of two indepen-

dent two-level systems. The rephasing amplitude 2D spectra are calculated by perturbatively solving the optical Bloch equations to the third order, which dictates the dynamics of the density matrix. The photon echo signal can be divided into contributions from 14 Liouville pathways, each represented by a double-sided Feynman diagram [3]. Two Feynman diagrams shown in Fig. 7.3a are employed to incorporate the intravalley exchange interaction between TA and TB. The initial condition for AB (BA) peak dynamics is set to be the pure population of the Ising excitons with the amplitude proportional to the initial AA (BB) peak intensity from experiments. The simulated rephasing amplitude 2D spectra are shown in Figs. 7.3c–h at different waiting times, and reproduce the experimental results (Figs. 7.2c–h) reasonably well. The ratios of the cross-peak amplitudes to the corresponding diagonal-peak amplitudes from the simulations are plotted in Fig. 7.3b together with the experimental data. The slight overestimation of the ratio AB/AA at $T > 500$ fs in the simulations is probably due to the neglect of phonon-assisted population down-conversion from TB to TA. For the same reason, the simulations show a slow rise in BA/BB whereas the experiment captured a fast rise within 200 fs.

7.3 Conclusion

In summary, the theoretically and experimentally unraveled intravalley exchange interaction reforms the exciton landscape in monolayer TMDCs. We reveal, for the first time in 2D materials, the role of the intravalley excitonic exchange interaction in mixing the excited transient states with opposite spin polarization. This interaction serves as an important channel for exciton spin relaxation, especially in the ultrafast time domain and therefore should be taken into account for the design of TMDC-based spintronics and valleytronics and in the exploration of novel 2D materials with valley-spin coupling.

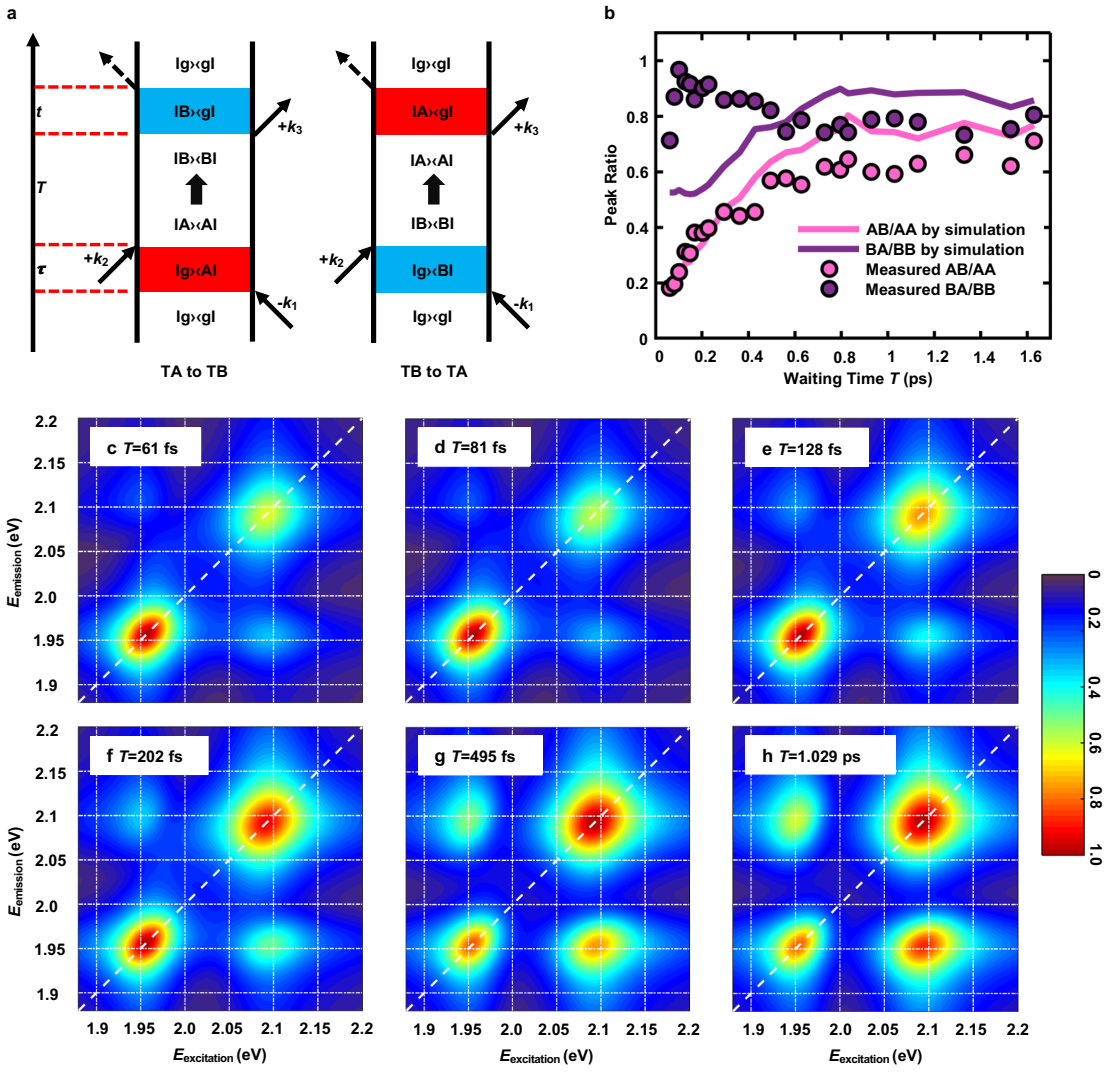

Figure 7.3: (a) Feynman diagrams with population transfer between TA and TB due to the intravalley exchange interaction. The left-hand diagram contributes to the cross peak AB whereas the right-hand one contributes to the cross peak BA. In the diagrams, g denotes the ground state before excitation. (b) Calculated peak ratios *versus* waiting time T compared with experimental results. (c–h) Simulated rephasing amplitude 2D spectra (normalized) at labeled waiting times, including the intravalley exchange interaction. The color scale in the simulated 2D spectra denotes values between 0 and 1 due to normalization.

Chapter 8

Physical origin of giant excitonic and magneto-optical responses in 2D ferromagnetic insulators

The magneto-optical (MO) effects, such as the magneto-optical Kerr effect (MOKE) and the Faraday effect (FE), have been intensively investigated experimentally in a variety of magnetic materials, serving as a highly sensitive probe for electronic and magnetic properties. Recent MOKE experiments have led to the discovery of 2D magnets, and demonstrated their rich magnetic behaviors [72, 73]. In particular, a giant Kerr response has been measured in ferromagnetic mono- and few-layer CrI_3 [73]. Magnetic circular dichroism (MCD) in photoabsorption has also been measured in ferromagnetic monolayer CrI_3 [12]. However, the exact microscopic origin of such large MO signals and MCD responses in 2D materials is still unclear, because treating accurately sizable SOC and excitonic effects that are essential for such an understanding in these systems has been very challenging for existing theoretical methods.

In this chapter, by performing first-principles full-spinor GW and GW-BSE calculations, we show that excitonic effects dominate the optical and MO responses in the prototypical 2D ferromagnetic insulator, CrI_3 [4]. We simulate the Kerr and Faraday effects in realistic experimental setups and based on which we predict the sensitive frequency- and substrate-dependence of MO responses. These findings provide a physical understanding of the phenomena as well as potential design principles for engineering MO and optoelectronic devices using 2D magnets.

8.1 Introduction

CrI_3, in its monolayer and few-layer form, is a prototypical 2D ferromagnetic insulator with an Ising-like magnetic behavior and a Curie temperature of about 45 K, exhibiting tremendous out-of-plane magnetic anisotropy [73]. Within one layer, the chromium atoms

form a honeycomb structure, with each chromium atom surrounded by six iodine atoms arranged in an octahedron (Figs. 8.1a, b), and the point group of the structure is S_6. The crystal field therefore splits the Cr $3d$ and I $5p$ ligand states into t_{2g} and e_g manifolds; the spin degeneracy of which are further lifted by the magnetic exchange interaction. Although the major-spin e_g states are delocalized due to strong p-d hybridization, the magnetic moment is approximately $3\mu_\mathrm{B}$ at each Cr site, in accordance with an atomic picture from the first Hund's rule [243].

With the full-spinor GW and GW-BSE methods, we show from first principles that the exceedingly large optical and MO responses in ferromagnetic monolayer CrI_3 arise *per se* from strongly bound exciton states consisting of spin-polarized electron-hole pairs that extend over several atoms. These exciton states are shown to have distinct characteristics compared with either the Frenkel excitons in ionic crystals and polymers, or Wannier excitons in other 2D semiconductors. By simulating realistic experimental setups, we further find that the substrate configuration and photon frequency strongly shape the MO signals. Our results provide the conceptual mechanism for the giant optical and MO responses, explaining quantitatively the recent experiments on CrI_3 [12, 73]. In addition, comparison between bulk and monolayer CrI_3 reveals the pivotal role of quantum confinement in enhancing the MO signals.

8.2 Results

8.2.1 Quasiparticle bandstructure

An accurate first-principles calculation of the electronic structure of CrI_3 should account for both the dielectric polarization from the ligand groups and the on-site Coulomb interactions among the localized spin-polarized electrons. We adopt the first-principles GW method at the G_0W_0 level implemented in the `BerkeleyGW` package [108] to describe the dielectric screening and quasiparticle excitations [9]. Through the screened Coulomb interaction W, the nonlocal and dynamical screening effects as well as the self-energy effects beyond the DFT Kohn-Sham orbital energies (within LSDA) are captured. Also, in previous studies, the method of LSDA with an on-site Hubbard potential (LSDA+U) has served as a reasonable mean-field starting point for G_0W_0 calculations in correlated systems to avoid the spurious p-d hybridization [149, 244]. In this work, we adopt an on-site Hubbard potential in the rotationally invariant formulation [245] with $U = 1.5$ eV and $J = 0.5$ eV, with fully relativistic pseudopotentials and a plane wave basis set, as implemented in the `Quantum ESPRESSO` package [106, 107]. The validity of this specific set of U and J has been carefully tested [4]. Throughout this work, the magnetization of ferromagnetic monolayer CrI_3 is taken to be along the $+z$ direction (Fig. 8.1b). As shown in Fig. 8.1c, our calculations reveal a strong self-energy correction to the quasiparticle band gap, due to the weak dielectric screening in reduced dimensions and the localized nature of the d states. The

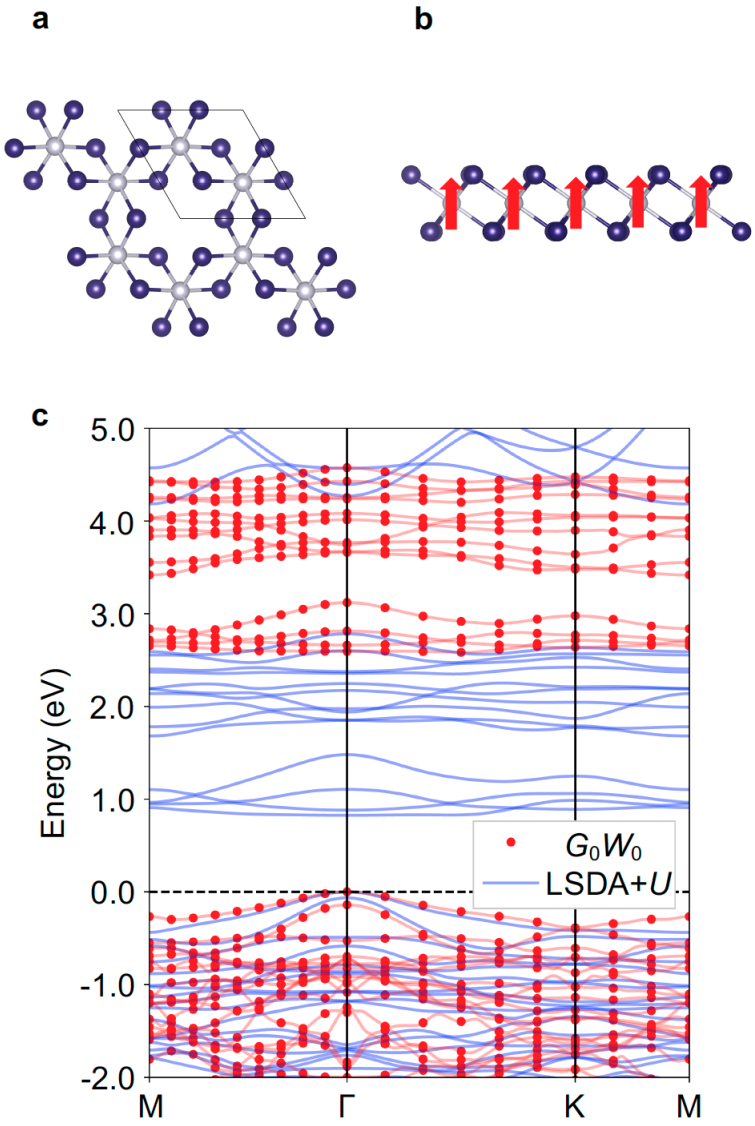

Figure 8.1: Crystal structure and electronic structure of ferromagnetic monolayer CrI_3. (a) Crystal structure (top view) of monolayer CrI_3. Chromium atoms are in gray while iodine atoms in purple. (b) Crystal structure (side view) of ferromagnetic monolayer CrI_3. Red arrows denote the out-of-plane magnetization, which is pointing along the $+z$ direction. (c) G_0W_0 (red dots) and LSDA+U (blue lines) bandstructures of ferromagnetic monolayer CrI_3. A rotationally invariant Hubbard potential is employed with $U = 1.5$ eV and $J = 0.5$ eV in the LSDA+U calculation, which is then used as the starting mean field for the G_0W_0 calculation. The G_0W_0 bandstructure is interpolated with spinor Wannier functions.

direct band gap is 0.82 eV at the Γ point at the LSDA+U level, whereas the direct G_0W_0 quasiparticle band gap including the self-energy effect is 2.59 eV, as shown in Fig. 8.1c. Throughout the calculations, we incorporate the SOC effect from the outset by employing spinor wavefunctions.

8.2.2 Exciton-dominant optical responses

The strong SOC strength and the ligand states strongly hybridizing with Cr d orbitals have decisive influences on the electronic structure and optical responses of ferromagnetic monolayer CrI_3. SOC significantly modifies the band gap and band dispersion near the valence band maximum (VBM) [243]. Figure 8.2a shows the G_0W_0 bandstructure together with each state's degree of spin polarization (with an out-of-plane quantization axis), of which the orbital and spin degeneracy are consistent with the above discussions. After solving the first-principles BSE, which describes the electron-hole interaction [34], with spinor wavefunctions, we find a series of strongly bound dark (optically inactive) and bright (optically active) exciton states with excitation energies (Ω_S) below the quasiparticle band gap, as shown in the plot of the exciton energy levels (Fig. 8.2b). As seen in Fig. 8.2c, the calculated linearly polarized absorption spectrum including the electron-hole interaction (i.e., with excitonic effects, red solid curve labeled GW-BSE) features three peaks at around 1.50, 1.85, and 2.35 eV (below the quasiparticle gap of 2.59 eV), which are composed of several bright exciton states in each peak and denoted as A, B, and C, respectively. This is in contrast to the calculated step-function-like noninteracting absorption spectrum (i.e., without excitonic effects, blue dashed curve labeled GW-RPA). The magnitude of the absorbance peak around 1.50 eV is deduced to be 0.7% from a previous differential reflectivity measurement (Fig. 8.2c, inset) [12], while our calculated absorbance with a broadening factor of 80 meV is around 0.6% at 1.50 eV. From our calculation (Fig. 8.2b), there are also two dark states (excitons D) with enormous binding energy of larger than 1.7 eV. The existence of two states of nearly the same energy comes from the fact that there are two Cr atoms in a unit cell. We plot the real-space exciton wavefunctions of these states, with the hole fixed on a Cr atom, in Figs. 8.2d–k. Unlike monolayer TMDCs where the bound excitons are of Wannier type with a diameter of several nanometers [85, 87], ferromagnetic monolayer CrI_3 hosts dark Frenkel-like excitons localized on a single Cr atom (Figs. 8.2d, e) and bright charge-transfer or Wannier excitons with wavefunctions extending over one to several primitive cells (Figs. 8.2f–k). These plots are consistent with the intuition that a smaller exciton binding energy is related to a larger exciton radius [87, 230]. Numerical calculations of the exciton radius further corroborate this conclusion (see Table 8.1).

In addition, ferromagnetism and broken time-reversal symmetry (TRS) play vital roles in determining the internal structure of the exciton states in ferromagnetic monolayer CrI_3, in contrast to the Frenkel/charge-transfer excitons determined solely by flat-band transitions in boron nitride systems [246, 247], organic materials [248, 249], or alkali halides [34]. The

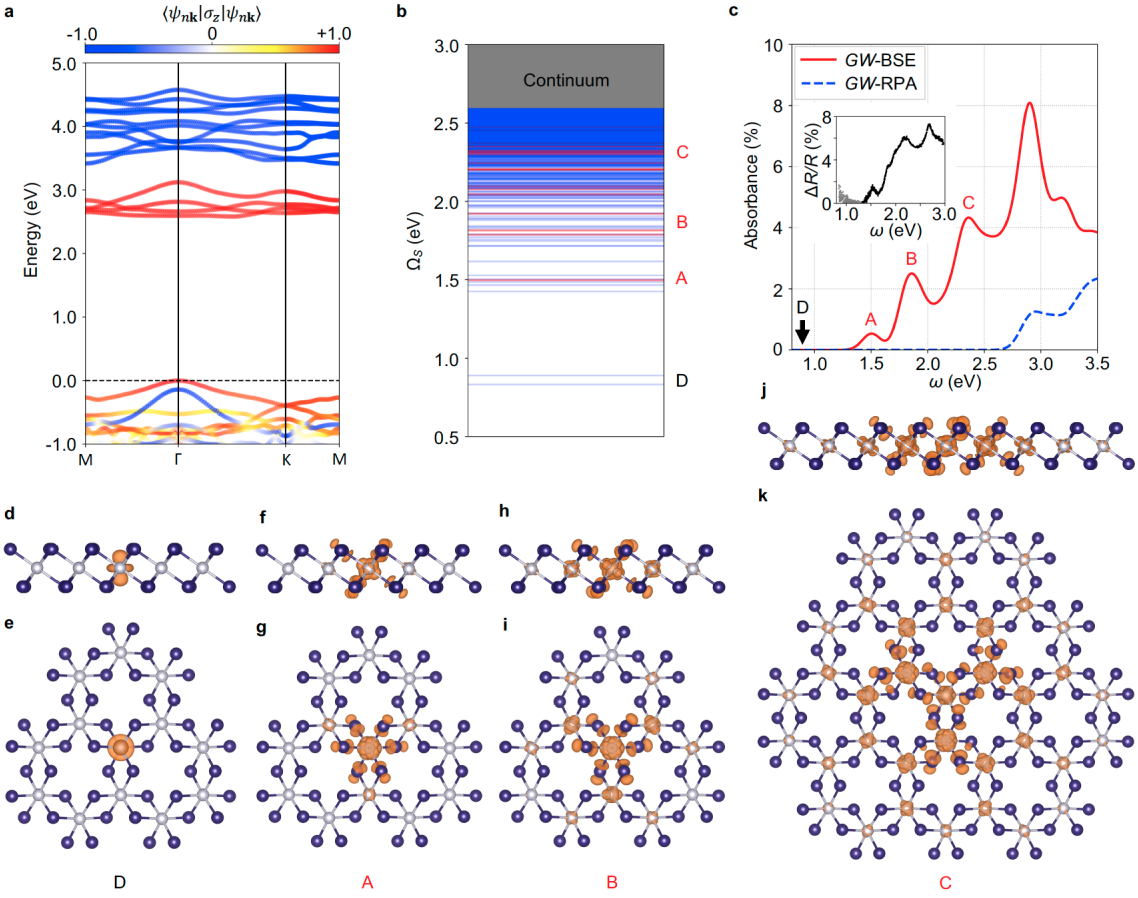

Figure 8.2: (a) G_0W_0 bandstructure of monolayer CrI_3 with colors denoting the magnitude of spin polarization along the out-of-plane direction. The red (blue) color denotes the major-spin (minor-spin) polarization. (b) Exciton energy levels of monolayer CrI_3. Bright states are in red while dark ones in blue. The bright states have at least two orders of magnitude stronger oscillator strength compared with the dark ones. We label the bound exciton states with D for the lowest-lying dark states and A–C for the higher-lying bright states. (c) Absorption spectrum of linearly polarized light with electron-hole interaction (GW-BSE, red solid line) and without electron-hole interaction (GW-RPA, blue dashed line). The inset data are extracted from Ref. [12] showing the experimental differential reflectivity, and the signals above 1.3 eV are shown in black for better comparison. (d–k) Exciton amplitudes in real space with the hole fixed on a Cr atom. Shown are iso-value surfaces of the amplitude square at 1% of the maximum value. Upper panel: side view. Lower panel: top view. (d, e) Dark exciton D ($\Omega_S = 0.89$ eV); (f, g) bright exciton A ($\Omega_S = 1.50$ eV); (h, i) bright exciton B ($\Omega_S = 1.82$ eV); (j, k) bright exciton C ($\Omega_S = 2.31$ eV). The dominant states (with the largest oscillator strength among the nearby states in the same group) are plotted.

	A	B^+	B^-	C
Ω_S (eV)	1.50	1.82	1.92	2.31
E_b (eV)	1.09	0.77	0.67	0.28
$\langle\lvert\boldsymbol{r}\rvert\rangle$ (Å)	2.33	3.55	5.36	6.99
$\sqrt{\langle r^2\rangle}$ (Å)	3.06	4.37	6.70	7.93

Table 8.1: The arithmetic mean radius $\langle|\boldsymbol{r}|\rangle$ and root mean square radius $\sqrt{\langle r^2\rangle}$ of selected bright exciton states. The excitation energy Ω_S and binding energy E_b are also included for reference.

eigenstate of an exciton is a coherent superposition of free electron-hole pairs at different k-points, and can be written as $|S\rangle = \sum_{cv\boldsymbol{k}} A^S_{cv\boldsymbol{k}}|cv,\boldsymbol{k}\rangle$, where $A^S_{cv\boldsymbol{k}}$ is the exciton envelope function in k-space [34]. Here c denotes conduction (electron) states and v denotes valence (hole) states. In Figs. 8.3a–d, we plot the module square of the exciton envelope function in k-space. As expected of highly localized Frenkel excitons in real space, the lowest-lying dark state D in Fig. 8.3a shows a uniform envelope function in k-space, whereas the bright states A ($\Omega_S = 1.50$ eV) and B^+ ($\Omega_S = 1.82$ eV) in Figs. 8.3b and 8.3c, respectively, have an envelope function localized around the Γ point with s characters. From Fig. 8.3d, an interesting hexagonal petal pattern with a node at the Γ point can be found for exciton B^- ($\Omega_S = 1.92$ eV). In Figs. 8.3e–h, we plot the distribution of the constituent free electron-hole pairs specified by (E_v, E_c) for selected exciton states, weighted by the module squared exciton envelope function for each specific interband transition. It is obvious that the electron-hole composition of exciton D is distinct from those of the bright states (A and B).

Because of broken TRS and strong SOC effect, the electron (hole) states that compose a given exciton in this system are from Bloch wavefunctions with spin polarization along different directions, giving rise to rich excitonic spin configurations. In fact, the lowest-lying bound exciton states are all formed by Kohn-Sham orbitals with particular spin polarization. Our calculations verify that the dark excitons D are dominated by (>99.5%) transitions between the major-spin valence bands and minor-spin conduction bands. The bright states (forming peaks A, B and C) in Figs. 8.2f–k and Figs. 8.3b–d, f–h, however, are all dominated by (>96%) transitions between the major-spin valence bands and major-spin conduction bands [4]. Ligand field theory can provide a qualitative understanding of the lowest-lying D and A exciton states of which the optical transitions mainly occur among the localized Cr d orbitals [12, 250]. However, ligand field theory is insufficient to evaluate the oscillator strength of the excitons quantitatively. In addition, the coexistence of Frenkel and Wannier excitons in our system poses significant challenges to ligand field theory, while this excitonic physics can be fully captured by the first-principles GW-BSE method.

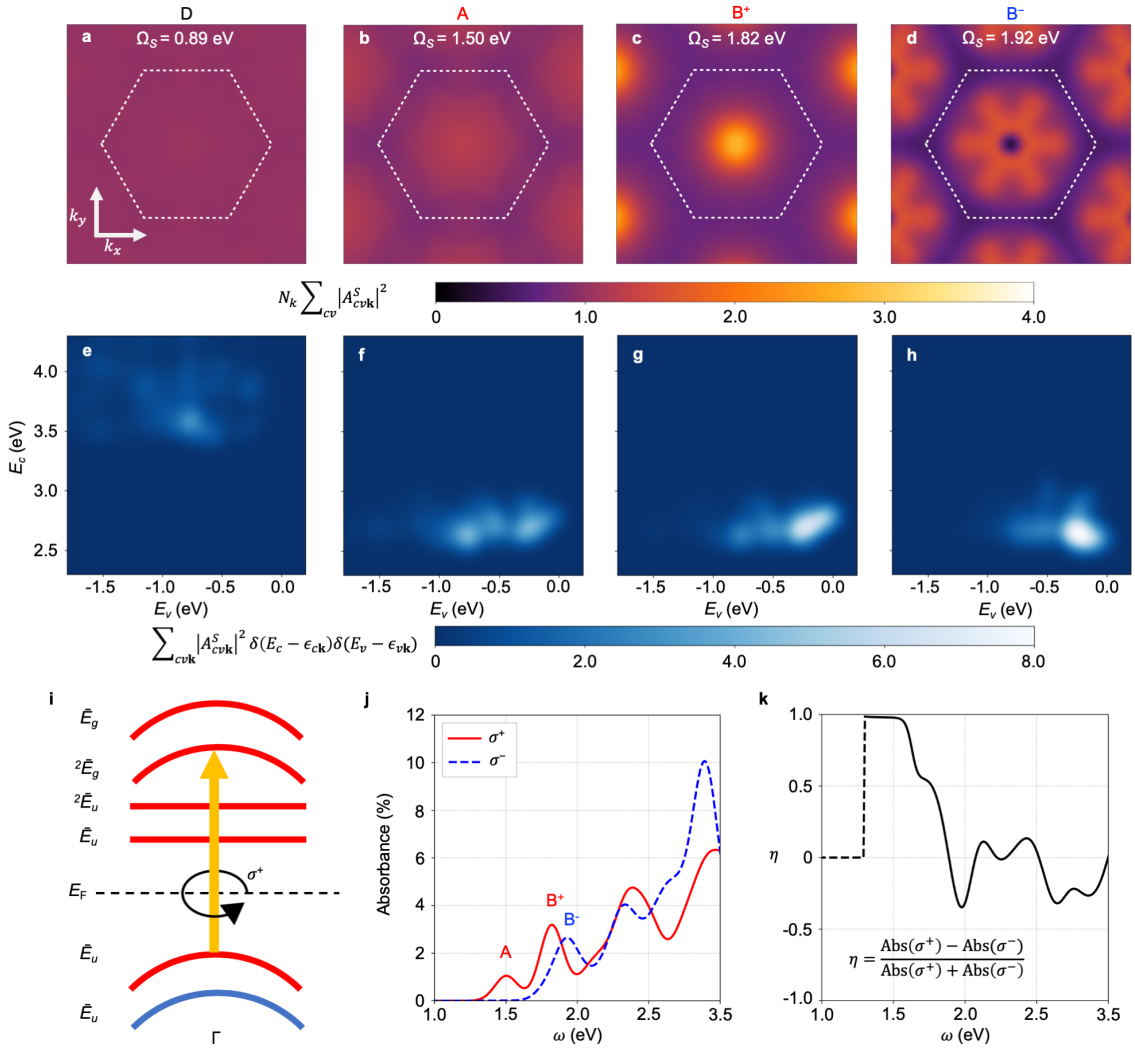

Figure 8.3: (a–d) Exciton envelope functions in k-space of excitons (a) D, (b) A, (c) B^+, and (d) B^-. The white dotted-line hexagon denotes the first Brillouin zone. The amplitudes are summed over band-pairs as given by $N_k \sum_{cv} |A^S_{cv\boldsymbol{k}}|^2$. (e–h) The distribution of free electron-hole pair with electron energy at E_c and hole energy at E_v for excitons (e) D, (f) A, (g) B^+, and (h) B^-, weighted by module squared exciton envelope function for each interband transition between states $|v\boldsymbol{k}\rangle$ and $|c\boldsymbol{k}\rangle$. All the band energies are measured with respect to the VBM energy. An 80 meV Gaussian broadening is used. (i) Schematic of interband transitions around the Γ point. Irreducible representations for states at Γ and the Fermi level are labeled. Only the indicated σ^+ transition is allowed. (j) Frequency-dependent circularly polarized absorbance of monolayer CrI_3 at normal incidence. The red solid (blue dashed) curve corresponds to the σ^+ (σ^-) circularly polarized light. (k) MCD of absorbance (η) as a function of the photon frequency. η is set to zero below 1.3 eV as shown by the dashed line.

8.2.3 Magneto-optical effects from first principles

The above-mentioned internal structures of the exciton states are essential for a deeper understanding of the MO responses. Note that all the irreducible representations of the double group S_6^D are one-dimensional, which facilitates our analysis of optical selection rules for circularly polarized lights around the Γ point, as shown in Fig. 8.3i. For 1s-like bright states A and B^+ wherein the transition mainly happens between the topmost valence band and the major-spin e_g manifold near the Γ point, only one σ^+ circularly polarized transition from the first valence band and the third conduction band is allowed among all the transitions. This conclusion is further confirmed by our first-principles circularly polarized absorption, as shown in Fig. 8.3j. The 2s-like exciton B^-, unlike A and B^+, is dominated by σ^- circularly polarized transitions. We quantify the MCD of absorbance by calculating the contrast, $\eta = (\mathsf{Abs}(\sigma^+) - \mathsf{Abs}(\sigma^-))/(\mathsf{Abs}(\sigma^+) + \mathsf{Abs}(\sigma^-))$, where $\mathsf{Abs}(\sigma^\pm)$ denotes the absorbance of σ^+ and σ^- circularly polarized lights, respectively. η is dominated by σ^+ circularly polarized light below 1.8 eV (Fig. 8.3k). If we flip the magnetization direction, η will also flip sign at all frequencies, which agrees with the measured MCD of photoluminescence signals [12].

In the following, we investigate the MO Kerr and Faraday effects of ferromagnetic monolayer CrI_3. Previous studies have shown that both SOC and the exchange splitting should be present to ensure non-zero MO effects in ferromagnets [118, 124, 127, 125, 126], and recent calculations within an independent-particle picture using DFT have been carried out for the MO responses of monolayer CrI_3 [251]. The essence of a theoretical modeling of the MO effects lies in accurately accounting for the diagonal and off-diagonal frequency-dependent macroscopic dielectric functions, which are readily available from our GW-BSE calculations with electron-hole interaction included. We find that the above-discussed giant excitonic effects in ferromagnetic monolayer CrI_3 strongly modify its MO responses, leading to significantly different behaviors going beyond those from a treatment considering only transitions between noninteracting Kohn-Sham orbitals [251]. Here we shall only consider the most physically relevant experiments for 2D ferromagnets, namely, polar MOKE (P-MOKE) and polar FE (P-FE), where both the sample magnetization and the wave vectors of light are along the normal of the surface. In accordance to typical, realistic experimental setup, we consider a device of ferromagnetic monolayer CrI_3 on top of a SiO_2/Si substrate (the thickness of SiO_2 layer is set to 285 nm, and Si is treated as semi-infinitely thick) [73], as shown in Fig. 8.4a. For insulating SiO_2 with a large band gap (8.9 eV), we use its dielectric constant $\varepsilon_{\mathsf{SiO_2}} = 3.9$ [252]. For silicon, we perform first-principles GW (at the G_0W_0 level) and GW-BSE calculations, and incorporate the frequency-dependence of the complex dielectric function [4]. Assuming a linearly polarized incident light, we calculate the Kerr (Faraday) signals by analyzing the polarization plane of the reflected (transmitted) light, which is in general elliptically polarized with a rotation angle θ_K (θ_F) and an ellipticity χ_K (χ_F), as discussed in Sec. 3.3.

We find that the MO signals are very sensitive to the thickness of SiO_2 and to the photon frequency. As shown in Figs. 8.4c, d, the thickness of SiO_2 layer will strongly affect the

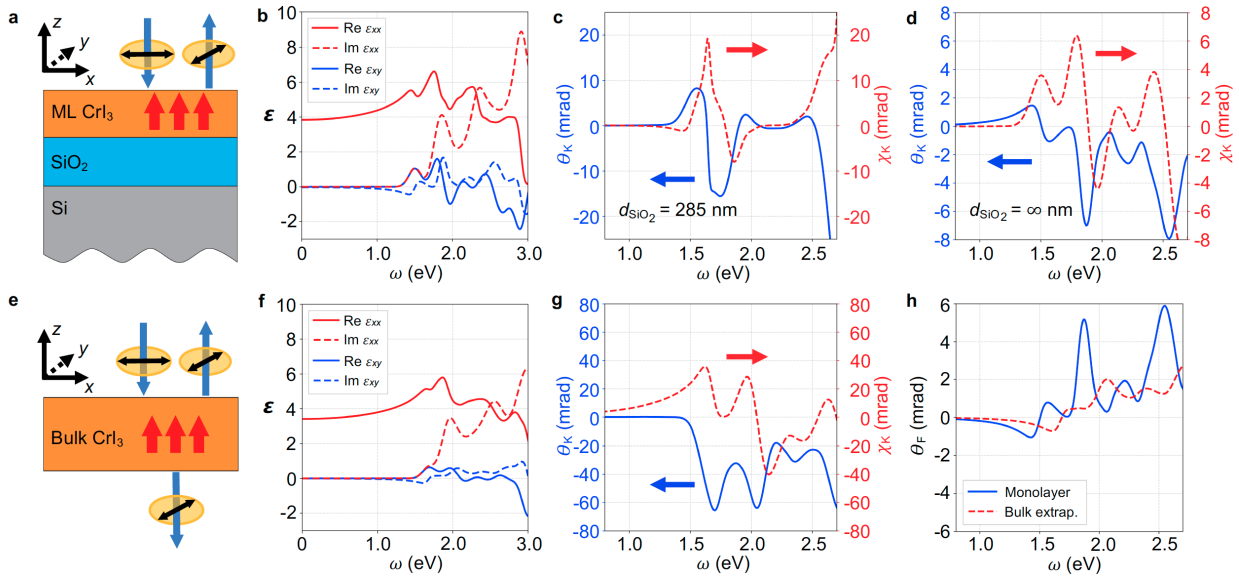

Figure 8.4: MO signals calculated from first-principles dielectric functions. (a) P-MOKE setup consisting of layers of vacuum, ferromagnetic monolayer CrI_3, SiO_2 film, and semi-infinitely thick Si. Red arrows denote the out-of-plane magnetization, which is pointing along the $+z$ direction. Blue arrows denote the propagation direction of light, and black double-headed arrows give the corresponding linear polarization direction. Each orange ellipse denotes a polarization plane of the electric field of light. (b) Calculated real part (solid lines) and imaginary part (dashed lines) of both the diagonal ε_{xx} (red) and off-diagonal ε_{xy} (blue) dielectric functions of ferromagnetic monolayer CrI_3, using a monolayer thickness $d = 6.6$ Å. (c) Kerr angle θ_K (left, blue solid) and Kerr ellipticity χ_K (right, red dashed) for the P-MOKE setup with a 285 nm SiO_2 layer. (d) Kerr angle θ_K (left, blue solid) and Kerr ellipticity χ_K (right, red dashed) for the P-MOKE setup in (a) with semi-infinitely thick SiO_2 layer. (e) P-MOKE and P-FE setup of a suspended ferromagnetic bulk CrI_3 layer with the directions of light propagation and magnetization similar to (a). (f) Calculated real part (solid lines) and imaginary part (dashed lines) of both the diagonal ε_{xx} (red) and off-diagonal ε_{xy} (blue) dielectric functions of ferromagnetic bulk CrI_3. (g) Kerr angle θ_K (left, blue solid) and Kerr ellipticity χ_K (right, red dashed) for the setup in (e) with infinitely thick ferromagnetic bulk CrI_3. (h) Comparison between Faraday angle θ_F of a suspended ferromagnetic monolayer CrI_3 and extrapolated bulk value down to the monolayer thickness (6.6 Å). An 80 meV energy broadening is applied.

MO signals, due to the interference of reflected lights from multiple interfaces [73]. Such interference has been accounted for with our three-interface setup in Fig. 8.4a. To analyze the relation between MO signals and dielectric functions, we also consider a simpler two-interface setup. For a two-interface setup with semi-infinitely thick SiO_2 layer, the Kerr angle θ_K (Fig. 8.4d, blue solid curve) is related to $\mathsf{Im}[\varepsilon_{xy}]$ (Fig. 8.4b, blue dashed curve) and therefore resonant with the exciton excitation energies; the Kerr ellipticity χ_K (Fig. 8.4d, red dashed curve), on the other hand, is proportional to $\mathsf{Re}[\varepsilon_{xy}]$ (Fig. 8.4b, blue solid curve). For a two-interface model, θ_K is also found to be proportional to $n_0/(n_2^2 - n_0^2)$, where n_0 (n_2) is the refractive index for the upper (lower) semi-infinitely thick medium. Moreover, θ_K and χ_K are connected through an approximate Kramers-Krönig relation, as expected from previous works [118, 253]. Because of this, close attention should be paid in interpreting MOKE experiments on 2D ferromagnets, where the substrate configuration significantly changes the behavior of the MOKE signals. The existing experimental data of θ_K, however, only have a few excitation frequencies of photons available, e.g., 5 ± 2 mrad at 1.96 eV for HeNe laser [73]. As shown in Fig. 8.4c, our simulations with a 285 nm SiO_2 layer in the three-interface setup achieve the same order of magnitude for θ_K around the MO resonance at $\sim$ 1.85 eV, in good agreement with experiment. Based on the simulations, we also predict a sign change of θ_K around 1.5 eV. For photon energies higher than the quasiparticle band gap, the plasmon resonance along with a vanishing ε_{xx} will nullify our assumption of continuous waves [253, 254]. It is also possible to achieve an in-plane ferromagnetic structure with an external magnetic field [255, 256]. However, due to the broken C_3 symmetry therein, we expect the system to have diminished values of MO signals (in the same polar configurations) but to remain having excitons with large binding energies, as confirmed by our first-principles calculations [4].

8.2.4 Effects of quantum confinement

To further understand the effects of quantum confinement in 2D magnets, we compare the MO properties of ferromagnetic bulk and monolayer CrI_3. Interestingly, the calculated optical properties of bulk CrI_3 are also dominated by strongly bound excitons with optical absorption edge starting from 1.5 eV (in good agreement with experiment [12]), while the quasiparticle indirect band gap is 1.89 eV and the direct band gap at the Γ point is 2.13 eV [4]. Within a one-interface model of semi-infinitely thick bulk CrI_3, θ_K reaches a magnitude of 60 mrad at the resonances at around 1.7 and 2.0 eV (Fig. 8.4g), proportional to $\mathsf{Re}[\varepsilon_{xy}]$ shown in Fig. 8.4f. To study the quantum confinement effects, we employ the P-FE setup shown in Fig. 8.4e, because P-FE in this setup is almost linear with respect to the ferromagnetic sample thickness and free from the substrate effects. Our calculated magnitude of the specific Faraday angle ($|\theta_\mathsf{F}|$) of bulk CrI_3 is $(1.3 \pm 0.3) \times 10^3$ rad cm^{-1} at the excitation frequency of 1.28 eV [4], in agreement with the experimental value of 1.9×10^3 rad cm^{-1} at the same excitation frequency [257]. By extrapolating the bulk θ_F to the monolayer thickness [140], and comparing with that of suspended ferromagnetic monolayer CrI_3 as shown in Fig. 8.4h,

we find that quantum confinement significantly enhances the MO responses by a factor of 2.5 near 2.0 eV and introduces a redshift of 0.2 eV.

8.3 Conclusion

In summary, from our first-principles *GW* and *GW*-BSE calculations, we discover that the optical and MO properties of ferromagnetic monolayer CrI_3 are dominated by strongly bound excitons of charge-transfer or Wannier characters. A systematic modeling framework for P-MOKE and P-FE experiments is also developed, where we have shown that the MO signals exhibit a sensitive dependence on photon frequency and substrate configuration. These findings of the exciton physics in 2D magnets should shed light on design principles for future MO and optoelectronic devices, such as photo-spin-voltaic devices [258] and spin-injecting electroluminescence [259, 260]. As a prototypical monolayer Ising magnetic insulator with a band gap in an easily accessible optical range, ferromagnetic monolayer CrI_3 is also expected to be useful in high-speed and high-density flexible MO drives using van der Waals homostructures or heterostructures [255, 261].

Bibliography

[1] Cao, T., Wu, M. & Louie, S. G. Unifying optical selection rules for excitons in two dimensions: Band topology and winding numbers. *Phys. Rev. Lett.* **120**, 087402 (2018).

[2] Li, Y. *et al.* Ultrasensitive tunability of the direct bandgap of 2D InSe flakes via strain engineering. *2D Mater.* **5**, 021002 (2018).

[3] Guo, L. *et al.* Exchange-driven intravalley mixing of excitons in monolayer transition metal dichalcogenides. *Nat. Phys.* **15**, 228–232 (2019).

[4] Wu, M., Li, Z., Cao, T. & Louie, S. G. Physical origin of giant excitonic and magneto-optical responses in two-dimensional ferromagnetic insulators. *Nat. Commun.* **10**, 2371 (2019).

[5] Nguyen, G. D. *et al.* Atomically precise graphene nanoribbon heterojunctions from a single molecular precursor. *Nat. Nanotechnol.* **12**, 1077 (2017).

[6] Rizzo, D. J. *et al.* Length-dependent evolution of type II heterojunctions in bottom-up-synthesized graphene nanoribbons. *Nano Lett.* **19**, 3221–3228 (2019).

[7] Chen, Y. *et al.* Strong correlations and orbital texture in single-layer 1T-$TaSe_2$. *Nat. Phys.* **16**, 218–224 (2020).

[8] Cohen, M. L. & Louie, S. G. *Fundamentals of Condensed Matter Physics* (Cambridge University Press, 2016).

[9] Hybertsen, M. S. & Louie, S. G. Electron correlation in semiconductors and insulators: Band gaps and quasiparticle energies. *Phys. Rev. B* **34**, 5390–5413 (1986).

[10] Cardona, M. & Thewalt, M. L. W. Isotope effects on the optical spectra of semiconductors. *Rev. Mod. Phys.* **77**, 1173–1224 (2005).

[11] Shao, M. *et al.* Low rank approximation in G_0W_0 calculations. *Sci. China Math.* **59**, 1593–1612 (2016).

[12] Seyler, K. L. *et al.* Ligand-field helical luminescence in a 2D ferromagnetic insulator. *Nat. Phys.* **14**, 277–281 (2018).

[13] Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871 (1964).

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

[15] Sólyom, J. *Fundamentals of the Physics of Solids: Volume 1: Structure and Dynamics* (Springer Science & Business Media, 2007).

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

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

[18] Perdew, J. P. & Zunger, A. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B* **23**, 5048–5079 (1981).

[19] Perdew, J. P. & Wang, Y. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B* **45**, 13244–13249 (1992).

[20] Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).

[21] Csonka, G. I. *et al.* Assessing the performance of recent density functionals for bulk solids. *Phys. Rev. B* **79**, 155107 (2009).

[22] Rajagopal, A. K. & Callaway, J. Inhomogeneous electron gas. *Phys. Rev. B* **7**, 1912–1919 (1973).

[23] von Barth, U. & Hedin, L. A local exchange-correlation potential for the spin polarized case. I. *J. Phys. C: Solid State Phys.* **5**, 1629 (1972).

[24] Vosko, S. H., Wilk, L. & Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: A critical analysis. *Can. J. Phys.* **58**, 1200–1211 (1980).

[25] Phillips, J. C. & Kleinman, L. New method for calculating wave functions in crystals and molecules. *Phys. Rev.* **116**, 287–294 (1959).

[26] Cohen, M. L. Application of the pseudopotential model to solids. *Annu. Rev. Mater. Sci.* **14**, 119–144 (1984).

[27] Martin, R. M. *Electronic Structure: Basic Theory and Practical Methods* (Cambridge University Press, 2020).

[28] Jones, R. O. Density functional theory: Its origins, rise to prominence, and future. *Rev. Mod. Phys.* **87**, 897–923 (2015).

[29] Giustino, F. *Materials Modelling Using Density Functional Theory: Properties and Predictions* (Oxford University Press, 2014).

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

[31] Louie, S. G. First-principles theory of electron excitation energies in solids, surfaces, and defects. In *Topics in Computational Materials Science*, 96–142 (World Scientific, 1998).

[32] Louie, S. G. & Hybertsen, M. S. Theory of quasiparticle energies: Band gaps and excitation spectra in solids. *Int. J. Quantum Chem.* **32**, 31–44 (1987).

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

[34] Rohlfing, M. & Louie, S. G. Electron-hole excitations and optical spectra from first principles. *Phys. Rev. B* **62**, 4927–4944 (2000).

[35] Rohlfing, M. & Louie, S. G. Electron-hole excitations in semiconductors and insulators. *Phys. Rev. Lett.* **81**, 2312–2315 (1998).

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

[37] Godby, R. W., Schlüter, M. & Sham, L. J. Accurate exchange-correlation potential for silicon and its discontinuity on addition of an electron. *Phys. Rev. Lett.* **56**, 2415–2418 (1986).

[38] Schwinger, J. On the Green's functions of quantized fields. I. *Proc. Natl. Acad. Sci. U.S.A* **37**, 452–455 (1951).

[39] Baym, G. Field-theoretic approach to the properties of the solid state. *Ann. Phys.* **14**, 1–42 (1961).

[40] Hedin, L. & Lundqvist, S. Effects of electron-electron and electron-phonon interactions on the one-electron states of solids. In *Solid State Physics*, vol. 23, 1–181 (Elsevier, 1970).

[41] Fetter, A. L. & Walecka, J. D. *Quantum Theory of Many-Particle Systems* (Courier Corporation, 2012).

[42] Mahan, G. D. *Many-Particle Physics* (Springer Science & Business Media, 2013).

[43] Nozières, P. *Theory of Interacting Fermi Systems* (Westview Press, 1997).

[44] Stefanucci, G. & van Leeuwen, R. *Nonequilibrium Many-Body Theory of Quantum Systems: A Modern Introduction* (Cambridge University Press, 2013).

[45] Strinati, G. Application of the Green's functions method to the study of the optical properties of semiconductors. *Riv. Nuovo Cimento* **11**, 1–86 (1988).

[46] Baym, G. & Kadanoff, L. P. Conservation laws and correlation functions. *Phys. Rev.* **124**, 287–299 (1961).

[47] Baym, G. Self-consistent approximations in many-body systems. *Phys. Rev.* **127**, 1391–1401 (1962).

[48] Adler, S. L. Quantum theory of the dielectric constant in real solids. *Phys. Rev.* **126**, 413–420 (1962).

[49] Wiser, N. Dielectric constant with local field effects included. *Phys. Rev.* **129**, 62–69 (1963).

[50] Green, M. S. Markoff random processes and the statistical mechanics of time-dependent phenomena. *J. Chem. Phys.* **20**, 1281–1295 (1952).

[51] Green, M. S. Markoff random processes and the statistical mechanics of time-dependent phenomena. II. Irreversible processes in fluids. *J. Chem. Phys.* **22**, 398–413 (1954).

[52] Kubo, R. Statistical-mechanical theory of irreversible processes. I. General theory and simple applications to magnetic and conduction problems. *J. Phys. Soc. Japan* **12**, 570–586 (1957).

[53] Kubo, R. The fluctuation-dissipation theorem. *Rep. Prog. Phys.* **29**, 255 (1966).

[54] Walter, J. P. & Cohen, M. L. Wave-vector-dependent dielectric function for Si, Ge, GaAs, and ZnSe. *Phys. Rev. B* **2**, 1821–1826 (1970).

[55] Louie, S. G., Chelikowsky, J. R. & Cohen, M. L. Local-field effects in the optical spectrum of silicon. *Phys. Rev. Lett.* **34**, 155–158 (1975).

[56] Cappellini, G., Del Sole, R., Reining, L. & Bechstedt, F. Model dielectric function for semiconductors. *Phys. Rev. B* **47**, 9892–9895 (1993).

[57] Levine, Z. H. & Louie, S. G. New model dielectric function and exchange-correlation potential for semiconductors and insulators. *Phys. Rev. B* **25**, 6310–6316 (1982).

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

[59] Salpeter, E. E. & Bethe, H. A. A relativistic equation for bound-state problems. *Phys. Rev.* **84**, 1232–1242 (1951).

[60] Sio, W. H., Verdi, C., Poncé, S. & Giustino, F. *Ab initio* theory of polarons: Formalism and applications. *Phys. Rev. B* **99**, 235139 (2019).

[61] Louie, S. G. Predicting materials and properties: Theory of the ground and excited state. *Contemp. Concepts Condens. Matter Sci.* **2**, 9–53 (2006).

[62] Qiu, D. Y., da Jornada, F. H. & Louie, S. G. Screening and many-body effects in two-dimensional crystals: Monolayer MoS_2. *Phys. Rev. B* **93**, 235435 (2016).

[63] Onida, G., Reining, L. & Rubio, A. Electronic excitations: Density-functional versus many-body Green's-function approaches. *Rev. Mod. Phys.* **74**, 601–659 (2002).

[64] Novoselov, K. S. *et al.* Electric field effect in atomically thin carbon films. *Science* **306**, 666–669 (2004).

[65] Novoselov, K. S. *et al.* Two-dimensional atomic crystals. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 10451–10453 (2005).

[66] Radisavljevic, B., Radenovic, A., Brivio, J., Giacometti, V. & Kis, A. Single-layer MoS_2 transistors. *Nat. Nanotechnol.* **6**, 147–150 (2011).

[67] Mak, K. F., Lee, C., Hone, J., Shan, J. & Heinz, T. F. Atomically thin MoS_2: A new direct-gap semiconductor. *Phys. Rev. Lett.* **105**, 136805 (2010).

[68] Li, L. *et al.* Black phosphorus field-effect transistors. *Nat. Nanotechnol.* **9**, 372 (2014).

[69] Bandurin, D. A. *et al.* High electron mobility, quantum Hall effect and anomalous optical response in atomically thin InSe. *Nat. Nanotechnol.* **12**, 223–227 (2017).

[70] Lee, G.-H. *et al.* Electron tunneling through atomically flat and ultrathin hexagonal boron nitride. *Appl. Phys. Lett.* **99**, 243114 (2011).

[71] Britnell, L. *et al.* Electron tunneling through ultrathin boron nitride crystalline barriers. *Nano Lett.* **12**, 1707–1710 (2012).

[72] Gong, C. *et al.* Discovery of intrinsic ferromagnetism in two-dimensional van der Waals crystals. *Nature* **546**, 265–269 (2017).

[73] Huang, B. *et al.* Layer-dependent ferromagnetism in a van der Waals crystal down to the monolayer limit. *Nature* **546**, 270–273 (2017).

[74] Deng, Y. *et al.* Gate-tunable room-temperature ferromagnetism in two-dimensional Fe_3GeTe_2. *Nature* **563**, 94–99 (2018).

[75] Qian, X., Liu, J., Fu, L. & Li, J. Quantum spin Hall effect in two-dimensional transition metal dichalcogenides. *Science* **346**, 1344–1347 (2014).

[76] Tang, S. *et al.* Quantum spin Hall state in monolayer 1T'-WTe_2. *Nat. Phys.* **13**, 683–687 (2017).

[77] Fei, Z. *et al.* Edge conduction in monolayer WTe_2. *Nat. Phys.* **13**, 677–682 (2017).

[78] Cao, Y. *et al.* Unconventional superconductivity in magic-angle graphene superlattices. *Nature* **556**, 43–50 (2018).

[79] Lu, J. *et al.* Evidence for two-dimensional Ising superconductivity in gated MoS_2. *Science* **350**, 1353–1357 (2015).

[80] Yu, Y. *et al.* High-temperature superconductivity in monolayer $Bi_2Sr_2CaCu_2O_{8+\delta}$. *Nature* **575**, 156–163 (2019).

[81] Cao, Y. *et al.* Correlated insulator behaviour at half-filling in magic-angle graphene superlattices. *Nature* **556**, 80–84 (2018).

[82] Bhimanapati, G. R. *et al.* Recent advances in two-dimensional materials beyond graphene. *ACS nano* **9**, 11509–11539 (2015).

[83] Mounet, N. *et al.* Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nat. Nanotechnol.* **13**, 246–252 (2018).

[84] Geim, A. K. & Grigorieva, I. V. Van der Waals heterostructures. *Nature* **499**, 419–425 (2013).

[85] Qiu, D. Y., da Jornada, F. H. & Louie, S. G. Optical spectrum of MoS_2: Many-body effects and diversity of exciton states. *Phys. Rev. Lett.* **111**, 216805 (2013).

[86] Qiu, D. Y., Cao, T. & Louie, S. G. Nonanalyticity, valley quantum phases, and lightlike exciton dispersion in monolayer transition metal dichalcogenides: Theory and first-principles calculations. *Phys. Rev. Lett.* **115**, 176801 (2015).

[87] Ye, Z. *et al.* Probing excitonic dark states in single-layer tungsten disulphide. *Nature* **513**, 214–218 (2014).

[88] Felipe, H., Xian, L., Rubio, A. & Louie, S. G. Universal slow plasmons and giant field enhancement in atomically thin quasi-two-dimensional metals. *Nat. Commun.* **11**, 1–10 (2020).

[89] Sakurai, J. J. *Advanced Quantum Mechanics* (Pearson Education India, 1967).

[90] Dresselhaus, G. Spin-orbit coupling effects in zinc blende structures. *Phys. Rev.* **100**, 580 (1955).

[91] Rashba, É. I. & Sheka, V. I. Symmetry of energy bands in crystals of wurtzite type II. Symmetry of bands with spin-orbit interaction included. *Fiz. Tverd. Tela: Collected Papers* **2**, 162–176 (1959).

[92] Bychkov, Y. A. & Rashba, É. I. Properties of a 2D electron gas with lifted spectral degeneracy. *JETP Lett.* **39**, 78 (1984).

[93] Bihlmayer, G., Rader, O. & Winkler, R. Focus on the Rashba effect. *New J. Phys.* **17**, 050202 (2015).

[94] Xiao, D., Liu, G.-B., Feng, W., Xu, X. & Yao, W. Coupled spin and valley physics in monolayers of MoS_2 and other group-VI dichalcogenides. *Phys. Rev. Lett.* **108**, 196802 (2012).

[95] Cao, T. *et al.* Valley-selective circular dichroism of monolayer molybdenum disulphide. *Nat. Commun.* **3**, 1–5 (2012).

[96] Haldane, F. D. M. Model for a quantum Hall effect without Landau levels: Condensed-matter realization of the "parity anomaly". *Phys. Rev. Lett.* **61**, 2015 (1988).

[97] Bernevig, B. A. & Zhang, S.-C. Quantum spin Hall effect. *Phys. Rev. Lett.* **96**, 106802 (2006).

[98] Bernevig, B. A., Hughes, T. L. & Zhang, S.-C. Quantum spin Hall effect and topological phase transition in HgTe quantum wells. *Science* **314**, 1757–1761 (2006).

[99] Kane, C. L. & Mele, E. J. Quantum spin Hall effect in graphene. *Phys. Rev. Lett.* **95**, 226801 (2005).

[100] Giannozzi, P. Notes on pseudopotential generation (2010). URL http://www.fisica.uniud.it/~giannozz.

[101] Shao, M. *et al.* A structure preserving Lanczos algorithm for computing the optical absorption spectrum. *SIAM J. Matrix Anal. Appl.* **39**, 683–711 (2018).

[102] Grüning, M., Marini, A. & Gonze, X. Exciton-plasmon states in nanoscale materials: Breakdown of the Tamm-Dancoff approximation. *Nano Lett.* **9**, 2820–2824 (2009).

[103] Kleinman, L. & Bylander, D. M. Efficacious form for model pseudopotentials. *Phys. Rev. Lett.* **48**, 1425–1428 (1982).

[104] Hamann, D. R. Optimized norm-conserving Vanderbilt pseudopotentials. *Phys. Rev. B* **88**, 085117 (2013).

[105] Schlipf, M. & Gygi, F. Optimization algorithm for the generation of ONCV pseudopotentials. *Comput. Phys. Commun.* **196**, 36–44 (2015).

[106] Giannozzi, P. *et al.* Advanced capabilities for materials modelling with Quantum ESPRESSO. *J. Phys. Condens. Matter* **29**, 465901 (2017).

[107] Giannozzi, P. *et al.* QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21**, 395502 (2009).

[108] Deslippe, J. *et al.* BerkeleyGW: A massively parallel computer package for the calculation of the quasiparticle and optical properties of materials and nanostructures. *Comput. Phys. Commun.* **183**, 1269–1289 (2012).

[109] Sze, S. M. & Ng, K. K. *Physics of Semiconductor Devices* (John Wiley & Sons, 2006).

[110] Zwerdling, S., Button, K. J., Lax, B. & Roth, L. M. Internal impurity levels in semiconductors: Experiments in p-type silicon. *Phys. Rev. Lett.* **4**, 173–176 (1960).

[111] Yu, Z., Huang, Y. X. & Shen, S. C. Spin-orbit splitting of the valence bands in silicon determined by means of high-resolution photoconductive spectroscopy. *Phys. Rev. B* **39**, 6287–6289 (1989).

[112] Rome, J. J. *et al.* Additional $p_{\frac{3}{2}}$ and $p_{\frac{1}{2}}$ infrared excited-state lines of gallium and indium in silicon. *Phys. Rev. B* **25**, 3615–3618 (1982).

[113] Fischer, D. W. & Rome, J. J. Additional structure in infrared excitation spectra of group-III acceptors in silicon. *Phys. Rev. B* **27**, 4826–4832 (1983).

[114] Faraday, M. Experimental researches in electricity. *Proc. R. Soc.* **5** (1845).

[115] Kerr, J. On rotation of the plane of polarisation by reflection from a magnetic pole. *Rep. Brit. Assoc. Adv. Sci.* **45** (1876).

[116] Kerr, J. On rotation of the plane of polarisation by reflection from the pole of a magnet. *Phil. Mag.* **3** (1877).

[117] Voigt, W. zur Theorie der magneto-optischen Erscheinungen. *Ann. Phys. u. Chem.* **67** (1899).

[118] Oppeneer, P. M. Magneto-optical Kerr spectra. *Handb. Magn. Mater.* **13**, 229–422 (2001).

[119] Mansuripur, M. *The Physical Principles of Magneto-Optical Recording* (Cambridge University Press, 1998).

[120] Liu, C., Moog, E. R. & Bader, S. D. Polar Kerr-effect observation of perpendicular surface anisotropy for ultrathin fcc Fe grown on Cu (100). *Phys. Rev. Lett.* **60**, 2422 (1988).

[121] Bader, S. D. SMOKE. *J. Magn. Magn. Mater.* **100**, 440–454 (1991).

[122] Schäfer, R. & Hubert, A. A new magnetooptic effect related to non-uniform magnetization on the surface of a ferromagnet. *Phys. Status Solidi A* **118**, 271–288 (1990).

[123] Hubert, A. & Schäfer, R. *Magnetic Domains: The Analysis of Magnetic Microstructures* (Springer Science & Business Media, 2008).

[124] Argyres, P. N. Theory of the Faraday and Kerr effects in ferromagnetics. *Phys. Rev.* **97**, 334 (1955).

[125] Misemer, D. K. The effect of spin-orbit interaction and exchange splitting on magneto-optic coefficients. *J. Magn. Magn. Mater.* **72**, 267–274 (1988).

[126] Oppeneer, P. M., Maurer, T., Sticht, J. & Kübler, J. *Ab initio* calculated magneto-optical Kerr effect of ferromagnetic metals: Fe and Ni. *Phys. Rev. B* **45**, 10924 (1992).

[127] Erskine, J. L. & Stern, E. A. Magneto-optic Kerr effect in Ni, Co, and Fe. *Phys. Rev. Lett.* **30**, 1329 (1973).

[128] Ebert, H. & Schütz, G. *Spin-Orbit-Influenced Spectroscopies of Magnetic Solids* (Springer, 1996).

[129] Ebert, H. Magneto-optical effects in transition metal systems. *Rep. Prog. Phys.* **59**, 1665 (1996).

[130] Guo, G.-Y. & Ebert, H. On the origins of the enhanced magneto-optical Kerr effect in ultrathin Fe and Co multilayers. *J. Magn. Magn. Mater.* **156**, 173–174 (1996).

[131] Guo, G.-Y. & Ebert, H. Band-theoretical investigation of the magneto-optical Kerr effect in Fe and Co multilayers. *Phys. Rev. B* **51**, 12633–12643 (1995).

[132] Feng, W., Guo, G.-Y., Zhou, J., Yao, Y. & Niu, Q. Large magneto-optical Kerr effect in noncollinear antiferromagnets Mn_3X (X=Rh, Ir, Pt). *Phys. Rev. B* **92**, 144426 (2015).

[133] Oppeneer, P. M. *et al.* First-principles study of the giant magneto-optical Kerr effect in MnBi and related compounds. *J. Appl. Phys.* **80**, 1099–1105 (1996).

[134] Yao, Y. *et al.* First-principles calculation of anomalous Hall conductivity in ferromagnetic bcc Fe. *Phys. Rev. Lett.* **92**, 037204 (2004).

[135] Wang, X., Antropov, V. P. & Harmon, B. N. First principles study of magneto-optical properties of half metallic Heusler alloys: NiMnSb and PtMnSb. *IEEE Trans. Magn.* **30**, 4458–4460 (1994).

[136] Onsager, L. Reciprocal relations in irreversible processes. I. *Phys. Rev.* **37**, 405 (1931).

[137] Hecht, E. *Optics* (Addison-Wesley, 2001).

[138] Collett, E. *Field Guide to Polarization*, vol. 15 (SPIE Press, 2005).

[139] Byrnes, S. J. Multilayer optical calculations. *arXiv:1603.02720* (2016).

[140] McGuire, M. A., Dixit, H., Cooper, V. R. & Sales, B. C. Coupling of crystal structure and magnetism in the layered, ferromagnetic insulator CrI_3. *Chem. Mater.* **27**, 612–620 (2015).

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

[142] García-González, P. & Godby, R. W. Self-consistent calculation of total energies of the electron gas using many-body perturbation theory. *Phys. Rev. B* **63**, 075112 (2001).

[143] Schöne, W.-D. & Eguiluz, A. G. Self-consistent calculations of quasiparticle states in metals and semiconductors. *Phys. Rev. Lett.* **81**, 1662 (1998).

[144] Golze, D., Dvorak, M. & Rinke, P. The *GW* compendium: A practical guide to theoretical photoemission spectroscopy. *Front. Chem.* **7**, 377 (2019).

[145] Bruneval, F. & Marques, M. A. L. Benchmarking the starting points of the *GW* approximation for molecules. *J. Chem. Theory Comput.* **9**, 324–329 (2013).

[146] Rinke, P., Qteish, A., Neugebauer, J., Freysoldt, C. & Scheffler, M. Combining *GW* calculations with exact-exchange density-functional theory: An analysis of valence-band photoemission for compound semiconductors. *New J. Phys.* **7**, 126 (2005).

[147] Fuchs, F., Furthmüller, J., Bechstedt, F., Shishkin, M. & Kresse, G. Quasiparticle band structure based on a generalized Kohn-Sham scheme. *Phys. Rev. B* **76**, 115109 (2007).

[148] Caruso, F., Dauth, M., van Setten, M. J. & Rinke, P. Benchmark of *GW* approaches for the *GW*100 test set. *J. Chem. Theory Comput.* **12**, 5076–5087 (2016).

[149] Shih, B.-C., Xue, Y., Zhang, P., Cohen, M. L. & Louie, S. G. Quasiparticle band gap of ZnO: High accuracy from the conventional G_0W_0 approach. *Phys. Rev. Lett.* **105**, 146401 (2010).

[150] Jiang, H., Gomez-Abal, R. I., Rinke, P. & Scheffler, M. First-principles modeling of localized *d* states with the *GW*@LDA+*U* approach. *Phys. Rev. B* **82**, 045108 (2010).

[151] Shishkin, M. & Kresse, G. Self-consistent *GW* calculations for semiconductors and insulators. *Phys. Rev. B* **75**, 235102 (2007).

[152] Lischner, J., Sharifzadeh, S., Deslippe, J., Neaton, J. B. & Louie, S. G. Effects of self-consistency and plasmon-pole models on *GW* calculations for closed-shell molecules. *Phys. Rev. B* **90**, 115130 (2014).

[153] Kotani, T., van Schilfgaarde, M., Faleev, S. V. & Chantis, A. Quasiparticle self-consistent *GW* method: A short summary. *J. Phys.: Condens. Matter* **19**, 365236 (2007).

[154] Faleev, S. V., van Schilfgaarde, M. & Kotani, T. All-electron self-consistent *GW* approximation: Application to Si, MnO, and NiO. *Phys. Rev. Lett.* **93**, 126406 (2004).

[155] van Schilfgaarde, M., Kotani, T. & Faleev, S. Quasiparticle self-consistent *GW* theory. *Phys. Rev. Lett.* **96**, 226402 (2006).

[156] Cao, H., Yu, Z., Lu, P. & Wang, L.-W. Fully converged plane-wave-based self-consistent *GW* calculations of periodic solids. *Phys. Rev. B* **95**, 035139 (2017).

[157] Delaney, K., García-González, P., Rubio, A., Rinke, P. & Godby, R. W. Comment on "Band-gap problem in semiconductors revisited: Effects of core states and many-body self-consistency". *Phys. Rev. Lett.* **93**, 249701 (2004).

[158] Del Sole, R., Reining, L. & Godby, R. W. $GW\Gamma$ approximation for electron self-energies in semiconductors and insulators. *Phys. Rev. B* **49**, 8024–8028 (1994).

[159] Reining, L., Olevano, V., Rubio, A. & Onida, G. Excitonic effects in solids described by time-dependent density-functional theory. *Phys. Rev. Lett.* **88**, 066404 (2002).

[160] Del Sole, R., Adragna, G., Olevano, V. & Reining, L. Long-range behavior and frequency dependence of exchange-correlation kernels in solids. *Phys. Rev. B* **67**, 045207 (2003).

[161] Bruneval, F., Sottile, F., Olevano, V., Del Sole, R. & Reining, L. Many-body perturbation theory using the density-functional concept: Beyond the *GW* approximation. *Phys. Rev. Lett.* **94**, 186402 (2005).

[162] Marini, A. & Rubio, A. Electron linewidths of wide-gap insulators: Excitonic effects in LiF. *Phys. Rev. B* **70**, 081103 (2004).

[163] Shishkin, M., Marsman, M. & Kresse, G. Accurate quasiparticle spectra from self-consistent *GW* calculations with vertex corrections. *Phys. Rev. Lett.* **99**, 246403 (2007).

[164] Grüneis, A., Kresse, G., Hinuma, Y. & Oba, F. Ionization potentials of solids: The importance of vertex corrections. *Phys. Rev. Lett.* **112**, 096401 (2014).

[165] Kosterlitz, J. M. & Thouless, D. J. Ordering, metastability and phase transitions in two-dimensional systems. *J. Phys. C: Solid State Phys.* **6**, 1181 (1973).

[166] Mermin, N. D. The topological theory of defects in ordered media. *Rev. Mod. Phys.* **51**, 591–648 (1979).

[167] Thouless, D. J., Kohmoto, M., Nightingale, M. P. & den Nijs, M. Quantized Hall conductance in a two-dimensional periodic potential. *Phys. Rev. Lett.* **49**, 405–408 (1982).

[168] Qi, X.-L. & Zhang, S.-C. Topological insulators and superconductors. *Rev. Mod. Phys.* **83**, 1057–1110 (2011).

[169] Hasan, M. Z. & Kane, C. L. Colloquium: Topological insulators. *Rev. Mod. Phys.* **82**, 3045–3067 (2010).

[170] Cao, T., Zhao, F. & Louie, S. G. Topological phases in graphene nanoribbons: Junction states, spin centers, and quantum spin chains. *Phys. Rev. Lett.* **119**, 076401 (2017).

[171] Dresselhaus, G. Effective mass approximation for excitons. *J. Phys. Chem. Sol.* **1**, 14–22 (1956).

[172] Elliott, R. J. Intensity of optical absorption by excitons. *Phys. Rev.* **108**, 1384 (1957).

[173] He, K. *et al.* Tightly bound excitons in monolayer WSe_2. *Phys. Rev. Lett.* **113**, 026803 (2014).

[174] Chernikov, A. *et al.* Exciton binding energy and nonhydrogenic Rydberg series in monolayer WS_2. *Phys. Rev. Lett.* **113**, 076802 (2014).

[175] Gross, E. F. Optical spectrum of excitons in the crystal lattice. *Il Nuovo Cimento (1955-1965)* **3**, 672–701 (1956).

[176] McCann, E. & Koshino, M. The electronic properties of bilayer graphene. *Rep. Prog. Phys.* **76**, 056503 (2013).

[177] Sui, M. *et al.* Gate-tunable topological valley transport in bilayer graphene. *Nat. Phys.* **11**, 1027–1031 (2015).

[178] Ju, L. *et al.* Topological valley transport at bilayer graphene domain walls. *Nature* **520**, 650–655 (2015).

[179] Zhang, F., MacDonald, A. H. & Mele, E. J. Valley Chern numbers and boundary modes in gapped bilayer graphene. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 10546–10551 (2013).

[180] Zhou, J., Shan, W.-Y., Yao, W. & Xiao, D. Berry phase modification to the energy spectrum of excitons. *Phys. Rev. Lett.* **115**, 166803 (2015).

[181] Srivastava, A. & Imamoğlu, A. Signatures of Bloch-band geometry on excitons: Nonhydrogenic spectra in transition-metal dichalcogenides. *Phys. Rev. Lett.* **115**, 166802 (2015).

[182] Bao, W. *et al.* Stacking-dependent band gap and quantum transport in trilayer graphene. *Nat. Phys.* **7**, 948–952 (2011).

[183] Lui, C. H., Li, Z., Mak, K. F., Cappelluti, E. & Heinz, T. F. Observation of an electrically tunable band gap in trilayer graphene. *Nat. Phys.* **7**, 944–947 (2011).

[184] Zhang, Y. *et al.* Direct observation of a widely tunable bandgap in bilayer graphene. *Nature* **459**, 820–823 (2009).

[185] Gorbachev, R. V. *et al.* Detecting topological currents in graphene superlattices. *Science* **346**, 448–451 (2014).

[186] Park, C.-H. & Louie, S. G. Tunable excitons in biased bilayer graphene. *Nano Lett.* **10**, 426–431 (2010).

[187] Mak, K. F., He, K., Shan, J. & Heinz, T. F. Control of valley polarization in monolayer MoS_2 by optical helicity. *Nat. Nanotechnol.* **7**, 494–498 (2012).

[188] Zeng, H., Dai, J., Yao, W., Xiao, D. & Cui, X. Valley polarization in MoS_2 monolayers by optical pumping. *Nat. Nanotechnol.* **7**, 490–493 (2012).

[189] Akinwande, D., Petrone, N. & Hone, J. Two-dimensional flexible nanoelectronics. *Nat. Commun.* **5**, 1–12 (2014).

[190] Lee, C., Wei, X., Kysar, J. W. & Hone, J. Measurement of the elastic properties and intrinsic strength of monolayer graphene. *Science* **321**, 385–388 (2008).

[191] Mohiuddin, T. M. G. *et al.* Uniaxial strain in graphene by Raman spectroscopy: G peak splitting, Grüneisen parameters, and sample orientation. *Phys. Rev. B* **79**, 205433 (2009).

[192] Conley, H. J. *et al.* Bandgap engineering of strained monolayer and bilayer MoS_2. *Nano Lett.* **13**, 3626–3630 (2013).

[193] He, K., Poole, C., Mak, K. F. & Shan, J. Experimental demonstration of continuous electronic structure tuning via strain in atomically thin MoS_2. *Nano Lett.* **13**, 2931–2936 (2013).

[194] Desai, S. B. *et al.* Strain-induced indirect to direct bandgap transition in multilayer WSe_2. *Nano Lett.* **14**, 4592–4597 (2014).

[195] Yagmurcukardes, M., Senger, R. T., Peeters, F. M. & Sahin, H. Mechanical properties of monolayer GaS and GaSe crystals. *Phys. Rev. B* **94**, 245407 (2016).

[196] Castellanos-Gomez, A. *et al.* Local strain engineering in atomically thin MoS_2. *Nano Lett.* **13**, 5361–5366 (2013).

[197] Rice, C. *et al.* Raman-scattering measurements and first-principles calculations of strain-induced phonon shifts in monolayer MoS_2. *Phys. Rev. B* **87**, 081307 (2013).

[198] Horzum, S. *et al.* Phonon softening and direct to indirect band gap crossover in strained single-layer $MoSe_2$. *Phys. Rev. B* **87**, 125415 (2013).

[199] Wang, Q. H., Kalantar-Zadeh, K., Kis, A., Coleman, J. N. & Strano, M. S. Electronics and optoelectronics of two-dimensional transition metal dichalcogenides. *Nat. Nanotechnol.* **7**, 699–712 (2012).

[200] Buscema, M. *et al.* Photocurrent generation with two-dimensional van der Waals semiconductors. *Chem. Soc. Rev.* **44**, 3691–3718 (2015).

[201] Bernardi, M., Palummo, M. & Grossman, J. C. Extraordinary sunlight absorption and one nanometer thick photovoltaics using two-dimensional monolayer materials. *Nano Lett.* **13**, 3664–3670 (2013).

[202] Wang, X. *et al.* Highly anisotropic and robust excitons in monolayer black phosphorus. *Nat. Nanotechnol.* **10**, 517–521 (2015).

[203] Qiu, D. Y., da Jornada, F. H. & Louie, S. G. Environmental screening effects in 2D materials: Renormalization of the bandgap, electronic structure, and optical spectra of few-layer black phosphorus. *Nano Lett.* **17**, 4706–4712 (2017).

[204] Li, L. *et al.* Direct observation of the layer-dependent electronic structure in phosphorene. *Nat. Nanotechnol.* **12**, 21 (2017).

[205] Castellanos-Gomez, A. *et al.* Isolation and characterization of few-layer black phosphorus. *2D Mater.* **1**, 025001 (2014).

[206] Bringuier, E., Bourdon, A., Piccioli, N. & Chevy, A. Optical second-harmonic generation in lossy media: Application to GaSe and InSe. *Phys. Rev. B* **49**, 16971 (1994).

[207] Mudd, G. W. *et al.* Tuning the bandgap of exfoliated InSe nanosheets by quantum confinement. *Adv. Mater.* **25**, 5714–5718 (2013).

[208] Tamalampudi, S. R. *et al.* High performance and bendable few-layered InSe photodetectors with broad spectral response. *Nano Lett.* **14**, 2800–2806 (2014).

[209] Lei, S. *et al.* Evolution of the electronic band structure and efficient photo-detection in atomic layers of InSe. *ACS Nano* **8**, 1263–1272 (2014).

[210] Sánchez-Royo, J. F. *et al.* Electronic structure, optical properties, and lattice dynamics in atomically thin indium selenide flakes. *Nano Res.* **7**, 1556–1568 (2014).

[211] Island, J. O. *et al.* Precise and reversible band gap tuning in single-layer $MoSe_2$ by uniaxial strain. *Nanoscale* **8**, 2589–2593 (2016).

[212] Wang, Y. *et al.* Strain-induced direct-indirect bandgap transition and phonon modulation in monolayer WS_2. *Nano Res.* **8**, 2562–2572 (2015).

[213] Quereda, J. *et al.* Strong modulation of optical properties in black phosphorus through strain-engineered rippling. *Nano Lett.* **16**, 2931–2937 (2016).

[214] Zhang, G. *et al.* Infrared fingerprints of few-layer black phosphorus. *Nat. Commun.* **8**, 1–9 (2017).

[215] Schmidt, R. *et al.* Reversible uniaxial strain tuning in atomically thin WSe_2. *2D Mater.* **3**, 021011 (2016).

[216] Yang, S. *et al.* Tuning the optical, magnetic, and electrical properties of $ReSe_2$ by nanoscale strain engineering. *Nano Lett.* **15**, 1660–1666 (2015).

[217] Lloyd, D. *et al.* Band gap engineering with ultralarge biaxial strains in suspended monolayer MoS_2. *Nano Lett.* **16**, 5836–5841 (2016).

[218] Signorello, G., Karg, S., Björk, M. T., Gotsmann, B. & Riel, H. Tuning the light emission from GaAs nanowires over 290 meV with uniaxial strain. *Nano Lett.* **13**, 917–924 (2013).

[219] Splendiani, A. *et al.* Emerging photoluminescence in monolayer MoS_2. *Nano Lett.* **10**, 1271–1275 (2010).

[220] Ugeda, M. M. *et al.* Giant bandgap renormalization and excitonic effects in a monolayer transition metal dichalcogenide semiconductor. *Nat. Mater.* **13**, 1091–1095 (2014).

[221] Rössler, U., Jorda, S. & Broido, D. Fine structure of quantum well excitons. *Solid State Commun.* **73**, 209–214 (1990).

[222] Andreani, L. C. & Bassani, F. Exchange interaction and polariton effects in quantum-well excitons. *Phys. Rev. B* **41**, 7536 (1990).

[223] Damen, T. C., Via, L., Cunningham, J. E., Shah, J. & Sham, L. J. Subpicosecond spin relaxation dynamics of excitons and free carriers in GaAs quantum wells. *Phys. Rev. Lett.* **67**, 3432 (1991).

[224] Maialle, M. Z., de Andrada e Silva, E. A. & Sham, L. J. Exciton spin dynamics in quantum wells. *Phys. Rev. B* **47**, 15776 (1993).

[225] Vinattieri, A. *et al.* Exciton dynamics in GaAs quantum wells under resonant excitation. *Phys. Rev. B* **50**, 10868 (1994).

[226] Yu, T. & Wu, M. W. Valley depolarization due to intervalley and intravalley electron-hole exchange interactions in monolayer MoS_2. *Phys. Rev. B* **89**, 205303 (2014).

[227] Jones, A. M. *et al.* Optical generation of excitonic valley coherence in monolayer WSe_2. *Nat. Nanotechnol.* **8**, 634–638 (2013).

[228] Hao, K. *et al.* Direct measurement of exciton valley coherence in monolayer WSe_2. *Nat. Phys.* **12**, 677–682 (2016).

[229] Yu, H., Liu, G.-B., Gong, P., Xu, X. & Yao, W. Dirac cones and Dirac saddle points of bright excitons in monolayer transition metal dichalcogenides. *Nat. Commun.* **5**, 1–7 (2014).

[230] Wu, F., Qu, F. & Macdonald, A. H. Exciton band structure of monolayer MoS_2. *Phys. Rev. B* **91**, 075310 (2015).

[231] Mai, C. *et al.* Many-body effects in valleytronics: Direct measurement of valley lifetimes in single-layer MoS_2. *Nano Lett.* **14**, 202–206 (2014).

[232] Singh, A. *et al.* Coherent electronic coupling in atomically thin $MoSe_2$. *Phys. Rev. Lett.* **112**, 216804 (2014).

[233] Hao, K. *et al.* Coherent and incoherent coupling dynamics between neutral and charged excitons in monolayer $MoSe_2$. *Nano Lett.* **16**, 5109–5113 (2016).

[234] Mak, K. F. *et al.* Tightly bound trions in monolayer MoS_2. *Nat. Mater.* **12**, 207–211 (2013).

[235] Brixner, T. *et al.* Two-dimensional spectroscopy of electronic couplings in photosynthesis. *Nature* **434**, 625–628 (2005).

[236] Panitchayangkoon, G. *et al.* Long-lived quantum coherence in photosynthetic complexes at physiological temperature. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 12766–12770 (2010).

[237] Li, X., Zhang, T., Borca, C. N. & Cundiff, S. T. Many-body interactions in semiconductors probed by optical two-dimensional Fourier transform spectroscopy. *Phys. Rev. Lett.* **96**, 057406 (2006).

[238] Graham, M. W., Calhoun, T. R., Green, A. A., Hersam, M. C. & Fleming, G. R. Two-dimensional electronic spectroscopy reveals the dynamics of phonon-mediated excitation pathways in semiconducting single-walled carbon nanotubes. *Nano Lett.* **12**, 813–819 (2012).

[239] Moody, G. *et al.* Intrinsic homogeneous linewidth and broadening mechanisms of excitons in monolayer transition metal dichalcogenides. *Nat. Commun.* **6**, 1–6 (2015).

[240] Czech, K. J. *et al.* Measurement of ultrafast excitonic dynamics of few-layer MoS_2 using state-selective coherent multidimensional spectroscopy. *ACS Nano* **9**, 12146–12157 (2015).

[241] Jones, A. M. *et al.* Excitonic luminescence upconversion in a two-dimensional semiconductor. *Nat. Phys.* **12**, 323–327 (2016).

[242] Yang, L. *et al.* Long-lived nanosecond spin relaxation and spin coherence of electrons in monolayer MoS_2 and WS_2. *Nat. Phys.* **11**, 830–834 (2015).

[243] Lado, J. L. & Fernández-Rossier, J. On the origin of magnetic anisotropy in two dimensional CrI_3. *2D Mater.* **4**, 035002 (2017).

[244] Jiang, H., Gomez-Abal, R. I., Rinke, P. & Scheffler, M. Localized and itinerant states in lanthanide oxides united by GW@LDA+U. *Phys. Rev. Lett.* **102**, 126403 (2009).

[245] Liechtenstein, A. I., Anisimov, V. I. & Zaanen, J. Density-functional theory and strong interactions: Orbital ordering in Mott-Hubbard insulators. *Phys. Rev. B* **52**, R5467 (1995).

[246] Lauret, J. S. *et al.* Optical transitions in single-wall boron nitride nanotubes. *Phys. Rev. Lett.* **94**, 037405 (2005).

[247] Wirtz, L., Marini, A. & Rubio, A. Excitons in boron nitride nanotubes: Dimensionality effects. *Phys. Rev. Lett.* **96**, 126104 (2006).

[248] Alvarado, S. F., Seidler, P. F., Lidzey, D. G. & Bradley, D. D. C. Direct determination of the exciton binding energy of conjugated polymers using a scanning tunneling microscope. *Phys. Rev. Lett.* **81**, 1082 (1998).

[249] Scholes, G. D. & Rumbles, G. Excitons in nanoscale systems. *Nat. Mater.* **5**, 683–696 (2006).

[250] Grant, P. M. & Street, G. B. Optical properties of the chromium trihalides in the region 1-11 eV. *Bull. Am. Phys. Soc.* **II**, 415 (1968).

[251] Gudelli, V. K. & Guo, G.-Y. Magnetism and magneto-optical effects in bulk and few-layer CrI_3: A theoretical GGA+U study. *New J. Phys.* **21**, 053012 (2019).

[252] Gray, P. R. & Meyer, R. G. *Analysis and Design of Analog Integrated Circuits* (Wiley, 2001).

[253] Bordács, S., Kézsmárki, I., Ohgushi, K. & Tokura, Y. Experimental band structure of the nearly half-metallic $CuCr_2Se_4$: An optical and magneto-optical study. *New J. Phys.* **12**, 053039 (2010).

[254] Feil, H. & Haas, C. Magneto-optical Kerr effect, enhanced by the plasma resonance of charge carriers. *Phys. Rev. Lett.* **58**, 65 (1987).

[255] Zhong, D. *et al.* Van der Waals engineering of ferromagnetic semiconductor heterostructures for spin and valleytronics. *Sci. Adv.* **3**, e1603113 (2017).

[256] Jiang, P., Li, L., Liao, Z., Zhao, Y. & Zhong, Z. Spin direction-controlled electronic band structure in two-dimensional ferromagnetic CrI_3. *Nano Lett.* **18**, 3844–3849 (2018).

[257] Weber, M. J. *Handbook of Optical Materials* (CRC Press, 2002).

[258] Ellsworth, D. *et al.* Photo-spin-voltaic effect. *Nat. Phys.* **12**, 861–866 (2016).

[259] Fiederling, R. *et al.* Injection and detection of a spin-polarized current in a light-emitting diode. *Nature* **402**, 787–790 (1999).

[260] Ohno, Y. *et al.* Electrical spin injection in a ferromagnetic semiconductor heterostructure. *Nature* **402**, 790–792 (1999).

[261] Jiang, S., Shan, J. & Mak, K. F. Electric-field switching of two-dimensional van der Waals magnets. *Nat. Mater.* **17**, 406–410 (2018).

[262] Kramers, H. A. Théorie générale de la rotation paramagnétique dans les cristaux. *Proc. R. Neth. Acad. Arts Sci.* **33**, 959–972 (1930).

Appendix A

Fourier transforms

For functions with one argument like the vector potential $\boldsymbol{A}(\boldsymbol{r},t)$, the following conventions of time and space Fourier transforms are adopted,

$$\begin{aligned}
\boldsymbol{A}(\boldsymbol{r},t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}\omega\, e^{-i\omega t}\boldsymbol{A}(\boldsymbol{r},\omega), && \text{(A.1)}\\
\boldsymbol{A}(\boldsymbol{r},\omega) &= \int_{-\infty}^{\infty} \mathrm{d}t\, e^{i\omega t}\boldsymbol{A}(\boldsymbol{r},t), && \text{(A.2)}\\
\boldsymbol{A}(\boldsymbol{r},\omega) &= \frac{1}{N_q\Omega}\sum_{\boldsymbol{qG}} e^{i(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}}\boldsymbol{A}_{\boldsymbol{G}}(\boldsymbol{q},\omega), && \text{(A.3)}\\
\boldsymbol{A}_{\boldsymbol{G}}(\boldsymbol{q},\omega) &= \int \mathrm{d}\boldsymbol{r}\, e^{-i(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}}\boldsymbol{A}(\boldsymbol{r},\omega). && \text{(A.4)}
\end{aligned}$$

For two-argument functions like the screened Coulomb interaction $W(\boldsymbol{r}_1,\boldsymbol{r}_2;t_1-t_2)$, time Fourier transforms are given by,

$$\begin{aligned}
W(\boldsymbol{r}_1,\boldsymbol{r}_2;t_1-t_2) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}\omega\, e^{-i\omega(t_1-t_2)}W(\boldsymbol{r}_1,\boldsymbol{r}_2;\omega), && \text{(A.5)}\\
W(\boldsymbol{r}_1,\boldsymbol{r}_2;\omega) &= \int_{-\infty}^{\infty} \mathrm{d}(t_1-t_2)\, e^{i\omega(t_1-t_2)}W(\boldsymbol{r}_1,\boldsymbol{r}_2;t_1-t_2). && \text{(A.6)}
\end{aligned}$$

And space Fourier transforms of W are given by,

$$W(\boldsymbol{r}_1,\boldsymbol{r}_2;\omega) = \tfrac{1}{N_q\Omega}\sum_{\boldsymbol{qq'GG'}} e^{i(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}_1}e^{-i(\boldsymbol{q'}+\boldsymbol{G'})\cdot\boldsymbol{r}_2}W(\boldsymbol{q}+\boldsymbol{G},\boldsymbol{q'}+\boldsymbol{G'};\omega), \quad \text{(A.7)}$$

$$W(\boldsymbol{q}+\boldsymbol{G},\boldsymbol{q'}+\boldsymbol{G'};\omega) = \tfrac{1}{N_q\Omega}\int \mathrm{d}\boldsymbol{r}_1\mathrm{d}\boldsymbol{r}_2\, e^{-i(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}_1}e^{i(\boldsymbol{q'}+\boldsymbol{G'})\cdot\boldsymbol{r}_2}W(\boldsymbol{r}_1,\boldsymbol{r}_2;\omega). \quad \text{(A.8)}$$

Based on the translational symmetry of crystals, we have a selection rule for W,

$$W(\boldsymbol{q}+\boldsymbol{G},\boldsymbol{q'}+\boldsymbol{G'};\omega) = \delta_{\boldsymbol{qq'}}W_{\boldsymbol{GG'}}(\boldsymbol{q};\omega). \quad \text{(A.9)}$$

Appendix B

Paramagnetic and diamagnetic currents

The continuity equation for a system of interacting and identical electrons under the influence of an external time-dependent electromagnetic field is given in the Heisenberg picture,

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{n}_e(\boldsymbol{r},t) = -\partial_\alpha \left\{ \sum_\sigma \left[\frac{-i\hbar}{2m_e} \left(\hat{\psi}^\dagger(\partial_\alpha \hat{\psi}) - (\partial_\alpha \hat{\psi}^\dagger)\hat{\psi} \right) - \frac{q_e}{m_e} A_\alpha \hat{\psi}^\dagger \hat{\psi} \right] \right\}. \tag{B.1}$$

Based on Eq. (B.1), we define the paramagnetic particle current density operator $\hat{\boldsymbol{j}}_{\mathsf{p}}(\boldsymbol{r})$ in the Schrödinger picture as,

$$\hat{\boldsymbol{j}}_{\mathsf{p}}(\boldsymbol{r}) \equiv \sum_\sigma \frac{-i\hbar}{2m_e} \left[\hat{\psi}^\dagger(\boldsymbol{x})(\nabla\hat{\psi}(\boldsymbol{x})) - (\nabla\hat{\psi}^\dagger(\boldsymbol{x}))\hat{\psi}(\boldsymbol{x}) \right]. \tag{B.2}$$

The diamagnetic particle current density operator $\hat{\boldsymbol{j}}_d(\boldsymbol{r},t)$ is defined as,

$$\hat{\boldsymbol{j}}_{\mathsf{d}}(\boldsymbol{r},t) \equiv -\frac{q_e}{m_e} \boldsymbol{A}(\boldsymbol{r},t)\hat{n}_e(\boldsymbol{r}). \tag{B.3}$$

Appendix C

Relation between time-ordered and retarded correlation functions

For a general pair of Hermitian bosonic operators $\hat{A}$ and $\hat{B}$, the time-ordered and retarded correlation functions are defined as,

$$C^{\mathsf{T}}_{AB}(12) \equiv \left(-\frac{i}{\hbar}\right) \frac{\langle \Psi_0 | T[\hat{A}'(1)\hat{B}'(2)] | \Psi_0 \rangle}{\langle \Psi_0 | \Psi_0 \rangle}, \tag{C.1}$$

$$C^{\mathsf{R}}_{AB}(12) \equiv \left(-\frac{i}{\hbar}\right) \Theta(t_1 - t_2) \frac{\langle \Psi_0 | [\hat{A}'(1), \hat{B}'(2)] | \Psi_0 \rangle}{\langle \Psi_0 | \Psi_0 \rangle}, \tag{C.2}$$

where $\hat{A}'$ and $\hat{B}'$ are the deviation operators of $\hat{A}$ and $\hat{B}$, respectively. The Lehmann representations of correlation functions in Eqs. (C.1) and (C.2) are given by,

$$C^{\mathsf{T}}_{AB}(\boldsymbol{x}_1, \boldsymbol{x}_2; \omega) = \sum_S \left[\frac{A_S(\boldsymbol{x}_1) B^*_S(\boldsymbol{x}_2)}{\hbar\omega + E_0 - E_s + i\eta} - \frac{A^*_S(\boldsymbol{x}_1) B_S(\boldsymbol{x}_2)}{\hbar\omega + E_s - E_0 - i\eta} \right], \tag{C.3}$$

$$C^{\mathsf{R}}_{AB}(\boldsymbol{x}_1, \boldsymbol{x}_2; \omega) = \sum_S \left[\frac{A_S(\boldsymbol{x}_1) B^*_S(\boldsymbol{x}_2)}{\hbar\omega + E_0 - E_s + i\eta} - \frac{A^*_S(\boldsymbol{x}_1) B_S(\boldsymbol{x}_2)}{\hbar\omega + E_s - E_0 + i\eta} \right], \tag{C.4}$$

where $\eta \to 0^+$, S denotes an excited state, and $A_S(\boldsymbol{x}) \equiv \langle \Psi_0 | \hat{A}'(\boldsymbol{x}) | \Psi_S \rangle$. For retarded correlation functions, we have,

$$C^{\mathsf{R}}_{AB}(\boldsymbol{x}_1, \boldsymbol{x}_2; -\omega) = (C^{\mathsf{R}}_{AB}(\boldsymbol{x}_1, \boldsymbol{x}_2; \omega))^*, \tag{C.5}$$

which implies that its real part is an even function and its imaginary part an odd function of ω. For time-ordered correlation functions, we have,

$$C^{\mathsf{T}}_{AB}(\boldsymbol{x}_1, \boldsymbol{x}_2; -\omega) = C^{\mathsf{T}}_{BA}(\boldsymbol{x}_2, \boldsymbol{x}_1; \omega). \tag{C.6}$$

The retarded and time-ordered correlation functions are related via the following relation,

$$C^{\mathsf{R}}_{AB}(\boldsymbol{x}_1, \boldsymbol{x}_2; \omega) = C^{\mathsf{T}}_{AB}(\boldsymbol{x}_1, \boldsymbol{x}_2; \omega), \quad \omega > 0. \tag{C.7}$$

Therefore, one can calculate $C^{\mathsf{T}}_{AB}(\boldsymbol{x}_1, \boldsymbol{x}_2; \omega)$ and then obtain $C^{\mathsf{R}}_{AB}(\boldsymbol{x}_1, \boldsymbol{x}_2; \omega)$ through Eqs. (C.5) and (C.7).

Appendix D

Ward identities

The Ward identities are derived from the gauge invariance of the theory. They are exact relations for any two-particle correlation function that satisfy the conserving approximation [43, 45, 47]. Detailed derivation can be found in Ref. [45]. Here we only present simplified Ward identities where the L's have been converted to $\boldsymbol{\chi}$'s. Note that all the variables are in the reciprocal space and frequency domain:

$$\omega\boldsymbol{\chi}_{\boldsymbol{G}_1\boldsymbol{G}_2} = \overrightarrow{\boldsymbol{\chi}}_{\boldsymbol{G}_1\boldsymbol{G}_2}\cdot(\boldsymbol{q}+\boldsymbol{G}_2) = (\boldsymbol{q}+\boldsymbol{G}_1)\cdot\overleftarrow{\boldsymbol{\chi}}_{\boldsymbol{G}_1\boldsymbol{G}_2}, \tag{D.1}$$

$$\omega\overleftarrow{\boldsymbol{\chi}}_{\boldsymbol{G}_1\boldsymbol{G}_2} = \overleftrightarrow{\boldsymbol{\chi}}_{\boldsymbol{G}_1\boldsymbol{G}_2}\cdot(\boldsymbol{q}+\boldsymbol{G}_2), \tag{D.2}$$

$$\omega\overrightarrow{\boldsymbol{\chi}}_{\boldsymbol{G}_1\boldsymbol{G}_2} = (\boldsymbol{q}+\boldsymbol{G}_1)\cdot\overleftrightarrow{\boldsymbol{\chi}}_{\boldsymbol{G}_1\boldsymbol{G}_2}. \tag{D.3}$$

Appendix E

Bloch-wave representation

In practice, it is very helpful to use the Bloch-wave representation instead of the coordinate representation for the single-particle Green's function G and the two-particle correlation function L. For Green's function in the frequency domain, the transforms between the Bloch-wave representation and the coordinate representation are given by,

$$G(\boldsymbol{x}_1, \boldsymbol{x}_1; \omega) = \sum_{\lambda\mu} \phi_\lambda(\boldsymbol{x}_1)\phi^*_\mu(\boldsymbol{x}_2)G_{\lambda\mu}(\omega), \tag{E.1}$$

and

$$G_{\lambda\mu}(\omega) = \int \mathrm{d}\boldsymbol{x}_1\mathrm{d}\boldsymbol{x}_2\, \phi^*_\lambda(\boldsymbol{x}_1)\phi_\mu(\boldsymbol{x}_2)G(\boldsymbol{x}_1, \boldsymbol{x}_2; \omega). \tag{E.2}$$

As for $L(\boldsymbol{x}_1, \boldsymbol{x}'_1; \boldsymbol{x}_2, \boldsymbol{x}'_2; \omega)$, we have,

$$L(\boldsymbol{x}_1, \boldsymbol{x}'_1; \boldsymbol{x}_2, \boldsymbol{x}'_2; \omega) = \sum_{\lambda\mu;\alpha\beta} \phi_\lambda(\boldsymbol{x}_1)\phi^*_\mu(\boldsymbol{x}'_1)\phi_\alpha(\boldsymbol{x}_2)\phi^*_\beta(\boldsymbol{x}'_2)L_{\lambda\mu;\alpha\beta}(\omega), \tag{E.3}$$

and

$$L_{\lambda\mu;\alpha\beta}(\omega) = \int \mathrm{d}\boldsymbol{x}_1\mathrm{d}\boldsymbol{x}'_1\mathrm{d}\boldsymbol{x}_2\mathrm{d}\boldsymbol{x}'_2\, \phi^*_\lambda(\boldsymbol{x}_1)\phi_\mu(\boldsymbol{x}'_1)\phi^*_\alpha(\boldsymbol{x}_2)\phi_\beta(\boldsymbol{x}'_2)L(\boldsymbol{x}_1, \boldsymbol{x}'_1; \boldsymbol{x}_2, \boldsymbol{x}'_2; \omega). \tag{E.4}$$

Appendix F

Kramers' theorem

According to Kramers' theorem [262], when the system has time-reversal symmetry (no external magnetic fields and no intrinsic magnetization), the band energies at $\boldsymbol{k}$ and $-\boldsymbol{k}$ are related by,

$$\epsilon_{n\boldsymbol{k}\sigma} = \epsilon_{n(-\boldsymbol{k})(-\sigma)}. \tag{F.1}$$

Furthermore, if the crystal is centrosymmetric (i.e., with spatial inversion symmetry), we have $\epsilon_{n\boldsymbol{k}\sigma} = \epsilon_{n(-\boldsymbol{k})\sigma}$. Combined with the Kramers' degeneracy in the case of time-reversal symmetry, we get $\epsilon_{n\boldsymbol{k}\sigma} = \epsilon_{n\boldsymbol{k}(-\sigma)}$, which means the spin degeneracy still holds when the system has both spatial inversion symmetry and time-reversal symmetry [8].

Appendix G

Dielectric function tensor with C_3 symmetry

Here we derive the shape of the dielectric function tensor in a material with C_3 symmetry with the rotational axis along the z direction. Let's focus on the xy-components of $\boldsymbol{\varepsilon}$,

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{xx} & \varepsilon_{xy} \\ \varepsilon_{yx} & \varepsilon_{yy} \end{pmatrix}. \tag{G.1}$$

The C_3 symmetry is given by the SO(2) rotational matrix,

$$\boldsymbol{D}(C_3) = \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}. \tag{G.2}$$

If we rotate the crystal with C_3, the dielectric tensor will transform as,

$$\begin{aligned} \tilde{\boldsymbol{\varepsilon}} &= \boldsymbol{D}(C_3) \cdot \boldsymbol{\varepsilon} \cdot \boldsymbol{D}^{-1}(C_3) \\ &= \frac{1}{4} \begin{pmatrix} \varepsilon_{xx} + \sqrt{3}(\varepsilon_{xy} + \varepsilon_{yx}) + 3\varepsilon_{yy} & -\sqrt{3}\varepsilon_{xx} + \varepsilon_{xy} - 3\varepsilon_{yx} + \sqrt{3}\varepsilon_{yy} \\ -\sqrt{3}\varepsilon_{xx} - 3\varepsilon_{xy} + \varepsilon_{yx} + \sqrt{3}\varepsilon_{yy} & 3\varepsilon_{xx} - \sqrt{3}(\varepsilon_{xy} + \varepsilon_{yx}) + \varepsilon_{yy} \end{pmatrix} \\ &\equiv \boldsymbol{\varepsilon}. \end{aligned} \tag{G.3}$$

Compare each entry of Eq. (G.1) and Eq. (G.3), and we will get the following relations,

$$\begin{aligned} \varepsilon_{xx} &= \varepsilon_{yy}, \\ \varepsilon_{yx} &= -\varepsilon_{xy}. \end{aligned} \tag{G.4}$$

Appendix H

Absorbance and transmittance of 2D materials

In this section, we derive the expression of optical absorbance and transmittance through an infinitely large 2D material with thickness d. Suppose a beam of monochromatic (with frequency ω) plane-wave light is shined on the material as a normal mode and at normal incidence, as shown in Fig. H.1. The corresponding frequency-dependent refractive index $n(\omega)$ of the 2D material is related to the relative dielectric function $\varepsilon(\omega)$ for the normal mode via $n^2(\omega) = \varepsilon(\omega)$. The magnetic permeability is assumed to be equal to that of the vacuum, i.e., $\mu = \mu_0$. Since d is very small, we assume that the charge current density J is uniform within the 2D material on a macroscopic scale. Apply the Maxwell's equations in this setup, and we get the following boundary conditions for amplitudes of the incident, reflected, and transmitted fields,

$$E^{\mathsf{i}} - E^{\mathsf{r}} - E^{\mathsf{t}} = 0, \tag{H.1}$$

$$B^{\mathsf{i}} + B^{\mathsf{r}} - B^{\mathsf{t}} = \mu_0 J d. \tag{H.2}$$

All the fields in Eqs. (H.1) and (H.2) refer to the macroscopic Fourier component at frequency ω with $\boldsymbol{G} = 0$ and $\boldsymbol{q} \to 0$. According to the Ohm's law, we have

$$J = \sigma E^{\mathsf{t}}, \tag{H.3}$$

where the frequency-dependent conductivity $\sigma(\omega)$ is related to the imaginary part of the relative dielectric function $\varepsilon_2(\omega)$ through the following relation,

$$\sigma(\omega) = \omega \varepsilon_0 \varepsilon_2(\omega). \tag{H.4}$$

In addition, the electric fields and magnetic fields in a plane-wave solution of electromagnetic waves in vacuum satisfy the following conditions,

$$B^{\mathsf{i}} = E^{\mathsf{i}}/c, \tag{H.5}$$

$$B^{\mathsf{r}} = E^{\mathsf{r}}/c, \tag{H.6}$$

$$B^{\mathsf{t}} = E^{\mathsf{t}}/c. \tag{H.7}$$

After combining Eqs. (H.1)–(H.7), we get the ratio between E^{t} and E^{i} as,

$$\frac{E^{\mathsf{t}}}{E^{\mathsf{i}}} = \frac{1}{1 + \frac{\varepsilon_2 \omega d}{2c}}. \tag{H.8}$$

where c is the speed of light and it satisfies, $c^2 = \frac{1}{\varepsilon_0 \mu_0}$. The transmittance T can then be calculated as,

$$T \equiv \left| \frac{E^{\mathsf{t}}}{E^{\mathsf{i}}} \right|^2 = \frac{1}{\left(1 + \frac{\varepsilon_2 \omega d}{2c}\right)^2} = \frac{1}{(1 + A_0/2)^2}, \quad A_0 = \frac{\varepsilon_2 \omega d}{c}. \tag{H.9}$$

And the absorbance A is given by,

$$A \equiv 1 - T = \frac{A_0(1 + A_0/4)}{(1 + A_0/2)^2}. \tag{H.10}$$

Usually $A_0 \ll 1$ for 2D materials within the optical range of the electromagnetic spectrum, and we can use the following approximate expressions of T and A,

$$T \approx 1 - A_0, \tag{H.11}$$

$$A \approx A_0. \tag{H.12}$$

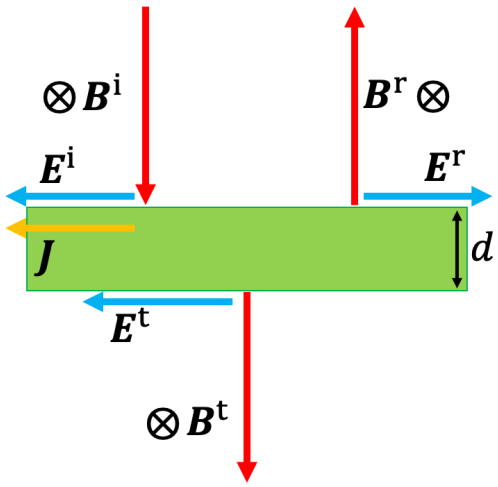

Figure H.1: Absorbance and transmittance of a 2D material. The green rectangle denotes the infinitely large 2D material with thickness d. The space above and below the 2D material are vacuum. Propagation directions of incident, reflected, and transmitted lights are denoted by the red arrows. Electric fields are denoted by the blue arrows, of which the direction is the positive direction for each. The positive direction of all the magnetic fields points into the page. The charge current density $\boldsymbol{J}$ is assumed to be uniform within the 2D material and denoted by the yellow arrow.