Solving the Bethe-Salpeter Equation in the Nonequilibrium Green's Function Formalism

Jiang Cao
Integrated Systems Laboratory
ETH Zurich
Zurich, Switzerland
jiacao@iis.ee.ethz.ch

Nicolas Vetsch
Integrated Systems Laboratory
ETH Zurich
Zurich, Switzerland
vetsch@

Vincent Maillou
Integrated Systems Laboratory
ETH Zurich
Zurich, Switzerland
vmaillou@iis.ee.ethz.ch

Anders Winka
Integrated Systems Laboratory
ETH Zurich
Zurich, Switzerland
awinka@iis.ee.ethz.ch

Alexander Maeder
Integrated Systems Laboratory
ETH Zurich
Zurich, Switzerland
almaeder@iis.ee.ethz.ch

Alexandros Nikolaos Ziogas
Integrated Systems Laboratory
ETH Zurich
Zurich, Switzerland
alziogas@iis.ee.ethz.ch

Mathieu Luisier
Integrated Systems Laboratory
ETH Zurich
Zurich, Switzerland
mluisier@iis.ee.ethz.ch

***Abstract*—Exciton-dominated optical responses have been observed in low-dimensional materials, opening up new avenues to realize high-speed, high-responsivity photo-detectors with low dark currents. To provide design guidelines for such excitonic devices, we present an *ab initio* computational framework for the Bethe-Salpeter equation (BSE) built on top of our previous NEGF-GW solver. We showcase its capability for a graphene nano-ribbon (GNR) photodiode containing 840 carbon atoms, uncovering the quantum dynamics of photo-excitation and exciton transport ultra-scaled nanostructures.**

***Index Terms*—Quantum transport, photo-detector, exciton.**

I. Introduction

Atomically precise bottom-up fabrication of graphene nano-ribbons (GNRs) has been recently demonstrated [1]. Unlike pristine graphene, which is gapless, armchair GNRs (AGNRs) exhibit semiconducting behavior, with a bandgap that varies systematically with ribbon width. This makes such structures particularly attractive for application such as field-effect transistors, photodetectors, and light-emitting devices. As device dimensions shrink into a few atoms wide, quantum confinement and reduced dielectric screening significantly enhance many-body effects. These one-dimensional (1-D) systems are expected to display pronounced excitonic effects that can be leveraged to realize photo-detectors with high responsivity, short response times, and operating under low electric fields.

Excitons are neutral, strongly correlated electron-hole pairs, as illustrated in Fig. 1. To accurately model these phenomena, it is necessary to go beyond single-particle approaches such as tight-binding or standard density functional theory (DFT) by including electron-electron (e-e) interactions. Applying GW corrections on the DFT leads to quasiparticle (QP) bandgap usually in good agreement with experiment. To capture excitonic resonances in the optical absorption spectrum and the

This work was supported by the Swiss National Science Foundation (SNSF) under grant No. 209358 (QuaTrEx). We acknowledge support from the Swiss National Supercomputing Centre (CSCS) under Project lp16.

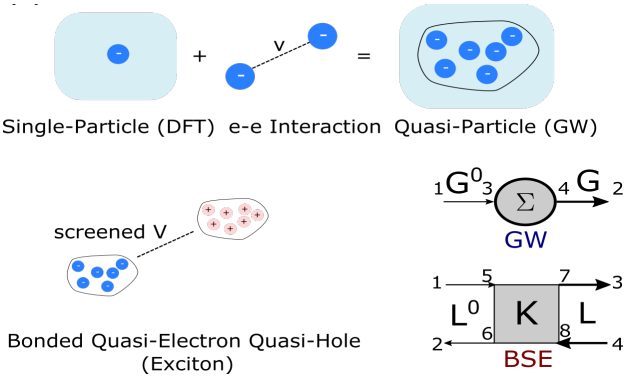

Fig. 1. Illustration of electron-electron (e-e) interactions and corresponding Feymann diagrams describing the Dyson equations for the GW correction and BSE.

electron energy loss spectrum (EELS), methods that explicitly include electron-hole interactions – e.g., the Bethe–Salpeter equation (BSE) on top of GW – are essential [2].

Several software packages, such as BerkleyGW [3], WEST [4], or the DFT code VASP [5], can solve the BSE on top of GW based on a plane-wave basis, with the plasmon-pole model or the full-frequency screening. Limited by the computational burden, the GW correction is usually done in single-shot (G_0W_0) or partially self-consistently (GW_0). The BSE takes the form of an eigenvalue problem for a two-particle effective Hamiltonian. The Tamm–Dancoff approximation is commonly imposed, which decouples the anti-resonant negative frequencies from the resonant positive frequencies, thus reducing the system size by half. The widely used VASP code conveniently integrates DFT, GW, and BSE capabilities within one framework. The WEST code is designed for large-scale GW calculations on thousands of atoms. It uses projective dielectric eigendecomposition (PDEP) to avoid explicit dielectric matrix storage of the empty states. A GPU-accelerated solution

of BSE has only recently been implemented in the WEST code for more than 1700 atoms [4].

The BSE is computationally very expensive, since it involves electron-hole pairs. Hence, the dimension of the BSE is approximately $N_{\mathrm{BSE}} = N_c \times N_v \times N_k$, where N_c (N_v) is the number of conduction (valence) bands and N_k the number of k points. If the BSE matrix is treated as dense, then the memory scales as $\mathcal{O}(N_{\mathrm{BSE}}^2)$, while the computation scales as $\mathcal{O}(N_{\mathrm{BSE}}^3)$ using full diagonalization. Practically, iterative solvers are used to reduce the cost when considering a few excitons.

Notably, none of the aforementioned existing codes can treat the BSE in devices driven out-of-equilibrium. They therefore fail at capturing, for example, the dissociation of excitons by an external voltage. For device-relevant system sizes, semi-classical models with rate equations are often employed to balance computational efficiency with physical fidelity. However, free (empirical) parameters are needed to fit experimental measurements. To shed light on exciton transport from first principles, we present here an extension of our previous device simulator [6], [7] capable of solving the NEGF-GW-BSE.

II. Method

In device simulation, a real-space approach is more suitable than a plane-wave (PW) representation. We thus convert PW-based DFT calculation into maximally localized Wannier functions (MLWF), obtaining a tight-binding-like lattice-periodic Hamiltonian H_{MLWF}, together with the associated Coulomb matrix V_{MLWF}. Then, we construct the device Hamiltonian H and Coulomb matrix V by scaling up H_{MLWF} and V_{MLWF}. These are the only inputs to our quantum transport solver. The NEGF-GW-BSE calculation workflow is described in Fig. 2.

In the first iteration, we follow Hedin's pentagon, ignoring the vertex contribution. This is the so-called GW approximation [8]. We successively compute the polarization P as

$$P(12) = -iG(12)G(21), \tag{1}$$

where G is the electron Green's function, while 1 and 2 are short-hand notation of space-time coordinates. This allows us to include dynamical effects into the dielectric screening. The Green's function of the screened Coulomb interaction is defined as

$$W(12) = V(12) + \int d(34)V(13)P(34)W(42). \tag{2}$$

By combining G and W, we obtain the self-energy Σ,

$$\Sigma(12) = iG(12)W(12), \tag{3}$$

which also gives the name of this approximation. We solve these GW equations in the Keldysh's non-equilibrium Green's function (NEGF) formalism, meaning that we need to solve the retarded, lesser and greater components of each Green's Function. These equations are defined more explicitly in our previous work [7].

In the second iteration, we enter the BSE solver. The central quantity is the two-electron correlation function

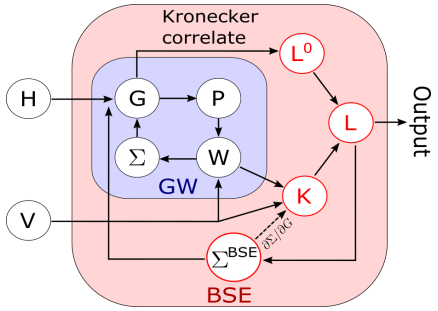

Fig. 2. Simulation workflow of NEGF+GW+BSE.

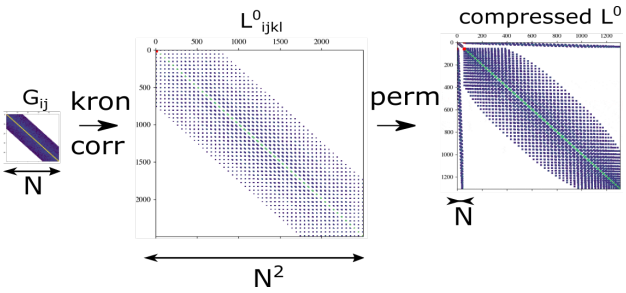

Fig. 3. Relationship between the G and L^0 matrices and their sparsity patterns.

$L(12;34)$, which is a Green's Function of 4 space-time coordinates. The non-interacting version can be simply constructed from the single-electron Green's function $L^0(12;34) = G(13)G(42)$. The polarization P is linked with L through $P(12) = L(11;22)$. Another quantity is the electron-hole interaction kernel $K(12,34)$, which includes a direct ($K_d(12;34) = -W(12)\delta(13)\delta(24)$) and exchange ($K_x(12;34) = V(13)\delta(12)\delta(34)$) term. The BSE is written in the integral form as Dyson equation

$$L(12;34) = L^0(12;34) + \int d(5678)L^0(12;56)K(56;78)L(78;34). \tag{4}$$

From L, we can extract the macroscopic dielectric function ε_M, which is an important linear-response quantity related to the optical absorption spectrum. We can also compute the self-energy more accurately

$$\Sigma^{\mathrm{BSE}}(12) = i \int d(34)G(13)W(14)\Gamma(34;2)d(34), \tag{5}$$

where Γ is the vertex function and $\Gamma(34;2) = L(34;22)$. Σ^{BSE} includes the vertex correction in Hedin's pentagon [2], thus accounting for correlation effects beyond the GW approximation. Then, we recalculate the Green's function G and iterate this loop. In this work, we stopped the calculation at this point due to the associated high computational intensity, but the self-consistent solution can be obtained by iterating the procedure until convergence.

Solving the BSE for a nano-device rapidly becomes prohibitively expensive due to the aforementioned unfavorable scaling. We adopt a cutoff truncation on the electron-electron

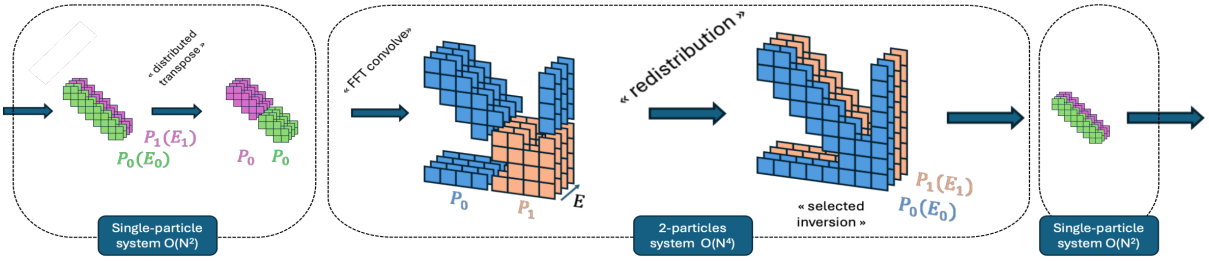

Fig. 4. Data distribution flow in our massively parallel BSE solver. P_0 and P_1 indicate processes. E is energy.

(e-e) interactions, which is motivated by their fast decay due to screening. This leads to a banded sparse G matrix of size N and bandwidth N_{diag}, as shown in Fig. 3. We represent the 4-D tensors in BSE as matrices by combining and flattening the first two and last two dimensions. As indicated in Fig. 3, after applying the cutoff, L^0 becomes highly sparse. We compute the sparsity pattern of BSE matrices based on the G matrix sparsity and memorize this information. Motivated by the relationship between P and L, and the K_d and K_x terms in K, we arrange the L^0 and L matrices into an exchange and direct part. $L = \begin{pmatrix} L_{xx} & L_{xd} \\ L_{dx} & L_{dd} \end{pmatrix}$, where the exchange part is defined by $L_{xx} = L_{ij,kl}\delta_{ij}\delta_{kl}$, similarly to K_x. The exchange part directly gives the (non-)interacting polarization since, by definition, $P_{ij} = -iL_{ii,jj}$. Therefore, we can selectively solve for the entries of L_{xx} and L_{dx} to obtain P and the vertex Γ, and avoid computing and storing the L_{dd} and L_{xd}. The L_{xx} term is of size N, thus orders of magnitude smaller than the L_{dd} term, which is of size $N \times (N-1)$, depicted in Fig. 3.

This permutation results in a compressed banded arrowhead (BA) matrix of bandwidth $\approx N_{\text{diag}}^2$, where N_{diag} is the bandwidth of G. All the zero entries of L^0 are permuted to the bottom right corner. The tip block on the top left corner corresponds to L_{xx}. To efficiently solve (4), we apply the recursive Green's function (RGF) algorithm, after special generalization for BA matrices [9]. The forward pass starts from the tail of the arrow and moves towards the tip. If we only need the tip block to compute P, then we can even stop at the end of the forward pass. If we need the vertex Γ, the backward pass is also needed to produce these entries. Such a selected solve approach reduces the complexity of the BSE calculation from $\mathcal{O}(N^6)$ for dense matrices to $\mathcal{O}(N \times N_{\text{diag}}^5)$ blocked BA ones.

Another major bottleneck lies in the computation of the non-interacting L^0 before solving the BSE. L^0 is obtained from an outer (Kronecker) product in the spatial coordinates and a correlation through energy. We apply the convolution theorem and transform the correlation into element-wise product in the Fourier space of energy, exploiting FFT for that purpose. This results in $\mathcal{O}(N_E \log(N_E))$ complexity rather than $\mathcal{O}(N_E^2)$, where N_E is the number of energies.

Due to the large number of non-zeros in L^0, we need to efficiently distribute its data and computation workload across multiple computing units. The data distribution flow is schematized in Fig. 4.

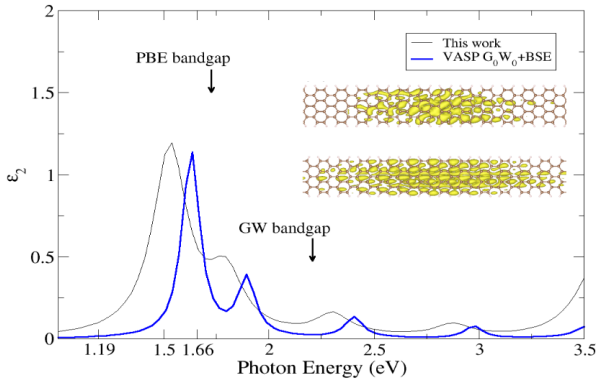

Fig. 5. Macroscopic dielectric function of the considered AGNR, as produced by VASP and by our solver (inset: lowest two exciton state wavefunctions).

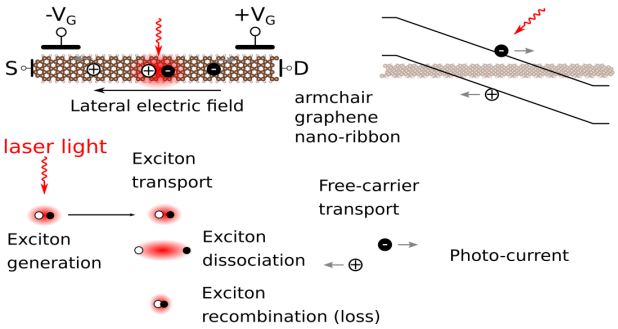

Fig. 6. Schematic of an AGNR photo-detector and physical processes involved in it.

III. Results

An AGNR-7 of length L=25.8 nm and containing 840 carbon atoms is considered as test example. The procedure starts with a PW DFT calculation of a representative unit cell made of 14 carbon atoms with edge passivation by hydrogen using the VASP code [5]. The results are converted into a set of MLWFs with Wannier90 [10] to construct the device H and V.

First, we compute the imaginary part of the macroscopic dielectric function (ε_2) with our implementation under flat-band equilibrium conditions and compare our result with a

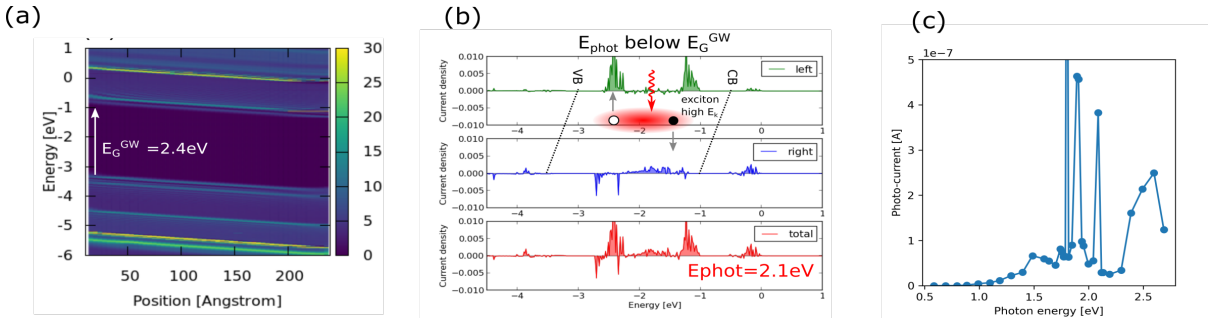

Fig. 7. (a) Local density-of-state of the simulated AGNR-7 when a linear potential drop is assumed. (b) Current density spectrum in the leads. (c) Photo-current vs. photon energy.

VASP G_0W_0 + BSE calculation in Fig. 5. Good agreement between both curves is achieved. ε_2 is proportional to the optical absorption. We observe two characteristic peaks below the GW QP bandgap, which originate from the excitonic resonances due to bound electron-hole pairs. The VASP ε_2 is overall slightly blue shifted with respect to our implementation. This comes from the difference in the QP bandgap. The PW basis used in VASP includes more bands than MLWFs, leading to a more accurate description of screening. However, the important excitonic resonances are correctly predicted with our implementation based on MLWFs, at a lower computational cost than VASP. The wavefunctions of the two lowest-energy excitonic states are computed and plotted together with the atomic structure in the inset. The second exciton state is more broader than the lowest one, as expected.

Next, a small linear potential drop and monochromatic illumination are applied to the AGNR-7 of length L=25.8 nm from before, as illustrated in Fig. 6. The optical absorption results in the generation of excitons. These excitons diffuse along the AGNR and dissociate into free electrons and holes under electric field, or annihilate without generating current. The local density-of-states (LDOS) accounting for the e-e interaction is plotted in Fig. 7(a). A band gap of 2.4 eV can be identified from the LDOS, which is larger than the PBE value and is close to the QP value obtained from VASP G_0W_0 calculations (2.42 eV). In addition, the energy difference between the first and second valence bands increases, consistent with previous report [1]. The energy spectrum of the photo-current collected at the left and right contact is displayed in Fig. 7(b) for photons with energy below the QP band gap. Non-zero spectral current density is observed in the QP bandgap. The band edge profiles along the AGNR are indicated by the two dashed lines in the figure, which are marked as CB (conduction band) and VB (valence band). This demonstrates the creation and dissociation of excitons that generate free carriers. Finally, we sweep the photon energy and record the photo-currents, as reported in Fig. 7(c). We observe current peaks at photon energies around 2 eV, below the QP bandgap and close to the second exciton resonance in the ε_2 shown in Fig. 5. The first exciton state has a larger binding energy, making it difficult to dissociate into free carriers.

IV. Conclusion

We developed an *ab initio* NEGF-GW-BSE solver and applied it to an AGNR photo-detector. In equilibrium (flat-band condition), we demonstrated good agreement with reference calculations from VASP and highlighted the presence of exciton transport and photo-current with below-bandgap photo-excitation in non-equilibrium. Our work opens the door for accurate investigations of devices exhibiting strong excitonic effects.

References

[1] R. Denk, M. Hohage, P. Zeppenfeld, J. Cai, C. A. Pignedoli, H. Söde, R. Fasel, X. Feng, K. Müllen, S. Wang, D. Prezzi, A. Ferretti, A. Ruini, E. Molinari, and P. Ruffieux, "Exciton-dominated optical response of ultra-narrow graphene nanoribbons," *Nature Communications*, vol. 5, no. 1, p. 4253, 2014.

[2] X. Blase, I. Duchemin, and D. Jacquemin, "The bethe–salpeter equation in chemistry: relations with td-dft, applications and challenges," *Chem. Soc. Rev.*, vol. 47, pp. 1022–1043, 2018.

[3] M. Rohlfing and S. G. Louie, "Electron-hole excitations and optical spectra from first principles," *Phys. Rev. B*, vol. 62, pp. 4927–4944, Aug 2000.

[4] V. W.-z. Yu, Y. Jin, G. Galli, and M. Govoni, "Gpu-accelerated solution of the bethe–salpeter equation for large and heterogeneous systems," *Journal of Chemical Theory and Computation*, vol. 20, pp. 10899–10911, 12 2024.

[5] G. Kresse and J. Furthmüller, "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set," *Phys. Rev. B*, vol. 54, pp. 11169–11186, Oct 1996.

[6] J. Cao, A. Ziogas, L. Deuschle, Q. Ding, N. Vetsch, A. Winka, V. Maillou, A. Maeder, and M. Luisier, "Ab initio quantum transport simulations of inas avalanche photo-diodes within the gw approximation," in *2023 International Electron Devices Meeting (IEDM)*, pp. 1–4, 2023.

[7] L. Deuschle, J. Cao, A. N. Ziogas, A. Winka, A. Maeder, N. Vetsch, and M. Luisier, "Electron-electron interactions in device simulation via nonequilibrium green's functions and the gw approximation," *Phys. Rev. B*, vol. 111, p. 195421, May 2025.

[8] M. Shishkin and G. Kresse, "Implementation and performance of the frequency-dependent *gw* method within the paw framework," *Phys. Rev. B*, vol. 74, p. 035101, Jul 2006.

[9] V. Maillou, L. Gaedke-Merzhaeuser, A. N. Ziogas, O. Schenk, and M. Luisier, "Serinv: A scalable library for the selected inversion of block-tridiagonal with arrowhead matrices," 2025.

[10] G. Pizzi, V. Vitale, R. Arita, S. Blügel, F. Freimuth, G. Géranton, M. Gibertini, D. Gresch, C. Johnson, T. Koretsune, J. Ibañez-Azpiroz, H. Lee, J.-M. Lihm, D. Marchand, A. Marrazzo, Y. Mokrousov, J. I. Mustafa, Y. Nohara, Y. Nomura, L. Paulatto, S. Poncé, T. Ponweiser, J. Qiao, F. Thöle, S. S. Tsirkin, M. Wierzbowska, N. Marzari, D. Vanderbilt, I. Souza, A. A. Mostofi, and J. R. Yates, "Wannier90 as a community code: new features and applications," *Journal of Physics: Condensed Matter*, vol. 32, p. 165902, jan 2020.