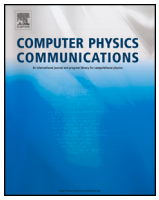
Computer Physics Communications 307 (2025) 109430

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc



Computational Physics

Fully coupled electron-phonon transport in two-dimensional-material-based devices using efficient FFT-based self-energy calculations ☆

Rutger Duflou [a,b,*], Gautam Gaddemane [a], Michel Houssa [a,b], Aryan Afzalian [a]

[a] imec, Kapeldreef 75, Leuven, 3001, Belgium
[b] Semiconductor Physics Laboratory, KU Leuven, Celestijnenlaan 200 D, Leuven, 3001, Belgium

ARTICLE INFO

Keywords:
Ab-initio
NEGF
2D materials
Self-heating

ABSTRACT

Self-heating can significantly degrade the performance in silicon nanoscale devices. In this work, the impact of self-heating is investigated in nanosheet transistors made of two-dimensional materials using ab-initio techniques. A new algorithm was developed to allow for efficient self-energy computations, achieving a ~500 times speedup. It is found that for the simple case of free-standing transition-metal dicalchogenides without explicit metal leads, electron-phonon scattering with room-temperature phonons dominates the device performance. For MoS_2, the effect of self-heating is negligible in comparison. For WS_2 and especially for WSe_2, self-heating effects demonstrate a further degradation of the ON-state current.

1. Introduction

The last two decades, two-dimensional (2D) materials have risen in interest as candidates for next-generation devices. They are predicted to show excellent electrostatic control, reducing short-channel effects, and to suffer less from device variation [1–4]. Ab-initio methods, such as Density Functional Theory (DFT) and the Non-Equilibrium Green's Function (NEGF) formalism [5], have been helpful tools in the research on 2D materials [1]. It has been shown that, to capture the correct behavior for 2D-material-based devices with NEGF, it is of great importance to incorporate electron-phonon interactions [5–7]. These dissipative simulations, however, usually assume electron-phonon interactions with phonons that are in equilibrium with a fixed temperature. Additionally, this fixed temperature is often assumed to be near room temperature. However, during device operation electron-phonon interactions also result in the generation of additional phonons, corresponding to a self-heating effect. These additional phonons can result in increased electron-phonon scattering, giving rise to hotspots with locally increased temperatures and device performance degradation. The simulation of such hotspots requires the incorporation of phonon transport simulations. These phonon transport simulations enable the computation of the phonon population through a balance of additional phonon creation and heat conduction through the device. A fully coupled scheme, including electron transport, phonon transport and the influence of electron-phonon interactions on both the electronic transport and the phonon population could thus be important to properly assess the performance of nanodevices [8]. It has been shown that in silicon devices, neglecting self-heating effects can result in an overestimation of the device performance [8,9]. A similar study for devices based on 2D materials has, to the authors' knowledge, not been performed. Our DFT-NEGF quantum transport solver, ATOMOS, allows for the simulation of devices with electron-phonon scattering by phonons at an equilibrium temperature [5,10] and for the simulation of ballistic phonon transport [11]. One aim of this work is to extend ATOMOS to allow for the simulation of fully coupled electron-phonon transport for 2D materials.

In the literature, coupled electron-phonon simulations are often at least partially based on the Boltzmann transport equation [8,12], not making use of the full quantum description as provided by the NEGF formalism. This is not surprising as fully coupled electron-phonon transport simulations using NEGF are usually characterized by exceedingly high computational costs [9], except for some very simple cases not corresponding to realistic materials [13,14]. These high computational costs are linked to the computation of the self-energy required to evaluate the electron-phonon scattering and the creation of additional phonons.

☆ The review of this paper was arranged by Prof. Weigel Martin.

* Corresponding author.
E-mail addresses: rutger.duflou@imec.be (R. Duflou), aryan.afzalian@imec.be (A. Afzalian).

https://doi.org/10.1016/j.cpc.2024.109430
Received 24 May 2024; Received in revised form 30 October 2024; Accepted 10 November 2024
Available online 15 November 2024

A second aim of this work is to present a new algorithm based on the Fast-Fourier-Transform (FFT) technique, which greatly reduces the computational cost of the self-energy calculation. This new algorithm could provide a way to more readily incorporate self-heating effects or even just regular electron-phonon scattering in future research on nanodevices with NEGF, while keeping the computation time tractable.

In Section 2, we discuss the theoretical foundations of our NEGF implementation of the electron and phonon Green's function. In Section 3, we describe the methods used to perform a device simulation and elaborate on the FFT-based self-energy calculation. In Section 4, we discuss the errors introduced by the approximations within this FFT-based self-energy computation and provide estimates for the gain in computational efficiency. Finally, in Section 5, we show the results of the simulation of a fully coupled 2D-material-based device with self-heating.

2. Theory

2.1. The DFT Hamiltonian

It can be shown that within the DFT formalism, a material is described by the following Hamiltonian [15],

$$\hat{H} = \sum_{n\mathbf{k}} e_{\mathbf{k}n}\hat{c}^{\dagger}_{\mathbf{k}n}\hat{c}_{\mathbf{k}n} + \sum_{\nu\mathbf{q}} \hbar\omega_{\mathbf{q}\nu}(\hat{a}^{\dagger}_{\mathbf{q}\nu}\hat{a}_{\mathbf{q}\nu} + \frac{1}{2}) + N_p^{-\frac{1}{2}} \sum_{\substack{\mathbf{kq}\\ mn\nu}} g_{mn\nu}(\mathbf{k},\mathbf{q})\hat{c}^{\dagger}_{\mathbf{k}+\mathbf{q}m}\hat{c}_{\mathbf{k}n}(\hat{a}_{\mathbf{q}\nu} + \hat{a}^{\dagger}_{-\mathbf{q}\nu}). \tag{1}$$

The description is in reciprocal space, with electronic band energies $e_{\mathbf{k}n}$, phonon energies $\hbar\omega_{\mathbf{q}\nu}$ and electron-phonon interaction parameters $g_{mn\nu}(\mathbf{k},\mathbf{q})$. Here, n (ν) denotes the band index (phonon mode) and $\mathbf{k}$ ($\mathbf{q}$) the k-point in the Brillouin zone for the electrons (phonons). $\hat{c}^{\dagger}_{\mathbf{k}n}$ and $\hat{c}_{\mathbf{k}n}$ ($\hat{a}^{\dagger}_{\mathbf{q}\nu}$ and $\hat{a}_{\mathbf{q}\nu}$) are the corresponding electron (phonon) creation and annihilation operators. However, these operators create and annihilate particles in reciprocal space. Device simulations typically require a real space description, which can be achieved by transforming the reciprocal space operators to real space, using a Wannier transformation [16],

$$\hat{c}^{\dagger}_{\mathbf{R}_e m} = \frac{1}{\sqrt{N_e}} \sum_{n\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{R}_e} U_{nm,\mathbf{k}} \hat{c}^{\dagger}_{\mathbf{k}n}, \tag{2}$$

where m denotes a Wannier function index, $\mathbf{R}_e$ denotes the primitive cell lattice point, N_e is the number of k-points and $U_{nm,\mathbf{k}}$ is a matrix built to maximize the real space localization of the electron. A similar real space transformation can be achieved for the localization of phonons [17],

$$\hat{a}^{\dagger}_{\mathbf{R}_p\kappa\alpha} = \frac{1}{\sqrt{N_p}} \sum_{\nu\mathbf{q}} e^{-i\mathbf{q}\cdot\mathbf{R}_p} e^{*}_{\kappa\alpha\nu,\mathbf{q}} \hat{a}^{\dagger}_{\mathbf{q}\nu}, \tag{3}$$

where $e_{\kappa\alpha\nu,\mathbf{q}}$ is the eigenvector of the dynamical matrix and κ, α and $\mathbf{R}_p$ denote an atom index, its polarization direction and its primitive cell lattice point, respectively. Note that there are minor differences compared to the conventions in Ref. [17].

In some cases, a mixed space description is beneficial. An exemplary case is a planar transistor, e.g., made from a 2D material. The transport direction requires a real space description to allow for the insertion of carriers at the source and their extraction at the drain. The out-of-plane direction, orthogonal to both the 2D material plane and the transport direction, is non-periodic and thus implies a real space description as well. The third direction, however, is typically very homogeneous and can thus be considered periodic. This periodicity allows for the subdivision of the system in a periodic part, which can be Fourier transformed to reciprocal space, and a remainder part, which is kept in real space. A schematic depiction for an hexagonal lattice is shown in Fig. 1.

This concept can be extended to other numbers of periodic directions, from 0 for nanowires to 2 for resistors or diodes. A more complete discussion of the transformations used is given in Appendix A. The final result is the following mixed space Hamiltonian

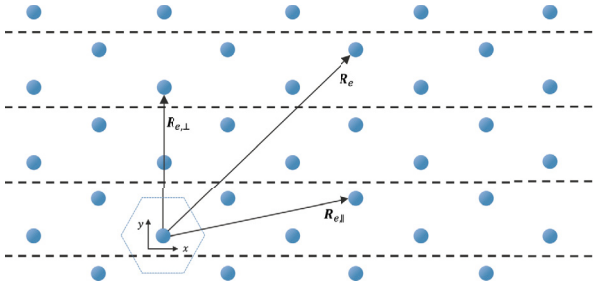


Fig. 1. Subdivision of the real space lattice in a periodic part orthogonal to the transport direction, which can be transformed to reciprocal space through a Fourier transformation, and a remainder part, which is kept in real space.

$$\hat{H} = \sum_{nn'\mathbf{k}_\perp} \bar{h}_{nn',\mathbf{k}_\perp} \hat{c}^{\dagger}_{\mathbf{k}_\perp n}\hat{c}_{\mathbf{k}_\perp n'} + \sum_{\nu\nu'\mathbf{q}_\perp} \bar{d}_{\nu\nu',\mathbf{q}_\perp} \hat{a}^{\dagger}_{\mathbf{q}_\perp\nu}\hat{a}_{\mathbf{q}_\perp\nu'} + N_\perp^{-\frac{1}{2}} \sum_{\substack{nn'\nu\\ \mathbf{k}_\perp\mathbf{q}_\perp}} \bar{g}_{nn'\nu,\mathbf{k}_\perp\mathbf{q}_\perp} \hat{c}^{\dagger}_{\mathbf{k}_\perp+\mathbf{q}_\perp n}\hat{c}_{\mathbf{k}_\perp n'}(\hat{a}_{\mathbf{q}_\perp\nu} + \hat{a}^{\dagger}_{-\mathbf{q}_\perp\nu}), \tag{4}$$

where the indices n and ν should not be confused with the band indices in (1), but represent a grouping of the indices $(m,\mathbf{R}_{e,\parallel})$ and $(\kappa,\alpha,\mathbf{R}_{p,\parallel})$ defined above and in Fig. 1. $N_\perp$ is equal to the number of orthogonal k-points. The third term in (4) represents the electron-phonon interactions and will be referred to as $\hat{H}_I$.

2.2. The NEGF formalism

The NEGF formalism relies on the definition of an electron and phonon Green's function [18,19]

$$i\underset{\mathbf{k}_\perp}{G}_{n,m}(t,t') = \frac{1}{\hbar}\left\langle T_c\left[\hat{c}_{\mathbf{k}_\perp n}(t)\hat{c}^{\dagger}_{\mathbf{k}_\perp m}(t')\right]\right\rangle, \tag{5}$$

$$i\underset{\mathbf{q}_\perp}{D}_{\nu,\mu}(t,t') = \frac{1}{\hbar}\left\langle T_c\left[(\hat{a}_{\mathbf{q}_\perp\nu}(t) + \hat{a}^{\dagger}_{-\mathbf{q}_\perp\nu}(t))(\hat{a}^{\dagger}_{\mathbf{q}_\perp\mu}(t') + \hat{a}_{-\mathbf{q}_\perp\mu}(t'))\right]\right\rangle, \tag{6}$$

where the creation and annihilation operators are given in the Heisenberg picture and are ordered on a two-branch contour. The averaging over the states is determined by a non-equilibrium occupation [19]. The indices n and m (ν and μ) can be understood as row and column indices, defining the Green's functions as matrices, $\boldsymbol{G}_{\mathbf{k}_\perp}(t,t')$ ($\boldsymbol{D}_{\mathbf{q}_\perp}(t,t')$). Although we refer to $\boldsymbol{D}_{\mathbf{q}_\perp}(t,t')$ as the phonon Green's function, according to its definition it is actually equal to the displacement-displacement correlation [15], but we will forgo this point for the sake of brevity.

The expressions in (5) and (6) cannot be solved exactly due to the electron-phonon interactions in $\hat{H}_I$ and due to the unknown occupation of states for devices not in equilibrium. A solution can be found through a perturbation expansion, of which the theory is well established. Here, we follow the derivation in Ref. [19]. The interacting system is subdivided into several non-interacting systems: the electrons and phonons in the device, a left lead and a right lead, as shown in Fig. 2. Additionally, each subsystem is divided into slabs such that every slab only interacts with its nearest neighbors [20]. Alternatively, one could interpret it as elements of (4) being grouped into matrices such that the total device Hamiltonian for the electron system and phonon system form two block tridiagonal matrices, respectively.

Each subsystem has a known one-particle occupation: the infinite leads are each characterized by Fermi-Dirac statistic functions for the electrons

$$f_i(\omega) = \frac{1}{\exp\frac{\hbar\omega - E_{f_i}}{k_B T_i} + 1} \tag{7}$$

and Bose-Einstein statistic functions for the phonons

$$N_i(\omega) = \frac{1}{\exp\frac{\hbar\omega}{k_B T_i} - 1} \tag{8}$$

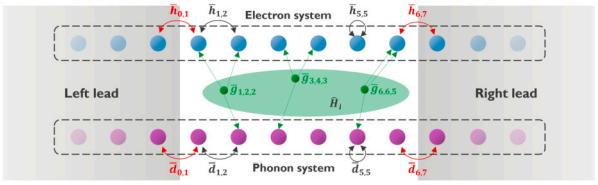


Fig. 2. Schematic representation of the coupled electron-phonon system within a device with two semi-infinite leads. The $\mathbf{k}_\perp$ and $\mathbf{q}_\perp$ subscripts are left out for the sake of brevity. Degrees of freedom of the electron and phonon system are grouped such that the on-site energies and coupling elements form matrices. The matrices that form the perturbation terms coupling the device to the left and right leads are denoted in red. The electron-phonon interactions are denoted in green. It should be noted that the representation here is only qualitatively true for phonons. A rigorous treatment can be found in the Supplemental Material [21]. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

with temperatures, T_1 and T_2, and chemical potentials for the electrons, E_{f_1} and E_{f_2}, for the left and right lead, respectively. The device itself can be kept empty before connection. The subsystems are connected by adiabatically switching on the interaction terms, $\hat{H}_I$, and the perturbation terms connecting the device to the leads. These perturbation terms can be grouped into matrices $\mathbf{U}_{\mathbf{k}_\perp}$ ($\mathbf{V}_{\mathbf{q}_\perp}$), related to $\{\bar{\mathbf{h}}_{i,j}, \bar{\mathbf{h}}_{j,i}\}$ ($\{\bar{\mathbf{d}}_{i,j}, \bar{\mathbf{d}}_{j,i}\}$) with $\mathbf{i}$, $\mathbf{j}$ equal to $\mathbf{0}$, $\mathbf{1}$ and $\mathbf{n}$, $\mathbf{n+1}$ [21]. Note that bold indices are used to indicate slab indices instead of individual degrees of freedom.

The switching on, in combination with Wick's theorem, leads to the following Dyson equations [19],

$$\mathbf{G}_{\mathbf{k}_\perp}(t,t') = \mathbf{G}^0_{\mathbf{k}_\perp}(t,t') + \int_C dt_1 \mathbf{G}^0_{\mathbf{k}_\perp}(t,t_1)\mathbf{U}_{\mathbf{k}_\perp}\mathbf{G}_{\mathbf{k}_\perp}(t_1,t') + \int_C dt_1 \int_C dt_2 \mathbf{G}^0_{\mathbf{k}_\perp}(t,t_1)\mathbf{\Sigma}^s_{\mathbf{k}_\perp}(t_1,t_2)\mathbf{G}_{\mathbf{k}_\perp}(t_2,t'), \tag{9}$$

$$\mathbf{D}_{\mathbf{q}_\perp}(t,t') = \mathbf{D}^0_{\mathbf{q}_\perp}(t,t') + \int_C dt_1 \mathbf{D}^0_{\mathbf{q}_\perp}(t,t_1)\mathbf{V}_{\mathbf{q}_\perp}\mathbf{D}_{\mathbf{q}_\perp}(t_1,t') + \int_C dt_1 \int_C dt_2 \mathbf{D}^0_{\mathbf{q}_\perp}(t,t_1)\mathbf{\Pi}^s_{\mathbf{q}_\perp}(t_1,t_2)\mathbf{D}_{\mathbf{q}_\perp}(t_2,t'), \tag{10}$$

where $\mathbf{G}^0_{\mathbf{k}_\perp}$ and $\mathbf{D}^0_{\mathbf{q}_\perp}$ are the Green's function solutions for the non-interacting non-connected subsystems, $\mathbf{U}_{\mathbf{k}_\perp}$ and $\mathbf{V}_{\mathbf{q}_\perp}$ are defined as above, $\mathbf{\Sigma}^s_{\mathbf{k}_\perp}$ and $\mathbf{\Pi}^s_{\mathbf{q}_\perp}$ are the self-energies related to electron-phonon scattering and the integrals are integrals over the two-branch contours. The contour-ordered Green's function can be resolved into lesser, greater, retarded and advanced Green's functions by confinement of the time arguments to specific branches and the contour integrals in (9) and (10) can be simplified to real axis integrals by using Langreth's theorem [19,22]. Finally, Fourier transformation of the integral equations to the energy domain results in the following well-known expressions for the electron Green's function [22,23],

$$\mathbf{G}^{R/A}_{\mathbf{k}_\perp}(\omega) = \left((\hbar\omega \pm i\eta)\mathbf{I} - \mathbf{H}_{\mathbf{k}_\perp} - \mathbf{\Sigma}^{R/A}_{\mathbf{k}_\perp}(\omega)\right)^{-1}, \tag{11}$$

$$\mathbf{G}^{\lessgtr}_{\mathbf{k}_\perp}(\omega) = \mathbf{G}^R_{\mathbf{k}_\perp}(\omega)\mathbf{\Sigma}^{\lessgtr}_{\mathbf{k}_\perp}(\omega)\mathbf{G}^A_{\mathbf{k}_\perp}(\omega), \tag{12}$$

with $\mathbf{G}^R_{\mathbf{k}_\perp}$, $\mathbf{G}^A_{\mathbf{k}_\perp}$, $\mathbf{G}^<_{\mathbf{k}_\perp}$ and $\mathbf{G}^>_{\mathbf{k}_\perp}$ the retarded, advanced, lesser and greater electron Green's function of the device, respectively, and $\mathbf{\Sigma}^{R/A}_{\mathbf{k}_\perp}$ and $\mathbf{\Sigma}^{\lessgtr}_{\mathbf{k}_\perp}$ their corresponding self-energies. $\mathbf{H}_{\mathbf{k}_\perp}$ is the device Hamiltonian with

$$\left(\mathbf{H}_{\mathbf{k}_\perp}\right)_{i,j} = \bar{h}_{ij}\,_{\mathbf{k}_\perp}. \tag{13}$$

Note that the matrices defined here only contain degrees of freedom within the device and not the degrees of freedom in the leads as before. The influence of the leads is introduced by the self-energies. The self-energies thus contain contributions from both the leads and the electron-phonon interactions

$$\mathbf{\Sigma} = \mathbf{\Sigma}^l + \mathbf{\Sigma}^s \tag{14}$$

where we dropped the ω and $\mathbf{k}_\perp$ dependency in the notation for the sake of brevity and where

$$\mathbf{\Sigma}^{l,R/A}_{\mathbf{1,1}} = \bar{\mathbf{h}}_{\mathbf{1,0}}\mathbf{G}^{0,R/A}_{\mathbf{0,0}}\bar{\mathbf{h}}_{\mathbf{0,1}}, \tag{15}$$

$$\mathbf{\Sigma}^{l,R/A}_{\mathbf{n,n}} = \bar{\mathbf{h}}_{\mathbf{n,n+1}}\mathbf{G}^{0,R/A}_{\mathbf{n+1,n+1}}\bar{\mathbf{h}}_{\mathbf{n+1,n}}, \tag{16}$$

$$\mathbf{\Sigma}^{l,<} = if_1\mathbf{\Gamma}^l_1 + if_2\mathbf{\Gamma}^l_2, \tag{17}$$

$$\mathbf{\Sigma}^{l,>} = -i(1-f_1)\mathbf{\Gamma}^l_1 - i(1-f_2)\mathbf{\Gamma}^l_2, \tag{18}$$

with

$$\left(\mathbf{\Gamma}^l_1\right)_{\mathbf{1,1}} = i\left(\mathbf{\Sigma}^{l,R}_{\mathbf{1,1}} - \mathbf{\Sigma}^{l,A}_{\mathbf{1,1}}\right), \tag{19}$$

$$\left(\mathbf{\Gamma}^l_2\right)_{\mathbf{n,n}} = i\left(\mathbf{\Sigma}^{l,R}_{\mathbf{n,n}} - \mathbf{\Sigma}^{l,A}_{\mathbf{n,n}}\right). \tag{20}$$

The retarded and advanced Green's function in the non-interacting non-connected leads, $\mathbf{G}^{0,R/A}$, are readily computed using the Sancho-Rubio algorithm [24].

Similar expressions can be found for the phonon Green's function [25,9],

$$\mathbf{D}^{R/A}_{\mathbf{q}_\perp}(\omega) = \left((\hbar^2\omega^2 \pm i\eta)\mathbf{I} - \mathbf{K}_{\mathbf{q}_\perp} - \mathbf{\Pi}^{R/A}_{\mathbf{q}_\perp}(\omega)\right)^{-1}, \tag{21}$$

$$\mathbf{D}^{\lessgtr}_{\mathbf{q}_\perp}(\omega) = \mathbf{D}^R_{\mathbf{q}_\perp}(\omega)\mathbf{\Pi}^{\lessgtr}_{\mathbf{q}_\perp}(\omega)\mathbf{D}^A_{\mathbf{q}_\perp}(\omega), \tag{22}$$

with $\mathbf{K}_{\mathbf{q}_\perp}$ the Fourier transform of $\mathbf{\Phi}$, the rescaled interatomic force constants matrix,

$$\mathbf{\Phi}_{i,j} = \frac{\hbar^2}{\sqrt{m_i m_j}}\frac{\partial^2 U}{\partial\tau_i\partial\tau_j}. \tag{23}$$

Here, U denotes the internal energy, the index i indicates a degree of freedom in the real space phonon system, i.e., an atom with mass m_i with a polarization direction along which it is displaced over a distance τ_i.

Similarly to the electron system, the effects of the leads and electron-phonon interactions are introduced through the self-energy

$$\mathbf{\Pi} = \mathbf{\Pi}^l + \mathbf{\Pi}^s \tag{24}$$

with

$$\mathbf{\Pi}^{l,R/A}_{\mathbf{1,1}} = \bar{\mathbf{k}}_{\mathbf{1,0}}\mathbf{D}^{0,R/A}_{\mathbf{0,0}}\bar{\mathbf{k}}_{\mathbf{0,1}}, \tag{25}$$

$$\mathbf{\Pi}^{l,R/A}_{\mathbf{n,n}} = \bar{\mathbf{k}}_{\mathbf{n,n+1}}\mathbf{D}^{0,R/A}_{\mathbf{n+1,n+1}}\bar{\mathbf{k}}_{\mathbf{n,n+1}}, \tag{26}$$

and

$$\mathbf{\Pi}^{l,<} = -iN_1\mathbf{\Delta}^l_1 - iN_2\mathbf{\Delta}^l_2, \tag{27}$$

$$\mathbf{\Pi}^{l,>} = -i(N_1+1)\mathbf{\Delta}^l_1 - i(N_2+1)\mathbf{\Delta}^l_2, \tag{28}$$

with

$$\left(\mathbf{\Delta}^l_1\right)_{\mathbf{1,1}} = i\left(\mathbf{\Pi}^{l,R}_{\mathbf{1,1}} - \mathbf{\Pi}^{l,A}_{\mathbf{1,1}}\right), \tag{29}$$

$$\left(\mathbf{\Delta}^l_2\right)_{\mathbf{n,n}} = i\left(\mathbf{\Pi}^{l,R}_{\mathbf{n,n}} - \mathbf{\Pi}^{l,A}_{\mathbf{n,n}}\right). \tag{30}$$

The similarity between (11)-(20) and (21)-(30) readily allows for the adaptation of electronic transport codes to phonon transport, as was done in Ref. [11]. However, the derivation of these expressions for the phonon system in the literature typically relies on different conventions and does not apply the same principles as used for the electron system [25,26]. This discrepancy complicates linking the electron and phonon Green's function for the self-energy calculation. Additionally, expressions are usually obtained for full real space [9,25,26] or reciprocal space [15], neglecting mixed space, which is useful for devices. We

therefore provide a derivation of the Green's function expressions provided above and their corresponding self-energies in Appendix B and Appendix C, respectively. The final results for the lesser and greater self-energies due to electron-phonon scattering are

$$\mathbf{\Sigma}^{s,\lessgtr}_{\mathbf{k}_\perp}(\omega) = \int_0^{+\infty} \frac{2i\hbar}{N_\perp} \sum_{\nu\mu\mathbf{q}_\perp} \mathbf{M}^{\nu}_{\mathbf{k}_\perp-\mathbf{q}_\perp,\mathbf{q}_\perp} \Big(\mathbf{G}^{\lessgtr}_{\mathbf{k}_\perp-\mathbf{q}_\perp}(\omega-\omega') D^{\lessgtr}_{\substack{\nu,\mu \\ \mathbf{q}_\perp}}(\omega') + \mathbf{G}^{\lessgtr}_{\mathbf{k}_\perp-\mathbf{q}_\perp}(\omega+\omega') D^{\gtrless}_{\substack{\mu,\nu \\ -\mathbf{q}_\perp}}(\omega') \Big) \mathbf{M}^{\mu}_{\mathbf{k}_\perp,-\mathbf{q}_\perp} \frac{d\omega'}{2\pi}, \quad (31)$$

$$\left(\mathbf{\Pi}^{s,\lessgtr}_{\mathbf{q}_\perp}(\omega)\right)_{\nu,\mu} = \int_{-\infty}^{+\infty} -\frac{2n_s i\hbar}{N_\perp} \times \sum_{\mathbf{k}_\perp} \mathrm{Tr}\left(\mathbf{M}^{\nu}_{\mathbf{k}_\perp,-\mathbf{q}_\perp} \mathbf{G}^{\lessgtr}_{\mathbf{k}_\perp}(\omega') \mathbf{M}^{\mu}_{\mathbf{k}_\perp-\mathbf{q}_\perp,\mathbf{q}_\perp} \mathbf{G}^{\gtrless}_{\mathbf{k}_\perp-\mathbf{q}_\perp}(\omega'-\omega) \right) \frac{d\omega'}{2\pi}, \quad (32)$$

where n_s is a spin degeneracy factor and Tr() denotes the trace. The matrix elements of $\mathbf{M}^{\nu}_{\mathbf{k}_\perp,\mathbf{q}_\perp}$ are given by

$$\left(\mathbf{M}^{\nu}_{\mathbf{k}_\perp,\mathbf{q}_\perp}\right)_{n,n'} = \bar{\bar{g}}_{\substack{nn'\nu \\ \mathbf{k}_\perp\mathbf{q}_\perp}}, \quad (33)$$

which are rescaled versions of the matrix elements in (4), as detailed in Appendix C. The row and column indices of $\mathbf{M}^{\nu}_{\mathbf{k}_\perp,\mathbf{q}_\perp}$ thus correspond to electron degrees of freedom and the matrix multiplications and trace operator in (31) and (32) are effectively summations over electron degrees of freedom. The phonon degrees of freedom cannot be readily linked to matrix multiplications and are thus written out explicitly. The total set of operations can, however, still be considered as tensorial products with $\mathbf{M}^{\nu}_{\mathbf{k}_\perp,\mathbf{q}_\perp}$, summing over both electron and phonon degrees of freedom.

The retarded and advanced self-energies are calculated from the following relation [22],

$$\mathbf{\Sigma}^{s,R/A}_{\mathbf{k}_\perp}(\omega) = \mathcal{P}\int_{-\infty}^{+\infty} \frac{\mathbf{\Gamma}^{s}_{\mathbf{k}_\perp}(\omega')}{\omega-\omega'} \frac{d\omega'}{2\pi} \mp \frac{i}{2}\mathbf{\Gamma}^{s}_{\mathbf{k}_\perp}(\omega), \quad (34)$$

$$\mathbf{\Pi}^{s,R/A}_{\mathbf{q}_\perp}(\omega) = \mathcal{P}\int_{-\infty}^{+\infty} \frac{\mathbf{\Delta}^{s}_{\mathbf{q}_\perp}(\omega')}{\omega-\omega'} \frac{d\omega'}{2\pi} \mp \frac{i}{2}\mathbf{\Delta}^{s}_{\mathbf{q}_\perp}(\omega), \quad (35)$$

with

$$\mathbf{\Gamma}^{s}_{\mathbf{k}_\perp}(\omega) = i\left(\mathbf{\Sigma}^{s,>}_{\mathbf{k}_\perp}(\omega) - \mathbf{\Sigma}^{s,<}_{\mathbf{k}_\perp}(\omega)\right), \quad (36)$$

$$\mathbf{\Delta}^{s}_{\mathbf{q}_\perp}(\omega) = i\left(\mathbf{\Pi}^{s,>}_{\mathbf{q}_\perp}(\omega) - \mathbf{\Pi}^{s,<}_{\mathbf{q}_\perp}(\omega)\right), \quad (37)$$

although the first terms in (34) and (35) are usually left out as the principal value integral is difficult to compute and merely results in an energy renormalization [22].

With the lesser and greater Green's functions available, we can compute the electron density [10], the current and the electron and phonon heat current [9] as

$$n_k = -\frac{in_s\hbar}{N_\perp} \sum_{\mathbf{k}_\perp} \int_{-\infty}^{+\infty} G^{<}_{\substack{k,k \\ \mathbf{k}_\perp}} \frac{d\omega}{2\pi}, \quad (38)$$

$$I_{i\to j} = -\frac{qn_s}{N_\perp} \sum_{\mathbf{k}_\perp} \int_{-\infty}^{+\infty} (\bar{h}_{\substack{ij \\ \mathbf{k}_\perp}} G^{<}_{\substack{j,i \\ \mathbf{k}_\perp}} - \bar{h}_{\substack{ji \\ \mathbf{k}_\perp}} G^{<}_{\substack{i,j \\ \mathbf{k}_\perp}}) \frac{d\omega}{2\pi}, \quad (39)$$

$$J_{el,i\to j} = \frac{n_s}{N_\perp} \sum_{\mathbf{k}_\perp} \int_{-\infty}^{+\infty} \hbar\omega(\bar{h}_{\substack{ij \\ \mathbf{k}_\perp}} G^{<}_{\substack{j,i \\ \mathbf{k}_\perp}} - \bar{h}_{\substack{ji \\ \mathbf{k}_\perp}} G^{<}_{\substack{i,j \\ \mathbf{k}_\perp}}) \frac{d\omega}{2\pi}, \quad (40)$$

$$J_{ph,i\to j} = \frac{1}{N_\perp} \sum_{\mathbf{q}_\perp} \int_0^{+\infty} \hbar\omega(\bar{k}_{\substack{ij \\ \mathbf{q}_\perp}} D^{<}_{\substack{j,i \\ \mathbf{q}_\perp}} - \bar{k}_{\substack{ji \\ \mathbf{q}_\perp}} D^{<}_{\substack{i,j \\ \mathbf{q}_\perp}}) \frac{d\omega}{2\pi}. \quad (41)$$

3. Methods

3.1. Material parameter extraction

The purpose of this work is to evaluate self-heating effects for devices based on conventional 2D materials. We therefore choose to focus on experimentally mature 2D materials, such as the transition-metal dicalchogenides (TMD) MoS_2, WS_2 and WSe_2. In contrast to graphene, these materials naturally demonstrate a bandgap without the need for confinement to nanoribbons. Additionally, this choice readily allows us to compare our results with previous work [6]. Another conventional 2D material evaluated in Ref. [6] is black phosphorus, but this material is predicted to suffer from strong current degradation due to electron-phonon scattering even for the case of equilibrium phonons at room temperature. We therefore limit ourselves to the three TMDs above in their most stable form, the 2H phase. For the sake of brevity, the model testing and error analysis was only performed for MoS_2, limiting our discussion of WS_2 and WSe_2 to their device performance with self-heating included.

The electronic band energies, phonon energies and electron-phonon matrix elements of all materials were extracted in reciprocal space using the QUANTUM ESPRESSO DFT code [27]. The structure was relaxed with an energy convergence criteria of 1×10^{-16} Ry between subsequent scf iteration steps and energy and force convergence criteria of 5×10^{-7} Ry and 5×10^{-6} Ry Bohr^{-1}, respectively, between subsequent ionic optimization steps. The PBE exchange-correlation functional was used with ultrasoft pseudopotentials, an energy cutoff of 70 Ry, 80 Ry and 90 Ry for MoS_2, WS_2 and WSe_2, respectively, and a k-mesh density of $16\times16\times1$. The latter two were set after a convergence test to see that this correspond to a relative energy variation of less than 1×10^{-6}, an absolute energy variation less than 1 mRy/atom and a variation of the lattice constant of less than 0.02 %. The obtained lattice constant for MoS_2 is 3.183 Å, which differs from the experimental result of 3.165 Å [28] due to the lack of van der Waals corrections. Relaxation with the Grimme DFT-D3 van der Waals correction [29] resulted in a lattice constant of 3.166 Å. However, this correction resulted in strong oscillations of the deformation potentials as a function of $\mathbf{q}$, which was deemed unrealistic. A vacuum of 15 Å and out-of-plane screening were used to block interactions between different layers. Spin-orbit coupling was neglected in all simulations as it significantly increases the computational cost of the NEGF simulations. The same k-mesh was used for the phonon calculation with a convergence threshold of 1×10^{-17}.

The reciprocal space parameters were converted to real space using the Wannier90 [16] and Perturbo [30] code. For the initial projections during the Wannierization process, 5 d-orbitals on Mo and W and 3 p-orbitals on S and Se were used. Perturbo provides the Hamiltonian elements in the Wannier basis, the interatomic force constants and atomic masses, which are readily combined to form $\mathbf{\Phi}$ in (23), and the real space deformation potentials, all in the HDF5 format [30]. To retrieve the matrix elements of (33) in real space, an additional transformation is required,

$$\tilde{\bar{g}}_{\substack{mn\kappa\alpha \\ \mathbf{R}_e,\mathbf{R}_p}} = \frac{\hbar}{\sqrt{2m_\kappa}} g^{Perturbo}_{mn\kappa\alpha}(\mathbf{R}_e,\mathbf{R}_p). \quad (42)$$

The details are provided in Appendix A and Appendix B.

3.2. Device simulation

The matrix elements extracted in the previous section are grouped into device Hamiltonians for both the electrons and phonons. These Hamiltonians are then Fourier transformed to mixed space by the ATOMOS quantum transport solver. 10 k-points were used for half of the mixed space Brillouin zone. The other half can be considered equal due to symmetry. The device is a dual-gate transistor, depicted in Fig. 3 for the case of MoS_2. The 42 nm long TMD sheet consists of source and drain extension regions, doped with a carrier concentration of 1.8×10^{13} cm^{-2},

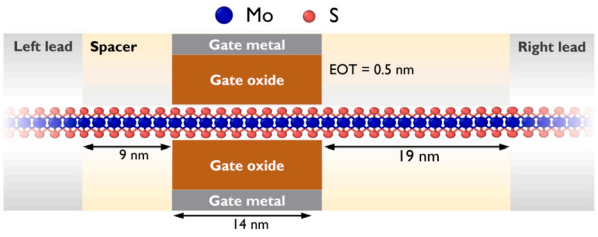


Fig. 3. Schematic representation of a MoS_2 dual-gate transistor with its dimensions.

and a 14 nm intrinsic channel region between a top and bottom gate. The doping is n-type for MoS_2 and WS_2 and p-type for WSe_2, consistent with their most conventional intrinsic doping type. The drain extension region is elongated compared to the source extension region to allow for better investigation and visualization of the thermalization of the carriers. Both gates have a corresponding gate oxide of 2 nm and relative permittivity of 15.6 for an effective oxide thickness of 0.5 nm. The bias between source and drain was set to 0.3 V. Unless specified otherwise, the source-gate potential was set to 0.6 V, corresponding to ON-state.

To verify the correctness of the matrix elements, the band structure and phonon dispersion are shown for MoS_2 in Fig. 4 (a) and (b), respectively. The deformation potentials as a function of $\mathbf{q}$ for $\mathbf{k}$ at Γ as provided by Perturbo are denoted with full lines for each phonon branch in Fig. 4 (c). However, incorporation of all the matrix elements required to reproduce Fig. 4 (c) in (31) and (32) results in computations that are prohibitively expensive. Indeed, self-energy computations often require neglecting certain matrix elements to keep the computation tractable [9]. In this work, only on-site interactions are considered, i.e., within (31) and (32), $\mathbf{\Sigma}^{\lessgtr}_{\mathbf{k}_\perp}$, $\mathbf{\Pi}^{\lessgtr}_{\mathbf{q}_\perp}$, $\mathbf{G}^{\lessgtr}_{\mathbf{k}_\perp}$, $\mathbf{D}^{\lessgtr}_{\mathbf{q}_\perp}$ and $\mathbf{M}^{\nu}_{\mathbf{k}_\perp,\mathbf{q}_\perp}$ are assumed to be diagonal matrices and the diagonal entries of $\mathbf{M}^{\nu}_{\mathbf{k}_\perp,\mathbf{q}_\perp}$ are only nonzero if the corresponding Wannier functions are located on the atom corresponding to ν. An estimate of the influence due to this approximation is obtained by computing deformation potentials with the same approximations, i.e., reciprocal space deformation potentials are computed according to the principles in Ref. [17], but matrix elements which would be neglected in our diagonal approach in (31) and (32) are set to zero in the computation of the deformation potentials as well. The resulting approximate deformation potentials are denoted by the dashed curves in Fig. 4 (c). One can clearly see that neglecting the non-local interactions has a large influence on the deformation potentials. First, the average deformation potential is significantly smaller due to neglecting non-local scattering processes. Second, the deformation potentials do not show any dispersion.

Concerning the decrease in average deformation potential, we have compensated for this in our computation by rescaling the on-site matrix elements with a scaling factor c,

$$\mathbf{M}^{\nu}_{\mathbf{k}_\perp,\mathbf{q}_\perp} \leftarrow c\, \mathbf{M}^{\nu}_{\mathbf{k}_\perp,\mathbf{q}_\perp}. \tag{43}$$

It was found that $c = 7.409$, 7.792 and 7.240 results in the same average deformation potentials as for the case when all interactions are included for MoS_2, WS_2 and WSe_2, respectively. This approach should ensure that the average strength of the electron-phonon interaction is preserved despite our approximation. Using the on-site matrix elements provides an attempt to preserve the relative sensitivity of different Wannier functions to atomic displacements. We would like to note that compensation for neglecting off-diagonal elements in electron-phonon scattering by introducing a scaling factor has been shown to reproduce correct device performance [31,32].

Concerning the lack of momentum dependence of the deformation potentials in our approximation, we claim that this does not severely affect our results, which we will verify when comparing our results with previous work in Section 5. There is one notable exception. The acoustic phonons in Fig. 4 (b) demonstrate a zero in the phonon energies

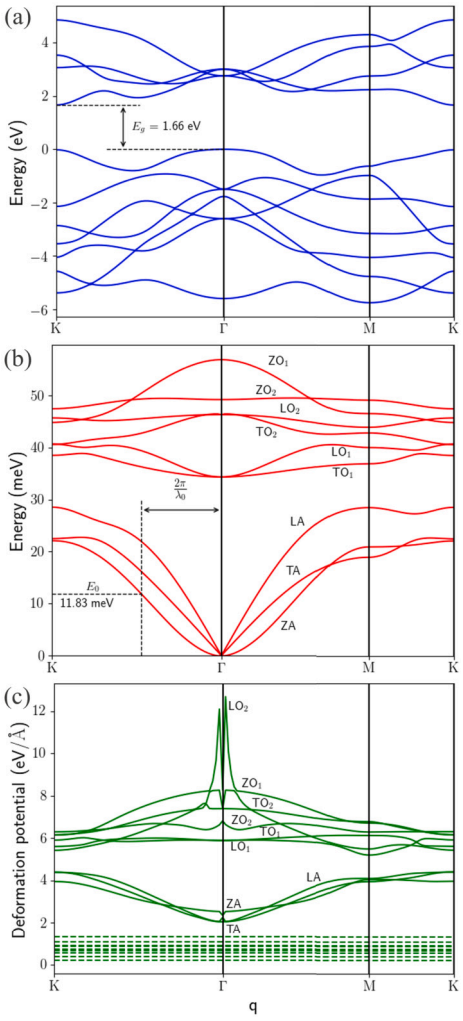


Fig. 4. Wannier interpolated band structure (a) and phonon dispersion (b) of MoS_2 obtained by ATOMOS using the matrix elements provided by Perturbo. The band gap energy, E_g, threshold energy for acoustic phonon mode damping, E_0, and the different phonon branches are denoted. (c) The deformation potentials for each phonon branch as a function of $\mathbf{q}$ for $\mathbf{k}$ at Γ as provided by Perturbo (full lines) and after removing interaction parameters to reduce the computational complexity in ATOMOS (dashed lines).

when $\mathbf{q}$ is at the Γ-point. This is expected as acoustic phonons at the Γ-point correspond to a mere lattice displacement. However, this zero phonon energy results in a singularity in the phonon Green's function due to both the displacement and the Bose-Einstein distribution function tending towards infinity for zero-energy phonons [33]. This singularity is negated by a zero in the deformation potential for intraband transitions. The deformation potentials for the acoustic phonons in Fig. 4 (c) do not demonstrate such a zero as the plotted deformation potentials are an average over all intraband and interband transitions. Differentiating between interband and intraband transitions results in the deformation potentials shown in Fig. 5.

It can be seen that the deformation potentials in Fig. 5 demonstrate abrupt jumps when varying $\mathbf{q}$. These jumps are related to electronic bands crossing at $\mathbf{k}+\mathbf{q}$, obfuscating the difference between intraband and interband transitions. The discussion should therefore be limited for $\mathbf{q}$ near the Γ-point. The intraband deformation potentials demonstrate zeros for 6 phonon branches at the Γ-point, including for the 3 acoustic phonon branches. The interband transitions do not demonstrate this symmetry. However, their contribution to the self-energy is prohibited by energy conservation. The difference in energy between different electronic bands and the fact that the phonon energy is zero

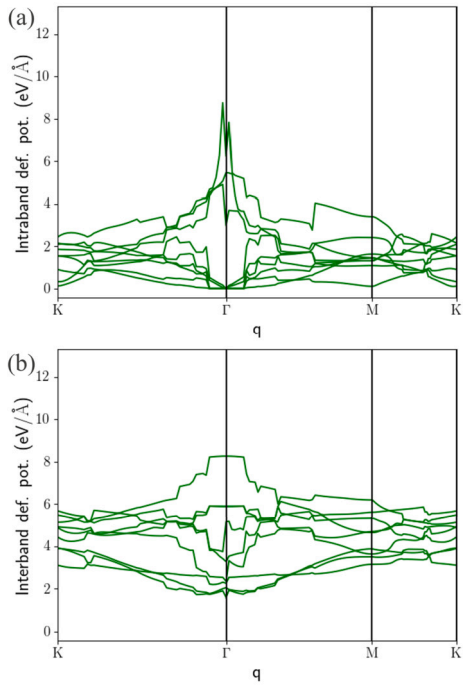


Fig. 5. The deformation potentials for each phonon branch of MoS_2 as a function of $\mathbf{q}$ for $\mathbf{k}$ at Γ as provided by Perturbo for intraband (a) and interband (b) transitions.

for acoustic phonons for $\mathbf{q}$ at the Γ-point asserts that these transitions do not happen. It is thus clear that, while one could claim that the momentum dependency of the deformation potentials in Fig. 4 (c) is not very strong except for the LO_2 phonons, the momentum dependency is essential for negating the singularity in the phonon Green's function for intraband transitions. However, the matrix elements between Wannier functions do not allow for distinguishing between intraband and interband transitions as every Wannier function has contributions from every band. Indeed, employing our approximation of the matrix elements without further consideration results in unstable simulations due to the singularity of the phonon Green's function at the Γ-point. A complete description including all matrix elements in (31) and (32) should preserve the negation of the singularity. However, including all matrix elements is prohibitively expensive and, in contrast to other work where the form of the Hamiltonian naturally provides a self-energy expression that negates the phonon singularity [9], it is unclear which matrix elements are essential.

A full comparison of different approximations is outside the scope of this work and we keep it for future research. Here, we choose to negate the singularity by damping the phonon Green's function at low energies. This is similar to how one can resolve the divergence of the scattering rate for out-of-plane transverse acoustic mode (ZA) phonons in 2D materials lacking out-of-plane mirror symmetry [33]. In this approach, ZA phonons are stiffened for wavelengths above a certain threshold λ_0. This threshold wavelength can be related to a threshold energy E_0 as indicated in Fig. 4 (b). Here, we choose a threshold $\lambda_0 = 1\,\text{nm}$, resulting in a value $E_0 = 11.83\,\text{meV}$, $10.26\,\text{meV}$ and $8.77\,\text{meV}$ for MoS_2, WS_2 and WSe_2, respectively. Concerning the damping of the phonon Green's function, we choose for the least intrusive procedure that negates the singularity. This is achieved by having a damping factor equal to 0 for $\hbar\omega = 0$ and equal to 1 for $\hbar\omega = E_0$. Additionally, we choose for a smooth transition between damped and undamped behavior, i.e., the slope of the damping factor was chosen to be zero at E_0. These requirements result in the following damping scheme

$$\mathbf{D}^{\lessgtr}_{\mathbf{q}_\perp}(\omega) \leftarrow \mathbf{D}^{\lessgtr}_{\mathbf{q}_\perp}(\omega)\left(1 - \left(1 - \frac{\hbar\omega}{E_0}\right)^2\right) \text{ for } \hbar\omega < E_0 \tag{44}$$

3.3. FFT-based implementation of the self-energy calculation

3.3.1. Premise

Despite the approximations made concerning the sparsity of the matrices in (31) and (32), the evaluation of these expressions still introduces a significant computational cost in the calculation. The energy integrals in Section 2.2 are evaluated by evaluating the Green's function on an energy grid and can hence be computed in $\mathcal{O}(N_E N_\perp)$ time, where N_E denotes the number of energy points of the energy grid. Hence, (31) and (32) also need to be evaluated $N_E N_\perp$ times, but a single evaluation itself scales as $\mathcal{O}(N_E N_\perp)$ due to the fact that the self-energies depend on the Green's function at all energies and k-points. The total cost of the self-energy calculation thus scales as $\mathcal{O}(N_E^2 N_\perp^2)$. Indeed, for the number of energy grid points and k-points required to converge to a sufficiently accurate result, a few hundreds and about 10, respectively, the computational cost of the self-energy calculation dwarfs the cost of the Green's function evaluation.

We propose an alternative to direct evaluation of (31) and (32). As $\mathbf{M}^{\nu}_{\mathbf{k}_\perp,\mathbf{q}_\perp}$ is independent of ω, (31) and (32) are essentially convolutions of two energy dependent functions, which is expressed on a discrete energy grid as

$$\Sigma^{\lessgtr}[k] = \sum_l G^{\lessgtr}[k-l]D^{\lessgtr}[l] + G^{\lessgtr}[k+l]D^{\gtrless}[l], \tag{45}$$

$$\Pi^{\lessgtr}[k] = \sum_l G^{\lessgtr}[l]G^{\gtrless}[l-k]. \tag{46}$$

For the sake of clarity, we dropped the subscripts in the notation and the multiplications are actually tensorial products involving the $\mathbf{M}^{\nu}_{\mathbf{k}_\perp,\mathbf{q}_\perp}$ matrix. However, it should be noted that this statement is entirely general and does not rely on our simplification on the matrices $\mathbf{\Sigma}^{\lessgtr}_{\mathbf{k}_\perp}$, $\mathbf{\Pi}^{\lessgtr}_{\mathbf{q}_\perp}$, $\mathbf{G}^{\lessgtr}_{\mathbf{k}_\perp}$, $\mathbf{D}^{\lessgtr}_{\mathbf{q}_\perp}$ and $\mathbf{M}^{\nu}_{\mathbf{k}_\perp,\mathbf{q}_\perp}$ in the discussion above. (45) and (46) merely rewrite (31) and (32) in a more dense format compatible with discrete energy grids, irrespective of which approximations are introduced on the matrices involved.

Convolutions of series with significant kernels can be evaluated efficiently by Fourier transforming both series, performing an element-wise multiplication and performing an inverse Fourier transform to obtain the final result. However, this technique requires that the energy grids on which the different functions are evaluated, are identical. Additionally, these grids are required to be equidistant. Both of these requirements are typically not met for (31) and (32). Shifts in energy are of no consequence as the integrals in (31) and (32) do not directly depend on the energy. However, the range of the energy windows that the electron and phonon Green's functions are evaluated on is usually also different, which typically results in different grid spacings. For instance, the energy window for the phonons of MoS_2 is usually no larger than 65 meV, but for the electrons the energy window can be 10 to 30 times larger depending on the potential in the device. Assuming an equidistant grid of $N_E = 1000$ points, an ON-state simulation of the device in Fig. 3 typically results in an energy grid spacing of 1 meV for electrons and 0.06 meV for phonons. Additionally, the Green's functions are characterized by Van Hove singularities, which require a dense energy grid to be evaluated accurately [34]. ATOMOS applies an adaptive grid strategy to locally refine the energy grid near singularities [35]. This allows the integrals in (38)-(41) to be evaluated efficiently and to extremely high accuracy without increasing the computational cost unnecessarily by also having a dense energy mesh where the Green's functions are smooth. This adaptive grid strategy, however, also implies that the energy grids are usually not equidistant.

The obstacles of having non-equidistant grids over different energy window sizes could be resolved by refining the energy step size everywhere in the electron and phonon energy grid to its most refined part and by extending the smaller energy window, usually the phonon energy window, to the larger energy window size. This would, however, result in much higher computation times and memory requirements. The extra

computation time related to evaluating additional energy points could be reduced by linearly interpolating the Green's functions between its evaluations on the nearest energy points. Indeed, interpolation is also used for direct evaluation of (31) and (32) [36]. The self-energy is calculated for every ω on the energy grid and uses every ω' on the energy grid. As the energy grids are not necessarily equal and equidistant, $\omega - \omega'$, $\omega + \omega'$ and $\omega' - \omega$ are not necessarily on the energy grid. However, as the energy grid is refined to capture all features of the Green's functions, it can be assumed that the Green's function can be obtained at these intermediate energies by interpolating the Green's function between its neighboring energy grid points. Likewise, extending the energy grid for the Green's function with the smaller energy window size is readily achieved by padding with zeros.

Interpolation and padding can thus reduce the cost of evaluating the Green's function on a large dense grid. However, storing these interpolated and padded Green's functions still gives rise to significant increases of the memory footprint and the Fourier transform and the evaluation of the Fourier transformed self-energies on a large dense grid still results in inflated computation times.

3.3.2. Difference in energy window

Let us first focus on the general difference in energy window sizes for the electron and phonon Green's function. For this, we assume that both Green's functions are evaluated on an equidistant grid with an equal number of energy points N_E. As mentioned above, the difference in relevant energy window sizes implies that despite the equal number of energy points, both grids are not equal. Without loss of generality, we can state that the electron energy window is a factor m larger than the phonon energy window, where m can be made an integer by increasing either energy window slightly if necessary. The discussion above indicates that generally $m \approx 17$. The conventional implementation of the self-energy computation is then given by Algorithm 1. The double sum gives rise to the quadratic time complexity. The unequal energy grids requires an interpolation step.

Algorithm 1 Conventional self-energy computation.

for k in $0..N_E$-1 **do**
  $\Sigma^{\lessgtr}[k] \leftarrow 0$
  for l in $0..N_E$-1 **do**
    $G^{\lessgtr}_{temp} \leftarrow \texttt{Interp}(G^{\lessgtr}[\lceil k - \frac{l}{m} \rceil], G^{\lessgtr}[\lfloor k - \frac{l}{m} \rfloor])$
    $\Sigma^{\lessgtr}[k] \leftarrow \Sigma^{\lessgtr}[k] + G^{\lessgtr}_{temp} D^{\lessgtr}[l]$
    $G^{\lessgtr}_{temp} \leftarrow \texttt{Interp}(G^{\lessgtr}[\lceil k + \frac{l}{m} \rceil], G^{\lessgtr}[\lfloor k + \frac{l}{m} \rfloor])$
    $\Sigma^{\lessgtr}[k] \leftarrow \Sigma^{\lessgtr}[k] + G^{\lessgtr}_{temp} D^{\gtrless}[l]$
  end for
  $\Pi^{\lessgtr}[k] \leftarrow 0$
  for l in $0..N_E$-1 **do**
    $G^{\lessgtr}_{temp} \leftarrow \texttt{Interp}(G^{\gtrless}[\lceil l - \frac{k}{m} \rceil], G^{\lessgtr}[\lfloor l - \frac{k}{m} \rfloor])$
    $\Pi^{\lessgtr}[k] \leftarrow \Pi^{\lessgtr}[k] + G^{\lessgtr}[l] G^{\gtrless}_{temp}$
  end for
end for

To enable an FFT-based computation of the self-energy computation, the energy grids must be made equal. The electron Green's function is thus interpolated $m-1$ times between every pair of evaluated energy points. The phonon Green's function is extended by appending $(m-1)N_E$ zeros. This is demonstrated in Algorithm 2 and shown for the evaluation of the first term in (45) in a schematic way in Fig. 6 (a) for $m = 3$.

As stated above, this results in a significant increase of the memory footprint and the time complexity of the self-energy computation. Now, mN_E instead of N_E Green's functions need to be stored. The Fourier transform and inverse Fourier transform is required on a grid of size mN_E, having time complexity $\mathcal{O}(mN_E \log(mN_E))$. The evaluation of the Fourier transformed self-energy has time complexity $\mathcal{O}(mN_E)$, which is arguably better than $\mathcal{O}(N_E^2)$ for the conventional convolution-based implementation, but still significant due to the tensorial product involving the $\mathbf{M}^{\nu}_{\mathbf{k}_\perp,\mathbf{q}_\perp}$ matrix.

Algorithm 2 Naive FFT-based self-energy computation.

for k in $0..mN_E - 1$ **do**
  $G^{\lessgtr}_{interp}[k] \leftarrow \texttt{Interp}(G^{\lessgtr}[\lceil \frac{k}{m} \rceil], G^{\lessgtr}[\lfloor \frac{k}{m} \rfloor])$
end for
for k in $N_E..mN_E - 1$ **do**
  $D^{\lessgtr}[k] \leftarrow 0$
end for
$g^{\lessgtr}[k] \leftarrow \texttt{FFT}(G^{\lessgtr}_{interp}[k])$ $\quad (k = 0..mN_E - 1)$
$d^{\lessgtr}[k] \leftarrow \texttt{FFT}(D^{\lessgtr}[k])$ $\quad (k = 0..mN_E - 1)$
for k in $1..mN_E - 1$ **do**
  $\sigma^{\lessgtr}[k] \leftarrow g^{\lessgtr}[k] d^{\lessgtr}[k] + g^{\lessgtr}[k] d^{\gtrless}[-k]$
  $\pi^{\lessgtr}[k] \leftarrow g^{\lessgtr}[k] g^{\gtrless}[-k]$
end for
$\Sigma^{\lessgtr}[k] \leftarrow \texttt{iFFT}(\sigma^{\lessgtr}[k])$ $\quad (k = 0..mN_E - 1)$
$\Pi^{\lessgtr}[k] \leftarrow \texttt{iFFT}(\pi^{\lessgtr}[k])$ $\quad (k = 0..mN_E - 1)$

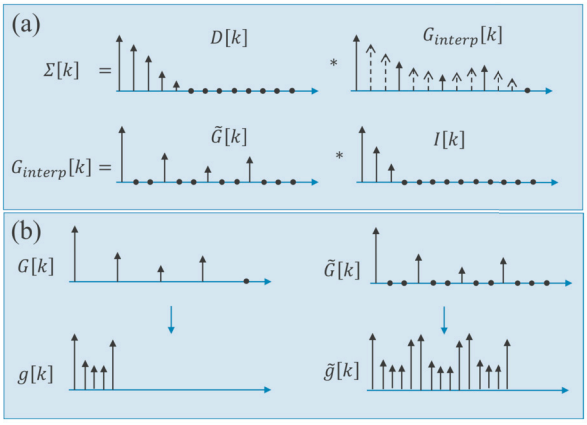


Fig. 6. (a) A schematic showing how the first term in the electron self-energy is a convolution of the phonon Green's function and the interpolated Green's function and how the interpolated Green's function is itself a convolution of the original Green's function evaluations only, $\tilde{G}[k]$, and a function $I[k]$. (b) A schematic showing how $\tilde{g}[k]$, the Fourier transform of $\tilde{G}[k]$, is a repetition of the Fourier transform of $G[k]$ on the original electron energy mesh.

We can alleviate these high computational requirements by performing the interpolation of the electron Green's function implicitly, as indicated by the second part of Fig. 6 (a). The interpolated Green's function can namely be written as a convolution of the original Green's function evaluations with intermediate zeros for the interpolation points, which we'll refer to as $\tilde{G}[k]$, and the function $I[k]$, with

$$I[k] = \begin{cases} 1 - \frac{|k|}{m} & \text{if } |k| <= m \\ 0 & \text{else} \end{cases} \tag{47}$$

assuming zero-based numbering. The first term in the self-energy in (45) is thus the result of a double convolution. The double sum in this double convolution can be reduced to an element-wise multiplication by Fourier transforming $\tilde{G}[k]$, $D[k]$ and $I[k]$ into $\tilde{g}[k]$, $d[k]$ and $i[k]$,

$$\begin{aligned} \Sigma[k] &= \sum_{l,k}^{mN_E} \tilde{G}[k-l-h] D[l] I[h] \\ &= \sum_{l,k}^{mN_E} \sum_{k'}^{mN_E} \frac{1}{mN_E} \tilde{g}[k'] e^{-\frac{2\pi i (k-l-h)k'}{mN_E}} \\ &\quad \sum_{l'}^{mN_E} \frac{1}{mN_E} d[l'] e^{-\frac{2\pi i l l'}{mN_E}} \sum_{h'}^{mN_E} \frac{1}{mN_E} i[h'] e^{-\frac{2\pi i h h'}{mN_E}} \\ &= \frac{1}{mN_E} \sum_{k'} \tilde{g}[k'] d[k'] i[k'] e^{-\frac{2\pi i k k'}{mN_E}}, \end{aligned} \tag{48}$$

where we used the identity

$$\sum_{l}^{mN_E} e^{-\frac{2\pi i l(l'-k')}{mN_E}} = mN_E\delta_{l'k'}. \tag{49}$$

Furthermore, as denoted in Fig. 6 (b), $\tilde{g}[k]$ merely consists of repetitions of $g[k]$, the Fourier transform of $G[k]$ on the original electron energy mesh. Additionally, the electron Green's function is only evaluated on every m'th grid point. $\Sigma[k]$ is therefore only required on every m'th grid point. These considerations allow for a further simplification,

$$\begin{aligned}\Sigma[mk] &= \frac{1}{mN_E}\sum_{k'}^{mN_E}\tilde{g}[k']d[k']i[k']e^{\frac{-2\pi imkk'}{mN_E}}\\ &= \frac{1}{mN_E}\sum_{k'}^{N_E}\sum_{n}^{m}\tilde{g}[k'+nN_E]d[k'+nN_E]i[k'+nN_E]e^{\frac{-2\pi ikk'}{N_E}}\\ &= \frac{1}{N_E}\sum_{k'}^{N_E}\tilde{g}[k']\tilde{d}[k']e^{\frac{-2\pi ikk'}{N_E}},\end{aligned} \tag{50}$$

with

$$\tilde{d}[k'] = \frac{1}{m}\sum_{n}^{m} d[k'+nN_E]i[k'+nN_E]. \tag{51}$$

The final line in (50) indicates that the first-term contribution to the electron self-energy in (45) can be computed as the inverse Fourier transform of a series of N_E elements. These N_E elements are the result of a product between the first N_E terms of $\tilde{g}[k]$, which can be obtained by Fourier transforming the original electron Green's function evaluations on a grid of N_E energy points, and a modified Fourier transformed phonon Green's function $\tilde{d}[k]$. We have thus bypassed having to store and Fourier transform mN_E electron Green's function as storage and Fourier transformation of N_E electron Green's functions is sufficient. Additionally, the expensive tensorial product involving the $\mathbf{M}^{\nu}_{\mathbf{k}_\perp,\mathbf{q}_\perp}$ matrix only has to be performed N_E times instead of mN_E times, albeit with a modified Fourier transformed phonon Green's function $\tilde{d}[k]$.

Computing $\tilde{d}[k]$ still involves mN_E multiplications, as indicated by (51). Additionally, obtaining the required $d[k]$ and $i[k]$ involves two Fourier transforms on a grid with mN_E points, once for $D[k]$ and once for $I[k]$. However, the mN_E multiplications in (51) are regular multiplications instead of expensive tensorial products as the values of $i[k]$ are mere numbers. Likewise, the computational cost of Fourier transforming and storing $I[k]$ is negligible due to the series consisting of mere numbers instead of Green's functions for a complete device. The only significant computational cost inflated by a factor m is therefore the Fourier transformation and storage of a phonon Green's function padded with zeros to a size of mN_E. Note, however, that $d[k]$ is never used directly, but only to compute $\tilde{d}[k]$, which has only N_E entries. As (51) involves regular multiplications, this can be done separately for every matrix element of $d[k]$. One can thus consider every matrix element $D_{i,j}[k]$ separately, pad it with zeros, Fourier transform it to $d_{i,j}[k]$, and use it to compute $\tilde{d}_{i,j}[k]$, before continuing to the next matrix element. There is therefore never a need to store mN_E complete Green's function matrices, implying that the memory footprint also need not be inflated by the factor m. The only increase in computational cost by having different energy windows is due to the fact that Fourier transforms of the phonon Green's function have to be performed on a grid with mN_E energy points.

This discussion is entirely general and can readily be extended to the second-term contribution to the electron self-energy in (45). The result is

$$\Sigma^{\lessgtr}[k] = \frac{1}{N_E}\sum_{k'}^{N_E}\left(\tilde{g}^{\lessgtr}[k']\tilde{d}^{\gtrless}[k'] + \tilde{g}^{\lessgtr}[k']\tilde{d}^{\gtrless}[-k']\right)e^{\frac{-2\pi ikk'}{N_E}}, \tag{52}$$

where k is an index on the original electron energy mesh.

A similar approach is possible for the phonon self-energy. (46) contains the electron Green's function twice. In the direct evaluation scheme, $G[l]$ corresponds to values on the original electron mesh and only $G[l-k]$ is interpolated,

$$\Pi[k] = \sum_{l,n}^{mN_E}\tilde{G}[l]\tilde{G}[l-k-n]I[n]. \tag{53}$$

This, however, treats the two electron Green's functions asymmetrically. Alternatively, one can interpolate both electron Green's functions on the refined mesh,

$$\Pi[k] = \frac{1}{m}\sum_{l,h,n}^{mN_E}\tilde{G}[l-h]\tilde{G}[l-k-n]I[h]I[n]. \tag{54}$$

The resulting expressions are

$$\Pi^{\lessgtr}[k] = \frac{1}{mN_E}\sum_{k'}^{mN_E}\left(\tilde{g}^{\lessgtr}[k']\tilde{g}^{\gtrless}[-k']i[k']\right)e^{\frac{-2\pi ikk'}{N_E}}, \tag{55}$$

and

$$\Pi^{\lessgtr}[k] = \frac{1}{m^2N_E}\sum_{k'}^{mN_E}\left(\tilde{g}^{\lessgtr}[k']\tilde{g}^{\gtrless}[-k']i[k']^2\right)e^{\frac{-2\pi ikk'}{N_E}}, \tag{56}$$

respectively.

As discussed above, the required functions $\tilde{g}[k]$ can be obtained with a Fourier transform of the original electron Green's function on an energy grid of size N_E. The expressions in (55) and (56) express a need for expensive tensorial products involving the $\mathbf{M}^{\nu}_{\mathbf{k}_\perp,\mathbf{q}_\perp}$ matrix on a grid with size mN_E. $\tilde{g}[k]$ is, however, periodic and only has N_E different entries. This implies that the expensive product will also only have N_E different entries and, hence, only needs to be computed N_E times.

$i[k]$ is not periodic and the multiplication with $i[k]$ needs to be performed mN_E times, but since $i[k]$ is a number, this multiplication is not expensive. The result of the multiplication with $i[k]$ needs to be stored and Fourier transformed to the phonon self-energy on an energy grid of mN_E points. Similarly to the creation of $\tilde{d}[k]$, this can be done for each matrix element separately. One can thus first compute the N_E entries of $\tilde{\pi}[k] = g^{\lessgtr}[k]g^{\gtrless}[-k]$, consider a series of single matrix elements $\tilde{\pi}_{i,j}[k]$, transform it to a grid of size mN_E by periodically multiplying it with $i[k]$, Fourier transform the result to $\Pi_{i,j}[k]$ and remove all but the first N_E entries before continuing to the next matrix element. This last step is allowed as the phonon Green's function is only evaluated on the first N_E grid points of the dense large mesh. There is therefore never a need to store mN_E complete self-energy matrices, confirming that the memory footprint need not be inflated by the factor m.

The considerations above are summarized in Algorithm 3, demonstrating that for equidistant grids, an efficient FFT-based calculation of the self-energies is possible, even when the electron and phonon energy mesh have significantly different ranges. We would like to note that none of these considerations depend on our simplifications of the matrices $\mathbf{\Sigma}^{\lessgtr}_{\mathbf{k}_\perp}$, $\mathbf{\Pi}^{\lessgtr}_{\mathbf{q}_\perp}$, $\mathbf{G}^{\lessgtr}_{\mathbf{k}_\perp}$, $\mathbf{D}^{\lessgtr}_{\mathbf{q}_\perp}$ and $\mathbf{M}^{\nu}_{\mathbf{k}_\perp,\mathbf{q}_\perp}$ in the discussion in Section 3.2. The assumed sparsity of these matrices does allow for a cheaper evaluation of the involved tensorial products, but otherwise has no effect on the analysis in this section.

Some additional considerations must be made, however. As a linear convolution is desired, both the electron and phonon Green's function must be padded to avoid a cyclic convolution. Additionally, this discussion is entirely general for the convolution of interpolated functions and need not be limited to the specific case of electron and phonon Green's functions and self-energies. The computation of $\Sigma^{\lessgtr}[k]$ from $D^{\lessgtr}[k]$ and $G^{\lessgtr}[k]$ in Algorithm 3 is readily extended to the case of obtaining a coarse grid convolution of a function evaluated on a dense grid and a function evaluated on the coarse grid and interpolated to the dense grid. Likewise, the computation of $\Pi^{\lessgtr}[k]$ from $G^{\lessgtr}[k]$ and $G^{\gtrless}[k]$ in Algorithm 3 is readily extended to the case of obtaining a dense grid

Algorithm 3 Final FFT-based self-energy computation.

$g^{\lessgtr}[k] \leftarrow \texttt{FFT}(G^{\lessgtr}[k])$ $(k = 0..N_E - 1)$
$i[k] \leftarrow \texttt{FFT}(I[k])$ $(k = 0..mN_E - 1)$
for i,j in $0..N_{dof} - 1$ **do**
  for k in $N_E..mN_E - 1$ **do**
    $D^{\lessgtr}_{i,j}[k] \leftarrow 0$
  end for
  $d^{\lessgtr}_{i,j}[k] \leftarrow \texttt{FFT}(D^{\lessgtr}_{i,j}[k])$ $(k = 0..mN_E - 1)$
  for k in $0..N_E - 1$ **do**
    $\tilde{d}^{\lessgtr}_{i,j}[k] \leftarrow 0$
    for n in $0..m - 1$ **do**
      $\tilde{d}^{\lessgtr}_{i,j}[k] \leftarrow \tilde{d}^{\lessgtr}_{i,j}[k] + \frac{d^{\lessgtr}_{i,j}[k+nN_E]i[k+nN_E]}{m}$
    end for
  end for
end for
for k in $0..N_E - 1$ **do**
  $\sigma^{\lessgtr}[k] \leftarrow g^{\lessgtr}[k]\tilde{d}^{\lessgtr}[k] + g^{\lessgtr}[k]\tilde{d}^{\gtrless}[-k]$
  $\tilde{\pi}^{\lessgtr}[k] \leftarrow g^{\lessgtr}[k]g^{\gtrless}[-k]$
end for
$\Sigma^{\lessgtr}[k] \leftarrow \texttt{iFFT}(\sigma^{\lessgtr}[k])$ $(k = 0..N_E - 1)$
for i,j in $0..N_{dof} - 1$ **do**
  for k in $0..N_E - 1$ **do**
    for n in $0..m - 1$ **do**
      $\pi^{\lessgtr}_{i,j}[k + nN_E] \leftarrow \tilde{\pi}^{\lessgtr}_{i,j}[k]i[k + nN_E]$
    end for
  end for
  $\Pi^{\lessgtr}_{i,j}[k] \leftarrow \texttt{iFFT}(\pi^{\lessgtr}_{i,j}[k])$ $(k = 0..mN_E - 1)$
end for

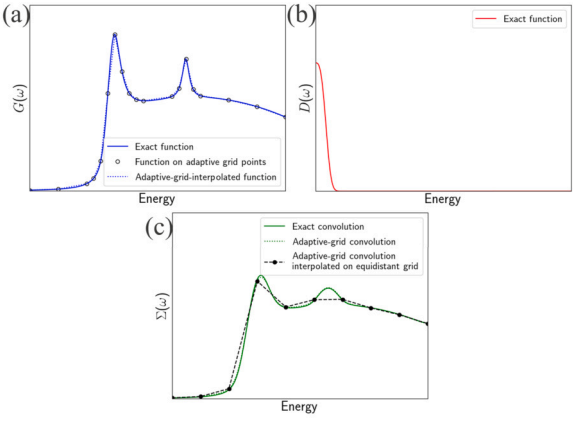


Fig. 7. Calculation of the self-energy for a toy-problem set of Green's functions. (a) shows an exemplary electron Green's function, its adaptive grid evaluations and the function interpolating these adaptive grid evaluations. (b) shows the phonon Green's function and (c) shows the convolution according to (45), both for the exact electron Green's function and the function interpolating the adaptive grid evaluations. Finally, it also shows the effect of interpolating the self-energy grid on an equidistant grid instead of direct evaluation on intermediate points.

convolution of two functions that are evaluated on a much coarser grid but interpolated on the dense grid for the convolution. Finally, during Fourier transformation of the energy mesh, Green's function entries of vastly different size are mixed, e.g., the electron Green's function in the band gap is mixed with the Green's function in the conduction band for an n-type transistor. After self-energy calculation and inverse Fourier transformation, the self-energy in the band gap will thus have a machine precision error relative to the self-energy in the conduction band, which can be many orders of magnitude higher than the self-energy in the band gap. We will refer to this error on the self-energy as the energy mixing error.

3.3.3. Non-equidistant grids

The limitation concerning non-equidistant energy grids is not resolved as readily. A possible solution is to abandon the adaptive grid strategy and apply a globally dense equidistant energy grid. However, it is shown in Appendix D that even for a relatively fine equidistant grid of $N_E = 1000$ grid points per k-point, the adaptive grid still tends to add a considerable number of additional points, implying that the integration error would otherwise still be significant. This result might be slightly inflated by the fact that it is based on ballistic data and that electron-phonon scattering tends to partially smoothen out the very sharp van Hove singularities. However, according to the authors' experience, even with electron-phonon scattering, the number of adaptively added grid points can be significant. Additionally, we would like to employ the full power of ATOMOS' adaptive grid strategy for optimal computational efficiency. We thus investigate the influence of using the full adaptive grid for the Green's function evaluations, while using an equidistant grid for the self-energy computation in order to enable the FFT-based implementation.

Fig. 7 shows an example of a non-equidistant energy grid being used to resolve two peaks in the electron Green's function. The function interpolating the adaptive grid evaluations is nearly superimposed on the function itself. The phonon Green's function has a significantly smaller energy window of relevance and is sampled on a dense grid, which is appended with zeros. The convolution of the two according to (45), effectively results in a smoothing of the electron Green's function. The results for the exact electron Green's function and the function interpolating the adaptive grid points are nearly identical. However, Fig. 7 shows that even if a methodology based on an equidistant grid could reproduce these results, an error will still be introduced as the results are only obtained on this equidistant grid. For intermediate points, interpolation of the self-energy is required, which does not properly capture the features of the self-energy at all energies. We will refer to this error as the self-energy interpolation error.

Additionally, when an equidistant grid is used instead of an adaptive grid, the computation of the self-energy is usually susceptible to errors, even on the equidistant grid points itself. This is illustrated in Fig. 8. Two strategies for reduction of the adaptive grid to an equidistant grid are shown. The first approach merely leaves out any grid refinements in the adaptive grid. The effect is that the total mass of certain peaks can be overestimated, underestimated, or even completely missed, with the corresponding effect on the self-energy. The second approach divides the energy window in sections, one for each equidistant grid point. The Green's function of each section is determined as the weighted average of the evaluations on the adaptive grid points in that section. This better preserves the total mass of features in the Green's function, and hence, the self-energy. However, the shapes of features in the energy profile of the self-energy are altered. We will refer to the error introduced by converting the Green's function evaluations on an adaptive grid to an equidistant grid, as the Green's function conversion error.

The influence of these three types of errors, the energy mixing error, the self-energy interpolation error and the Green's function conversion error, on macroscopic properties such as the current, heat current and charge, are the focus of Section 4.1.

4. Model testing

4.1. Error estimate of FFT-based self-energy calculation

In Section 3, we introduced definitions for different types of errors made by the approximations in the FFT-based self-energy calculation. They are summarized here as

- The energy mixing error, arising during Fourier transformation because the Green's function at all energies in the energy window are mixed. Some of these entries are orders of magnitude larger than others. Machine precision errors on a large-value Green's function,

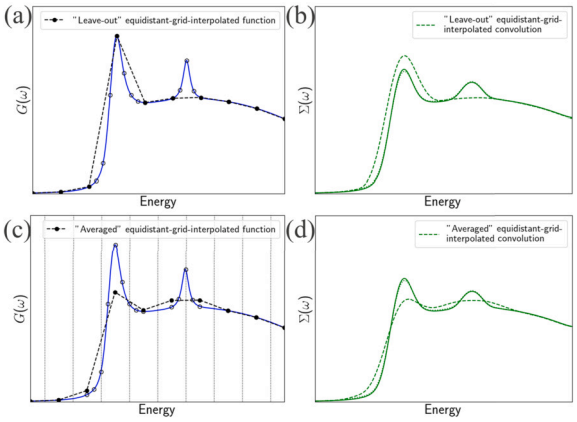


Fig. 8. The effect on the self-energy of using an electron Green's function approximation interpolating an equidistant mesh. The equidistant mesh can be achieved by leaving out any intermediate adaptive grid points (a) or by averaging the nearest adaptive grid points (c). The effects on the self-energy are shown in (b) and (d), respectively.

e.g., at the bottom of the conduction band, thus result in large relative errors on small-value Green's functions, e.g., in the middle of the band gap. The same is true for the self-energies during the inverse Fourier transformation. In the conventional implementation of (31) and (32), the Green's functions are only mixed with entries relatively close in energy due to the limited energy window of the phonon Green's function. In the FFT-based implementation, all energies are mixed, which can amplify this error.

- The self-energy interpolation error, arising because the self-energy computed by the FFT-based implementation is only provided on an equidistant grid. The adaptive grid also requires the self-energy on intermediate points, which are obtained by interpolation.
- The Green's function conversion error, arising because the Green's function evaluations on a non-equidistant adaptive grid need to be converted to an equidistant grid for Fourier transformation. Fig. 8 showed two ways of achieving this: either by leaving out any evaluations on adaptively added grid points or by averaging the adaptive grid evaluations.

These errors have to be compared with the integration error, introduced by performing the integral in (38)-(41) using a finite set of integration points. An estimate of this integration error can be obtained by comparing the obtained results with the results for a strongly refined energy grid. The integration error is not related to the FFT-based implementation of the self-energy calculation, but is always present, even with ATOMOS' adaptive grid. Additionally, an estimate of the integration error determines how the adaptive grid is refined. ATOMOS starts from an initial equidistant grid, which is then further refined in order to reduce the error below a certain threshold. Here, we impose a relative error threshold of at most 1% on certain macroscopic parameters such as the electron current, phonon heat current and electronic charge density. However, a minimal initial grid density is required for the adaptive grid refinement to work. We show in Appendix D that this minimal initial grid density corresponds to ∼100 equidistant initial grid points. We also show that a significantly denser equidistant grid is required if we desire the errors introduced by the FFT-based computation to be of the same magnitude as the imposed threshold.

The energy mixing error is several orders of magnitude smaller than the integration error for all energy grid densities. The self-energy interpolation error and the Green's function conversion error require an equidistant grid of ∼1000 and ∼500 grid points, respectively, to achieve a relative error of 1% on the macroscopic electronic parameters. Additionally, the "averaged" approach should be used to convert the Green's functions on the non-equidistant grid to an equidistant grid. Finally, even for the rather dense initial equidistant grid of ∼1000 grid points, an error of several percent can be present on the phonon heat current.

This error of several percent on the phonon heat current is due to the Green's function conversion error creating phonons at slightly shifted energies. It can be seen from Fig. 8, that the "leave-out" approach has the tendency to locally over- and underestimate the self-energy as a function of energy and that the "averaged" approach has the tendency to get the total mass of the self-energy correct, but that the mass is shifted slightly in energy. The self-energy is the cause of the creation and annihilation of particles. The total mass of the self-energy thus determines how many particles are injected and removed. The "averaged" approach therefore achieves a correct number of particles that are injected and removed. Indeed, it was verified that the number of phonons at 1000 initial equidistant grid points was not altered more than 1% by the "averaged" approach. The particles are, however, created at slightly shifted energy. Since both the electron and phonon self-energy depend on the electron Green's function, the shift in energy is at most the equidistant energy grid spacing for the electrons, i.e., usually 1 meV as stated above. Such a shift does not have a large effect on the electronic macroscopic parameters such as the current and the charge, especially since the electron current and electron charge do not depend on the energy of the electrons. Hence, introducing electrons at a slightly wrong energy has no direct influence on the computed electronic properties. For phonons, the energy shift is much more significant as 1 meV corresponds to a more significant part of the phonon energy window. Additionally, the phonon heat current does depend directly on the phonon energy. Hence, introducing phonons at a slightly wrong energy has a much stronger influence on the related computed property. We would like to note, however, that electronic scattering depends on the number of phonons, which is correct, and that the shift in energy is small compared to the features in the electronic Green's functions. This implies that the effect on the electron scattering will be small, which is the main interest in our research on the self-heating.

4.2. Computation time reduction by FFT-based self-energy calculation

In Section 4.1, it was shown that the self-energy can be computed using an FFT-based implementation at the cost of increasing the number of initial grid points, from ∼100 to ∼1000, and introducing a 1% error on the electronic output parameters of the device and a few percent error on the heat current. Note that this tenfold increase of the number of initial grid points does not give rise to a tenfold increase of the Green's function computation time. As shown in Appendix D, the number of adaptively added points does not increase accordingly or even decreases. Timing tests on a single Intel Xeon Gold 6132 processor showed an increase of the computation time by only 60% by this tenfold increase of the initial of the initial grid. The reduction in the computation time of the self-energy is, however, significant, as shown in Fig. 9.

Fig. 9 (a) shows the computation time to compute the self-energies using the conventional convolution-based implementation for a single non-self-consistent iteration on a single Intel Xeon Gold 6132 processor. Note that no quadratic dependency can be observed as the computation time depends on the total number of grid points, which is not linearly dependent on the number of initial grid points. The computation time is dominated by the electron self-energy as this depends on the phonon Green's function, which is characterized by more grid points than the electron Green's function.

Fig. 9 (b) shows the computation time for the self-energies using the FFT-based implementation. The computation of the Fourier transformed self-energies scales linearly with the number of equidistant initial grid points, as expected. The averaging approach to convert the adaptive grid to an equidistant grid scales slower than linearly as it scales with the number of adaptive grid points. The majority of the computation time is related to the Fourier transform and the inverse Fourier transform, which scales as $\mathcal{O}(N_E \log N_E)$. Note that, despite the higher required initial grid size of ∼1000 energy points, the total computation time is

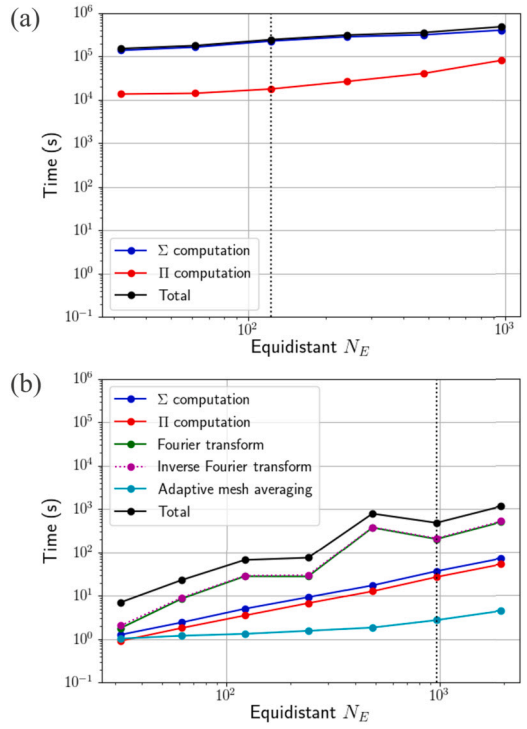


Fig. 9. The computation time of calculating the self-energies for a single non-self-consistent iteration of the system in Section 3.2 as a function of the number of initial equidistant grid energy points. (a) and (b) show the computation times for the conventional convolution-based implementation and the FFT-based implementation, respectively. The dotted lines indicate the required number of initial energy grid points for an estimated relative error of 1% on the electronic properties.

still significantly lower than for the conventional implementation with ∼100 initial grid points, differing by a factor of ∼500.

With the conventional implementation, the self-energy computation takes 98% of the total computation time for a single Born iteration, when performed on a single processor. With the FFT-based implementation, this reduces to 9%. Additionally, the Fourier transform computation time scales linearly with the number of k-points and the computation can be done in parallel for every degree of freedom of the system, i.e., for the Green's function on every orbital or atom. The FFT-based implementation thus readily allows for an increase of the number of k-points or parallelization of the self-energy calculation. Also the computation of the Fourier transformed self-energies can be parallelized more easily than the conventional convolution-based self-energy calculation. As can be seen from (52) and (55), the Fourier transformed self-energy at grid point k' only depends on the Fourier transformed Green's functions on grid points k' and $-k'$, in contrast to the conventional convolution-based computation, which requires the Green's functions on every grid point. Parallelization of the FFT-based self-energy computation thus requires at most a doubling of the memory requirements, whereas for the conventional convolution-based implementation, the memory requirements scale with the number of cores. Fig. 10 shows the parallel efficiency of our parallel implementation for a single non-self-consistent self-energy calculation as a function of the number of cores for the FFT-based implementation, where the parallel efficiency is defined as

$$\text{Efficiency}(N) = \frac{\text{Time on 1 core}}{N \cdot \text{Time on } N \text{cores}}. \tag{57}$$

4.3. Choice of phonon self-energy interpolation scheme

In the FFT-based self-energy computation, there are two choices for the interpolation scheme of the electron Green's function during the calculation of the phonon self-energy, corresponding to (55) and (56), respectively. In the discussion above, (55) was used in all simulations with the FFT-based implementation. This choice was not well-founded. However, Fig. 11 demonstrates that the difference between these two choices is significantly lower than the equidistant-grid errors discussed above.

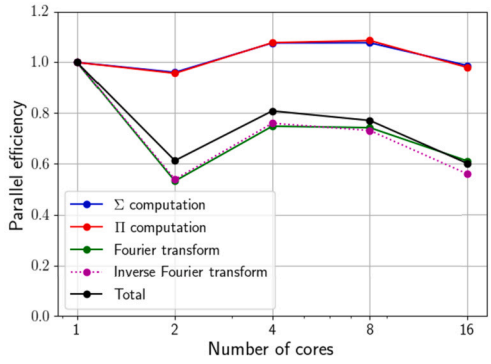


Fig. 10. The parallel efficiency of calculating the self-energies as a function of the number of cores for a single non-self-consistent iteration of the system in Section 3.2 with a parallelized FFT-based implementation with 1000 equidistant grid points.

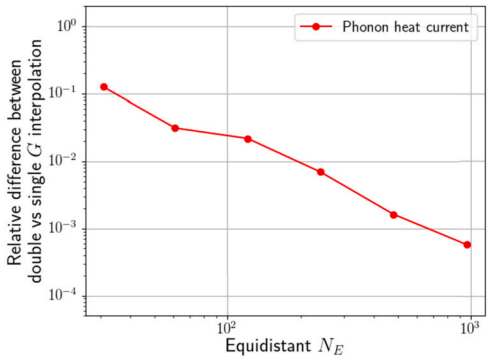


Fig. 11. The relative difference between the phonon heat currents computed with (55) and (56) for the system in Section 3.2 as a function of the number of initial equidistant grid energy points.

5. Self-consistent coupled electron phonon transport in a device

The discussion in Section 4 showed that the FFT-based self-energy computation can provide an efficient way to perform fully coupled transport simulations. We now extend that discussion to more than a single Born iteration. The number of Born iterations that is required to achieve a converged current, charge and phonon heat current as well as current conservation in the channel depends on the gate voltage, but on average 150 Born iterations were required for each bias point, which would make the conventional convolution-based implementation exceedingly expensive. The resulting electron local density of states and current and phonon heat current spectra for a converged simulation of the MoS_2 device in Section 3.2 are given in Fig. 12. Fig. 12 (a) shows the electron density of states as a function of the position in the device, nicely demonstrating the variation of the bottom of the conduction band with the potential. Fig. 12 (b) shows the electron current spectrum. It can be seen that the current behaves quasi-ballistic until it reaches the drain extension region. The availability of lower-energy states then allows for heat dissipation as the electrons lose energy to generate additional phonons. This is confirmed by the phonon heat current spectrum in Fig. 12 (c). Phonons are generated in the drain extension region and travel away in both directions. Peaks in the phonon heat current can be distinguished, which are attributed to peaks in the phonon density of

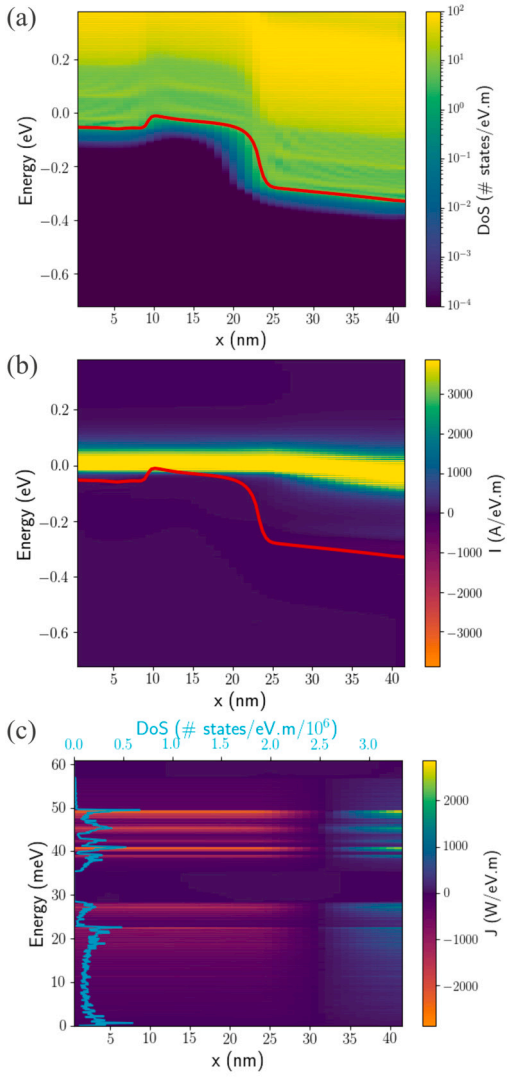


Fig. 12. The electron local density of states (a), the electron current spectrum (b) and the phonon heat current spectrum (c) for a converged simulation of the system in Section 3.2. The spectra are given as a function of energy and the position along the transport direction. The red line in the electron local density of states and current spectrum denotes the bottom of the conduction band. The cyan line in the phonon heat current spectrum denotes the phonon density of states.

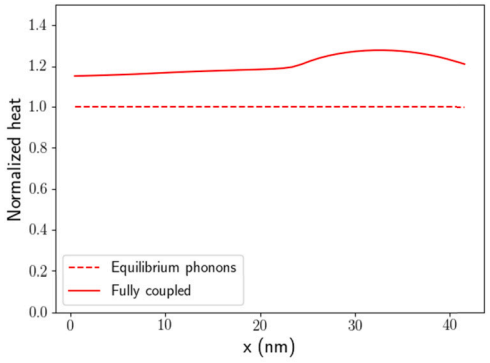


Fig. 13. The heat related to the phonon Green's function for a converged fully coupled simulation of the system in Section 3.2, and a converged dissipative simulation with phonons kept in equilibrium. Both curves are given as a function of the position along the transport direction and are normalized by the equilibrium phonon heat in the middle of the device.

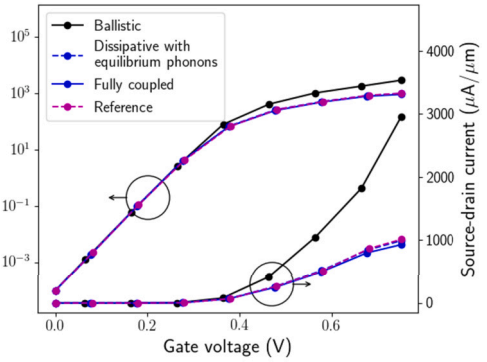


Fig. 14. The source-drain current for converged simulations of the system in Section 3.2 as a function of the gate potential. The current is shown for the ballistic case, the case with scattering by phonons in equilibrium, the fully coupled case and a reference case to validate our model.

states, which is uniform throughout the device as it is not influenced by potential variations. Fig. 13 compares this heat generation with a simulation with equilibrium phonons. The heat of the equilibrium phonons is, as expected, independent of the position. Compared to the equilibrium phonon case, the heat of the fully coupled simulation is higher everywhere in the device. It reaches a maximum in the drain extension region, where the phonons are generated, and decreases further away.

The source-drain current as a function of the gate bias is shown in Fig. 14. Four curves can be distinguished: the results for a ballistic simulation, where the electron self-energy is kept at zero, a simulation with electron-phonon scattering with phonons at equilibrium, achieved by computing the electron self-energy but keeping the phonon self-energy at zero, and a fully-coupled simulation, where both the electron and phonon self-energy are computed. The fourth curve corresponds to a reference simulation based on an established model used in previous work [6], but adapted to our device geometry and bias conditions. This established model is based on phonons in equilibrium with a fixed temperature and can, hence, not be extended to incorporate self-heating. However, the model does not make use of the approximations on the matrix elements discussed in Section 3.2. It can therefore act as a reference for our simulation with electron-phonon scattering with phonons at equilibrium and validate the approximations that we made. It is seen in Fig. 14 that the influence of scattering on the subthreshold regime is limited. The ON-state current, however, is reduced by 66%, consistent with the conclusions from previous work that electron-phonon scattering should be incorporated to correctly predict the device performance. Additionally, the results for our model with electron-phonon scattering with phonons at equilibrium are nearly superimposed on the results of the reference simulation. We interpret this as an indication that our model is qualitatively correct and that rescaling of the matrix elements and damping of the phonon Green's function correctly compensates for neglecting the long-range interactions. Finally, the results for a fully coupled simulation and a simulation with equilibrium phonons are nearly identical, differing by a mere 6% in the ON-state current. For the MoS_2-based device of Section 3.2, self-heating effects thus have negligible influence on the device performance.

Similar device simulations for WS_2 and WSe_2-based devices give slightly different results, as shown in Fig. 15. Fig. 15 (a) shows the results for the c-factors computed in Section 3.2. While qualitatively similar behavior is observed, the results do not match equally well with the reference simulation as for the MoS_2 case. Additionally, the influence of self-heating is more pronounced for these materials and WSe_2 demonstrates a slightly stronger influence of the effect of electron-phonon scattering altogether. For WS_2, incorporating electron-phonon scattering results in a decrease of the ON-state current of 66% and incorporating self-heating decreases it further with 16%. For WSe_2, electron-phonon scattering degrades the ON-state current with 82% and self-heating decreases it further with 32%. In both cases, the additional degradation

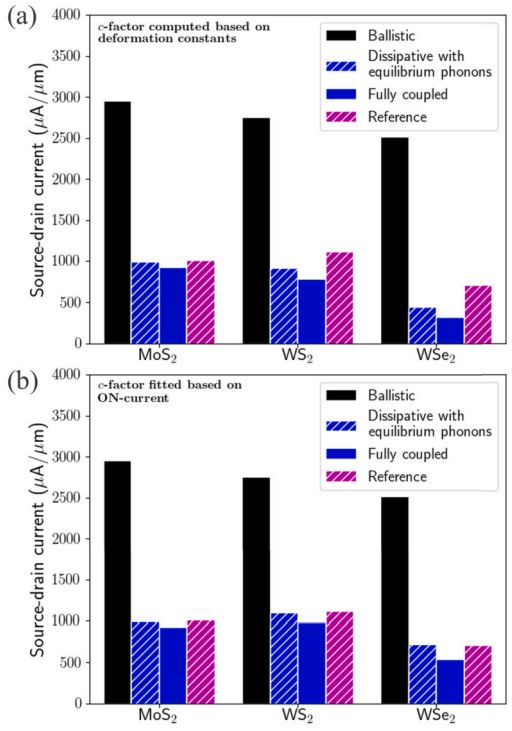


Fig. 15. The source-drain current in ON-state for devices based on the system in Section 3.2 for MoS_2, WS_2 and WSe_2. The current is shown for the ballistic case, the case with scattering by phonons in equilibrium and the fully coupled case and a reference simulation. (a) shows the results for the values of c computed in Section 3.2. (b) shows the results for values of c obtained to ensure matching results between the equilibrium phonon case and the reference simulation.

by self-heating appears to be on the order of the mismatch with the reference simulation, which makes it difficult to interpret whether these predicted self-heating effects are accurate. We therefore performed a second set of simulations with fitted c-factors in order to ensure matching results with the reference simulation, based on the principles used to compensate for neglecting off-diagonal elements in previous work [32]. The values found for WS_2 and WSe_2 are $c = 6.7327$ and 4.9643, respectively. The results are shown in Fig. 15 (b). In this case, for WS_2, incorporating electron-phonon scattering results in a decrease of the ON-state current of 60% and incorporating self-heating decreases it further with 11%. For WSe_2, electron-phonon scattering degrades the ON-state current with 71% and self-heating decreases it further with 29%. As expected, both the direct effect of electron-phonon scattering, as well as the effect of self-heating is reduced for the lower values of c. The self-heating effects are, however, still more pronounced for WS_2 and especially for WSe_2 than for MoS_2.

A proper investigation of the approximations made in this work by neglecting long-range interactions might be useful. A comparison of simulations with different levels of approximations in the matrix elements would be exceedingly expensive with the conventional convolution-based method of computing the self-energy, but might be feasible with the FFT-based implementation introduced in this work. This is nevertheless considered outside the scope of this work and we leave it for future work. Here, we limit ourselves to the observation that the effect of room-temperature phonons clearly dominates the device performance in all cases and neglecting self-heating effects does give a qualitatively correct description. In some cases, such as for WSe_2, the effect of self-heating is not completely negligible and should be included if a highly accurate prediction of the device performance is required.

It should be noted that Fig. 12 (a) shows that the electrons are not fully thermalized, despite the elongated drain extension region. This is also in line with previous work on scattering with equilibrium phonons. To achieve full thermalization, a significantly longer drain extension region is required. However, we feel that this would make the device geometry deviate too much from realistic devices. The conclusions would, hence, not be appropriate to assess the device performance in realistic devices. A consequence of non-thermalized carriers is that the remaining heat is lost in the metal leads. This thermalization in the leads can give rise to increases in the temperature, corresponding to a globally larger phonon population, which can affect device performances. A full incorporation of the metal contacts and the computation of a global temperature increase is, however, outside the scope of this work. Here, we focused on the presence of local hotspots near the channel region.

6. Conclusion and future work

In this work, we showed how real space matrix elements provided by readily available codes such as QUANTUM ESPRESSO and Perturbo can be used to perform fully coupled electron-phonon transport simulations using NEGF. Additionally, we showed how the computationally demanding self-energy calculation step can be made significantly more efficient by using an FFT-based implementation. For a serial computation, a ∼500 times speedup for the self-energy calculation step was achieved by using this FFT-based implementation, going from 98% of the total computation time to a mere 9%. Additionally, we showed that the FFT-based implementation is readily parallelized, without the significant additional memory requirements coming with parallelization of the conventional convolution-based implementation.

For equidistant energy grids, the error introduced by this FFT-based implementation is limited to an energy mixing error, which is predicted to be multiple orders smaller than the integration error. For adaptive non-equidistant energy grids, the relative error on macroscopic electronic properties, such as the charge density and current, can be made less than 1% while the error on the heat current in the device can be limited to a few percent.

Additionally, our simulations confirm that including scattering by electron-phonon interactions is important to correctly predict the device performance of devices based on 2D materials. The influence of self-heating effects is less pronounced in comparison and can be neglected for MoS_2-based devices. Devices based on WS_2, and especially devices based on WSe_2 demonstrate a slightly stronger susceptibility to the incorporation of self-heating effects. This conclusion is focused on hotspots near the channel region and is, hence, based on the fact that further thermalization of electrons in the metallic leads does not introduce significant heating effects as the metal leads were not included in our simulations.

Finally, the simulations in this work rely on approximations concerning the self-energy computation. Certain electron-phonon matrix elements and Green's function elements were neglected to reduce the significant computational cost of the full self-energy computation. To compensate for this, the remaining electron-phonon matrix elements were rescaled and the phonon Green's function was damped at low energies. Our simulations result in qualitatively similar device behavior compared to a reference model based on the literature, validating these approximations. However, some discrepancies are found and further research on these approximations could be of interest. An exact self-energy calculation, using all electron-phonon matrix elements and Green's function elements, is prohibitively expensive with the conventional convolution-based implementation. The FFT-based implementation could, however, provide a means to estimate the influence of these currently neglected elements.

CRediT authorship contribution statement

Rutger Duflou: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Gautam Gaddemane:** Writing – review & editing, Resources. **Michel Houssa:** Writing – review & editing, Supervision. **Aryan Afzalian:** Writing – review

& editing, Supervision, Software, Project administration, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Rutger Duflou reports financial support was provided by Research Foundation Flanders. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was funded by the FWO as part of the PhD fellowship 1100321N.

Appendix A. Reciprocal-real space transformations

We elaborate on the different reciprocal-real space and real-mixed space transformations of Section 2.1. The reciprocal-real space transformation of the electron creation operator in (2), repeated here for the sake of clarity,

$$\hat{c}^{\dagger}_{\mathbf{R}_e m} = \frac{1}{\sqrt{N_e}} \sum_{n\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{R}_e} U_{nm,\mathbf{k}} \hat{c}^{\dagger}_{\mathbf{k}n}, \tag{A.1}$$

can be derived from the fact that

$$|n\mathbf{k}\rangle = \hat{c}^{\dagger}_{\mathbf{k}n} |0\rangle , \tag{A.2}$$

$$|m\mathbf{R}_e\rangle = \hat{c}^{\dagger}_{\mathbf{R}_e m} |0\rangle , \tag{A.3}$$

and that

$$|m\mathbf{R}_e\rangle = \frac{1}{\sqrt{N_e}} \sum_{n\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{R}_e} U_{nm,\mathbf{k}} |n\mathbf{k}\rangle . \tag{A.4}$$

Note that the normalization convention is different than the ones in Ref. [17,16]. The states $|n\mathbf{k}\rangle$ and $|m\mathbf{R}_e\rangle$ are normalized over all of space. This is required to let both $\hat{c}^{\dagger}_{\mathbf{k}n}$ and $\hat{c}^{\dagger}_{\mathbf{R}_e m}$ be proper creation operators. Taking the Hermitian conjugate of (2), we namely find the transformation expression of the electron annihilation operators,

$$\hat{c}_{\mathbf{R}_e m} = \frac{1}{\sqrt{N_e}} \sum_{n\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{R}_e} U^{\dagger}_{mn,\mathbf{k}} \hat{c}_{\mathbf{k}n}. \tag{A.5}$$

For these operators to be proper creation and annihilation operators, they need to obey the anticommutation relations,

$$\begin{aligned} &\{\hat{c}_{\mathbf{k}n}, \hat{c}^{\dagger}_{\mathbf{k}'n'}\} = \delta_{n,n'}\delta_{\mathbf{k},\mathbf{k}'}, \\ &\{\hat{c}^{\dagger}_{\mathbf{k}n}, \hat{c}^{\dagger}_{\mathbf{k}'n'}\} = \{\hat{c}_{\mathbf{k}n}, \hat{c}_{\mathbf{k}'n'}\} = 0, \\ &\{\hat{c}_{\mathbf{R}_e m}, \hat{c}^{\dagger}_{\mathbf{R}'_e m'}\} = \delta_{m,m'}\delta_{\mathbf{R}_e,\mathbf{R}'_e}, \\ &\{\hat{c}^{\dagger}_{\mathbf{R}_e m}, \hat{c}^{\dagger}_{\mathbf{R}'_e m'}\} = \{\hat{c}_{\mathbf{R}_e m}, \hat{c}_{\mathbf{R}'_e m'}\} = 0. \end{aligned} \tag{A.6}$$

This is only true for both the reciprocal and real space operators for the normalization convention of (A.1).

In a similar fashion, the reciprocal-real space transformation for the phonon operators can be found as

$$\hat{a}^{\dagger}_{\mathbf{R}_p\kappa\alpha} = \frac{1}{\sqrt{N_p}} \sum_{\nu\mathbf{q}} e^{-i\mathbf{q}\cdot\mathbf{R}_p} e^{*}_{\kappa\alpha\nu,\mathbf{q}} \hat{a}^{\dagger}_{\mathbf{q}\nu}, \tag{A.7}$$

$$\hat{a}_{\mathbf{R}_p\kappa\alpha} = \frac{1}{\sqrt{N_p}} \sum_{\nu\mathbf{q}} e^{i\mathbf{q}\cdot\mathbf{R}_p} e_{\kappa\alpha\nu,\mathbf{q}} \hat{a}_{\mathbf{q}\nu}. \tag{A.8}$$

Again, a different convention from Ref. [17] is used. Similarly to the electron case, the normalization with the number of k-points is symmetric over the two transformations. Additionally, no mass rescaling is performed. For these operators to be proper bosonic operators, they need to obey the commutation relations,

$$\begin{aligned} &[\hat{a}_{\mathbf{q}\nu}, \hat{a}^{\dagger}_{\mathbf{q}'\nu'}] = \delta_{\nu,\nu'}\delta_{\mathbf{q},\mathbf{q}'}, \\ &[\hat{a}^{\dagger}_{\mathbf{q}\nu}, \hat{a}^{\dagger}_{\mathbf{q}'\nu'}] = [\hat{a}_{\mathbf{q}\nu}, \hat{a}_{\mathbf{q}'\nu'}] = 0, \\ &[\hat{a}_{\mathbf{R}_p\nu}, \hat{a}^{\dagger}_{\mathbf{R}'_p\nu'}] = \delta_{\nu,\nu'}\delta_{\mathbf{R}_p,\mathbf{R}'_p} \\ &[\hat{a}^{\dagger}_{\mathbf{R}_p\nu}, \hat{a}^{\dagger}_{\mathbf{R}'_p\nu'}] = [\hat{a}_{\mathbf{R}_p\nu}, \hat{a}_{\mathbf{R}'_p\nu'}] = 0. \end{aligned} \tag{A.9}$$

This, and the fact that the creation and annihilation operator are each other's Hermitian conjugate, can only satisfied for the normalization conventions of (A.7) and (A.8).

The inverse transformations of (A.1), (A.5), (A.7) and (A.8) are given by

$$\hat{c}^{\dagger}_{\mathbf{k}n} = \frac{1}{\sqrt{N_e}} \sum_{m\mathbf{R}_e} e^{i\mathbf{k}\cdot\mathbf{R}_e} U^{\dagger}_{mn,\mathbf{k}} \hat{c}^{\dagger}_{\mathbf{R}_e m}, \tag{A.10}$$

$$\hat{c}_{\mathbf{k}n} = \frac{1}{\sqrt{N_e}} \sum_{m\mathbf{R}_e} e^{-i\mathbf{k}\cdot\mathbf{R}_e} U_{nm,\mathbf{k}} \hat{c}_{\mathbf{R}_e m}, \tag{A.11}$$

$$\hat{a}^{\dagger}_{\mathbf{q}\nu} = \frac{1}{\sqrt{N_p}} \sum_{\kappa\alpha\mathbf{R}_p} e^{i\mathbf{q}\cdot\mathbf{R}_p} e_{\kappa\alpha\nu,\mathbf{q}} \hat{a}^{\dagger}_{\mathbf{R}_p\kappa\alpha}, \tag{A.12}$$

$$\hat{a}_{\mathbf{q}\nu} = \frac{1}{\sqrt{N_p}} \sum_{\kappa\alpha\mathbf{R}_p} e^{-i\mathbf{q}\cdot\mathbf{R}_p} e^{*}_{\kappa\alpha\nu,\mathbf{q}} \hat{a}_{\mathbf{R}_p\kappa\alpha}. \tag{A.13}$$

Inserting these relations into (1), we obtain

$$\begin{aligned} \hat{H} = &\sum_{\substack{mm' \\ \mathbf{R}_e\mathbf{R}'_e}} \underset{\mathbf{R}'_e-\mathbf{R}_e}{\tilde{h}_{mm'}} \hat{c}^{\dagger}_{\mathbf{R}_e m} \hat{c}_{\mathbf{R}'_e m'} \\ &+ \sum_{\substack{\kappa\alpha\kappa'\alpha' \\ \mathbf{R}_p\mathbf{R}'_p}} \underset{\mathbf{R}'_p-\mathbf{R}_p}{\tilde{d}_{\kappa\alpha\kappa'\alpha'}} \hat{a}^{\dagger}_{\mathbf{R}_p\kappa\alpha} \hat{a}_{\mathbf{R}'_p\kappa'\alpha'} + H_{zero} \\ &+ \sum_{\substack{m'n'\kappa\alpha \\ \mathbf{R}_e\mathbf{R}'_e\mathbf{R}_p}} \underset{\substack{\mathbf{R}'_e-\mathbf{R}_e \\ \mathbf{R}_p-\mathbf{R}_e}}{\tilde{g}_{m'n'\kappa\alpha}} \hat{c}^{\dagger}_{\mathbf{R}_e m'} \hat{c}_{\mathbf{R}'_e n'} (\hat{a}_{\mathbf{R}_p\kappa\alpha} + \hat{a}^{\dagger}_{\mathbf{R}_p\kappa\alpha}) \end{aligned} \tag{A.14}$$

with H_{zero} the zero-point energy of the phonon system, which will be neglected from here forth.

$$H_{zero} = \sum_{\nu\mathbf{q}} \frac{\hbar\omega_{\mathbf{q}\nu}}{2} \tag{A.15}$$

The matrix elements in (A.14) are defined as

$$\underset{\mathbf{R}'_e-\mathbf{R}_e}{\tilde{h}_{mm'}} = \frac{1}{N_e} \sum_{n\mathbf{k}} e^{-i\mathbf{k}\cdot(\mathbf{R}'_e-\mathbf{R}_e)} U^{\dagger}_{mn,\mathbf{k}} e_{\mathbf{k}n} U_{nm',\mathbf{k}}, \tag{A.16}$$

$$\underset{\mathbf{R}'_p-\mathbf{R}_p}{\tilde{d}_{\kappa\alpha\kappa'\alpha'}} = \frac{1}{N_p} \sum_{\nu\mathbf{q}} e^{-i\mathbf{q}\cdot(\mathbf{R}'_p-\mathbf{R}_p)} e_{\kappa\alpha\nu,\mathbf{q}} \hbar\omega_{\mathbf{q}\nu} e^{*}_{\kappa'\alpha'\nu,\mathbf{q}}, \tag{A.17}$$

$$\begin{aligned} \underset{\substack{\mathbf{R}'_e-\mathbf{R}_e \\ \mathbf{R}_p-\mathbf{R}_e}}{\tilde{g}_{m'n'\kappa\alpha}} = &\frac{1}{N_p N_e} \sum_{\substack{mn\mathbf{k} \\ \mathbf{q}\nu}} e^{-i\mathbf{k}\cdot(\mathbf{R}'_e-\mathbf{R}_e)-i\mathbf{q}\cdot(\mathbf{R}_p-\mathbf{R}_e)} \\ &U^{\dagger}_{m'm,\mathbf{k}+\mathbf{q}} g_{mn\nu}(\mathbf{k},\mathbf{q}) U_{nn',\mathbf{k}} e^{*}_{\kappa\alpha\nu,\mathbf{q}}. \end{aligned} \tag{A.18}$$

The electron Hamiltonian elements of (A.16) are equal to the ones computed by the Wannier90 code [16] and presented in the flexible HDF5 format by Perturbo [30].

ATOMOS does not use the phonon Hamiltonian elements, but the interatomic force constants. Grouping the indices (κ, α, $\mathbf{R}_p$) into a single row or column index, $\bar{\mathbf{d}}$ can be considered a matrix. The square of this matrix can be linked to the required interatomic force constants,

$$\sum_{\kappa'\alpha'\mathbf{R}'_p} \underset{\mathbf{R}'_p-\mathbf{R}_p}{\tilde{d}_{\kappa\alpha\kappa'\alpha'}} \underset{\mathbf{R}''_p-\mathbf{R}'_p}{\tilde{d}_{\kappa'\alpha'\kappa''\alpha''}} = \frac{\hbar^2}{N_p} \sum_{\nu\mathbf{q}} e^{-i\mathbf{q}\cdot(\mathbf{R}''_p-\mathbf{R}_p)} e_{\kappa\alpha\nu,\mathbf{q}} \omega^2_{\mathbf{q}\nu} e^{*}_{\kappa''\alpha''\nu,\mathbf{q}}, \tag{A.19}$$

where we used that

$$\sum_{\mathbf{R}'_p} e^{-i(\mathbf{q}-\mathbf{q}')\cdot\mathbf{R}'_p} = N_p \delta_{\mathbf{q},\mathbf{q}'}, \tag{A.20}$$

and used the orthonormality of $e_{\kappa\alpha\nu,\mathbf{q}}$ [17],

$$\sum_{\kappa'\alpha'} e^*_{\kappa'\alpha'\nu,\mathbf{q}} e_{\kappa'\alpha'\nu',\mathbf{q}} = \delta_{\nu,\nu'}. \tag{A.21}$$

The right-hand side of (A.19) is a rescaled version of interatomic force constant defined in (23) [17,15],

$$\sum_{\kappa'\alpha'\mathbf{R}'_p} \tilde{d}_{\substack{\kappa\alpha\kappa'\alpha' \\ \mathbf{R}'_p-\mathbf{R}_p}} \tilde{d}_{\substack{\kappa'\alpha'\kappa''\alpha'' \\ \mathbf{R}''_p-\mathbf{R}'_p}} = \frac{\hbar^2}{\sqrt{m_\kappa m_{\kappa''}}} \frac{\partial^2 U}{\partial\tau_{\kappa\alpha p}\partial\tau_{\kappa''\alpha'' p''}}. \tag{A.22}$$

The interatomic force constants are also provided in the HDF5 format by Perturbo [30].

As will be detailed in Appendix B and Appendix C, ATOMOS does not use the matrix elements in (A.18), but requires a modification,

$$\tilde{\tilde{g}}_{\substack{m'n'\kappa\alpha \\ \mathbf{R}'_e-\mathbf{R}_e \\ \mathbf{R}_p-\mathbf{R}_e}} = \frac{1}{N_p N_e} \sum_{\substack{mn\mathbf{k} \\ \mathbf{q}\nu}} e^{-i\mathbf{k}\cdot(\mathbf{R}'_e-\mathbf{R}_e)-i\mathbf{q}\cdot(\mathbf{R}_p-\mathbf{R}_e)} U^\dagger_{m'm,\mathbf{k}+\mathbf{q}} g^{def}_{mn\nu}(\mathbf{k},\mathbf{q}) U_{nn',\mathbf{k}} e^*_{\kappa\alpha\nu,\mathbf{q}} \tag{A.23}$$

with

$$g^{def}_{mn\nu}(\mathbf{k},\mathbf{q}) = \sqrt{\hbar\omega_{\mathbf{q}\nu}}\, g_{mn\nu}(\mathbf{k},\mathbf{q}). \tag{A.24}$$

For the interpretation of this new constant, we compare the original definition of our matrix elements $g_{mn\nu}(\mathbf{k},\mathbf{q})$ in (1) (Equations 21 and 31-38 in Ref. [15]) to the original work on Wannierization of matrix elements that Perturbo is based upon (Equations 6 and 17 in Ref. [17]). We find that

$$g^{def}_{mn\nu}(\mathbf{k},\mathbf{q}) = \frac{\hbar}{\sqrt{2m_0}} g^{Perturbo}_{mn\nu}(\mathbf{k},\mathbf{q}). \tag{A.25}$$

Next, comparing the Perturbo reciprocal-real space transformation (Equation 24 in Ref. [17]) with (A.23), we find a difference in the normalization convention related to N_e and an additional mass rescaling. Concerning the normalization convention, Perturbo is consistent with our convention (Equation 27 in Ref. [30]), which was verified by reproducing the deformation potentials as provided by Perturbo. The difference in mass rescaling can be compensated for, resulting in

$$\tilde{\tilde{g}}_{\substack{mn\kappa\alpha \\ \mathbf{R}_e,\mathbf{R}_p}} = \frac{\hbar}{\sqrt{2m_\kappa}} g^{Perturbo}_{mn\kappa\alpha}(\mathbf{R}_e,\mathbf{R}_p). \tag{A.26}$$

The full real space Hamiltonian can be further transformed to mixed space by Fourier transforming the periodic directions using the following transformations,

$$\hat{c}^\dagger_{\mathbf{k}_\perp\mathbf{R}_{e,\parallel}n} = \frac{1}{\sqrt{N_\perp}} \sum_{\mathbf{R}_{e,\perp}} e^{i\mathbf{k}_\perp\cdot\mathbf{R}_{e,\perp}} \hat{c}^\dagger_{\mathbf{R}_{e,\perp}\mathbf{R}_{e,\parallel}n}, \tag{A.27}$$

$$\hat{c}_{\mathbf{k}_\perp\mathbf{R}_{e,\parallel}n} = \frac{1}{\sqrt{N_\perp}} \sum_{\mathbf{R}_{e,\perp}} e^{-i\mathbf{k}_\perp\cdot\mathbf{R}_{e,\perp}} \hat{c}_{\mathbf{R}_{e,\perp}\mathbf{R}_{e,\parallel}n}, \tag{A.28}$$

$$\hat{a}^\dagger_{\mathbf{q}_\perp\mathbf{R}_{p,\parallel}\kappa\alpha} = \frac{1}{\sqrt{N_\perp}} \sum_{\mathbf{R}_{p,\perp}} e^{i\mathbf{q}_\perp\cdot\mathbf{R}_{p,\perp}} \hat{a}^\dagger_{\mathbf{R}_{p,\perp}\mathbf{R}_{p,\parallel}\kappa\alpha}, \tag{A.29}$$

$$\hat{a}_{\mathbf{q}_\perp\mathbf{R}_{p,\parallel}\kappa\alpha} = \frac{1}{\sqrt{N_\perp}} \sum_{\mathbf{R}_{p,\perp}} e^{-i\mathbf{q}_\perp\cdot\mathbf{R}_{p,\perp}} \hat{a}_{\mathbf{R}_{p,\perp}\mathbf{R}_{p,\parallel}\kappa\alpha}. \tag{A.30}$$

Currently no assumptions are made about the number of periodic directions and the number of periodic k-points, $N_\perp$, except that this number is equal for the electron and phonon system. The number of periodic directions can range from 0 for fully real space nanowires with $N_\perp = 1$ and the systems being unchanged, to 1 for planar or 2 for diode-like or resistor-like systems. The inverse transformations are given by

$$\hat{c}^\dagger_{\mathbf{R}_{e,\perp}\mathbf{R}_{e,\parallel}n} = \frac{1}{\sqrt{N_\perp}} \sum_{\mathbf{k}_\perp} e^{-i\mathbf{k}_\perp\cdot\mathbf{R}_{e,\perp}} \hat{c}^\dagger_{\mathbf{k}_\perp\mathbf{R}_{e,\parallel}n}, \tag{A.31}$$

$$\hat{c}_{\mathbf{R}_{e,\perp}\mathbf{R}_{e,\parallel}n} = \frac{1}{\sqrt{N_\perp}} \sum_{\mathbf{k}_\perp} e^{i\mathbf{k}_\perp\cdot\mathbf{R}_{e,\perp}} \hat{c}_{\mathbf{k}_\perp\mathbf{R}_{e,\parallel}n}, \tag{A.32}$$

$$\hat{a}^\dagger_{\mathbf{R}_{p,\perp}\mathbf{R}_{p,\parallel}\kappa\alpha} = \frac{1}{\sqrt{N_\perp}} \sum_{\mathbf{q}_\perp} e^{-i\mathbf{q}_\perp\cdot\mathbf{R}_{p,\perp}} \hat{a}^\dagger_{\mathbf{q}_\perp\mathbf{R}_{p,\parallel}\kappa\alpha}, \tag{A.33}$$

$$\hat{a}_{\mathbf{R}_\perp\mathbf{R}_{p,\parallel}\kappa\alpha} = \frac{1}{\sqrt{N_\perp}} \sum_{\mathbf{q}_\perp} e^{i\mathbf{q}_\perp\cdot\mathbf{R}_{p,\perp}} \hat{a}_{\mathbf{q}_\perp\mathbf{R}_{p,\parallel}\kappa\alpha}. \tag{A.34}$$

It is readily verified that the mixed space operators obey the required commutation and anticommutations relations. Insertion in (A.14) results in

$$\begin{aligned}\hat{H} = &\sum_{\substack{mn\mathbf{k}_\perp \\ \mathbf{R}_{e,\parallel}\mathbf{R}'_{e,\parallel}}} \bar{h}_{\substack{mn\mathbf{k}_\perp \\ \mathbf{R}_{e,\parallel}\mathbf{R}'_{e,\parallel}}} \hat{c}^\dagger_{\mathbf{k}_\perp\mathbf{R}_{e,\parallel}m} \hat{c}_{\mathbf{k}_\perp\mathbf{R}'_{e,\parallel}n} \\ &+ \sum_{\substack{\kappa\alpha\kappa'\alpha'\mathbf{q}_\perp \\ \mathbf{R}_{p,\parallel}\mathbf{R}'_{p,\parallel}}} \bar{d}_{\substack{\kappa\alpha\kappa'\alpha'\mathbf{q}_\perp \\ \mathbf{R}_{p,\parallel}\mathbf{R}'_{p,\parallel}}} \hat{a}^\dagger_{\mathbf{q}_\perp\mathbf{R}_{p,\parallel}\kappa\alpha} \hat{a}_{\mathbf{q}_\perp\mathbf{R}'_{p,\parallel}\kappa'\alpha'} \\ &+ N_\perp^{-\frac{1}{2}} \sum_{\substack{mn\kappa\alpha \\ \mathbf{R}_{e,\parallel}\mathbf{R}'_{e,\parallel} \\ \mathbf{k}_\perp\mathbf{q}_\perp\mathbf{R}_{p,\parallel}}} \bar{g}_{\substack{mn\kappa\alpha \\ \mathbf{R}_{e,\parallel}\mathbf{R}'_{e,\parallel} \\ \mathbf{k}_\perp\mathbf{q}_\perp\mathbf{R}_{p,\parallel}}} \hat{c}^\dagger_{\mathbf{k}_\perp+\mathbf{q}_\perp\mathbf{R}_{e,\parallel}m} \hat{c}_{\mathbf{k}_\perp\mathbf{R}'_{e,\parallel}n} (\hat{a}_{\mathbf{q}_\perp\mathbf{R}_{p,\parallel}\kappa\alpha} + \hat{a}^\dagger_{-\mathbf{q}_\perp\mathbf{R}_{p,\parallel}\kappa\alpha}).\end{aligned} \tag{A.35}$$

Grouping the real space indices into a single index results in (4). The matrix element transformations are performed by ATOMOS and correspond to

$$\bar{h}_{\substack{mn\mathbf{k}_\perp \\ \mathbf{R}_{e,\parallel}\mathbf{R}'_{e,\parallel}}} = \sum_{\mathbf{R}_{e,\perp}} \tilde{h}_{\substack{mn \\ \mathbf{R}'_{e,\parallel}-\mathbf{R}_{e,\parallel}}} e^{i\mathbf{k}_\perp\cdot\mathbf{R}_{e,\perp}}, \tag{A.36}$$

$$\bar{d}_{\substack{\kappa\alpha\kappa'\alpha'\mathbf{q}_\perp \\ \mathbf{R}_{p,\parallel}\mathbf{R}'_{p,\parallel}}} = \sum_{\mathbf{R}_{p,\perp}} \tilde{d}_{\substack{\kappa\alpha\kappa'\alpha' \\ \mathbf{R}_{p,\perp} \\ \mathbf{R}'_{p,\parallel}-\mathbf{R}_{p,\parallel}}} e^{i\mathbf{q}_\perp\cdot\mathbf{R}_{p,\perp}}, \tag{A.37}$$

$$\bar{g}_{\substack{mn\kappa\alpha \\ \mathbf{R}_{e,\parallel}\mathbf{R}'_{e,\parallel} \\ \mathbf{k}_\perp\mathbf{q}_\perp\mathbf{R}_{p,\parallel}}} = \sum_{\mathbf{R}_{e,\perp}\mathbf{R}_{p,\perp}} \tilde{g}_{\substack{mn\kappa\alpha \\ \mathbf{R}_{e,\perp} \\ \mathbf{R}'_{e,\parallel}-\mathbf{R}_{e,\parallel} \\ \mathbf{R}_{p,\perp} \\ \mathbf{R}_{p,\parallel}-\mathbf{R}_{e,\parallel}}} e^{i\mathbf{k}_\perp\cdot\mathbf{R}_{e,\perp}+i\mathbf{q}_\perp\cdot\mathbf{R}_{p,\perp}}. \tag{A.38}$$

Appendix B. Green's function expressions

We provide a full derivation of the electron and phonon Green's functions expression introduced in Section 2.2. We start with the electron Green's function, despite derivations for these expressions being readily available in the literature [22,23], for the sake of completeness and to provide a reference for the phonon case. Consider the definition of the retarded, advanced, greater and lesser Green's function

$$G^R_{\substack{n,m \\ \mathbf{k}_\perp}}(t,t') = \frac{-i}{\hbar}\theta(t-t')\left\langle \hat{c}_{\mathbf{k}_\perp n}(t)\hat{c}^\dagger_{\mathbf{k}_\perp m}(t') + \hat{c}^\dagger_{\mathbf{k}_\perp m}(t')\hat{c}_{\mathbf{k}_\perp n}(t)\right\rangle, \tag{B.1}$$

$$G^A_{\substack{n,m \\ \mathbf{k}_\perp}}(t,t') = \frac{i}{\hbar}\theta(t'-t)\left\langle \hat{c}_{\mathbf{k}_\perp n}(t)\hat{c}^\dagger_{\mathbf{k}_\perp m}(t') + \hat{c}^\dagger_{\mathbf{k}_\perp m}(t')\hat{c}_{\mathbf{k}_\perp n}(t)\right\rangle, \tag{B.2}$$

$$G^>_{\substack{n,m \\ \mathbf{k}_\perp}}(t,t') = \frac{-i}{\hbar}\left\langle \hat{c}_{\mathbf{k}_\perp n}(t)\hat{c}^\dagger_{\mathbf{k}_\perp m}(t')\right\rangle, \tag{B.3}$$

$$G^<_{\substack{n,m \\ \mathbf{k}_\perp}}(t,t') = \frac{i}{\hbar}\left\langle \hat{c}^\dagger_{\mathbf{k}_\perp m}(t')\hat{c}_{\mathbf{k}_\perp n}(t)\right\rangle. \tag{B.4}$$

When an expression for the contour-ordered Green's function is known involving convolution integrals on a two-branch contour, e.g., (9), then Langreth's theorem can be used to find equivalent expressions for the retarded, advanced, greater and lesser Green's functions involving only real-time convolutions [19]. Convolutions in the time domain become multiplications in the frequency domain. Hence, using the Fourier transform

$$\mathbf{G}_{\mathbf{k}_\perp}(\omega) = \int_{-\infty}^{+\infty} \mathbf{G}_{\mathbf{k}_\perp}(t,t') e^{i\omega(t-t')} d(t-t'), \tag{B.5}$$

$$\mathbf{G}_{\mathbf{k}_\perp}(t,t') = \int_{-\infty}^{+\infty} \mathbf{G}_{\mathbf{k}_\perp}(\omega) e^{-i\omega(t-t')} \frac{d\omega}{2\pi}, \tag{B.6}$$

with the Green's function denoting any of the Green's functions in (B.1)-(B.4), results in

$$\begin{aligned}\mathbf{G}^{R/A}_{\mathbf{k}_\perp}(\omega) =& \mathbf{G}^{0,R/A}_{\mathbf{k}_\perp}(\omega) + \mathbf{G}^{0,R/A}_{\mathbf{k}_\perp}(\omega)\mathbf{U}_{\mathbf{k}_\perp}\mathbf{G}^{R/A}_{\mathbf{k}_\perp}(\omega) \\ &+ \mathbf{G}^{0,R/A}_{\mathbf{k}_\perp}(\omega)\boldsymbol{\Sigma}^{s,R/A}_{\mathbf{k}_\perp}(\omega)\mathbf{G}^{R/A}_{\mathbf{k}_\perp}(\omega),\end{aligned} \tag{B.7}$$

$$\begin{aligned}\mathbf{G}^{\gtrless}_{\mathbf{k}_\perp}(\omega) =& \mathbf{G}^{0,\gtrless}_{\mathbf{k}_\perp}(\omega) + \mathbf{G}^{0,\gtrless}_{\mathbf{k}_\perp}(\omega)\mathbf{U}_{\mathbf{k}_\perp}\mathbf{G}^{A}_{\mathbf{k}_\perp}(\omega) \\ &+ \mathbf{G}^{0,R}_{\mathbf{k}_\perp}(\omega)\mathbf{U}_{\mathbf{k}_\perp}\mathbf{G}^{\gtrless}_{\mathbf{k}_\perp}(\omega) \\ &+ \mathbf{G}^{0,R}_{\mathbf{k}_\perp}(\omega)\boldsymbol{\Sigma}^{s,R}_{\mathbf{k}_\perp}(\omega)\mathbf{G}^{\gtrless}_{\mathbf{k}_\perp}(\omega) \\ &+ \mathbf{G}^{0,R}_{\mathbf{k}_\perp}(\omega)\boldsymbol{\Sigma}^{s,\gtrless}_{\mathbf{k}_\perp}(\omega)\mathbf{G}^{A}_{\mathbf{k}_\perp}(\omega) \\ &+ \mathbf{G}^{0,\gtrless}_{\mathbf{k}_\perp}(\omega)\boldsymbol{\Sigma}^{s,A}_{\mathbf{k}_\perp}(\omega)\mathbf{G}^{A}_{\mathbf{k}_\perp}(\omega).\end{aligned} \tag{B.8}$$

For the electrons, the matrix $\mathbf{U}_{\mathbf{k}_\perp}$ contains the elements $\bar{h}_{mn\mathbf{k}\perp}$ connecting the device with the leads, as denoted in Fig. 2. The remaining elements form

$$\hat{H}_0^{el} = \sum_{\substack{nm \\ \mathbf{k}_\perp}} \bar{h}_{nm} \hat{c}^\dagger_{\mathbf{k}_\perp n} \hat{c}_{\mathbf{k}_\perp m}, \tag{B.9}$$

with the summation leaving out interaction elements connecting the device and the leads. Note that all leftmost factors in (B.7) and (B.8) are non-interacting, which implies that for the sake of these Green's functions, the Heisenberg picture operators in their definitions are equivalent to interaction picture operators. The time-dependency of interaction picture operators is readily found as [18]

$$\begin{aligned} i\hbar\frac{\partial}{\partial t}\hat{c}_{\mathbf{k}_\perp n}(t) &= \left[\hat{c}_{\mathbf{k}_\perp n}(t), \hat{H}_0^{el}\right] \\ &= \sum_m \bar{h}_{nm}\hat{c}_{\mathbf{k}_\perp m}(t). \end{aligned} \tag{B.10}$$

The time-derivative of the non-interacting retarded Green's functions is then given by

$$i\hbar\frac{\partial}{\partial t} G^{0,R}_{n,m\,\mathbf{k}_\perp}(t,t') = \delta(t-t')\delta_{n,m} + \sum_{n'} \bar{h}_{nn'\,\mathbf{k}_\perp} G^{0,R}_{n',m\,\mathbf{k}_\perp}(t,t') \tag{B.11}$$

or expressed as a matrix equation,

$$i\hbar\frac{\partial}{\partial t}\mathbf{G}^{0,R}_{\mathbf{k}_\perp}(t,t') = \delta(t-t')\mathbf{I} + \mathbf{H}_{\mathbf{k}_\perp}\mathbf{G}^{0,R}_{\mathbf{k}_\perp}(t,t'). \tag{B.12}$$

Inserting (B.6) and using $\delta(t-t') = \int_{-\infty}^{+\infty} e^{-i\omega(t-t')}\frac{d\omega}{2\pi}$, we find

$$\left(\hbar\omega\mathbf{I} - \mathbf{H}_{\mathbf{k}_\perp}\right)\mathbf{G}^{0,R}_{\mathbf{k}_\perp}(\omega) = \mathbf{I}. \tag{B.13}$$

The time-derivative of the lesser and greater Green's function is found in a similar fashion,

$$i\hbar\frac{\partial}{\partial t} G^{0,\gtrless}_{n,m\,\mathbf{k}_\perp}(t,t') = \sum_{n'} \bar{h}_{nn'\,\mathbf{k}_\perp} G^{0,\gtrless}_{n',m\,\mathbf{k}_\perp}(t,t') \tag{B.14}$$

or expressed as a matrix expression

$$i\hbar\frac{\partial}{\partial t}\mathbf{G}^{0,\gtrless}_{\mathbf{k}_\perp}(t,t') = \mathbf{H}_{\mathbf{k}_\perp}\mathbf{G}^{0,\gtrless}_{\mathbf{k}_\perp}(t,t') \tag{B.15}$$

which, after Fourier transformation, becomes

$$\left(\hbar\omega\mathbf{I} - \mathbf{H}_{\mathbf{k}_\perp}\right)\mathbf{G}^{0,\gtrless}_{\mathbf{k}_\perp}(\omega) = \mathbf{0}. \tag{B.16}$$

Leaving out the ω and $\mathbf{k}_\perp$ dependency in the notation for the sake of brevity and left-multiplying (B.7) and (B.8) with $(\hbar\omega\mathbf{I} - \mathbf{H})$, results in

$$(\hbar\omega\mathbf{I} - \mathbf{H})\,\mathbf{G}^R = \mathbf{I} + \mathbf{U}\mathbf{G}^R + \boldsymbol{\Sigma}^{s,R}\mathbf{G}^R, \tag{B.17}$$

$$(\hbar\omega\mathbf{I} - \mathbf{H})\,\mathbf{G}^{\gtrless} = \mathbf{U}\mathbf{G}^{\gtrless} + \boldsymbol{\Sigma}^{s,R}\mathbf{G}^{\gtrless} + \boldsymbol{\Sigma}^{s,\gtrless}\mathbf{G}^A. \tag{B.18}$$

Inserting (B.7) into the second term on the right-hand side of (B.17), we find

$$\begin{aligned}(\hbar\omega\mathbf{I} - \mathbf{H})\,\mathbf{G}^R = \mathbf{I} &+ \mathbf{U}\mathbf{G}^{0,R} + \mathbf{U}\mathbf{G}^{0,R}\mathbf{U}\mathbf{G}^R \\ &+ \mathbf{U}\mathbf{G}^{0,R}\boldsymbol{\Sigma}^{s,R}\mathbf{G}^R + \boldsymbol{\Sigma}^{s,R}\mathbf{G}^R.\end{aligned} \tag{B.19}$$

Note that the matrices so far contain row and column indices corresponding to all degrees of freedom in both the device itself and the left and right leads. As the two leads are infinite, this prohibits building and solving for these matrices in practice. However, to find the currents and charges in the device, only the Green's function elements with both row and column indices within the device are required. We therefore confine the matrices to their subblocks pertaining to the device itself. Note that this does not introduce complications for the multiplication with $\mathbf{H}$ and $\boldsymbol{\Sigma}^{s,R}$. $\mathbf{H}$ has a disconnected subblock corresponding to the device as the connections with the leads are contained in $\mathbf{U}$. $\boldsymbol{\Sigma}^{s,R}$ is non-zero only within the device as can be seen from its definition in (31) and (34) and the fact that the electron-phonon subsystems are only connected within the device as demonstrated in Fig. 2. This is not the case for the term $\mathbf{U}\mathbf{G}^0$. $\mathbf{G}^0$ corresponds to the disconnected device and thus has disconnected blocks for the device and the leads. $\mathbf{U}$, on the other hand, only contains cross terms, connecting the device and the leads. Their matrix product will hence, also only contain cross terms connecting device to leads. When both row and column indices of the expressions are thus confined to the device, this cross term disappears. The same is true for the product $\mathbf{U}\mathbf{G}^0\boldsymbol{\Sigma}^{s,R}\mathbf{G}^R$ as $\boldsymbol{\Sigma}^{s,R}$ is only non-zero within the device. Leaving out these cross terms, we find

$$\left(\hbar\omega\mathbf{I} - \mathbf{H} - \boldsymbol{\Sigma}^{s,R} - \mathbf{U}\mathbf{G}^{0,R}\mathbf{U}\right)\mathbf{G}^R = \mathbf{I}. \tag{B.20}$$

The term $\mathbf{U}\mathbf{G}^{0,R}\mathbf{U}$ connects the device to the leads and is only non-zero in the subblocks corresponding to the leftmost and rightmost slab, i.e., slab **1** and **n**,

$$\left(\mathbf{U}\mathbf{G}^{0,R}\mathbf{U}\right)_{\mathbf{1,1}} = \bar{\mathbf{h}}_{\mathbf{1,0}}\,\mathbf{G}^{0,R/A}_{\mathbf{0,0}}\,\bar{\mathbf{h}}_{\mathbf{0,1}}, \tag{B.21}$$

$$\left(\mathbf{U}\mathbf{G}^{0,R}\mathbf{U}\right)_{\mathbf{n,n}} = \bar{\mathbf{h}}_{\mathbf{n,n+1}}\,\mathbf{G}^{0,R/A}_{\mathbf{n+1,n+1}}\,\bar{\mathbf{h}}_{\mathbf{n+1,n}}. \tag{B.22}$$

Defining

$$\boldsymbol{\Sigma}^l = \mathbf{U}\mathbf{G}^0\mathbf{U}, \tag{B.23}$$

we can merge these two self-energies with (14) and obtain

$$\mathbf{G}^R = \left(\hbar\omega\mathbf{I} - \mathbf{H} - \boldsymbol{\Sigma}^R\right)^{-1}. \tag{B.24}$$

The only difference between (11) and (B.24) is the presence of the convergence term $i\eta$, usually introduced to ensure the convergence of the Fourier transform [37]. It can be readily achieved by adding a factor $e^{-\eta(t-t')}$ to the definition of the retarded Green's function in (B.1) to make it absolutely integrable. A similar procedure is possible for the advanced Green's function using a factor $e^{\eta(t-t')}$, which will result in a convergence term $-i\eta$. We forgo this point and just add it retroactively.

For the lesser and greater Green's function, we insert (B.8) into the first term on the right-hand side of (B.18),

$$\begin{aligned}\mathbf{U}\mathbf{G}^{\gtrless} = &\mathbf{U}\mathbf{G}^{0,\gtrless} + \mathbf{U}\mathbf{G}^{0,\gtrless}\mathbf{U}\mathbf{G}^A + \mathbf{U}\mathbf{G}^{0,R}\mathbf{U}\mathbf{G}^{\gtrless} \\ &+ \mathbf{U}\mathbf{G}^{0,R}\boldsymbol{\Sigma}^{s,R}\mathbf{G}^{\gtrless} + \mathbf{U}\mathbf{G}^{0,R}\boldsymbol{\Sigma}^{s,\gtrless}\mathbf{G}^A \\ &+ \mathbf{U}\mathbf{G}^{0,\gtrless}\boldsymbol{\Sigma}^{s,A}\mathbf{G}^A.\end{aligned} \tag{B.25}$$

The first term and last three terms in (B.25) are cross terms, which become zero when the matrices are limited to their device subblocks. We thus have from (B.18),

$$\left(\hbar\omega\mathbf{I} - \mathbf{H} - \boldsymbol{\Sigma}^{s,R} - \mathbf{U}\mathbf{G}^{0,R}\mathbf{U}\right)\mathbf{G}^{\gtrless} = \left(\boldsymbol{\Sigma}^{s,\gtrless} + \mathbf{U}\mathbf{G}^{0,\gtrless}\mathbf{U}\right)\mathbf{G}^A. \tag{B.26}$$

Using (B.23) and (14), we obtain

$$\left(\hbar\omega\mathbf{I}-\mathbf{H}-\mathbf{\Sigma}^{R}\right)\mathbf{G}^{\lessgtr}=\mathbf{\Sigma}^{\lessgtr}\mathbf{G}^{A}. \tag{B.27}$$

Finally, left-multiplication with $\mathbf{G}^{R}$ results in

$$\mathbf{G}^{\lessgtr}=\mathbf{G}^{R}\mathbf{\Sigma}^{\lessgtr}\mathbf{G}^{A}, \tag{B.28}$$

which is identical to (12). The expressions for the lead self-energies, (17)-(20), are obtained by noting that the density of states in the non-connected leads is given by [23]

$$\boldsymbol{A}^{0}=i\left(\mathbf{G}^{0,R}-\mathbf{G}^{0,A}\right), \tag{B.29}$$

and the fact that the lesser (greater) Green's function is linked to the density of electrons (holes). In the non-connected leads, the electrons are in equilibrium and fill the density of states according to the Fermi-Dirac statistics function (7)

$$\mathbf{G}^{0,<}_{\mathbf{0,0}}=if_1\mathbf{A}^{0}_{\mathbf{0,0}}, \tag{B.30}$$

$$\mathbf{G}^{0,<}_{\mathbf{n+1,n+1}}=if_2\mathbf{A}^{0}_{\mathbf{n+1,n+1}}, \tag{B.31}$$

$$\mathbf{G}^{0,>}_{\mathbf{0,0}}=-i(1-f_1)\mathbf{A}^{0}_{\mathbf{0,0}}, \tag{B.32}$$

$$\mathbf{G}^{0,>}_{\mathbf{n+1,n+1}}=-i(1-f_2)\mathbf{A}^{0}_{\mathbf{n+1,n+1}}. \tag{B.33}$$

We now perform the same procedure for the phonon Green's function. Making use of the commuting behavior of the phonon operators for the contour-ordering operator [18] and the definition of the retarded, advanced, lesser and greater Green's function [19], we obtain

$$\begin{aligned}D^{R}_{\substack{\nu,\mu\\ \mathbf{q}_\perp}}(t,t')=&-\frac{i}{\hbar}\theta(t-t')\Big\langle(\hat{a}_{\mathbf{q}_\perp\nu}(t)+\hat{a}^{\dagger}_{-\mathbf{q}_\perp\nu}(t))(\hat{a}^{\dagger}_{\mathbf{q}_\perp\mu}(t')+\hat{a}_{-\mathbf{q}_\perp\mu}(t'))\\&-(\hat{a}^{\dagger}_{\mathbf{q}_\perp\mu}(t')+\hat{a}_{-\mathbf{q}_\perp\mu}(t'))(\hat{a}_{\mathbf{q}_\perp\nu}(t)+\hat{a}^{\dagger}_{-\mathbf{q}_\perp\nu}(t))\Big\rangle,\end{aligned} \tag{B.34}$$

$$\begin{aligned}D^{A}_{\substack{\nu,\mu\\ \mathbf{q}_\perp}}(t,t')=&\frac{i}{\hbar}\theta(t'-t)\Big\langle(\hat{a}_{\mathbf{q}_\perp\nu}(t)+\hat{a}^{\dagger}_{-\mathbf{q}_\perp\nu}(t))(\hat{a}^{\dagger}_{\mathbf{q}_\perp\mu}(t')+\hat{a}_{-\mathbf{q}_\perp\mu}(t'))\\&-(\hat{a}^{\dagger}_{\mathbf{q}_\perp\mu}(t')+\hat{a}_{-\mathbf{q}_\perp\mu}(t'))(\hat{a}_{\mathbf{q}_\perp\nu}(t)+\hat{a}^{\dagger}_{-\mathbf{q}_\perp\nu}(t))\Big\rangle,\end{aligned} \tag{B.35}$$

$$D^{>}_{\substack{\nu,\mu\\ \mathbf{q}_\perp}}(t,t')=-\frac{i}{\hbar}\left\langle(\hat{a}_{\mathbf{q}_\perp\nu}(t)+\hat{a}^{\dagger}_{-\mathbf{q}_\perp\nu}(t))(\hat{a}^{\dagger}_{\mathbf{q}_\perp\mu}(t')+\hat{a}_{-\mathbf{q}_\perp\mu}(t'))\right\rangle, \tag{B.36}$$

$$D^{<}_{\substack{\nu,\mu\\ \mathbf{q}_\perp}}(t,t')=-\frac{i}{\hbar}\left\langle(\hat{a}^{\dagger}_{\mathbf{q}_\perp\mu}(t')+\hat{a}_{-\mathbf{q}_\perp\mu}(t'))(\hat{a}_{\mathbf{q}_\perp\nu}(t)+\hat{a}^{\dagger}_{-\mathbf{q}_\perp\nu}(t))\right\rangle. \tag{B.37}$$

Using Langreth's theorem on (10) and Fourier transforming results in expressions very similar to (B.7) and (B.8), with all leftmost factors being non-interacting,

$$\begin{aligned}\mathbf{D}^{R/A}_{\mathbf{q}_\perp}(\omega)=&\mathbf{D}^{0,R/A}_{\mathbf{q}_\perp}(\omega)+\mathbf{D}^{0,R/A}_{\mathbf{q}_\perp}(\omega)\mathbf{V}_{\mathbf{q}_\perp}\mathbf{D}^{R/A}_{\mathbf{q}_\perp}(\omega)\\&+\mathbf{D}^{0,R/A}_{\mathbf{q}_\perp}(\omega)\mathbf{\Pi}^{s,R/A}_{\mathbf{q}_\perp}(\omega)\mathbf{D}^{R/A}_{\mathbf{q}_\perp}(\omega),\end{aligned} \tag{B.38}$$

$$\begin{aligned}\mathbf{D}^{\gtrless}_{\mathbf{q}_\perp}(\omega)=&\mathbf{D}^{0,\gtrless}_{\mathbf{q}_\perp}(\omega)+\mathbf{D}^{0,\gtrless}_{\mathbf{q}_\perp}(\omega)\mathbf{V}_{\mathbf{q}_\perp}\mathbf{D}^{A}_{\mathbf{q}_\perp}(\omega)\\&+\mathbf{D}^{0,R}_{\mathbf{q}_\perp}(\omega)\mathbf{V}_{\mathbf{q}_\perp}\mathbf{D}^{\gtrless}_{\mathbf{q}_\perp}(\omega)\\&+\mathbf{D}^{0,R}_{\mathbf{q}_\perp}(\omega)\mathbf{\Pi}^{s,R}_{\mathbf{q}_\perp}(\omega)\mathbf{D}^{\gtrless}_{\mathbf{q}_\perp}(\omega)\\&+\mathbf{D}^{0,R}_{\mathbf{q}_\perp}(\omega)\mathbf{\Pi}^{s,\gtrless}_{\mathbf{q}_\perp}(\omega)\mathbf{D}^{A}_{\mathbf{q}_\perp}(\omega)\\&+\mathbf{D}^{0,\gtrless}_{\mathbf{q}_\perp}(\omega)\mathbf{\Pi}^{s,A}_{\mathbf{q}_\perp}(\omega)\mathbf{D}^{A}_{\mathbf{q}_\perp}(\omega).\end{aligned} \tag{B.39}$$

For the phonons, Fig. 2 is only qualitatively correct. To obtain a Dyson equation of the form (B.8), and hence (B.38) and (B.39), a slightly different subdivision in device and connecting terms is required. The full derivation is, however, cumbersome and has been moved to the Supporting Material [21]. The final result is that the elements,

$$\left(\mathbf{V}_{\mathbf{q}_\perp}\right)_{\nu,\mu}=\underset{\mathbf{q}_\perp}{\upsilon_{\nu\mu}}, \tag{B.40}$$

are the result of a perturbation Hamiltonian,

$$\hat{H}^{ph}_{P}=\sum_{\substack{\nu\mu\\ \mathbf{q}_\perp}}\underset{\mathbf{q}_\perp}{\upsilon_{\nu\mu}}\left(\hat{a}^{\dagger}_{\mathbf{q}_\perp\nu}\hat{a}_{\mathbf{q}_\perp\mu}+\frac{1}{2}\left(\hat{a}^{\dagger}_{\mathbf{q}_\perp\nu}\hat{a}^{\dagger}_{-\mathbf{q}_\perp\mu}+\hat{a}_{-\mathbf{q}_\perp\nu}\hat{a}_{\mathbf{q}_\perp\mu}\right)\right), \tag{B.41}$$

and that the non-interacting Hamiltonian for the phonons is given by

$$\hat{H}^{ph}_{0}=\sum_{\substack{\nu\mu\\ \mathbf{q}_\perp}}(\underset{\mathbf{q}_\perp}{\bar{d}_{\nu\mu}}-\underset{\mathbf{q}_\perp}{\upsilon_{\nu\mu}})\hat{a}^{\dagger}_{\mathbf{q}_\perp\nu}\hat{a}_{\mathbf{q}_\perp\mu}-\frac{\upsilon_{\mathbf{q}_\perp\nu\mu}}{2}\left(\hat{a}^{\dagger}_{\mathbf{q}_\perp\nu}\hat{a}^{\dagger}_{-\mathbf{q}_\perp\mu}+\hat{a}_{-\mathbf{q}_\perp\nu}\hat{a}_{\mathbf{q}_\perp\mu}\right). \tag{B.42}$$

The summation here extends over all matrix elements, also elements connecting the leads with the device. Note that despite the rather unintuitive form of $\hat{H}^{ph}_{P}$ and $\hat{H}^{ph}_{0}$, their sum is still equal to the ballistic phonon Hamiltonian in (4), implying that the material properties are unchanged.

The time-dependency of the non-interacting Green's functions in (B.38) and (B.39) can be evaluated explicitly. However, this time the time-dependency of both the annihilation and creation operator is required,

$$\begin{aligned}i\hbar\frac{\partial}{\partial t}\hat{a}_{\mathbf{q}_\perp\nu}(t)&=\left[\hat{a}_{\mathbf{q}_\perp\nu}(t),\hat{H}^{ph}_{0}\right]\\&=\sum_{\mu}(\underset{\mathbf{q}_\perp}{\bar{d}_{\nu\mu}}-\underset{\mathbf{q}_\perp}{\upsilon_{\nu\mu}})\hat{a}_{\mathbf{q}_\perp\mu}(t)-\frac{\underset{\mathbf{q}_\perp}{\upsilon_{\nu\mu}}+\underset{-\mathbf{q}_\perp}{\upsilon_{\mu\nu}}}{2}\hat{a}^{\dagger}_{-\mathbf{q}_\perp\mu}(t)\end{aligned} \tag{B.43}$$

and

$$\begin{aligned}i\hbar\frac{\partial}{\partial t}\hat{a}^{\dagger}_{\mathbf{q}_\perp\nu}(t)&=\left[\hat{a}^{\dagger}_{\mathbf{q}_\perp\nu}(t),\hat{H}^{ph}_{0}\right]\\&=\sum_{\mu}-(\underset{\mathbf{q}_\perp}{\bar{d}_{\mu\nu}}-\underset{\mathbf{q}_\perp}{\upsilon_{\mu\nu}})\hat{a}^{\dagger}_{\mathbf{q}_\perp\mu}(t)+\frac{\underset{\mathbf{q}_\perp}{\upsilon_{\mu\nu}}+\underset{-\mathbf{q}_\perp}{\upsilon_{\nu\mu}}}{2}\hat{a}_{-\mathbf{q}_\perp\mu}(t).\end{aligned} \tag{B.44}$$

As $e_{\kappa\alpha\nu,\mathbf{q}}=e^{*}_{\kappa\alpha\nu,-\mathbf{q}}$ [17], it follows from (A.17) that $\underset{\mathbf{R}'_p-\mathbf{R}_p}{\tilde{d}_{\kappa\alpha\kappa'\alpha'}}=\underset{\mathbf{R}_p-\mathbf{R}'_p}{\tilde{d}_{\kappa'\alpha'\kappa\alpha}}$. This, in combination with (A.37), in turn implies that $\underset{\mathbf{q}_\perp}{\bar{d}_{\mu\nu}}=\underset{-\mathbf{q}_\perp}{\bar{d}_{\nu\mu}}$. We can choose to impose the same symmetry on $\underset{\mathbf{q}_\perp}{\upsilon_{\mu\nu}}$. We thus have

$$i\hbar\frac{\partial}{\partial t}\hat{a}_{\mathbf{q}_\perp\nu}(t)=\sum_{\mu}(\underset{\mathbf{q}_\perp}{\bar{d}_{\nu\mu}}-\underset{\mathbf{q}_\perp}{\upsilon_{\nu\mu}})\hat{a}_{\mathbf{q}_\perp\mu}(t)-\underset{\mathbf{q}_\perp}{\upsilon_{\nu\mu}}\hat{a}^{\dagger}_{-\mathbf{q}_\perp\mu}(t) \tag{B.45}$$

and

$$i\hbar\frac{\partial}{\partial t}\hat{a}^{\dagger}_{-\mathbf{q}_\perp\nu}(t)=\sum_{\mu}-(\underset{\mathbf{q}_\perp}{\bar{d}_{\nu\mu}}-\underset{\mathbf{q}_\perp}{\upsilon_{\nu\mu}})\hat{a}^{\dagger}_{-\mathbf{q}_\perp\mu}(t)+\underset{\mathbf{q}_\perp}{\upsilon_{\nu\mu}}\hat{a}_{\mathbf{q}_\perp\mu}(t). \tag{B.46}$$

The sign change by the time-derivative of the creation operator has the effect that the first time-derivative of the retarded phonon Green's function is not readily expressed as a function of the Green's function itself. We therefore take the second time-derivative,

$$\begin{aligned}&-\hbar^2\frac{\partial^2}{\partial t^2}D^{0,R}_{\substack{\nu,\mu\\ \mathbf{q}_\perp}}(t,t')\\=&i\hbar\delta(t-t')'\Big\langle(\hat{a}_{\mathbf{q}_\perp\nu}(t)+\hat{a}^{\dagger}_{-\mathbf{q}_\perp\nu}(t))(\hat{a}^{\dagger}_{\mathbf{q}_\perp\mu}(t')+\hat{a}_{-\mathbf{q}_\perp\mu}(t'))\\&\qquad-(\hat{a}^{\dagger}_{\mathbf{q}_\perp\mu}(t')+\hat{a}_{-\mathbf{q}_\perp\mu}(t'))(\hat{a}_{\mathbf{q}_\perp\nu}(t)+\hat{a}^{\dagger}_{-\mathbf{q}_\perp\nu}(t))\Big\rangle\\&+\sum_{\nu'}\underset{\mathbf{q}_\perp}{\bar{d}_{\nu\nu'}}2\delta(t-t')\Big\langle(\hat{a}_{\mathbf{q}_\perp\nu'}(t)-\hat{a}^{\dagger}_{-\mathbf{q}_\perp\nu'}(t))(\hat{a}^{\dagger}_{\mathbf{q}_\perp\mu}(t')+\hat{a}_{-\mathbf{q}_\perp\mu}(t'))\\&\qquad-(\hat{a}^{\dagger}_{\mathbf{q}_\perp\mu}(t')+\hat{a}_{-\mathbf{q}_\perp\mu}(t'))(\hat{a}_{\mathbf{q}_\perp\nu'}(t)-\hat{a}^{\dagger}_{-\mathbf{q}_\perp\nu'}(t))\Big\rangle\\&+\sum_{\nu'\nu''}(\underset{\mathbf{q}_\perp}{\bar{d}_{\nu\nu'}}\underset{\mathbf{q}_\perp}{\bar{d}_{\nu'\nu''}}-2\underset{\mathbf{q}_\perp}{\bar{d}_{\nu\nu'}}\underset{\mathbf{q}_\perp}{\upsilon_{\nu'\nu''}})D^{0,R}_{\substack{\nu'',\mu\\ \mathbf{q}_\perp}}(t,t').\end{aligned} \tag{B.47}$$

The time-derivative of the delta function $\delta(t-t')'$ is ill-defined except inside an integral. We therefore perform a Fourier transform by multiplying with $e^{i\omega(t-t')}$ and integrating over time. Integration by parts of the first term on the right-hand side of (B.47) results in

$$\begin{aligned}&\int_{-\infty}^{+\infty}i\hbar\delta(t-t')'\Big\langle\hat{a}_{\mathbf{q}_\perp\nu}(t)+\hat{a}^{\dagger}_{-\mathbf{q}_\perp\nu}(t))(\hat{a}^{\dagger}_{\mathbf{q}_\perp\mu}(t')+\hat{a}_{-\mathbf{q}_\perp\mu}(t'))\\&\qquad-(\hat{a}^{\dagger}_{\mathbf{q}_\perp\mu}(t')+\hat{a}_{-\mathbf{q}_\perp\mu}(t'))(\hat{a}_{\mathbf{q}_\perp\nu}(t)+\hat{a}^{\dagger}_{-\mathbf{q}_\perp\nu}(t))\Big\rangle e^{i\omega(t-t')}d(t-t')\\&=-\sum_{\nu'}\underset{\mathbf{q}_\perp}{\bar{d}_{\nu\nu'}}\left(\left[\hat{a}_{\mathbf{q}_\perp\nu'}(t),\hat{a}^{\dagger}_{\mathbf{q}_\perp\mu}(t)\right]+\left[\hat{a}_{-\mathbf{q}_\perp\mu}(t),\hat{a}^{\dagger}_{-\mathbf{q}_\perp\nu'}(t)\right]\right)\\&\qquad+\hbar\omega\left(\left[\hat{a}_{\mathbf{q}_\perp\nu'}(t),\hat{a}^{\dagger}_{\mathbf{q}_\perp\mu}(t)\right]-\left[\hat{a}_{-\mathbf{q}_\perp\mu}(t),\hat{a}^{\dagger}_{-\mathbf{q}_\perp\nu'}(t)\right]\right)\\&=-2\sum_{\nu'}\underset{\mathbf{q}_\perp}{\bar{d}_{\nu\nu'}}\delta_{\nu'\mu}=-2\underset{\mathbf{q}_\perp}{\bar{d}_{\nu\mu}}.\end{aligned} \tag{B.48}$$

Similarly, it can be shown that the second term on the right-hand side of (B.47) is equal to $4\bar{d}_{\nu\mu}_{\mathbf{q}_\perp}$. In total we thus have

$$\hbar^2\omega^2 D^{0,R}_{\nu,\mu\,\mathbf{q}_\perp}(\omega) = 2\bar{d}_{\nu\mu\,\mathbf{q}_\perp} + \sum_{\nu'\nu''} \bar{d}_{\nu\nu'\,\mathbf{q}_\perp}\bar{d}_{\nu'\nu''\,\mathbf{q}_\perp} D^{0,R}_{\nu'',\mu\,\mathbf{q}_\perp}(\omega) - \sum_{\nu'\nu''} 2\bar{d}_{\nu\nu'\,\mathbf{q}_\perp} \upsilon_{\nu'\nu''\,\mathbf{q}_\perp} D^{0,R}_{\nu'',\mu\,\mathbf{q}_\perp}(\omega). \tag{B.49}$$

The sum $\sum_{\nu'} \bar{d}_{\nu\nu'\,\mathbf{q}_\perp}\bar{d}_{\nu'\nu''\,\mathbf{q}_\perp}$ corresponds to a matrix squaring. It is readily verified from the definition in (A.37), that the square of a Fourier transformed matrix is equal to the Fourier transform of the square. Additionally, (A.22) showed that the square of $\tilde{\mathbf{d}}$ is equal to the interatomic force constants matrix. We can thus state that $\bar{\mathbf{d}}^2_{\mathbf{q}_\perp} = \mathbf{K}^{tot}_{\mathbf{q}_\perp}$ or, alternatively,

$$\left(\hbar^2\omega^2\mathbf{I} - \mathbf{K}^{tot}_{\mathbf{q}_\perp} + 2\bar{\mathbf{d}}_{\mathbf{q}_\perp}\mathbf{V}_{\mathbf{q}_\perp}\right)\mathbf{D}^{0,R}_{\mathbf{q}_\perp}(\omega) = 2\bar{\mathbf{d}}_{\mathbf{q}_\perp}. \tag{B.50}$$

In a similar fashion, it can be shown that

$$\hbar^2\frac{\partial^2}{\partial t^2} D^{0,\lessgtr}_{\nu,\mu\,\mathbf{q}_\perp}(t,t') = \sum_{\nu'\nu''}(\bar{d}_{\nu\nu'\,\mathbf{q}_\perp}\bar{d}_{\nu'\nu''\,\mathbf{q}_\perp} - 2\bar{d}_{\nu\nu'\,\mathbf{q}_\perp}\upsilon_{\nu'\nu''\,\mathbf{q}_\perp}) D^{0,\lessgtr}_{\nu'',\mu\,\mathbf{q}_\perp}(t,t'), \tag{B.51}$$

and hence,

$$\left(\hbar^2\omega^2 - \mathbf{K}^{tot}_{\mathbf{q}_\perp} + 2\bar{\mathbf{d}}_{\mathbf{q}_\perp}\mathbf{V}_{\mathbf{q}}\right)\mathbf{D}^{0,\lessgtr}_{\mathbf{q}_\perp}(\omega) = \mathbf{0}. \tag{B.52}$$

The forms of (B.50) and (B.52) differ significantly from (B.13) and (B.16). First, as the superscript denotes, $\mathbf{K}^{tot}_{\mathbf{q}_\perp}$ contains all matrix elements, including elements connecting the leads with the device, whereas $\mathbf{H}_{\mathbf{k}_\perp}$ did not. Second, there is an extra term $2\bar{\mathbf{d}}_{\mathbf{q}_\perp}\mathbf{V}_{\mathbf{q}_\perp}$ and the right-hand side is not unity for (B.50). These differences will prevent us from eliminating cross terms when we limit the degrees of freedom to the device. To resolve this, we propose the following transformations,

$$\bar{\mathbf{D}}_{\mathbf{q}_\perp} = \left(2\bar{\mathbf{d}}_{\mathbf{q}_\perp}\right)^{-\frac{1}{2}}\mathbf{D}_{\mathbf{q}_\perp}\left(2\bar{\mathbf{d}}_{\mathbf{q}_\perp}\right)^{-\frac{1}{2}}, \tag{B.53}$$

$$\bar{\mathbf{V}}_{\mathbf{q}_\perp} = \left(2\bar{\mathbf{d}}_{\mathbf{q}_\perp}\right)^{\frac{1}{2}}\mathbf{V}_{\mathbf{q}_\perp}\left(2\bar{\mathbf{d}}_{\mathbf{q}_\perp}\right)^{\frac{1}{2}}, \tag{B.54}$$

$$\bar{\mathbf{\Pi}}_{\mathbf{q}_\perp} = \left(2\bar{\mathbf{d}}_{\mathbf{q}_\perp}\right)^{\frac{1}{2}}\mathbf{\Pi}_{\mathbf{q}_\perp}\left(2\bar{\mathbf{d}}_{\mathbf{q}_\perp}\right)^{\frac{1}{2}}. \tag{B.55}$$

Applying (B.53)-(B.54) to (B.50) and (B.52), we obtain

$$\left(\hbar^2\omega^2\mathbf{I} - \mathbf{K}^{tot}_{\mathbf{q}_\perp} + \bar{\mathbf{V}}_{\mathbf{q}_\perp}\right)\mathbf{D}^{0,R}_{\mathbf{q}_\perp}(\omega) = \mathbf{I}, \tag{B.56}$$

$$\left(\hbar^2\omega^2\mathbf{I} - \mathbf{K}^{tot}_{\mathbf{q}_\perp} + \bar{\mathbf{V}}_{\mathbf{q}_\perp}\right)\mathbf{D}^{0,\lessgtr}_{\mathbf{q}_\perp}(\omega) = \mathbf{0}. \tag{B.57}$$

We can now choose the matrix elements $\upsilon_{\nu\mu\,\mathbf{q}_\perp}$ such that $\bar{\mathbf{V}}_{\mathbf{q}_\perp}$ contains the elements of $\mathbf{K}^{tot}_{\mathbf{q}_\perp}$ connecting the leads with the device. We can then define $\mathbf{K}_{\mathbf{q}_\perp} = \mathbf{K}^{tot}_{\mathbf{q}_\perp} - \bar{\mathbf{V}}_{\mathbf{q}_\perp}$, which has the same block matrix structure as $\mathbf{H}_{\mathbf{k}_\perp}$. (B.56) and (B.57) then have a formally equivalent structure to (B.13) and (B.16). Additionally, the transformations (B.53)-(B.55) leave the Dyson equations (B.38) and (B.39) unchanged. The remainder of the derivation is thus identical to the electron case. Leaving out the ω and $\mathbf{k}_\perp$ dependency in the notation, defining

$$\bar{\mathbf{\Pi}}^l = \bar{\mathbf{V}}\bar{\mathbf{D}}^0\bar{\mathbf{V}} \tag{B.58}$$

and merging the self-energies with (24), we obtain

$$\bar{\mathbf{D}}^R = \left(\hbar^2\omega^2\mathbf{I} - \mathbf{K} - \bar{\mathbf{\Pi}}\right)^{-1}, \tag{B.59}$$

$$\bar{\mathbf{D}}^{\lessgtr} = \bar{\mathbf{D}}^R\bar{\mathbf{\Pi}}^{\lessgtr}\bar{\mathbf{D}}^A. \tag{B.60}$$

The expressions in (21) and (B.59) and (22) and (B.60) differ only by the presence of the convergence term $i\eta$, which we will again add retroactively, and by the bar notation. It is hereby shown that the expressions for the phonon Green's function usually found in literature and derived from semi-classical principles do not pertain to the Green's functions defined in (6), but that additional transformations, (B.53) and (B.55), are required. As will be shown in Appendix C, the transformations (B.53) and (B.55) will result in a modification of the electron-phonon matrix elements in (A.38).

Appendix C. Self-energy expressions

We provide a derivation of the self-energy expressions related to electron-phonon scattering in (31) and (32). The self-energy expressions due to the leads were obtained in Appendix B so from here on out the influence of the leads is neglected. It can be shown that when interactions are introduced, the electron Green's function becomes [18,19]

$$iG_{n,m}(t,t') = \frac{1}{\hbar}\langle T_c\left[e^{\frac{-i}{\hbar}\int_C \hat{H}_I(t_1)dt_1}\hat{c}_n(t)\hat{c}^\dagger_m(t')\right]\rangle, \tag{C.1}$$

where $\hat{H}_I$ is defined in Section 2.1 and the integral is taken over the two-branch contour, defined in Section 2.2. Averaging here is done according to the occupation of the non-interacting and non-contacted and, hence, one-particle states described in Section 2.2 and the operators are described in the interaction picture. Note that an exact treatment actually requires a three-branch contour and that $\hat{H}_I$ requires a separate definition on this third branch [38]. The influence of this third branch is the incorporation of correlation effects after switching on the interactions and contacts. However, these correlations can usually be neglected in steady state [19]. Limiting (C.1) to a second order expansion results in

$$\begin{aligned} iG_{n,m}(t,t') = {} & \frac{1}{\hbar}\langle T_c\left[\hat{c}_n(t)\hat{c}^\dagger_m(t')\right]\rangle \\ & + \langle T_c\left[\frac{-i}{\hbar^2}\int_C \hat{H}_I(t_1)dt_1\hat{c}_n(t)\hat{c}^\dagger_m(t')\right]\rangle \\ & + \langle T_c\left[\frac{-1}{2\hbar^3}\int_C\int_C \hat{H}_I(t_1)\hat{H}_I(t_2)dt_1dt_2\hat{c}_n(t)\hat{c}^\dagger_m(t')\right]\rangle. \end{aligned} \tag{C.2}$$

The first term is just the non-interacting Green's function $iG^0_{n,m}$ defined in (9). The second term can be neglected due to an odd number of phonon creation or annihilation operators. Since averaging is done over non-interacting one-particle states and the operators are non-interacting operators, Wick's theorem can be used to write the third term as [38]

$$\begin{aligned} & \frac{-1}{2\hbar^3 N_\perp}\sum_{\substack{m_1n_1\nu_1\\ \mathbf{k}_{\perp,1}\mathbf{q}_{\perp,1}}} \bar{g}^{\,m_1n_1\nu_1}_{\mathbf{k}_{\perp,1}\mathbf{q}_{\perp,1}}\sum_{\substack{m_2n_2\nu_2\\ \mathbf{k}_{\perp,2}\mathbf{q}_{\perp,2}}} \bar{g}^{\,m_2n_2\nu_2}_{\mathbf{k}_{\perp,2}\mathbf{q}_{\perp,2}} \\ & \times\int_C dt_1\int_C dt_2\,\langle(\hat{a}_{\mathbf{q}_{\perp,1}\nu_1}(t_1) + \hat{a}^\dagger_{-\mathbf{q}_{\perp,1}\nu_1}(t_1))(\hat{a}_{\mathbf{q}_{\perp,2}\nu_2}(t_2) + \hat{a}^\dagger_{-\mathbf{q}_{\perp,2}\nu_2}(t_2))\rangle \\ & \Big(\langle\hat{c}_{\mathbf{k}_{\perp,1}n_1}(t_1)\hat{c}^\dagger_{\underset{m_1}{\mathbf{k}_{\perp,1}+\mathbf{q}_{\perp,1}}}(t_1)\rangle\langle\hat{c}_{\mathbf{k}_{\perp,2}n_2}(t_2)\hat{c}^\dagger_{\underset{m_2}{\mathbf{k}_{\perp,2}+\mathbf{q}_{\perp,2}}}(t_2)\rangle\langle\hat{c}_{\mathbf{k}_\perp n}(t)\hat{c}^\dagger_{\mathbf{k}_\perp m}(t')\rangle \\ & -\langle\hat{c}_{\mathbf{k}_{\perp,2}n_2}(t_2)\hat{c}^\dagger_{\underset{m_1}{\mathbf{k}_{\perp,1}+\mathbf{q}_{\perp,1}}}(t_1)\rangle\langle\hat{c}_{\mathbf{k}_{\perp,1}n_1}(t_1)\rangle\hat{c}^\dagger_{\underset{m_2}{\mathbf{k}_{\perp,2}+\mathbf{q}_{\perp,2}}}(t_2)\langle\hat{c}_{\mathbf{k}_\perp n}(t)\hat{c}^\dagger_{\mathbf{k}_\perp m}(t')\rangle \\ & -\langle\hat{c}_{\mathbf{k}_\perp n}(t)\hat{c}^\dagger_{\underset{m_1}{\mathbf{k}_{\perp,1}+\mathbf{q}_{\perp,1}}}(t_1)\rangle\langle\hat{c}_{\mathbf{k}_{\perp,2}n_2}(t_2)\hat{c}^\dagger_{\underset{m_2}{\mathbf{k}_{\perp,2}+\mathbf{q}_{\perp,2}}}(t_2)\rangle\langle\hat{c}_{\mathbf{k}_{\perp,1}n_1}(t_1)\hat{c}^\dagger_{\mathbf{k}_\perp m}(t')\rangle \\ & -\langle\hat{c}_{\mathbf{k}_{\perp,1}n_1}(t_1)\hat{c}^\dagger_{\underset{m_1}{\mathbf{k}_{\perp,1}+\mathbf{q}_{\perp,1}}}(t_1)\rangle\langle\hat{c}_{\mathbf{k}_\perp n}(t)\hat{c}^\dagger_{\underset{m_2}{\mathbf{k}_{\perp,2}+\mathbf{q}_{\perp,2}}}(t_2)\rangle\langle\hat{c}_{\mathbf{k}_{\perp,2}n_2}(t_2)\hat{c}^\dagger_{\mathbf{k}_\perp m}(t')\rangle \\ & +\langle\hat{c}_{\mathbf{k}_\perp n}(t)\hat{c}^\dagger_{\underset{m_1}{\mathbf{k}_{\perp,1}+\mathbf{q}_{\perp,1}}}(t_1)\rangle\langle\hat{c}_{\mathbf{k}_{\perp,1}n_1}(t_1)\hat{c}^\dagger_{\underset{m_2}{\mathbf{k}_{\perp,2}+\mathbf{q}_{\perp,2}}}(t_2)\rangle\langle\hat{c}_{\mathbf{k}_{\perp,2}n_2}(t_2)\hat{c}^\dagger_{\mathbf{k}_\perp m}(t')\rangle \\ & +\langle\hat{c}_{\mathbf{k}_\perp n}(t)\hat{c}^\dagger_{\underset{m_2}{\mathbf{k}_{\perp,2}+\mathbf{q}_{\perp,2}}}(t_2)\rangle\langle\hat{c}_{\mathbf{k}_{\perp,2}n_2}(t_2)\hat{c}^\dagger_{\underset{m_1}{\mathbf{k}_{\perp,1}+\mathbf{q}_{\perp,1}}}(t_1)\rangle\langle\hat{c}_{\mathbf{k}_{\perp,1}n_1}(t_1)\hat{c}^\dagger_{\mathbf{k}_\perp m}(t')\rangle\Big). \end{aligned} \tag{C.3}$$

The first two terms correspond to disconnected diagrams and can be neglected. The third and fourth term result in an exchange of zero en-

ergy and momentum and can thus also be neglected. The last two terms are identical except for an exchange of the indices and thus cancel the factor of 2 in the denominator. Applying momentum conservation and substituting in (C.3) results in

$$G^{n,m}_{\mathbf{k}_\perp}(t,t') = G^{0}_{n,m\,\mathbf{k}_\perp}(t,t') + \sum_{\substack{m_1 n_1 m_2 n_2 \\ \nu_1 \nu_2 \mathbf{q}_\perp}} \bar{g}^{\,m_1 n_1 \nu_1}_{\mathbf{k}_\perp - \mathbf{q}_\perp, \mathbf{q}_\perp} \bar{g}^{\,m_2 n_2 \nu_2}_{\mathbf{k}_\perp, -\mathbf{q}_\perp} \int_C dt_1 \int_C dt_2 \frac{i\hbar}{N_\perp} G^{0}_{n,m_1\,\mathbf{k}_\perp}(t,t_1) G^{0}_{n_1,m_2\,\mathbf{k}_\perp - \mathbf{q}_\perp}(t_1,t_2) D^{0}_{\nu_1,\nu_2\,\mathbf{q}_\perp}(t_1,t_2) G^{0}_{n_2,m\,\mathbf{k}_\perp}(t_2,t'). \tag{C.4}$$

Note that the summation over indices corresponds to a matrix multiplication. Higher-order terms of the perturbation expansion in (C.3) can be obtained by turning this into a Dyson equation [18]. Replacing all but the leftmost non-interacting Green's functions with interacting Green's functions, we obtain (9) with

$$\left(\mathbf{\Sigma}^s_{\mathbf{k}_\perp}(t_1,t_2)\right)_{n,m} = \frac{i\hbar}{N_\perp} \sum_{\substack{n_1 m_2 \\ \nu_1 \nu_2 \mathbf{q}_\perp}} \bar{g}^{\,m_1 n_1 \nu_1}_{\mathbf{k}_\perp - \mathbf{q}_\perp, \mathbf{q}_\perp} G^{\,n_1,m_2}_{\mathbf{k}_\perp - \mathbf{q}_\perp}(t_1,t_2) D_{\nu_1,\nu_2\,\mathbf{q}_\perp}(t_1,t_2) \bar{g}^{\,m_2 n_2 \nu_2}_{\mathbf{k}_\perp, -\mathbf{q}_\perp}. \tag{C.5}$$

We are interested in an expression in terms of $\tilde{\mathbf{D}}$, which requires an additional transformation according to (B.53). This introduces an extra factor of 2. Additionally, it can be shown from (A.17), (A.18), (A.37), (A.38) and (A.23) that

$$\sum_\nu \bar{g}^{\,nm\nu}_{\mathbf{k}_\perp \mathbf{q}_\perp} (\mathbf{d}^{-}_{\mathbf{q}_\perp}{}^{\frac{1}{2}})_{\nu\nu'} = \bar{\bar{g}}_{\substack{nm\nu' \\ \mathbf{k}_\perp \mathbf{q}_\perp}}, \tag{C.6}$$

$$\sum_\nu \bar{g}^{\,nm\nu}_{\mathbf{k}_\perp, -\mathbf{q}_\perp} (\mathbf{d}^{-}_{\mathbf{q}_\perp}{}^{\frac{1}{2}})_{\nu'\nu} = \bar{\bar{g}}_{\substack{nm\nu' \\ \mathbf{k}_\perp, -\mathbf{q}_\perp}}. \tag{C.7}$$

This, together with (33), results in

$$\mathbf{\Sigma}^s_{\mathbf{k}_\perp}(t_1,t_2) = \frac{2i\hbar}{N_\perp} \sum_{\nu\mu\mathbf{q}_\perp} \mathbf{M}^{\nu}_{\mathbf{k}_\perp - \mathbf{q}_\perp, \mathbf{q}_\perp} \mathbf{G}_{\mathbf{k}_\perp - \mathbf{q}_\perp}(t_1,t_2) \bar{D}_{\nu,\mu\,\mathbf{q}_\perp}(t_1,t_2) \mathbf{M}^{\mu}_{\mathbf{k}_\perp, -\mathbf{q}_\perp}. \tag{C.8}$$

The lesser and greater Green's function can be extracted from the contour-ordered Green's function by confining the time arguments to specific branches. The same is true for the self-energy. Fourier transformation according to (B.6) then gives

$$\mathbf{\Sigma}^{s,\lessgtr}_{\mathbf{k}_\perp}(\omega) = \int_{-\infty}^{\infty} \frac{2i\hbar}{N_\perp} \sum_{\nu\mu\mathbf{q}_\perp} \mathbf{M}^{\nu}_{\mathbf{k}_\perp - \mathbf{q}_\perp, \mathbf{q}_\perp} \mathbf{G}^{\lessgtr}_{\mathbf{k}_\perp - \mathbf{q}_\perp}(\omega - \omega') \bar{D}^{\lessgtr}_{\nu,\mu\,\mathbf{q}_\perp}(\omega') \mathbf{M}^{\mu}_{\mathbf{k}_\perp, -\mathbf{q}_\perp} \frac{d\omega'}{2\pi}. \tag{C.9}$$

Finally, we drop the bar notation on $\bar{D}_{\nu,\mu}$ and use the fact that

$$D^{\lessgtr}_{\nu,\mu\,\mathbf{q}_\perp}(\omega) = D^{\gtrless}_{\mu,\nu\,-\mathbf{q}_\perp}(-\omega) \tag{C.10}$$

to confine the time integral to the positive axis, obtaining (31). Similarly, we can find a perturbation expansion of the phonon Green's function,

$$\begin{aligned} iD_{\nu,\mu\,\mathbf{q}_\perp}(t,t') &= \frac{1}{\hbar} \langle T_c \left[(\hat{a}_{\mathbf{q}_\perp \nu}(t) + \hat{a}^\dagger_{-\mathbf{q}_\perp \nu}(t))(\hat{a}^\dagger_{\mathbf{q}_\perp \mu}(t') + \hat{a}_{-\mathbf{q}_\perp \mu}(t')) \right] \rangle \\ &+ \langle T_c \left[\frac{-i}{\hbar^2} \int_C \hat{H}_I(t_1) dt_1 (\hat{a}_{\mathbf{q}_\perp \nu}(t) + \hat{a}^\dagger_{-\mathbf{q}_\perp \nu}(t))(\hat{a}^\dagger_{\mathbf{q}_\perp \mu}(t') + \hat{a}_{-\mathbf{q}_\perp \mu}(t')) \right] \rangle \\ &+ \left\langle T_c \left[\frac{-1}{2\hbar^3} \int_C \int_C \hat{H}_I(t_1) \hat{H}_I(t_2) dt_1 dt_2 (\hat{a}_{\mathbf{q}_\perp \nu}(t) + \hat{a}^\dagger_{-\mathbf{q}_\perp \nu}(t))(\hat{a}^\dagger_{\mathbf{q}_\perp \mu}(t') + \hat{a}_{-\mathbf{q}_\perp \mu}(t')) \right] \right\rangle. \end{aligned} \tag{C.11}$$

The first term is just the non-interacting phonon Green's function and the second term is zero due to having an odd number of creation or annihilation operators. The third term can be written as the following Wick decomposition,

$$\begin{aligned} &\frac{-1}{2\hbar^3 N_\perp} \sum_{\substack{m_1 n_1 \nu_1 \\ \mathbf{k}_{\perp,1} \mathbf{q}_{\perp,1}}} \bar{g}^{\,m_1 n_1 \nu_1}_{\mathbf{k}_{\perp,1} \mathbf{q}_{\perp,1}} \sum_{\substack{m_2 n_2 \nu_2 \\ \mathbf{k}_{\perp,2} \mathbf{q}_{\perp,2}}} \bar{g}^{\,m_2 n_2 \nu_2}_{\mathbf{k}_{\perp,2} \mathbf{q}_{\perp,2}} \int_C dt_1 \int_C dt_2 \\ &\Big(\langle \hat{c}_{\mathbf{k}_{\perp,1} n_1}(t_1) \hat{c}^\dagger_{\mathbf{k}_{\perp,1} + \mathbf{q}_{\perp,1}\, m_1}(t_1) \rangle \langle \hat{c}_{\mathbf{k}_{\perp,2} n_2}(t_2) \hat{c}^\dagger_{\mathbf{k}_{\perp,2} + \mathbf{q}_{\perp,2}\, m_2}(t_2) \rangle \\ &- \langle \hat{c}_{\mathbf{k}_{\perp,1} n_1}(t_1) \hat{c}^\dagger_{\mathbf{k}_{\perp,2} + \mathbf{q}_{\perp,2}\, m_2}(t_2) \rangle \langle \hat{c}_{\mathbf{k}_{\perp,2} n_2}(t_2) \hat{c}^\dagger_{\mathbf{k}_{\perp,1} + \mathbf{q}_{\perp,1}\, m_1}(t_1) \rangle \Big) \\ &\Big(\langle (\hat{a}_{\mathbf{q}_{\perp,1} \nu_1}(t_1) + \hat{a}^\dagger_{-\mathbf{q}_{\perp,1} \nu_1}(t_1))(\hat{a}_{\mathbf{q}_{\perp,2} \nu_2}(t_2) + \hat{a}^\dagger_{-\mathbf{q}_{\perp,2} \nu_2}(t_2)) \rangle \\ &\quad \langle (\hat{a}_{\mathbf{q}_\perp \nu}(t) + \hat{a}^\dagger_{-\mathbf{q}_\perp \nu}(t))(\hat{a}^\dagger_{\mathbf{q}_\perp \mu}(t') + \hat{a}_{-\mathbf{q}_\perp \mu}(t')) \rangle \\ &+ \langle (\hat{a}_{\mathbf{q}_\perp \nu}(t) + \hat{a}^\dagger_{-\mathbf{q}_\perp \nu}(t))(\hat{a}_{\mathbf{q}_{\perp,1} \nu_1}(t_1) + \hat{a}^\dagger_{-\mathbf{q}_{\perp,1} \nu_1}(t_1)) \rangle \\ &\quad \langle (\hat{a}_{\mathbf{q}_{\perp,2} \nu_2}(t_2) + \hat{a}^\dagger_{-\mathbf{q}_{\perp,2} \nu_2}(t_2))(\hat{a}^\dagger_{\mathbf{q}_\perp \mu}(t') + \hat{a}_{-\mathbf{q}_\perp \mu}(t')) \rangle \\ &+ \langle (\hat{a}_{\mathbf{q}_\perp \nu}(t) + \hat{a}^\dagger_{-\mathbf{q}_\perp \nu}(t))(\hat{a}_{\mathbf{q}_{\perp,2} \nu_2}(t_2) + \hat{a}^\dagger_{-\mathbf{q}_{\perp,2} \nu_2}(t_2)) \rangle \\ &\quad \langle (\hat{a}_{\mathbf{q}_{\perp,1} \nu_1}(t_1) + \hat{a}^\dagger_{-\mathbf{q}_{\perp,1} \nu_1}(t_1))(\hat{a}^\dagger_{\mathbf{q}_\perp \mu}(t') + \hat{a}_{-\mathbf{q}_\perp \mu}(t')) \rangle \Big). \end{aligned} \tag{C.12}$$

The first term in the electron part corresponds to an exchange of zero energy and momentum and will be neglected. The first term of the phonon part corresponds to a disconnected diagram and will be neglected as well. The last two terms of the phonon part are equal except for an exchange of indices and can thus be summed to compensate for the factor of 2 in the denominator. Applying momentum conservation, adding a factor due to spin degeneracy and substituting in (C.11) results in

$$D_{\nu,\mu\,\mathbf{q}_\perp}(t,t') = D^{0}_{\nu,\mu\,\mathbf{q}_\perp}(t,t') - \frac{n_s i\hbar}{N_\perp} \sum_{\substack{m_1 n_1 \nu_1 \\ m_2 n_2 \nu_2 \\ \mathbf{k}_\perp}} \bar{g}^{\,m_1 n_1 \nu_1}_{\mathbf{k}_\perp, -\mathbf{q}_\perp} \bar{g}^{\,m_2 n_2 \nu_2}_{\mathbf{k}_\perp - \mathbf{q}_\perp, \mathbf{q}_\perp} \int_C dt_1 \int_C dt_2 G^{0}_{n_1,m_2\,\mathbf{k}_\perp}(t_1,t_2) G^{0}_{n_2,m_1\,\mathbf{k}_\perp - \mathbf{q}_\perp}(t_2,t_1) D^{0}_{\nu,\nu_1\,\mathbf{q}_\perp}(t,t_1) D^{0}_{\nu_2,\mu\,\mathbf{q}_\perp}(t_2,t'). \tag{C.13}$$

Converting to a Dyson equation to include higher-order perturbation terms, we obtain (10) with

$$\left(\mathbf{\Pi}^s_{\mathbf{q}_\perp}(t_1,t_2)\right)_{\nu,\mu} = -\frac{n_s i\hbar}{N_\perp} \sum_{\substack{m_1 n_1 \\ m_2 n_2 \mathbf{k}_\perp}} \bar{g}^{\,m_1 n_1 \nu}_{\mathbf{k}_\perp, -\mathbf{q}_\perp} G^{\,n_1,m_2}_{\mathbf{k}_\perp}(t_1,t_2) \bar{g}^{\,m_2 n_2 \mu}_{\mathbf{k}_\perp - \mathbf{q}_\perp, \mathbf{q}_\perp} G^{\,n_2,m_1}_{\mathbf{k}_\perp - \mathbf{q}_\perp}(t_2,t_1). \tag{C.14}$$

We want an expression for $\bar{\mathbf{\Pi}}$ however. Using (B.55), (C.6), (C.7) and (33), it can be shown that

$$\left(\bar{\mathbf{\Pi}}^s_{\mathbf{q}_\perp}(t_1,t_2)\right)_{\nu,\mu} = -\frac{2n_s i\hbar}{N_\perp} \sum_{\mathbf{k}_\perp} \mathrm{Tr}\left(\mathbf{M}^{\nu}_{\mathbf{k}_\perp, -\mathbf{q}_\perp} \mathbf{G}_{\mathbf{k}_\perp}(t_1,t_2) \mathbf{M}^{\mu}_{\mathbf{k}_\perp - \mathbf{q}_\perp, \mathbf{q}_\perp} \mathbf{G}_{\mathbf{k}_\perp - \mathbf{q}_\perp}(t_2,t_1) \right). \tag{C.15}$$

Confining the time arguments to specific branches and Fourier transforming according to (B.6) then gives

$$\begin{aligned} \left(\bar{\mathbf{\Pi}}^{s,\lessgtr}_{\mathbf{q}_\perp}(\omega)\right)_{\nu,\mu} &= \int_{-\infty}^{+\infty} -\frac{2n_s i\hbar}{N_\perp} \\ &\times \sum_{\mathbf{k}_\perp} \mathrm{Tr}\left(\mathbf{M}^{\nu}_{\mathbf{k}_\perp, -\mathbf{q}_\perp} \mathbf{G}^{\lessgtr}_{\mathbf{k}_\perp}(\omega') \mathbf{M}^{\mu}_{\mathbf{k}_\perp - \mathbf{q}_\perp, \mathbf{q}_\perp} \mathbf{G}^{\gtrless}_{\mathbf{k}_\perp - \mathbf{q}_\perp}(\omega' - \omega) \right) \frac{d\omega'}{2\pi}. \end{aligned} \tag{C.16}$$

Dropping the bar notation on $\left(\bar{\mathbf{\Pi}}_{\mathbf{q}_\perp}(t_1,t_2)\right)_{\nu,\mu}$ results in (32).

Appendix D. Error estimates for the FFT-based self-energy computation

D.1. The integration error

The grid refinement strategy and corresponding integration error are illustrated by the results shown in Fig. D.16. First, we obtained a potential profile within the device described in Section 3.2 with a fine initial energy grid. Then, the macroscopic parameters of interest were obtained with a single ballistic non-self-consistent iteration for several initial grid sizes. Fig. D.16 (a) shows the relative difference compared to a single ballistic iteration with a fine initial grid of 2000 energy points for both the electrons and phonons as an estimate of the integration error. Fig. D.16 (b) shows the number of adaptively added grid points as a function of the initial grid size. Note that significantly more points are added to the phonon energy grid as the full phonon dispersion needs to be integrated. For the electrons only the bottom of the conduction band is of interest. Additionally, it can be seen that a minimum of initial grid points is required to achieve accurate integration despite the adaptive grid. This can be understood from Fig. 7 (a) where the second peak is nearly missed by the initial grid. The adaptive grid will not refine itself around features that are too fine to be captured by the initial grid. From Fig. D.16 (b) we can see that the number of adaptively added grid points increases slightly up to initial grid sizes of 100-250 points. For higher number of initial grid sizes, the number of adaptively added points does not increase or even decreases as some of the adaptively added points are now already introduced by the initial grid. Fig. D.16 (a) shows that an initial grid size of ∼100 points is sufficient to reach the 1% error threshold on the macroscopic properties of interest. Further refinement of the initial grid does reduce the error further, but not by introducing more adaptive grid points as the error threshold is already reached.

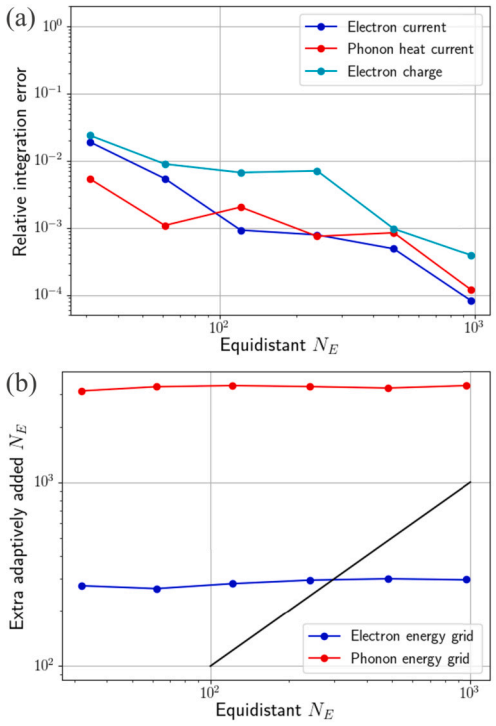


Fig. D.16. (a) The relative integration error on the electron current, phonon heat current and charge density for a ballistic simulation of the system in Section 3.2 as a function of the number of equidistant grid energy points. The error is obtained as the difference with a fine-initial-grid simulation with 2000 initial grid points. A temperature difference of 0.1 K was applied to provide a net heat current through the device. (b) The number of adaptively added points as a function of the number of equidistant initial grid points, for both electrons and phonons. The black line denotes the cross-over line where the adaptive grid points dominate the computational cost.

D.2. The energy mixing error

An estimate of the energy mixing error is shown in Fig. D.17, which shows the self-energies of the system described in Section 3.2 after one non-self-consistent iteration. The potential was obtained with a fine initial energy grid, after which the energy grids were limited to an equidistant grid with 32 points with no further adaptive grid refinements and a single k-point for the calculation of the self-energy. The removal of adaptive grid refinements was to eliminate the self-energy interpolation error and the Green's function conversion error. Fig. D.17 (a) and (b) clearly shown the band gap and the bottom of the conduction band. Fig. D.17 (c) and (d) show that the lesser self-energy in the middle of the band gap does not have a single correct digit.

An estimate of the errors on the macroscopic device properties for a simulation with 10 k-points are shown in Fig. D.18 as a function of the number of equidistant energy grid points. It can be seen that the energy mixing error increases with the number of energy points. However, for the number of energy points considered here, the energy mixing error is still multiple orders of magnitude smaller than the integration error in Fig. D.16 (a). The large relative error on the self-energy thus does not pose problems for obtaining correct macroscopic parameters. This can be attributed to the fact that the large relative error is present at energy ranges where the density of states is low, and hence, at energies which do not have a large effect on the macroscopic properties.

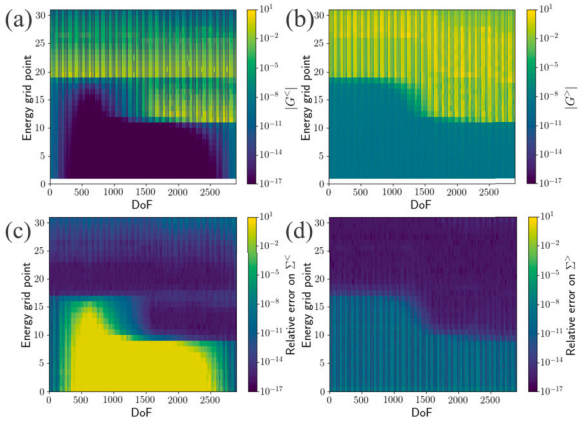


Fig. D.17. Green's function and self-energy results after a single self-consistent Born iteration for the system in Section 3.2. Only a single k-point and 32 energy grid points were used without further energy grid refinements. (a) and (b) show the lesser and greater Green's function as a function of energy and the degree of freedom, i.e., atom orbital index, of the device. (c) and (d) show the relative error on the lesser and greater self-energy, calculated as the difference between the self-energy calculated with the conventional convolution-based implementation and the FFT-based implementation.

D.3. The self-energy interpolation error

An estimate of the self-energy interpolation error is shown in Fig. D.19 as a function of the number of initial equidistant energy grid points. To reach a relative error below 1% on all macroscopic parameters, ∼1000 initial equidistant grid energy points are required. The self-energy interpolation energy thus requires more equidistant initial grid points than the integration error.

D.4. The Green's function conversion error

An estimate of the Green's function conversion errors are shown in Fig. D.20 (a) and (b) for the "leave-out" and "averaged" approach, respectively. Fig. D.20 (a) demonstrates that the "leave-out" approach shows no or very slow convergence of the macroscopic parameters with

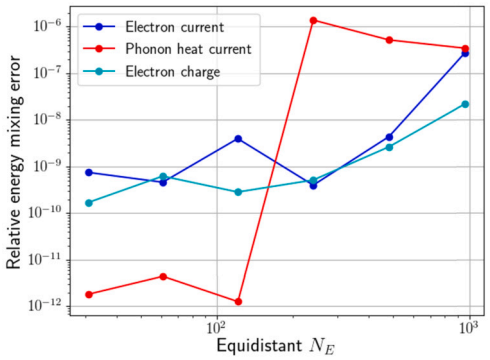


Fig. D.18. The relative energy mixing error on the electron current, phonon heat current and charge density for the system in Section 3.2 as a function of the number of equidistant grid energy points without further refinements. The error is calculated as the difference between the calculation using the conventional convolution-based implementation and the FFT-based implementation.

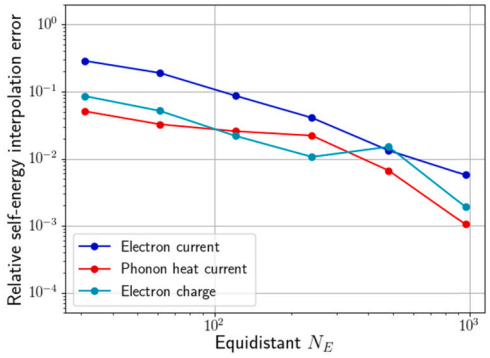


Fig. D.19. The relative self-energy interpolation error on the electron current, phonon heat current and charge density for the system in Section 3.2 as a function of the number of initial equidistant grid energy points. The error is calculated as the difference between the results when the self-energy is calculated with the conventional convolution-based implementation for every energy grid point and when it is done only for the initial grid with the rest obtained through interpolation.

increasing initial equidistant grid sizes. For the "averaged" approach, the electronic properties of the system converge and reach a relative error below 1% for ∼500 initial equidistant grid energy points. The phonon heat current, however, demonstrates no or very slow convergence, resulting in a relative error of a few percent at high numbers of initial equidistant grid energy points. Although unfortunate, this is not surprising nor does it invalidate the FFT-based implementation.

As indicated by Fig. 8, the "averaged" approach has the effect of shifting the mass of the spectra to different energies. The shift in energy is at most the distance between equidistant grid points. For the electron self-energy, this does not pose any problems as the shift in energy will be on the same order of magnitude as the electron grid, for the electron Green's function, or much smaller, for the phonon Green's function. For the phonon self-energy, however, this does introduce a slightly larger error as the shift in energy happens on the electron energy grid and can thus be significantly larger than the distance between phonon energy grid points. Additionally, a shift in energy also results in a shift in the energy of the phonons that are created in the system, which has a direct influence on the phonon heat current. The electronic properties, on the other hand, are less dependent on the energy of the electrons.

Appendix E. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cpc.2024.109430.

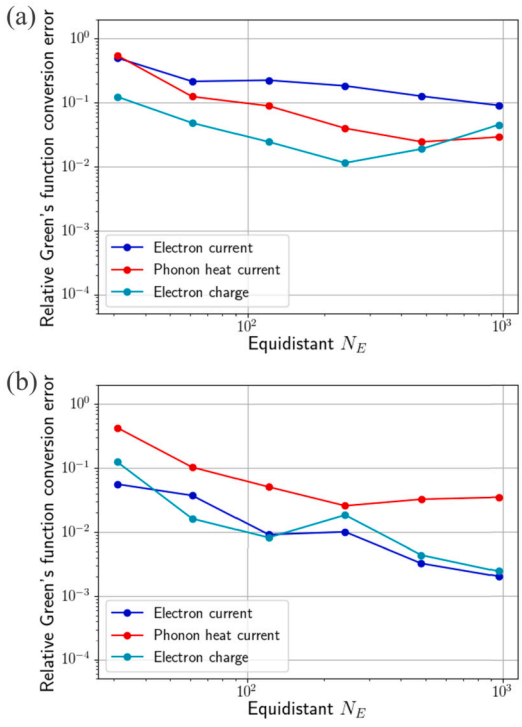


Fig. D.20. The relative Green's function conversion error on the electron current, phonon heat current and charge density for the system in Section 3.2 as a function of the number of initial equidistant grid energy points. The error is calculated as the difference between the results when the calculation is calculated with the conventional convolution-based implementation using the full adaptive grid and the FFT-based implementation using a converted equidistant grid. In both cases the self-energies are only calculated on the initial equidistant grid and the self-energies for adaptively added points is obtained by interpolation. (a) shows the results for the leave-out strategy of Fig. 8 (a) and (b) shows the results for the averaged strategy of Fig. 8 (b).

Data availability

The data that has been used is confidential.

References

[1] F. Schwierz, J. Pezoldt, R. Granzner, Two-dimensional materials and their prospects in transistor electronics, Nanoscale 7 (18) (2015) 8261–8283.
[2] D. Logoteta, Q. Zhang, G. Fiori, What can we really expect from 2d materials for electronic applications?, in: 72nd Device Research Conference, IEEE, 2014.
[3] M. Chhowalla, D. Jena, H. Zhang, Two-dimensional semiconductors for transistors, Nat. Rev. Mater. 1 (11) (2016) 16052.
[4] J. Kang, W. Cao, X. Xie, D. Sarkar, W. Liu, K. Banerjee, Graphene and beyond-graphene 2d crystals for next-generation green electronics, in: Micro-and Nanotechnology Sensors, Systems, and Applications VI, vol. 9083, International Society for Optics and Photonics, 2014, p. 908305.
[5] A. Afzalian, Ab initio perspective of ultra-scaled CMOS from 2d-material fundamentals to dynamically doped transistors, npj 2D Mater. Appl. 5 (1) (2021), https://doi.org/10.1038/s41699-020-00181-1.
[6] A. Afzalian, G. Pourtois, Atomos: an atomistic modelling solver for dissipative dft transport in ultra-scaled hfs2 and black phosphorus mosfets, in: 2019 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), 2019, pp. 1–4.
[7] R. Duflou, M. Houssa, A. Afzalian, Electron-phonon scattering in cold-metal contacted two-dimensional semiconductor devices, in: 2021 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), 2021, pp. 94–97.
[8] J.A. Rowlette, K.E. Goodson, Fully coupled nonequilibrium electron–phonon transport in nanometer-scale silicon fets, IEEE Trans. Electron Devices 55 (1) (2008) 220–232, https://doi.org/10.1109/TED.2007.911043.
[9] R. Rhyner, M. Luisier, Atomistic modeling of coupled electron-phonon transport in nanowire transistors, Phys. Rev. B 89 (2014) 235311, https://doi.org/10.1103/PhysRevB.89.235311.
[10] A. Afzalian, E. Akhoundi, G. Gaddemane, R. Duflou, M. Houssa, Advanced dft–negf transport techniques for novel 2-d material and device exploration includ-

ing hfs2/wse2 van der Waals heterojunction tfet and wte2/ws2 metal/semiconductor contact, IEEE Trans. Electron Devices 68 (11) (2021) 5372–5379, https://doi.org/10.1109/TED.2021.3078412.
[11] R. Duflou, M. Houssa, A. Afzalian, Ballistic heat transport in mos2 monolayers, in: 2023 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), 2023.
[12] L. Chen, Z. Yan, S. Kumar, Coupled electron-phonon transport and heat transfer pathways in graphene nanostructures, Carbon 123 (2017) 525–535, https://doi.org/10.1016/j.carbon.2017.07.095.
[13] J.T. Lü, J.-S. Wang, Coupled electron and phonon transport in one-dimensional atomic junctions, Phys. Rev. B 76 (2007) 165418, https://doi.org/10.1103/PhysRevB.76.165418.
[14] Y. Kajiwara, N. Mori, Nonequilibrium Green function simulation of coupled electron–phonon transport in one-dimensional nanostructures, Jpn. J. Appl. Phys. 58 (SD) (2019) SDDE05, https://doi.org/10.7567/1347-4065/ab0df3.
[15] F. Giustino, Electron-phonon interactions from first principles, Rev. Mod. Phys. 89 (1) (Feb 2017), https://doi.org/10.1103/revmodphys.89.015003.
[16] N. Marzari, A.A. Mostofi, J.R. Yates, I. Souza, D. Vanderbilt, Maximally localized Wannier functions: theory and applications, Rev. Mod. Phys. 84 (2012) 1419–1475, https://doi.org/10.1103/RevModPhys.84.1419.
[17] F. Giustino, M.L. Cohen, S.G. Louie, Electron-phonon interaction using Wannier functions, Phys. Rev. B 76 (2007) 165108, https://doi.org/10.1103/PhysRevB.76.165108.
[18] A.L. Fetter, J.D. Walecka, Quantum Theory of Many-Particle Systems, McGraw-Hill, Boston, 1971.
[19] J. Maciejko, An Introduction to Nonequilibrium Many-Body Theory, Lecture Notes, Springer, 2007.
[20] A. Svizhenko, M. Anantram, T. Govindan, B. Biegel, R. Venugopal, Two-dimensional quantum mechanical modeling of nanotransistors, J. Appl. Phys. 91 (4) (2002) 2343–2354.
[21] See the Supplemental Material for a rigorous treatment of the interaction terms between the phonons in the leads and the device.
[22] R. Lake, R. Pandey, Non-equilibrium green functions in electronic device modeling, preprint, arXiv:cond-mat/0607219, 08 2006.
[23] S. Datta, Quantum Transport: Atom to Transistor, Cambridge University Press, 2005.
[24] M.P.L. Sancho, J.M.L. Sancho, J.M.L. Sancho, J. Rubio, Highly convergent schemes for the calculation of bulk and surface green functions, J. Phys. F, Met. Phys. 15 (4) (1985) 851–858, https://doi.org/10.1088/0305-4608/15/4/009.
[25] N. Mingo, L. Yang, Phonon transport in nanowires coated with an amorphous material: an atomistic Green's function approach, Phys. Rev. B 68 (2003) 245406, https://doi.org/10.1103/PhysRevB.68.245406.
[26] N. Mingo, Thermal nanosystems and nanomaterials, in: Ch. Green's Function Methods for Phonon Transport Through Nano-Contacts, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 63–94.
[27] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G.L. Chiarotti, M. Cococcioni, I. Dabo, A.D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A.P. Seitsonen, A. Smogunov, P. Umari, R.M. Wentzcovitch, QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials, J. Phys. Condens. Matter 21 (39) (2009) 395502, https://doi.org/10.1088/0953-8984/21/39/395502.
[28] A. Al-Hilli, B. Evans, The preparation and properties of transition metal dichalcogenide single crystals, J. Cryst. Growth 15 (2) (1972) 93–101, https://doi.org/10.1016/0022-0248(72)90129-7.
[29] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu, J. Chem. Phys. 132 (15) (2010) 154104, https://doi.org/10.1063/1.3382344.
[30] J.-J. Zhou, J. Park, I.-T. Lu, I. Maliyov, X. Tong, M. Bernardi, Perturbo: a software package for ab initio electron–phonon interactions, charge transport and ultrafast dynamics, Comput. Phys. Commun. 264 (2021) 107970, https://doi.org/10.1016/j.cpc.2021.107970.
[31] Y. Lee, S. Fiore, M. Luisier, Ab initio mobility of single-layer mos2 and ws2: comparison to experiments and impact on the device characteristics, in: 2019 IEEE International Electron Devices Meeting (IEDM), 2019, pp. 24.4.1–24.4.4.
[32] A. Afzalian, F. Ducry, Pushing the limits of ab-initio-negf transport using efficient dissipative mode-space algorithms for realistic simulations of low-dimensional semiconductors including their oxide interfaces, in: 2023 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), 2023, pp. 305–308.
[33] M.V. Fischetti, W.G. Vandenberghe, Mermin-Wagner theorem, flexural modes, and degraded carrier mobility in two-dimensional crystals with broken horizontal mirror symmetry, Phys. Rev. B 93 (2016) 155413, https://doi.org/10.1103/PhysRevB.93.155413.
[34] Y. Chu, P. Sarangapani, J. Charles, G. Klimeck, T. Kubis, Explicit screening full band quantum transport model for semiconductor nanodevices, J. Appl. Phys. 123 (24) (2018) 244501, https://doi.org/10.1063/1.5031461.
[35] A. Afzalian, J.-P. Colinge, D. Flandre, Physics of gate modulated resonant tunneling (rt)-fets: multi-barrier mosfet for steep slope and high on-current, Solid-State Electron. 59 (1) (2011) 50–61, https://doi.org/10.1016/j.sse.2011.01.016.
[36] A. Afzalian, G. Doornbos, T.-M. Shen, M. Passlack, J. Wu, A high-performance inas/gasb core-shell nanowire line-tunneling tfet: an atomistic mode-space negf study, IEEE J. Electron. Devices Soc. 7 (2019) 88–99, https://doi.org/10.1109/JEDS.2018.2881335.
[37] S. Datta, Quantum Transport: Atom to Transistor, Cambridge University Press, 2005.
[38] M. Wagner, Expansions of nonequilibrium Green's functions, Phys. Rev. B 44 (12) (1991) 6104.