Electronic structure of van der Waals ferromagnet math xmlns="http://www.w3.org/1998/Math/MathML">msub>mi>CrI/mi>mn>3/mn>/msub>/math> from self-consistent vertex corrected math xmlns="http://www.w3.org/1998/Math/MathML">mrow>mi>G/mi>mi>W/mi>/mrow>/math> approaches

Andrey L. Kutepov

Phys. Rev. Materials **5**, 083805 — Published 27 August 2021

DOI: 10.1103/PhysRevMaterials.5.083805

Electronic structure of van der Waals ferromagnet CrI_3 from self consistent vertex corrected GW approaches

Andrey L. Kutepov*
Condensed Matter Physics and Materials Science Department, Brookhaven National Laboratory, Upton, NY 11973

Electronic structure of layered van der Waals ferromagnet CrI_3 is studied with self consistent diagrammatic approaches beyond GW approximation. Considerable improvement in the calculated band gap as compared to the non-self-consistent G0W0 results has been found. Certain spectral features in the valence bands discovered recently by the angle resolved photoemission spectroscopy, are reproduced better when we use full frequency dependent self energy. Density functional theory and quasiparticle self-consistent GW method which are based on frequency-independent self energy are unable to resolve these features. Non-locality effects in the diagrams beyond GW approximation are large for both polarizability and self energy. This finding can potentially have an impact on the development of methods like GW+DMFT.

PACS numbers:

Introduction

Magnetic van der Waals material CrI_3 represents considerable interest in view of its promising applications in spintronics. It possesses some remarkable properties which include, for instance, the preservation of magnetic order down to a single layer.[1,2] The bi-layer of this material shows anti-ferromagnetic ordering whereas its mono-layer, three-layer and bulk are all ordered ferromagnetically.[1] It is important to understand these (and other) properties from the theoretical point of view in order to be able to explain already known properties or even to predict new ones in this class of materials. The key to understand them is their electronic structure.

Electronic structure of CrI_3 was studied both experimentally[3,4] and theoretically.[5–10] As it seems, there is a general consensus that basic features of it (such as band gap) are similar in bulk material and in thin film.[3,6] However, there is still no consensus on the reasons of apparent inconsistency between experimental and theoretical values of the band gap in CrI_3.

In the bulk CrI_3, optical measurement[4] resulted in the optical gap of 1.24 eV. Recent ARPES (angle resolved photoemission spectroscopy) measurements[3] reported the electronic band gap of about 1.3 eV. Normally, one would think that optical gap should be a bit smaller than electronic because of the excitonic effects. Therefore, the above two values are consistent if we assume that the exciton binding energies are on the scale of 0.1 eV. In theory, there are issues on the larger scale. In density functional theory (DFT) calculations, the band gap is 0.78 eV.[10] This value corresponds exactly to what one would expect from DFT: underestimation of the gap by 30-50%. The problem reveals itself when we try to improve DFT band gap. Routinely, it is done by applying the so called one-shot (non self consistent) GW approximation (G0W0). In a vast majority of semiconductors, G0W0 improves the DFT band gap considerably[11] with remaining small underestimation up to 10-15%. However, when applied to the mono-layer of CrI_3, G0W0 results in the band gap 2.59–2.76 eV.[9,12] It is important to note that reported G0W0 calculations of CrI_3 monolayer used DFT+U as a starting point. If we assume that bulk and mono-layer band gaps of CrI_3 are not very different, the reported G0W0 results for the mono-layer exceed considerably the experimental value. Which, most likely, is the case because authors of both works also reported very strong excitonic effects with exciton binding energies up to 1.5 eV. Formally, the presence of strong excitons could explain the value of the optical gap but it doesn't explain the value of the electronic gap. Nor, does it explain the small difference between optical and electronic gaps in experiments. However, it suggests that the electronic gap obtained in G0W0 calculations should be a subject of a strong renormalization if one includes diagrams beyond GW approximation in the evaluation of the electronic gap. For instance, if one uses Bethe-Salpeter equation (BSE) instead of random phase approximation (RPA) in the evaluation of polarizability and then applies the corresponding screened interaction W in the evaluation of the GW diagram, G0W0 band gap might be much smaller. Thus, the results obtained in Refs. [9,12] suggest to study the effect of higher order diagrams (vertex corrections) on the electronic structure of CrI_3.

An important step forward in elucidating the electronic structure of CrI_3 (and related materials) was done by Lee et al.[10] In their work, the hybrid method QSGW80[13] was used. The QSGW80 approach consists in empirical mixing of QSGW (quasiparticle self-consistent GW) self energy and LDA (local density approximation) exchange-correlation potential: $\Sigma_{QSGW80} = 0.8\Sigma_{QSGW} + 0.2V^{xc}_{LDA}$. As authors of Ref. [10] argue, the mixing effectively corrects the underestimation of screening in QSGW method. Formally, the QSGW80 approach should be considered as a semiempirical one but it allows to improve the

*e-mail: akutepov@bnl.gov

calculated electronic structure of simple semiconductors considerably.[13,14] For CrI_3, application of QSGW80 without spin-orbit coupling (SOC) resulted in the band gap 2.23 eV,[10] whereas calculations with perturbative (after the self-consistency was reached) inclusion of SOC resulted in the band gap 1.68 eV. Thus, SOC renormalization of the electronic structure of CrI_3 is noticeable. Unfortunately, authors of Ref. [10] do not report the gap value obtained with standard QSGW, i.e. without admixture of LDA exchange-correlation potential. So, it is hard to say about the actual effect of it. QSGW80 is constructed in such a way that it empirically enhances the screening which is underestimated by QSGW. So, the mere fact that Lee et al. use QSGW80 instead of QSGW suggests an importance of higher order diagrams which would directly (instead of empirically) address the issue of insufficient screening in QSGW.

Authors of Ref. [10] also make an interesting research into the importance of nonlocality of self energy. Namely, by direct comparison of DFT+U and QSGW80 calculations they observe that DFT+U approach cannot mimic the QSGW80 results because of single-site approximation inherent to DFT+U. Obviously, this analysis of nonlocality of self energy in CrI_3 (and related materials) makes direct impact on the validity of other methods based on the single site approximation (like DFT plus dynamical mean field theory (DMFT)) when applied to this class of materials.

Motivated by the above cited works, this study focuses on application of the diagrammatic approaches which go beyond GW approximation, i.e. directly (and self-consistently) include vertex corrections. In this way, we estimate step by step the effect of the first order vertex correction and then the effect of replacing the first order diagram for polarizability by solving BSE for it. We also apply QSGW and, by doing this, we answer the question (though using different codes) on the difference between QSGW and QSGW80. Also, the effect of the SOC is studied directly. Namely, fully relativistic (FR) approach (Dirac's equation based) is used along with the scalar-relativistic (SR) approach in order to estimate SOC effect directly and compare it with the perturbative estimate made in Ref. [10]. We extend the study of non-local effects conducted by Lee et al. in [10] by investigating non-local contribution of the diagrams beyond GW. It is done by directly evaluating them using a full setup (all functions are **k**-dependent) and a simplified setup where we assume the local (single site) approximation. Our study, therefore, has an explicit impact on the development of the methods like GW+DMFT[15–21] where one assumes the single site approximation for the DMFT part.

The paper begins with a brief discussion of the distinctive features of the methods used in this work and the setup parameters for the calculations (the first section). The second section provides principal results obtained for the electronic structure of CrI_3. The third section presents the results of the investigation into the importance of non-local effects for higher order diagrams. The conclusions are given afterwards.

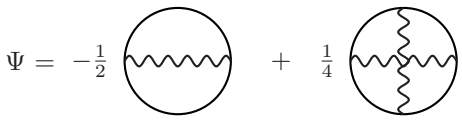

FIG. 1: Diagrammatic representation of Ψ-functional which includes the simplest non-trivial vertex. First diagram on the right hand side stands for scGW approximation, whereas total expression corresponds to sc(GW+G3W2) approximation.

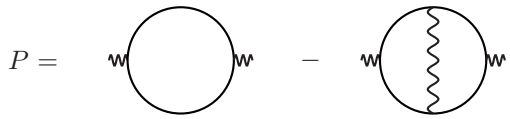

FIG. 2: Diagrammatic representation of irreducible polarizability in the simplest vertex corrected scheme sc(GW+G3W2).

Methods and calculation setups

All calculations in this work were performed using code FlapwMBPT.[22] Recently, a few updates were implemented in the code.[23,24] For DFT calculations, we used the local density approximation (LDA) as parametrized by Perdew and Wang.[25] In this study we use scGW method and two self-consistent vertex corrected schemes (see below). They are based on the L. Hedin's theory.[26] ScGW and one of the vertex corrected schemes, sc(GW+G3W2)[27], can also be defined using Ψ-functional formalism of Almbladh et al.[28] Corresponding Ψ-functional which includes vertex corrections is shown in Fig. 1. In Fig. 1, the first diagram corresponds to GW approximation, whereas the sum of the first and the second diagram represents sc(GW+G3W2) approximation. Diagrammatic representations for irreducible polarizability (Fig. 2) and for self energy (Fig. 3) in scGW and in sc(GW+G3W2) follow from the chosen approximation for Ψ-functional.

The second vertex corrected scheme which we use in this work is the scheme G according to the classification introduced in Ref. [29]. This scheme differs from sc(GW+G3W2) in the evaluation of polarizability: Bethe-Salpeter equation is used in the scheme G. In this case, the second term on the right hand side of Fig. 2 is replaced with an infinite sequence of diagrams (ladder diagrams) so that the vertex correction to polarizability can be represented as in Fig. 4. Diagrammatic representation of self energy is the same in both vertex corrected schemes used in this work. For convenience, let

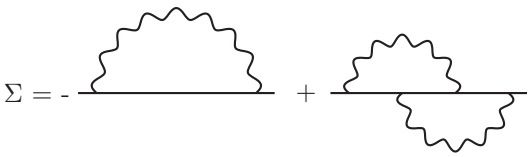

FIG. 3: Diagrammatic representation of self energy in the simplest vertex corrected scheme sc(GW+G3W2).

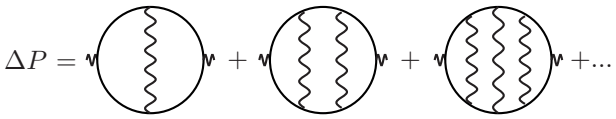

FIG. 4: Ladder sequence of diagrams for the vertex correction to polarizability in sc(BSE:P@GW+G3W2) approach.

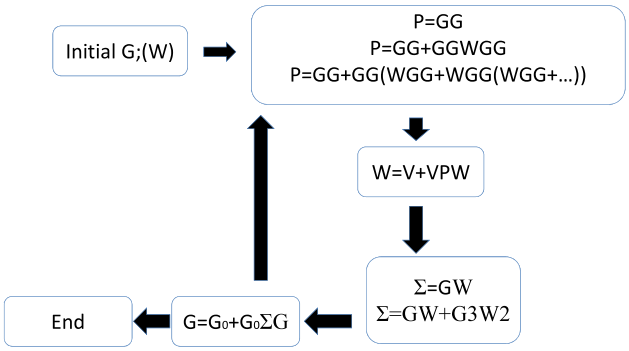

FIG. 5: Flowchart of scGW, sc(GW+G3W2), and sc(BSE:P@GW+G3W2) calculations. All equations are presented using symbolic notations. In the expressions for polarizability, first equation corresponds to scGW, second equation is used in sc(GW+G3W2), and the third one in sc(BSE:P@GW+G3W2). In the expressions for self energy, first equation corresponds to scGW, and the second one to both sc(GW+G3W2) and sc(BSE:P@GW+G3W2). G_0 stands for Green's function in Hartree approximation. Any calculation begins with self-consistent DFT iterations where the basis set is formed and the initial approach for G is generated. Iterations of scGW method use this initial Green's function as an input in order to start. During scGW iterations, G is updated and screened interaction W is generated. Both G and W serve as an input to start iterations of sc(GW+G3W2) or sc(BSE:P@GW+G3W2) approaches. sc(BSE:P@GW+G3W2), being computationally most demanding, can be run after a few iterations of sc(GW+G3W2), which can save computer time. In spin-polarized calculations, an external magnetic field is applied at the first iteration to create initial spin splitting.

us here introduce a new abbreviation for the scheme G: sc(BSE:P@GW+G3W2). In this abbreviation, the part after the symbol @ stands for diagrammatic representation of self energy, whereas the part before the symbol @ says that polarizability is evaluated from BSE. The rational of using sc(BSE:P@GW+G3W2) is to directly check the relative importance of excitonic effects on the evaluated electronic band structure. It is important to mention that our implementation[29] of BSE uses full frequency dependence of screened interaction W opposite to a common approximation[30,31] where one uses static (frequency independent and taken at zero frequency) screened interaction W. As one can deduce from its construction, scheme sc(BSE:P@GW+G3W2) is not Ψ-derivable (as opposed to scGW or sc(GW+G3W2)) and, therefore, is not conserving. However, evaluation of polarizability in sc(BSE:P@GW+G3W2) follows (at least approximately) its definition as being a functional derivative of electronic density with respect to full electrostatic potential, which is the foundation of BSE. Therefore, scheme sc(BSE:P@GW+G3W2) also has certain strong principle built in its construction. As it is evidenced in Ref. [32] it usually results in better band gaps as compared to sc(GW+G3W2). More details about properties of vertex corrected schemes one can find in Refs. [29,33].

Technical details of the GW part were described in Refs. [34,35]. Detailed account of the algorithms for sc(GW+G3W2), sc(BSE:P@GW+G3W2), and also for other vertex corrected schemes can be found in Refs. [27,29,32,33]. Brief account of the implementation of BSE also is provided in Appendix A. Figure 5 presents the flowchart of the calculations which gives a general idea of how the calculations are organized. The diagrammatic (GW and the diagramms beyond GW) parts of the FlapwMBPT code take full advantage of the fact that certain diagrams can more efficiently be evaluated in reciprocal (and frequency) space whereas other diagrams are easier to evaluate in real (and time) space. As a result, GW part of the code scales as $N_k N_\omega N_b^3$ where N_k is the number of **k**-points in the Brillouin zone, N_ω is the number of Matsubara frequencies, and N_b stands for the size of the basis set. The vertex part of the code scales as $N_k^2 N_\omega^2 N_b^4$. For comparison, if one uses naive (all in reciprocal space and frequency) implementation then GW part scales as $N_k^2 N_\omega^2 N_b^4$ (i.e. exactly as the vertex part when the implementation is efficient), and the vertex part scales as $N_k^3 N_\omega^3 N_b^5$. Besides of efficiency of the implementation, we have to mention two more factors which make the use of the diagrams beyond GW feasible. First is the fact that the higher order diagrams converge much faster than the GW diagram with respect to the basis set size and to the number of **k**-points.[29,32] Second is that the higher order diagrams are very well suited for massive parallelization.

We also use quasiparticle self consistent GW (QSGW) approach. Similar to scGW, sc(GW+G3W2), and sc(BSE:P@GW+G3W2) approaches, it is based on the

finite temperature (Matsubara) formalism and in this respect it is different from the well known QSGW implementation by Kotani et al.[36] Quasiparticle approximation includes linearization of self energy near the zero frequency (see for details Refs. [34,35]) and, therefore, the method is reliable only not very far from the Fermi level - usually within a few electron-volts. Effective self energy is static (frequency independent) and the method is not diagrammatic. However, as it was explained by Kotani et al.[36], QSGW satisfies the zero frequency and long wave limit of the Ward Identity (WI) because of the so called Z-factor cancellation. This fact makes it often quite accurate, especially in simple metals and semiconductors where the above mentioned limit is important. Considering the differences between QSGW and the above introduced approaches, together they represent a good set of methods to study new materials.

Principal difference between fully relativistic calculations (FR) ans scalar relativistic (SR) calculations consists in the fact that we use Dirac-Kohn-Sham equations to generate LAPW+LO basis set in the FR case (see Ref. [23] for the implementation in the FlapwMBPT code) instead of scalar-relativistic Kohn-Sham equations.[37] Generalization of the evaluation of diagrams to the FR case is relatively straightforward: one just replaces the SR basis functions with FR basis functions in the evaluation of matrix elements (see for instance the generalization of scGW and QSGW to fully relativistic variant in Ref. [34]).

Let us now specify the setup parameters used in the calculations. In order to make presentation more compact, principal structural parameters for the studied solids have been collected in Table I and the most important set up parameters have been collected in Table II. All calculations have been performed for the electronic temperature $600K$. In all calculations we assumed the ferromagnetic (FM) ordering. The calculations (excluding the vertex part) were performed with the $4 \times 4 \times 4$ mesh of **k**-points in the Brillouin zone. 500 band states (1000 in the FR case) were used to expand Green's function and self energy. Product basis (PB) consisted of approximately 3100 functions (depending on **k**-point). The diagrams beyond GW approximation were evaluated using $2 \times 2 \times 2$ mesh of **k**-points in the Brillouin zone and with about 40 (80 in the FR case) bands (closest to the Fermi level). With the above mentioned faster convergence of higher order diagrams with respect to these parameters, this choice represented a reasonable compromise between the accuracy and the computational cost. Most important convergence tests are presented in Tables III, IV, and V. As one can deduce from the convergence tests, the remaining uncertainty of the band gap obtained in fully relativistic sc(BSE:P@GW+G3W2) calculations could be at the level of 0.1–0.2 eV. Also, most likely effect of further refining of the computational setup would be a reduction of the calculated band gap.

TABLE I: Structural parameters of the solids studied in this work. Lattice parameters are in Angstroms, MT radii are in atomic units (1 Bohr radius), and atomic positions are given relative to the three primitive translation vectors. Experimental structural data from Ref. [38] are used.

Solid	Space group	a	c	Atomic positions	R_{MT}
CrI_3	148	6.867	19.807	Cr: 1/3;2/3;0.33299	2.471
				I: 0.31677;0.33453;0.4123	2.667

TABLE II: Principal setup parameters of the studied solids are given. The following abbreviations are introduced: Ψ is for wave functions, ρ is for the electronic density, V is for Kohn-Sham potential, and PB is for the product basis.

Solid	Core states	Semicore	L_{max} $\Psi/\rho, V$	L_{max} PB	RK_{max}
CrI_3	Cr: [Ne]	3s,3p	6/6	6	6.0
	I: [Kr]	5s,4d	6/6	6	

Results

We begin the presentation of results by showing in Table VI the band gaps and magnetic moments (on chromium sites) obtained using different approximations. Magnetic moments do not show any noticeable dependence on the method and are in accordance with other calculations.[7] They also depend slightly on the choice of the muffin-tin radii and, correspondingly, are given here just for the reference. Calculated band gaps, however, show remarkable dependence on the approximation used. As usual, LDA underestimates the band gap by about 30-50% depending on how one approximates the relativistic effects. Both QSGW and scGW seriously overestimate the experimental band gap (by about factor of two). QSGW does not show improvement in the calculated band gap of CrI_3 as compared to scGW, which one would expect in small gap sp semiconductors.[40] From this fact, one can conclude that the presence of Cr 3d electrons makes this material somewhat different from the simple semiconductors. Noticeable improvement in the evaluated band gap happens when we include first order vertex correction, i.e. when we switch from scGW to sc(GW+G3W2). Further improvement, i.e. when we switch from sc(GW+G3W2) to sc(BSE:P@GW+G3W2), is a bit smaller. The effect of inclusion/neglecting the SOC is approximately of the same amplitude as the effect of using BSE when we consider the SOC effect at sc(BSE:P@GW+G3W2) level. At this level it is about twice smaller than in Ref. [10] which means that the self-consistent inclusion of the SOC makes some difference. At the level of scGW/QSGW, however, the effect of SOC is somewhat larger. It is interesting that the best

TABLE III: Convergence of the band gaps obtained in scalar relativistic G0W0 calculations with respect to the number of high energy local orbitals (HELO) included in the LAPW+LO basis set. Local orbitals associated with semicore states are not included. Numbers after orbital character indicate how many LO's are included with a given orbital character. The results presented in the main text correspond to the second row (i.e. s2p1d2/s1p2d1).

High energy LO		
Cr	I	Band gap (eV)
s1d1	p1	2.09
s2p1d2	s1p2d1	2.07
s2p1d2f1	s1p2d1f1	2.07
s3p2d3f2	s2p3d2f2	2.09
s3p3d4f3	s3p4d3f3	2.10

TABLE IV: Dependence of the calculated band gap of CrI_3 on the $\mathbf{k}$-grid $N_{\mathbf{k}}$ in G0W0 calculations. Scalar relativistic approach has been used.

$N_{\mathbf{k}}$	Band gap
2^3	2.31
3^3	2.16
4^3	2.07
5^3	2.09
6^3	2.06

(and the most sophisticated) result for the band gap in our study (1.57 eV) is quite close to the result 1.68 eV obtained in [10] using empirical enhancement of the screening. Thus, if we assume that there is no big differences in QSGW between this study and Ref. [10], we can state that QSGW80 works rather well for this material.

Our final result for the band gap (1.57 eV) still is a bit larger as compared to the experimental 1.3 eV obtained in ARPES studies.[3] One can name a few possible reasons for this remaining disagreement: i) numerical cutoffs (especially in the vertex part); ii)higher order diagrams not included in this study; iii) electroh-phonon interaction. All three reasons, normally, should result in some reduction of the calculated band gap bringing it in even better agreement with the experiment. But even at the present level, the error is already small enough and allows us to state that this material is a weakly correlated one and can be described using ab-initio diagrammatic methods.

In Fig. 7 we show partial density of states (atom and orbital resolved) of CrI_3 obtained in LDA calculations. Besides a little shrinkage of the band gap in fully relativistic case, there is very little difference between scalar relativistic and fully relativistic results. As one can see, principal spectral features around the Fermi level are almost completely defined by Cr 3d and I 5p states. In this respect, one can point out to a certain disagreement with the experimental ARPES data obtained by Kundu et al.[3] Namely, in experiments, valence band maximum (VBM) is formed by I 5p states only and Cr 3d states are shifted downward by about 0.6 eV. However, there is no such separation between I 5p and Cr 3d states in LDA calculations. Thus, we can conclude that LDA not only underestimates the band gap by almost 50% but

TABLE V: Dependence of the calculated band gap of CrI_3 on the calculation setup for the diagrams beyond GW. Scalar relativistic sc(GW+G3W2) approach has been used. N_{bnd}^{vrt} means the number of band states included in the evaluation of the beyond-GW diagrams. $N_{\mathbf{k}}^{vrt}$ means the $\mathbf{k}$-grid used for the evaluation of the beyond-GW diagrams. Dependence on the N_{bnd}^{vrt} was studied with fixed grid of $\mathbf{k}$-points: $4\times4\times4$ for GW part and $2\times2\times2$ for vertex part. Dependence on the $N_{\mathbf{k}}^{vrt}$ was studied with fixed grid of $\mathbf{k}$-points $6\times6\times6$ for GW part and with $N_{bnd}^{vrt}=40$. "Saturation" of the band gap when N_{bnd}^{vrt} reaches 40 is related to the fact that all important band states, i.e. Cr 3d and I 5p bands, are included.

Parameter	Setup	Band gap
N_{bnd}^{vrt}	20	2.91
	30	2.72
	40	2.25
	50	2.19
	60	2.16
$N_{\mathbf{k}}^{vrt}$	1^3	2.49
	2^3	2.25
	3^3	2.27

TABLE VI: Band gaps (eV) and magnetic moments (μ_B, Chromium site) of CrI_3 obtained at different levels of theory. SR stands for scalar-relativistic approximation, and FR is for fully relativistic approach. The positions of the peaks in $\mathbf{k}$-resolved spectral functions have been used to measure the band gaps. This is demonstarted in Fig. 6. Two variants of G0W0 differ by starting point: PerdewBurkeErnzerhof (PBE) functional[39] and Hartree-Fock (HF) approximation.

	Band gap		Moment	
Approximation	SR	FR	SR	FR
LDA	0.85	0.66	2.95	3.06
G0W0(PBE)	2.07	1.99	NA	NA
G0W0(HF)	4.22	3.74	NA	NA
QSGW	3.11	2.64	3.08	3.11
scGW	3.03	2.51	3.23	3.35
sc(GW+G3W2)	2.25	1.97	3.21	3.32
sc(BSE:P@GW+G3W2)	1.86	1.57	3.20	3.31
Experiment:				
Optical gap [4]	1.24			
ARPES [3]	1.3			

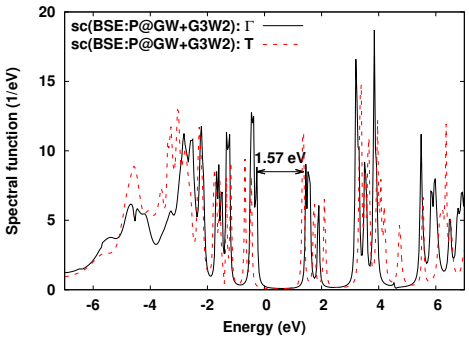

FIG. 6: Spectral function of CrI_3 at Γ and T points in the Brillouin zone as obtained in fully relativistic sc(BSE:P@GW+G3W2) approach. The value of the band gap defined as the difference in the positions of peaks is shown.

also predicts incorrect distribution of the orbital character among the valence bands.

In Fig. 8 we present partial spectral functions for Green's function based methods as obtained in scalar relativistic approximation. Similar results obtained in fully relativistic approach are shown in Fig. 9. Similar to the DFT case, there is no considerable difference between SR and FR results. So, our discussion is relevant to both figures equally. Firstly, we point out that QSGW approximation does not show a shift between Cr 3d and I 5p states. In this respect it is in a disagreement with the ARPES (as LDA is). Its difference with LDA is only in the considerable overestimation of the band gap. The rest of methods (scGW, sc(GW+G3W2), and sc(BSE:P@GW+G3W2)) clearly show the separation between Cr 3d and I 5p states. In these three methods VBM is formed solely by I 5p orbitals (as in experiments) and the onset of Cr 3d states is shifted downward from the VBM by 0.5–1.0 eV in agreement with the separation 0.6 eV found in the ARPES measurements.[3] The difference between QSGW and other three methods is that self energy is static (frequency independent) in QSGW whereas three other methods take full frequency dependence of self energy into account. Obviously, this frequency dependence is crucial for CrI_3. Another qualitative feature missing in QSGW consists in breaking the Cr 3d states in the conduction bands into two groups. Figures 8 and 9 also show gradual reduction of the band gap, but this was already discussed above.

Important comment about second order (in W) vertex correction to self energy has to be given. The problem of negative spectral weight appearance (when one uses this correction) was discussed and certain measures were taken to circumvent the issue.[43–45] Particularly, it was stated that it is impossible to perform self-consistent calculations which include G3W2 correction.[43] However, as it appears, sc(GW+G3W2) calculations can definitely be performed for CrI_3. They were also performed for a number of other systems[27,32,46] and also for electron gas[33] where sufficiently high convergence can be achieved. Besides of considerable increase in computer time needed, sc(GW+G3W2) calculations did not show any additional problems as compared to scGW calculations. Author of this work does not know the explanation of why the issue does not reveal itself. May be the reason is that all sc(GW+G3W2) (as well as scGW) calculations are performed using Matsubara's frequency axis and this fact somehow conceals the problem. Or, may be the self consistence itself, in fact, cures the problem because the sc(GW+G3W2) approach is Ψ-derivable and therefore is conserving.

As it follows from the above discussion, basic features of the electronic structure known from experiments (the band gap and Cr 3d/I 5p separation) can quite accurately be described using ab-initio diagrammatic methods. Thus, there is no need to apply the methods with adjustable parameters (DFT+U or DFT+DMFT) to study CrI_3 and, most likely, other materials from this class.

Nonlocal effects

In order to check the quality of the local (single site) approximation we also performed simplified calculations at sc(GW+G3W2) level (scalar relativistic) and compared the results with the corresponding calculations which, however, take full non-locality into account. Instead of the **k**-dependent band states as a basis set in full calculations, we used a set of orbitals confined inside their muffin tin spheres as a basis in our simplified calculations. We have to point out that our simplified (single site) basis set was still slightly extended as compared to what normally would be used in, for instance, GW+DMFT study. Namely, for Cr sites, we included in the basis set not only 3d orbitals but also their energy derivatives as they naturally appear in the linearized augmented plane wave (LAPW) method. We also included 5p and their energy derivatives in the basis set on I sites. Single site approximation makes drastic effect on the performance: vertex corrections in this case take practically zero time to be evaluated. However, as we discuss below, the calculations performed with the single site approximation are not free from some issues.

Quite predictably, the most problematic for the local approximation quantity is the "head" of polarizability $P^{\mathbf{q}}_{\mathbf{G}=\mathbf{G}'=0}$, where vectors **G** and **G**' represent reciprocal lattice translations. Polarizability is an intrinsically nonlocal function in real space. In reciprocal space, the momentum dependence of its "head" at small momenta is $P^{\mathbf{q}}_{\mathbf{G}=\mathbf{G}'=0} = Bq^2$ in exact theory. This behavior cancels the $1/q^2$ divergence of the bare Coulomb potential at small momenta. In self consistent diagrammatic approaches we normally have $P^{\mathbf{q}}_{\mathbf{G}=\mathbf{G}'=0} = A + Bq^2$ with A being small and negative. Its absolute value is normally much smaller than the absolute value of the "head" at all **q**-points on our **q**-mesh with non-zero momenta. In practice, we evaluate (by fitting) the coefficients A and

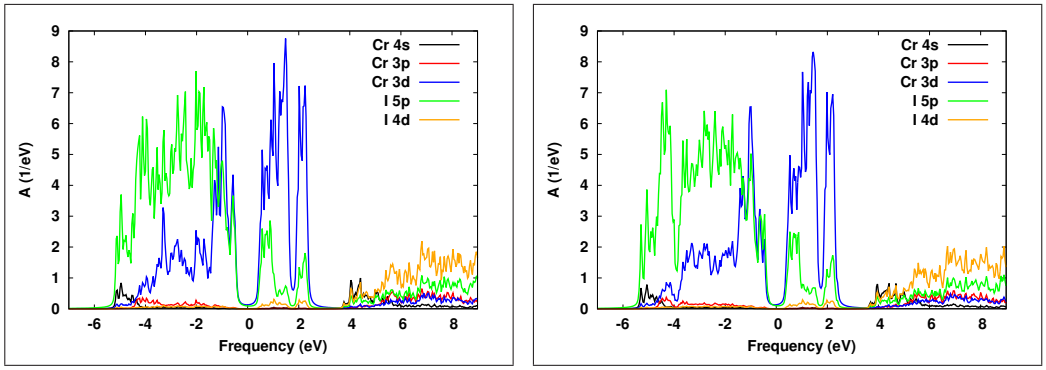

FIG. 7: Total and partial (atom and orbital resolved) spectral functions of CrI_3 obtained in LDA calculations. Scalar relativistic results are in the left window. Fully relativistic results are in the right window. Sums of spin-up and spin-down quantities in the SR case, and sums of spin-orbit components (i.e. $p_{1/2} + p_{3/2}$ and $d_{3/2} + d_{5/2}$) in the FR case are given.

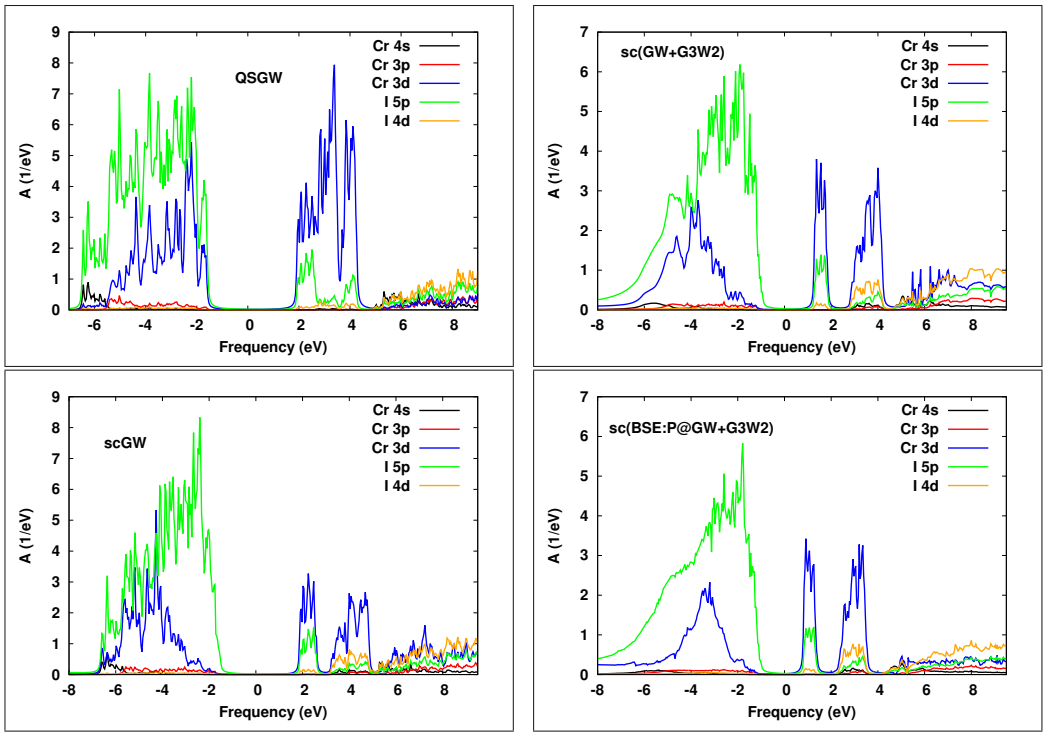

FIG. 8: Total and partial (atom and orbital resolved) spectral functions of CrI_3 obtained in Green's function based methods. Scalar relativistic results. Sums of spin-up and spin-down quantities are given. Analytical continuation of self energy[41,42] was used to get Green's function on the real frequency axis. The curves become smoother in the sequence QSGW-scGW-sc(GW+G3W2)-sc(BSE:P@GW+G3W2) primarily because of increase in the many-body effects (incoherence).

B and use only the Bq^2 part to proceed. The A coefficient becomes smaller when the number of the diagrams is increased (order by order or by using the BSE). To a certain degree its value also depends on the numerical approximations (cutoffs) within the same diagrammatic approach. In this respect, it is important to use **q**-dependent functions in the evaluation of polarizability. If, however, we accept the local approximation for the vertex part, the "head" of the correction to polarizability becomes momentum independent with very large A coefficient for total polarizability.

Figure 10 illustrates the above discussion. In the full calculation, the "head" is slightly positive at $q = 0$ which is to compensate the negative value obtained from the first diagram in Fig. 2 (GG part). As one can see from the right window of Fig. 10 where the "head" of total polarizability is shown, the compensation is not complete because of the numerical approximations and the limited number of diagrams. The correction to the "head" of polarizability obtained in local approximation is essentially a constant (momentum independent) and it looks as if it approximates the average over the Brillouin zone

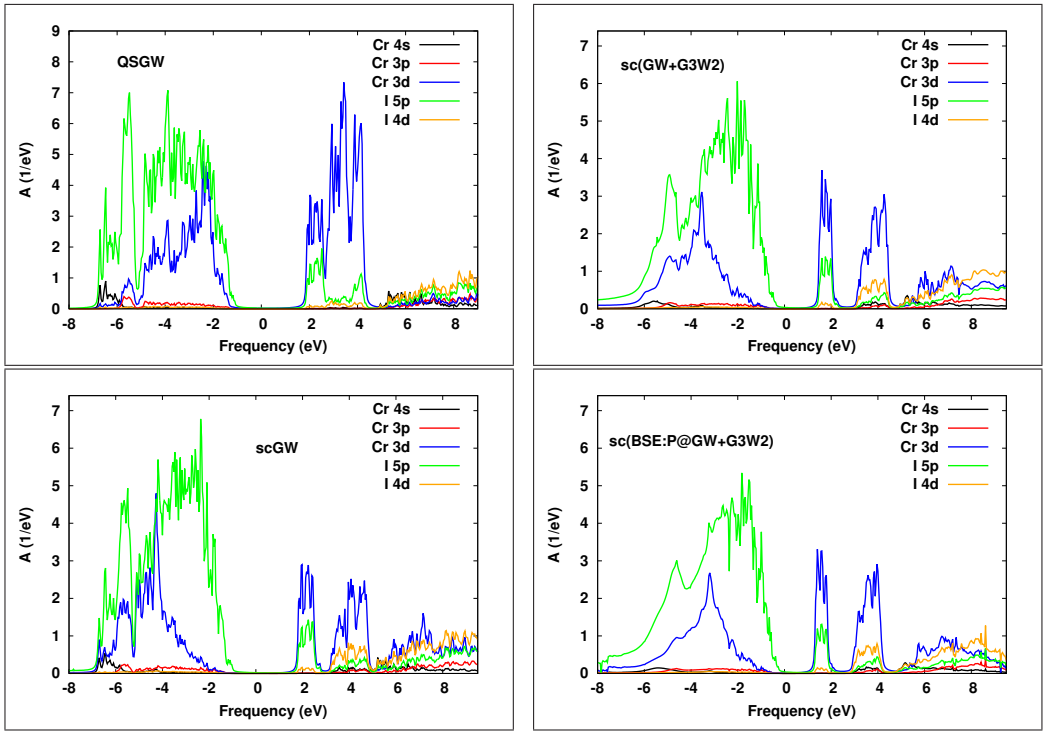

FIG. 9: Total and partial (atom and orbital resolved) spectral functions of CrI_3 obtained in Green's function based methods. Fully relativistic results. Sums of spin-orbit components (i.e. $p_{1/2} + p_{3/2}$ and $d_{3/2} + d_{5/2}$) are given. Analytical continuation of self energy[41,42] was used to get Green's function on the real frequency axis.

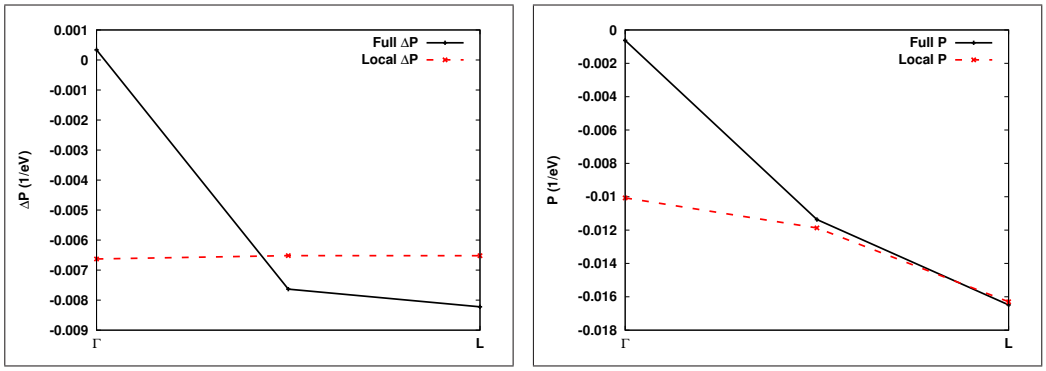

FIG. 10: The components $P^{\mathbf{q}}_{\mathbf{G}=\mathbf{G}'=0}(\nu = 0)$ of the calculated irreducible polarizability as functions of the momentum $\mathbf{q}$ along the direction Γ-L in the Brillouin zone. Vectors $\mathbf{G}$ and $\mathbf{G}$' represent reciprocal lattice translations. Left window shows the vertex correction, and in the right window one can see the full polarizability.

value. It is large compared to the GG part which makes total polarizability a poor approximation to the correct function.

Another important function for comparison is self energy. An example of it for the VBM is shown in Fig. 11. In the full calculation, the effects of interference make the vertex correction to self energy relatively small and very well localized in frequency space. It approximates zero when frequency is about 100 eV. The vertex correction to self energy obtained in local approximation looks quite differently. It is larger in absolute value and it is very slowly decaying function in frequency space. One can speculate that slow diminishing of the amplitude of self energy (local approximation) at high frequencies is somehow related to the truncation of screened interaction W. Truncation of W is most dangerous at high frequencies when it approaches bare Coulomb interaction and, therefore, is of long ranged nature. Thus, at least for CrI_3, the interference effects which are neglected in local approximation are quite important. Total self energy (right window in Fig. 11) shows that differences in the vertex correction part make the total functions also quite different. It is important to point out that the difference in total self energy is a combined effect of the difference in

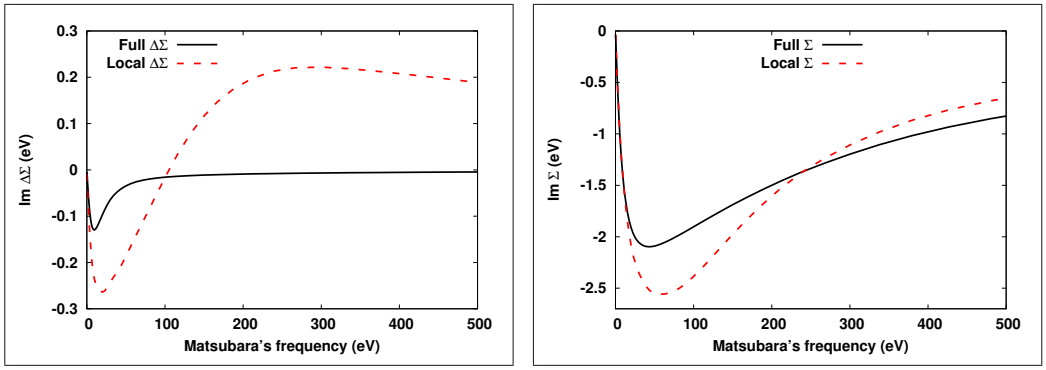

FIG. 11: Imaginary part of self energy at $k = (0; 0; 0)$ as a function of Matsubara's frequency. Diagonal matrix element for the VBM band is used for plotting. Left window shows the vertex correction, and in the right window one can see full self energy.

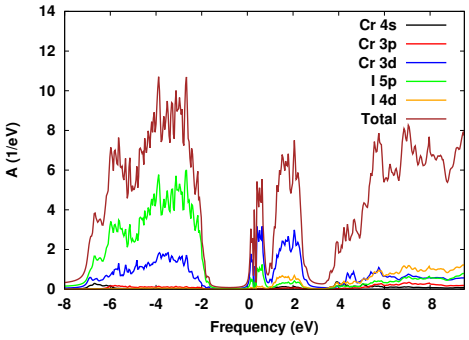

FIG. 12: Partial and total spectral functions of CrI_3 obtained in sc(GW+G3W2) calculation assuming local approximation for the vertex part. Scalar relativistic results. Sums of spin-up and spin-down quantities are given. Analytical continuation of self energy[41,42] was used to get Green's function on the real frequency axis.

vertex correction to self energy and the self-consistency effect which affects also the GW part of it.

In the evaluation of the band gap, the issues with the local approximation become hidden to a certain degree, as we integrate over the Brillouin zone a few times during every self-consistency iteration. Still, the band gap evaluated in the single site approximation (1.87 eV), tells us that the corresponding correction to the GW value is almost 25% larger than the correction obtained without using the local approximation (where the gap is 2.25 eV). The effect of the vertex correction is smaller in the full case because of the interference effects which are neglected in the local approximation. If we forget for a moment about the issues with polarizability and self energy detailed above, the final band gap obtained in the single site approximation might seem reasonable. Partial and total spectral functions obtained with local approximation and shown in Fig. 12 show some differences with the corresponding spectral functions obtained without using the local approximation (Fig. 8, upper right window) but those differences are not dramatic. However, considering the problems with this approximation at the intermediate steps of the calculation, one can conclude that the local approximation (even for the diagrams beyond GW level) represents a poor alternative to the methods which treat the non-local effects systematically. Whereas the quantitative effects are, most likely, material dependent, there is no reason to think that this conclusion will be different for the majority of materials. Considering the importance of this conclusion for GW+DMFT (and related) method, more studies of this kind are needed. As a remedy for the most problematic situations, where both the non-locality effects beyond GW and the strong correlations beyond sc(GW+G3W2) are important, one can suggest an extension of GW+DMFT, for instance sc(GW+G3W2)+DMFT method, which, at least formally, can be implemented along the same lines as GW+DMFT. In this method, DMFT would only be used for evaluation of the diagrams not included in sc(GW+G3W2) approach.

Conclusions

In conclusion, we have applied two self consistent diagrammatic approaches beyond GW approximation to study the electronic structure of the layered van der Waals ferromagnet CrI_3. Considerable overestimation of the band gap obtained in other works when using G0W0 approach was shown to be remedied by applying the vertex corrections. The important correction comes from the first order vertex function used in both polarizability and self energy. Application of Bethe-Salpeter equation for polarizability further improves the band gap. Inclusion of SOC is important, but its effect is smaller than the effect of vertex corrections.

We also studied the non-locality effects in the diagrams beyond GW approximation and found them as sufficiently large. This can have an impact on development of the methods like GW+DMFT.

As an interesting venue for future work on the subject one can consider studying optical properties of CrI_3 and other materials using vertex-corrected GW calculations as a starting point for a standard implementation

of BSE. "Standard implementation" here means using static (taken at zero frequency) screened interaction W in the kernel of BSE. In standard implementation, one casts BSE in an effective eigen value problem from which the exciton spectra can be directly obtained. Recently, it was shown how it can be done in the context of self-consistent QSGW calculations.[47]

Acknowledgments

This work was supported by the U.S. Department of energy, Office of Science, Basic Energy Sciences as a part of the Computational Materials Science Program.

Appendix A: Details of the Bethe-Salpeter Equation implementation

As it was mentioned in section , our implementation of BSE uses full frequency dependence of screened interaction W opposite to a common approximation[30,31] where one uses static (frequency independent and taken at zero frequency) screened interaction W. As a result, BSE is solved iteratively in this study. Each iteration adds one more diagram from an infinite sequence shown in Fig. 4 into the vertex correction to polarizability ΔP. In this Appendix we give the steps of iterations with some details on how frequency/time dependence is handled. Full (and rather lengthy) account of the implementation was published in Ref. [29] which includes the details of the basis sets, **k**-dependencies, and handling of time-to-frequency and frequency-to-time transformations. In this brief account, space arguments of all functions are represented by digits. Integration over repeated space arguments (if they are only on the right hand side of equations) is assumed. Below we use auxiliary functions K_0, K, $\triangle K$ and $\triangle\Gamma$ which are defined by the corresponding equations. Before the iterations we evaluate K_0:

$$K_0(123;\omega,\nu) = -G(13;\omega)G(32;\omega-\nu), \tag{A1}$$

and assign $\triangle K = 0$. ω and ν are fermionic and bosonic Matsubara's frequencies correspondingly. Also we transform $K_0(123;\tau,\nu) = \frac{1}{\beta}\sum_\omega e^{-i\omega\tau}K_0(123;\omega,\nu)$ where τ is Matsubara's time and $\beta = 1/T$.

During each iteration we perform the following steps (A2-A6):

$$K(123;\tau,\nu) = K_0(123;\tau,\nu) + \triangle K(123;\tau,\nu), \tag{A2}$$

$$\triangle\Gamma(123;\tau,\nu) = W(21;\tau)K(123;\tau,\nu), \tag{A3}$$

$$\triangle\Gamma(123;\omega,\nu) = \int d\tau e^{i\omega\tau}\triangle\Gamma(123;\tau,\nu), \tag{A4}$$

$$\triangle K(123;\omega,\nu) = \\ -G(14;\omega)\triangle\Gamma(453;\omega,\nu)G(52;\omega-\nu), \tag{A5}$$

$$\triangle K(123;\tau,\nu) = \frac{1}{\beta}\sum_\omega e^{-i\omega\tau}\triangle K(123;\omega,\nu). \tag{A6}$$

The above steps are repeated a specific number of times (iterations). In the end of iterations we evaluate vertex correction to polarizability:

$$\triangle P(12;\nu) = -\triangle K(112;\tau=0,\nu). \tag{A7}$$

For weakly correlated semicondictors the iterations (A2-A6) converge very fast (see for instance Fig. 7 in Ref. [32]). In case of CrI_3 we also found that 4 iterations were quite sufficient.

[1] B. Huang, G. Clark, E. Navarro-Moratalla, D. R. Klein, R. Cheng, K. L. Seyler, D. Zhong, E. Schmidgall, M. A. McGuire, D. H. Cobden, W. Yao, D. Xiao, P. Jarillo-Herrero and X. Xu, Nature **546**, 270 (2017).

[2] Y. Liu, L. Wu, X. Tong, J. Li, J. Tao, Y. Zhu and C. Petrovic, Sci. Rep. **9**, 13599 (2019).

[3] A. K. Kundu, Y. Liu, C. Petrovic and T. Valla, Sci. Rep. **10**, 15602 (2020).

[4] J. F. Dillon, Jr. and C. E. Olson, J. of Appl. Phys. **36**, 1259 (1965).

[5] S. W. Jang, M. Y. Jeong, H. Yoon, S. Ryee, and M. J. Han, Phys. Rev. Materials **3**, 031001(R) (2019).

[6] W. -B. Zhang, Q. Qu, P. Zhua and C. -H. Lam, J. of Mater. Chem. C **3**, 12457 (2015).

[7] V. K. Gudelli1 and G. -Y. Guo, New J. of Phys. **21**, 053012 (2019).

[8] P. Jiang, L. Li, Z. Liao, Y. X. Zhao, and Z. Zhong, New J. of Phys. **18**, 3844 (2018).

[9] M. Wu, Z. Li, T. Cao, and S. G. Louie, Nature Comm. **10**, 2371 (2019).

[10] Y. Lee, T. Kotani, and L. Ke, Phys. Rev. B **101**, 241409(R) (2020).

[11] H. Jiang, and P. Blaha, Phys. Rev. B **93**, 115203 (2016).

[12] A. Molina-Sanchez, G. Catarina, D. Sangallib and J. Fernandez-Rossier, J.Mater. Chem. C **8**, 8856 (2020).

[13] D. Deguchi, K. Sato, H. Kino, and T. Kotani,

Japan. J. of Appl. Phys. **55**, 051201 (2016).
[14] C. Bhandari, M. van Schilfgaarde, T. Kotani, W. R. L. Lambrecht, Phys. Rev. Materials **2**, 013807 (2018).
[15] S. Biermann, F. Aryasetiawan, and A. Georges, Phys. Rev. Lett. **90**, 086402 (2003).
[16] L. Boehnke, F. Nilsson, F. Aryasetiawan, and P. Werner, Phys. Rev. B **94**, 201106(R) (2016).
[17] F. Nilsson, L. Boehnke, P. Werner, and F. Aryasetiawan, Phys. Rev. Materials **1**, 043803 (2017).
[18] F. Petocchi, F. Nilsson, F. Aryasetiawan, and P. Werner, Phys. Rev. Research **2**, 013191 (2020).
[19] F. Petocchi, V. Christiansson, F. Nilsson, F. Aryasetiawan, and P. Werner, Phys. Rev. X **10**, 041047 (2020).
[20] S. Choi, A. Kutepov, K. Haule, M. van Schilfgaarde, and G. Kotliar, NPJ Quantum Materials **1**, 16001 (2016).
[21] S. Choi, P. Semon, B. Kang, A. Kutepov, G. Kotliar, Comp. Phys. Comm. **244**, 277 (2019).
[22] The latest publicly available version of the FlapwMBPT code (FlapwMBPT2106) can be downloaded from the website https://github.com/andreykutepov65/FlapwMBPT.
[23] A. L. Kutepov, Phys. Rev. B **103**, 165101 (2021).
[24] A. L. Kutepov, J. Phys.: Condens. Matter **33**, 235503 (2021).
[25] J. P. Perdew and Y. Wang, Phys. Rev.B **45**, 13244 (1992).
[26] L. Hedin, Phys. Rev. **139**, A796 (1965).
[27] A. L. Kutepov, arXiv.cond.mat.:2105.03770 (2021).
[28] C.-O. Almbladh, U. von Barth and R. van Leeuwen, Int. J. of Mod.Phys. B **13**, 535 (1999).
[29] A. L. Kutepov, Phys. Rev. B **94**, 155101 (2016).
[30] S. Albrecht, L. Reining, R. Del Sole and G. Onida, Phys. Rev. Lett. **80**, 4510 (1998).
[31] F. Fuchs, C. Rödl, A. Schleife, and F. Bechstedt, Phys. Rev. B **78**, 085103 (2008).
[32] A. L. Kutepov, Phys. Rev. B **95**, 195120 (2017).
[33] A. L. Kutepov and G. Kotliar, Phys. Rev. B **96**, 035108 (2017).
[34] A. Kutepov, K. Haule, S. Y. Savrasov, and G. Kotliar, Phys. Rev. B **85**, 155129 (2012).
[35] A. L. Kutepov, V. S. Oudovenko, G. Kotliar, Comp. Phys. Comm. **219**, 407 (2017).
[36] T. Kotani and M. van Schilfgaarde, S. V. Faleev, Phys. Rev.B **76**, 165106 (2007).
[37] T. Takeda, Z. Physik B **32**, 43 (1978).
[38] M. A. McGuire, H. Dixit, V. R. Cooper, and B. C. Sales, Chem. Mater. **27**, 612 (2015).
[39] J.P. Perdew, K. Burke and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[40] M. Grumet, P. Liu, M. Kaltak, J. Klimes, and G. Kresse, Phys. Rev. B **98**, 155143 (2018).
[41] H. J. Vidberg and J. W. Serene, J. Low Temp. Phys. **29**, 179 (1977).
[42] A. L. Kutepov, Comp. Phys. Commun. **257**, 107502 (2020).
[43] G. Stefanucci, Y. Pavlyukh, A. -M. Uimonen, R. van Leeuwen, Phys. Rev. B **90**, 115134 (2014).
[44] A. -M. Uimonen, G. Stefanucci, Y. Pavlyukh, and R. van Leeuwen, Phys. Rev. B **91**, 115104 (2015).
[45] Y. Pavlyukh, G. Stefanucci, R. van Leeuwen, Phys. Rev. B **102**, 045121 (2020).
[46] A. L. Kutepov, arXiv.cond.mat.:2106.03800 (2021).
[47] S. K. Radha, W. R. L. Lambrecht, B. Cunningham, M. Grüning, D. Pashov and M. van Schilfgaarde, arXiv.cond.mat.:2106.09137 (2021).