First-Principles Simulation of Nonequilibrium Coupled Electron–Phonon Dynamics: Algorithms, Acceleration, and Coherent Phenomena

Thesis by
Jia Yao

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2026
Defended January 22, 2026

Jia Yao
ORCID: 0000-0002-3250-6132

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Marco Bernardi, for his guidance, patience, and steadfast support throughout my Ph.D. Marco's mentorship played a central role in shaping both this thesis. He consistently provided detailed and thoughtful feedback on my work, encouraging clarity, rigor, and physical interpretation at every stage of the research process. I am particularly grateful for his leniency and understanding during challenging periods. Beyond scientific guidance, Marco has been supportive of my career development, offering invaluable encouragement, perspective, and advice.

I am grateful to my collaborators at Lawrence Livermore National Laboratory, including Carol Woodward, David Gardner, and Daniel Reynolds, for their collaboration, technical expertise, and many productive discussions. Their contributions were essential to this thesis. I would also like to thank my thesis committee members, David Hsieh, Scott Cushing, and Gil Refael, for their insightful feedback, thoughtful questions, and guidance.

I would like to thank the members of the Bernardi research group for creating a collaborative and supportive research environment. I am especially grateful to Ivan Maliyov, Shiyu Peng, Sergei Kliavinek, and Yao Luo for their close collaborations and important contributions to the core projects of this thesis. I also thank other current and former group members – Tommaso Chiarotti, Khoa Le, Miles Johnson, Ina Sørensen, Dhruv Desai, Laurie Tan, David Abramovitch, Jinsoo Park, and Jinjian Zhou for their feedback during group meetings, talks, and informal discussions, which consistently improved the quality of my work. I am grateful to John Leung, a SURF student, for his contributions to parts of the thesis.

Finally, I would like to express my deepest thanks to my parents. Being overseas for many years and unable to return home frequently, I have relied on their understanding, patience, and encouragement from afar. I am also grateful to my friends at Caltech and beyond for their companionship and support. In particular, I thank my friend and former roommate, Meichen Fang, for her friendship and meaningful academic and personal exchanges during my time at Caltech, and Priscilla Chan for meaningful conversations outside of my PhD. Lastly, I would like to thank my fiancé, Samuel Lee, for his patience, understanding, and unwavering support. His encouragement and belief in me made this long and challenging path possible.

ABSTRACT

Studying nonequilibrium charge and heat transport is central to the design and control of modern electronic, optoelectronic, and energy materials. On ultrafast timescales, these processes, such as carrier relaxation, lattice heating, and lattice-driven changes of material properties are governed by coupled electron-phonon dynamics. While first-principles methods based on density functional theory enable quantitative predictions of electron-phonon and phonon-phonon (ph-ph) interactions, their extension to real-time, fully coupled nonequilibrium simulations remains computationally challenging, particularly for complex materials and long simulation times.

This thesis develops and advances a comprehensive first-principles framework for simulating nonequilibrium coupled electron-phonon dynamics within the real-time Boltzmann transport equation (rt-BTE). Several algorithmic and computational strategies are introduced to overcome the prohibitive cost of such simulations. First, adaptive and multirate time-integration schemes are developed to efficiently resolve disparate electronic and phononic timescales, enabling accurate simulations of coupled dynamics in complex materials with anharmonic ph-ph interactions. Second, dynamic mode decomposition is applied to extrapolate long-time behavior from short-time simulations, providing efficient access to steady-state and transient transport regimes. Third, GPU parallelization and algorithm optimization are implemented to accelerate the evaluation of collision integrals, yielding substantial performance improvements on modern high-performance computing architectures. Fourth, tensor-learning and compression techniques are introduced to reduce the computational and memory costs associated with ph-ph interactions, further enabling simulations previously inaccessible due to system size or interaction complexity.

Beyond algorithmic acceleration, this work extends the theoretical description of nonequilibrium lattice dynamics from the rt-BTE by introducing a framework accounting for coherently driven phonon dynamics, bridging the gap between incoherent transport and coherent ultrafast phenomena observed in modern pump-probe experiments. Together, these developments expand the efficiency and scope of first-principles simulations of nonequilibrium electron-phonon dynamics, providing new tools for studying transport, relaxation, and coherent control in quantum materials.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] Yao, J., Maliyov, I., Gardner, D. J., Woodward, C. S. & Bernardi, M. Advancing simulations of coupled electron and phonon nonequilibrium dynamics using adaptive and multirate time integration. *npj Comput. Mater.* **11**, 256 (2025). https://www.nature.com/articles/s41524-025-01738-8.
J.Y. participated in conception of the project, conducted computational research, and wrote the manuscript.

[2] Yao, J., Gardner, D. J., Reynolds, D. R., Woodward, C. S. & Bernardi, M. Towards optimal adaptive multirate time integration for coupled electron and phonon nonequilibrium dynamics simulations. *In preparation* (2026).
J.Y. contributed to conception of the project, conducted computational research, and participated in manuscript preparation.

[3] Peng, S., Pinkston, D., Yao, J., Kliavinek, S., Maliyov, I. & Bernardi, M. Efficient GPU parallelization of electronic transport and nonequilibrium dynamics from electron-phonon interactions in the perturbo code (2025). https://arxiv.org/abs/2511.03683.
J.Y. contributed to algorithm design, simulation setup, and writing of the manuscript.

[4] Luo, Y., Mangtani, D., Peng, S., Yao, J., Kliavinek, S. & Bernardi, M. Tensor learning and compression of n-phonon interactions. *Phys. Rev. Lett.* **135**, 126101 (2025). https://link.aps.org/doi/10.1103/nmgj-yq1g.
J.Y. contributed to the calculation setup and software development.

[5] Maliyov, I., Yin, J., Yao, J., Yang, C. & Bernardi, M. Dynamic mode decomposition of nonequilibrium electron-phonon dynamics: accelerating the first-principles real-time boltzmann equation. *npj Comput. Mater.* **10**, 123 (2024). https://www.nature.com/articles/s41524-024-01308-4.
J.Y. participated in the computational research and manuscript preparation.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Overview

The study of charge and heat transport lies at the core of materials physics and electronic technologies. In crystalline solids, the motion of charge carriers determines electrical conductivity, carrier mobility, and optical response, and thus underpins the design of electronic and optoelectronic devices. Heat transport, in contrast, is dominated in non-metallic solids by lattice vibrations, which control thermal conductivity and energy dissipation, impacting device performance and reliability.

As charge carriers propagate through crystals, lattice vibrations perturb the periodic potential experienced by the carriers, leading to scattering processes that redistribute carrier energy and momentum, while simultaneously mediating the transfer of energy from the electronic system to the lattice. These collective modes of lattice vibration can be described in terms of phonons, which provide a quantum-mechanical framework for understanding classical anomalies in heat capacity and thermal conductivity. Such scattering processes, described by electron-phonon (*e*-ph) interactions, are responsible for a wide range of many-body phenomena [1]. They set the intrinsic limits on carrier mobility [2, 3], determine temperature-dependent transport and optical linewidths [4, 5], and govern high-field transport [6, 7]. Beyond phonon-mediated superconductivity [8], they give rise to charge density waves in low-dimensional materials [9], where lattice distortions couple strongly to Fermi surface nesting. In polar materials, *e*-ph interactions can also lead to the formation of polarons and renormalize electronic band structures [10].

On the timescales of femtoseconds to hundreds of picoseconds, nonequilibrium dynamics with *e*-ph interactions provide additional avenues for engineering material properties [11, 12]. This broad range of applications includes ultrafast switching in memory materials, heat management and thermoelectrics, enhancement of carrier extraction in photovoltaic and optoelectronic devices, and strain-engineered functionality in next-generation electronic materials. Recent experimental advances have demonstrated controlled carrier thermalization [13], light-induced superconductivity [14, 15] and phase transitions [16–18], spin dynamics driven by optically excited phonons [19, 20], as well as driven nonlinear lattice dynamics [21–23].

The nonequilibrium carrier dynamics described above can be probed using a variety of experimental techniques. Among the most widely-used are transient absorption spectroscopies, which measure optical or X-ray absorption signals arising from electronic state filling, band renormalization, and changes in the dielectric function [24–26]. From these measurements, one can extract information on carrier occupations, recombination dynamics, and lattice heating. Over the past decade, time- and angle-resolved photoemission spectroscopy (tr-ARPES) has emerged as an established tool for studying nonequilibrium electronic structure [27–32]. In a typical tr-ARPES experiment, the material is first excited by an ultrafast pump pulse and subsequently probed by a time-delayed photoemission pulse that enables reconstruction of the transient electronic structure. While optical spectroscopies provide momentum-integrated access to carrier and lattice dynamics, tr-ARPES offers direct, momentum-resolved insight into nonequilibrium electronic structure, scattering rates, and carrier relaxation pathways.

While e-ph interactions govern the initial stages of carrier relaxation, subsequent lattice dynamics and thermalization over the timescale of picoseconds to nanoseconds are dominated by the phonon subsystem. In the harmonic approximation, lattice vibrations are treated as independent phonon modes with infinite lifetimes, providing a useful starting point for systems near equilibrium. However, anharmonicity in real solids gives rise to phonon-phonon (ph-ph) interactions, which enable energy exchange between different phonon modes and lead to finite phonon lifetimes [33, 34]. The ph-ph interactions are responsible for thermal expansion, temperature-dependent shifts in phonon frequencies, and setting the intrinsic limits on lattice thermal conductivity [35, 36]. In nonequilibrium scenarios, ph-ph interactions govern the redistribution of energy among phonon modes following initial excitation by hot carriers or external perturbations, ultimately leading to thermalization of the lattice subsystem. Phonon dynamics under nonequilibrium conditions in turn influence carrier relaxation, creating effects such as hot phonon bottlenecks that can prolong carrier cooling times [37].

Several classes of materials display strong phonon anharmonicity over a wide temperature range that must be explicitly modeled to accurately describe thermal transport and carrier cooling rates. Prominent examples include IV–VI compounds such as PbTe and SnSe, whose exceptionally low thermal conductivity and soft transverse optical phonons make them attractive candidates for thermoelectric applications [38, 39]. Lead halide perovskites ($APbX_3$) likewise exhibit pronounced

anharmonicity, which gives rise to slow carrier cooling through acoustic phonon upconversion and hot-phonon bottlenecks [40]. Reduced dimensionality also enhances anharmonic effects in monolayer transition-metal dichalcogenides and related two-dimensional materials, where low energy flexural modes coexist with tunable electronic structure, strong spin–orbit coupling (SOC), and other interesting many-body phenomena such as excitonic resonances [41]. In addition, systems like GeTe and $SrTiO_3$ undergo structural phase transitions associated with the softening of phonon modes [42, 43]. Beyond these widely studied cases, strong ph-ph interactions also play a central role in a broader range of materials, including complex oxides, skutterudites, layered chalcogenides, and more.

Experimental methods for probing nonequilibrium phonon dynamics include Raman spectroscopy for zone-center phonons [44], ultrafast electron diffraction (UED) [45], transient thermal grating [46], time-resolved X-ray diffraction [47, 48], and inelastic X-ray scattering (IXS) [49, 50]. In particular, IXS provides momentum-resolved information on phonon dispersions and lifetimes, enabling extraction of phonon mode softening, linewidth broadening, and other anharmonic effects from ph-ph interactions at elevated temperatures or under nonequilibrium conditions. By probing both electronic and lattice degrees of freedom, these experimental approaches together provide a comprehensive picture of the coupled electron-phonon nonequilibrium dynamics.

In parallel to the ongoing experiments, theoretical study of coupled electron-phonon dynamics has progressed from simplified model Hamiltonians to large-scale computational approaches based on first-principles methods. Early work relied on phenomenological models such as deformation potential theory, Fröhlich or Holstein-type Hamiltonians [10, 51–53]. While these models provide qualitative intuitions, they rely on parameter fitting and restrictive assumptions such as dispersionless phonons and simple screening, which limit their predictive power for real materials.

Modern first-principles methods based on density functional theory (DFT) and density functional perturbation theory (DFPT) [54–56] enable quantitative predictions of transport as well as nonequilibrium dynamics by directly computing phonon dispersions and e-ph scattering matrix elements in place of fitted parameters [57, 58]. For nonequilibrium dynamics, different first-principles theoretical approaches are suited to distinct regimes of timescales and interactions. Time-dependent density functional theory (TDDFT) [59, 60] captures coherent femtosecond dynamics and memory effects, where electronic degrees of freedom are treated quantum

mechanically while nuclear motions are described classically. A related mixed quantum-classical framework is nonadiabatic molecular dynamics (NAMD), which incorporates electronic state transitions with explicit propagation of nuclear trajectories. NAMD and its momentum-space formulation ($NAMD_k$) have recently shown promising results in simulating electron relaxation on picosecond timescales [61–63]. In contrast, nonequilibrium Green's function (NEGF) [64, 65] provides a fully quantum description for out-of-equilibrium many-body systems, naturally unifying coherent dynamics, dissipation and memory effects. However, its high computational cost limits its practical application to small systems and short simulation times. Most NEGF implementations rely on approximations, such as lower-order self-energy truncations, generalized Kadanoff-Baym ansatz, Markovian limits, and quasiparticle approximations [66, 67].

As nonequilibrium dynamics evolve from coherent to incoherent regimes, the explicit treatment of quantum coherence becomes less critical, while computational efficiency and scalability become increasingly important. In this limit, semiclassical approaches offer a favorable balance between accuracy and computational cost. The two-temperature model (TTM) [68] describes macroscopic energy exchange between electrons and phonons, while Langevin and Fokker-Planck equations use stochastic forces to model thermal fluctuations and dissipation averaged over individual scattering events.

A more microscopic description is achieved by the Boltzmann transport equation (BTE) [1, 69, 70], which treats carriers as well-defined quasiparticles whose distribution functions evolve under the influence of external fields and scattering processes. The scattering matrix elements in the BTE can be obtained directly from DFT and DFPT calculations. It has demonstrated quantitative predictive power for various transport properties such as carrier mobility, electrical and thermal conductivity, phonon linewidths, and Seebeck coefficients [3, 71, 72].

The time-dependent formulation of the BTE, referred to here as the real-time BTE (rt-BTE), extends this framework to track time-dependent carrier distributions, enabling simulations such as carrier relaxation, high-field transport, ultrafast spectroscopic responses [7, 73]. More recently, the rt-BTE has been generalized to include anharmonic ph-ph interactions, enabling simulations of fully coupled e-ph dynamics [74–76]. In a broader context, the Bernardi group has performed extensive work expanding the predictive scope of the rt-BTE, including defect and impurity scattering, spin decoherence, exciton-phonon coupling, and magnon-phonon

interactions [77–80].

However, adding the anharmonic phonon interactions to the rt-BTE significantly increases computational cost. While electronic distributions can often be evolved efficiently, solving for phonon distributions over the entire Brillouin zone (BZ) with anharmonic ph-ph interactions is significantly more expensive. As a result, most rt-BTE works treat phonons as a thermal bath at fixed temperature without considering anharmonicity. Only a few studies include ph-ph interactions and are limited to two-dimensional (2D) materials with simple unit cells.

These limitations highlight two central challenges in advancing real-time, first-principles simulations of coupled electron-phonon dynamics: (1) the need for efficient algorithms to solve the rt-BTE with anharmonic phonon interactions to enable simulations of increasingly complex materials, and (2) an absence of efficient theoretical and computational framework to describe coherently driven dynamics, which are central to many modern ultrafast experiments but are not captured within the standard BTE formulations. Current methods incorporating quantum corrections, such as Wigner functions or density-matrix formulations, involve a significantly higher cost, and efficient, fully first-principles theoretical frameworks and their implementations are still an active subject of research [81–83].

1.2 Electron-phonon, phonon-phonon interactions from first principles

We first compute the electronic ground state and band structure using DFT. In the Kohn-Sham (KS) formulation [54], the interacting many-electron system is mapped onto an effective single-particle system in which the charge density is constructed from KS orbitals. The KS equations are solved self-consistently using plane-wave basis sets and pseudopotentials to account for the electron-ion interactions. Exchange-correlation effects are treated using standard approximations to the KS potential [84]. All DFT calculations in this work are performed using either Quantum ESPRESSO [56] or VASP [85, 86]. The resulting KS energies $\varepsilon_{n\mathbf{k}}$ and eigenstates $\psi_{n\mathbf{k}}$ are obtained on uniform grids of electron crystal momenta $\mathbf{k}$ in the BZ.

The perturbation potential, $\partial_{\mathbf{q}_c,\kappa\alpha}V^{\mathrm{KS}}$, and the dynamical matrix, $D(\mathbf{q}_c)$, are computed on a coarse phonon momentum grid $\mathbf{q}_c$ in the BZ. Here, $\kappa\alpha$ denotes the displacement of atom κ in direction α. These quantities are obtained within DFPT, which describes the linear response of the electronic system to small atomic displacements and is implemented in Quantum ESPRESSO. The phonon frequencies

$\omega_{\nu\mathbf{q}}$ and eigenvectors $e^{\kappa\alpha}_{\nu\mathbf{q}}$ at momentum $\mathbf{q}$ and mode ν are obtained by diagonalizing the dynamical matrix.

These phonon modes and perturbation potentials form the basis of calculating e-ph scattering matrix elements, which are defined as

$$g_{mn\nu}(\mathbf{k},\mathbf{q}) = \sqrt{\frac{\hbar}{2\omega_{\nu\mathbf{q}}}} \sum_{\kappa\alpha} \frac{e^{\kappa\alpha}_{\nu\mathbf{q}}}{\sqrt{\mu_\kappa}} \langle \psi_{m\mathbf{k}+\mathbf{q}} | \partial_{\mathbf{q},\kappa\alpha} V^{\mathrm{KS}} | \psi_{n\mathbf{k}} \rangle, \tag{1.1}$$

where $\psi_{n\mathbf{k}}$ is the KS Bloch state with band index n and crystal momentum $\mathbf{k}$, and μ_κ is the mass of atom κ. It describes the probability amplitude for an electron to scatter from an initial state $|\psi_{n\mathbf{k}}\rangle$ to a final state $|\psi_{m\mathbf{k}+\mathbf{q}}\rangle$ by emitting or absorbing a phonon of mode ν and momentum $\mathbf{q}$.

To save computational cost, the e-ph matrix elements are first transformed into the basis of maximally localized Wannier functions [87] and interpolated to finer electron and phonon momentum grids using our in-house code Perturbo [88]. For phonons, we can similarly transform the dynamical matrix to real space to obtain the interatomic force constants (IFCs) between atoms in the unit cell and atoms in neighboring cells. The dynamical matrix at any desired $\mathbf{q}$ can then be obtained by Fourier transforming the IFCs back to reciprocal space.

The IFCs can be computed using several complementary approaches. In principle, the n-th order IFCs are defined as the n-th derivatives of the Born-Oppenheimer potential energy surface (PES) with respect to atomic displacements. For example, the second- and third-order IFCs, $\Phi_{\alpha\beta}(l\kappa; l'\kappa')$ and $\Psi_{\alpha\beta\gamma}(l\kappa; l'\kappa'; l''\kappa'')$, are computed using finite-difference or DFPT by expanding the potential energy U with respect to small atomic displacements $\mathbf{u}$ [34, 89]:

$$\begin{aligned} U = U_0 &+ \frac{1}{2!} \sum_{ll'} \sum_{\kappa\kappa'} \sum_{\alpha\beta} \Phi_{\alpha\beta}(l\kappa; l'\kappa') u^{\alpha}_{l\kappa} u^{\beta}_{l'\kappa'} \\ &+ \frac{1}{3!} \sum_{ll'l''} \sum_{\kappa\kappa'\kappa''} \sum_{\alpha\beta\gamma} \Psi_{\alpha\beta\gamma}(l\kappa; l'\kappa'; l''\kappa'') u^{\alpha}_{l\kappa} u^{\beta}_{l'\kappa'} u^{\gamma}_{l''\kappa''} + \dots . \end{aligned} \tag{1.2}$$

Here, U_0 is the total potential energy for atoms in equilibrium, and the vector $\mathbf{u}_{l\kappa}$ is the displacement from equilibrium of atom κ in the l-th unit cell, while α, β, and γ label Cartesian coordinates.

In practice, these derivatives are computed by fitting forces of displaced supercells. As implemented in ALAMODE [90], finite-displacement approaches explicitly extract anharmonic IFCs by expanding the PES around a reference structure. An

approach for obtaining finite-temperature IFCs is to use temperature-dependent effective potentials (TDEP) [91], which compute effective, temperature renormalized IFCs by fitting an effective Hamiltonian to forces sampled from finite-temperature molecular dynamics (MD) trajectories. Stochastic self-consistent harmonic approximation [92] is another method that provides temperature-dependent, renormalized harmonic force constants and captures anharmonic effects nonperturbatively; however, it does not yield explicit higher-order IFC tensors required for perturbative phonon scattering calculations. In all works in the thesis, IFCs are computed using either ALAMODE or TDEP.

The 3rd-order ph-ph scattering matrix elements $\Psi_{\nu\nu'\nu''}(\mathbf{q},\mathbf{q}',\mathbf{q}'')$ in the momentum space - in this thesis, ph-ph matrix elements refer only to third-order interactions unless specified - are then defined as a Fourier transform of the 3rd-order IFCs:

$$\begin{aligned}\Psi_{\nu\nu'\nu''}(\mathbf{q},\mathbf{q}',\mathbf{q}'') = \frac{1}{6}\sqrt{\frac{\hbar^3}{8\omega_{\nu\mathbf{q}}\omega_{\nu'\mathbf{q}'}\omega_{\nu''\mathbf{q}''}}}\Delta(\mathbf{q}+\mathbf{q}'+\mathbf{q}'') \\ \times\sum_{l'l''}\sum_{\kappa\kappa'\kappa''}\sum_{\alpha\beta\gamma}\exp\left[i(\mathbf{q}'\cdot\mathbf{R}_{l'}+\mathbf{q}''\cdot\mathbf{R}_{l''})\right] \\ \frac{e^{\kappa\alpha}_{\nu\mathbf{q}}e^{\kappa'\beta}_{\nu'\mathbf{q}'}e^{\kappa''\gamma}_{\nu''\mathbf{q}''}}{\sqrt{\mu_\kappa\mu_{\kappa'}\mu_{\kappa''}}}\Psi_{\alpha\beta\gamma}(0\kappa;l'\kappa';l''\kappa'')\end{aligned} \tag{1.3}$$

where $\mathbf{R}_l$ is the lattice vector of the unit cell l, and the delta function $\Delta(\mathbf{q}+\mathbf{q}'+\mathbf{q}'')$ conserves crystal momentum, ensuring $\mathbf{q}+\mathbf{q}'+\mathbf{q}''$ is equal to zero or a reciprocal lattice vector. As the formula suggests, the ph-ph matrix elements associate with three phonon modes, $\nu\mathbf{q}$, $\nu'\mathbf{q}'$, and $\nu''\mathbf{q}''$, in contrast to the e-ph matrix elements $g_{mn\nu}(\mathbf{k},\mathbf{q})$, which involve one phonon and two electronic states.

In these works, the e-ph scattering matrix elements are computed using DFPT based on the perturbation potential and phonon dispersions of harmonic phonons at 0 K, whereas the phonon dispersions in nonequilibrium dynamics and the corresponding anharmonic ph-ph interactions are computed using finite-temperature IFCs obtained from TDEP. This mixed approach assumes that anharmonic effects primarily renormalize the lattice dynamics, while the electronic response to lattice vibrations can still be accurately captured by linear response theory at ground state. Such a separation is commonly employed in finite-temperature e-ph calculations [71]. A fully consistent treatment based on MD sampling or higher-order perturbation theory remains an important direction for future work.

1.3 rt-BTE for coupled electron-phonon dynamics

The rt-BTE describes the time evolution of electronic distributions under the influence of external fields and scattering processes. In its most general form, the rt-BTE for electrons in real space under external electric field $\mathbf{E}(\mathbf{r}, t)$ reads:

$$\frac{\partial f_{n\mathbf{k}}(\mathbf{r}, t)}{\partial t} + \mathbf{v}_{n\mathbf{k}} \cdot \nabla_{\mathbf{r}} f_{n\mathbf{k}}(\mathbf{r}, t) + \frac{e\mathbf{E}(\mathbf{r}, t)}{\hbar} \cdot \nabla_{\mathbf{k}} f_{n\mathbf{k}}(\mathbf{r}, t) = \mathcal{I}^{e\text{-ph}}[f_{n\mathbf{k}}(\mathbf{r}, t), N_{\nu\mathbf{q}}(\mathbf{r}, t)], \tag{1.4}$$

where the electron occupations, hereon referred to as populations, $f_{n\mathbf{k}}(\mathbf{r}, t)$ depend on position $\mathbf{r}$, time t, band index n, and crystal momentum $\mathbf{k}$. The electron group velocity is given by $\mathbf{v}_{n\mathbf{k}} = \frac{1}{\hbar}\nabla_{\mathbf{k}}\varepsilon_{n\mathbf{k}}$, where $\varepsilon_{n\mathbf{k}}$ is the electronic band energy. e is the electron charge. The e-ph collision integral, $\mathcal{I}^{e\text{-ph}}$, describes how electron populations change due to scattering with phonons.

In the absence of external fields and spatial inhomogeneities, the temporal evolution of electronic populations depend only on collision integrals. For coupled electron-phonon dynamics, the rt-BTE in momentum space can be written as a set of coupled integro-differential equations for both electron and phonons:

$$\begin{aligned} \frac{\partial f_{n\mathbf{k}}(t)}{\partial t} &= \mathcal{I}^{e\text{-ph}}[f_{n\mathbf{k}}(t), N_{\nu\mathbf{q}}(t)], \\ \frac{\partial N_{\nu\mathbf{q}}(t)}{\partial t} &= \mathcal{I}^{\text{ph-}e}[f_{n\mathbf{k}}(t), N_{\nu\mathbf{q}}(t)] + \mathcal{I}^{\text{ph-ph}}[N_{\nu\mathbf{q}}(t)]. \end{aligned} \tag{1.5}$$

The electron populations, $f_{n\mathbf{k}}(t)$, are labeled by the electronic band index n and crystal momentum $\mathbf{k}$ (hole carriers can be simulated with populations $1 - f_{n\mathbf{k}}(t)$), and the phonon populations, $N_{\nu\mathbf{q}}(t)$, by mode index ν and wave-vector $\mathbf{q}$. The e-ph, ph-e and the ph-ph collision integrals, $\mathcal{I}^{e\text{-ph}}$, $\mathcal{I}^{\text{ph-}e}$, and $\mathcal{I}^{\text{ph-ph}}$, are all computed with ab initio e-ph and ph-ph scattering matrix elements. The workflow for propagating the above equations is implemented as an extension to in the current version of the PERTURBO [74, 88] code.

In Eq. 1.5, the collision integral $\mathcal{I}^{e\text{-ph}}[f_{n\mathbf{k}}, N_{\nu\mathbf{q}}]$ can be computed by [53, 74]:

$$\begin{aligned} \mathcal{I}^{e\text{-ph}}[f_{n\mathbf{k}}, N_{\nu\mathbf{q}}] = &-\frac{2\pi}{\hbar}\frac{1}{\mathcal{N}_{\mathbf{q}}}\sum_{m\nu\mathbf{q}}|g_{mn\nu}(\mathbf{k}, \mathbf{q})|^2 \\ &\times \Big[F_{\text{em}} \times \delta(\varepsilon_{n\mathbf{k}} - \hbar\omega_{\nu\mathbf{q}} - \varepsilon_{m\mathbf{k}+\mathbf{q}}) \\ &+ F_{\text{abs}} \times \delta(\varepsilon_{n\mathbf{k}} + \hbar\omega_{\nu\mathbf{q}} - \varepsilon_{m\mathbf{k}+\mathbf{q}})\Big], \end{aligned} \tag{1.6}$$

where $\mathcal{N}_{\mathbf{q}}$ is the total number of phonon momentum grid points. The emission and absorption factors are defined as

$$F_{\text{em}} = f_{n\mathbf{k}}(1 - f_{m\mathbf{k}+\mathbf{q}})(1 + N_{\nu\mathbf{q}}) - (1 - f_{n\mathbf{k}})f_{m\mathbf{k}+\mathbf{q}}N_{\nu\mathbf{q}}, \tag{1.7}$$

and

$$F_{\text{abs}} = f_{n\mathbf{k}}(1 - f_{m\mathbf{k+q}})N_{\nu\mathbf{q}} - (1 - f_{n\mathbf{k}})f_{m\mathbf{k+q}}(1 + N_{\nu\mathbf{q}}). \tag{1.8}$$

The time dependence of these terms is implicitly assumed in the notations.

Similarly,

$$\mathcal{I}^{\text{ph-}e}[f_{n\mathbf{k}}, N_{\nu\mathbf{q}}] = -\frac{4\pi}{\hbar}\frac{1}{\mathcal{N}_{\mathbf{k}}}\sum_{mn\mathbf{k}}|g_{mn\nu}(\mathbf{k},\mathbf{q})|^2 \times F_{\text{abs}} \times \delta(\varepsilon_{n\mathbf{k}} + \hbar\omega_{\nu\mathbf{q}} - \varepsilon_{m\mathbf{k+q}}). \tag{1.9}$$

The ph-ph collision integral can be written as

$$\begin{aligned}\mathcal{I}^{\text{ph-ph}}[N_{\nu\mathbf{q}}] = -\frac{36\pi}{\hbar}\frac{1}{\mathcal{N}_{\mathbf{q}}}\sum_{\nu'\nu''}\sum_{\mathbf{q}'\mathbf{q}''}|\Psi_{\nu\nu'\nu''}(\mathbf{q},\mathbf{q}',\mathbf{q}'')|^2 \\ \times \Big[G_{\text{em}} \times \delta(\hbar\omega_{\nu'\mathbf{q}'} - \hbar\omega_{\nu\mathbf{q}} + \hbar\omega_{\nu''\mathbf{q}''}) \\ + 2\,G_{\text{abs}} \times \delta(\hbar\omega_{\nu\mathbf{q}} - \hbar\omega_{\nu'\mathbf{q}'} + \hbar\omega_{\nu''\mathbf{q}''})\Big],\end{aligned} \tag{1.10}$$

where

$$G_{\text{em}} = N_{\nu\mathbf{q}}(N_{\nu'\mathbf{q}'} + 1)(N_{\nu''\mathbf{q}''} + 1) - (N_{\nu\mathbf{q}} + 1)N_{\nu'\mathbf{q}'}N_{\nu''\mathbf{q}''}, \tag{1.11}$$

and

$$G_{\text{abs}} = N_{\nu\mathbf{q}}(N_{\nu'\mathbf{q}'} + 1)N_{\nu''\mathbf{q}''} - (N_{\nu\mathbf{q}} + 1)N_{\nu'\mathbf{q}'}(N_{\nu''\mathbf{q}''} + 1). \tag{1.12}$$

At equilibrium, quasiparticle lifetimes are obtained by linearizing the BTE around equilibrium and solving for steady-state solutions. (See Appendix B for details).

Computational challenges

Three major computational challenges are present in solving the coupled rt-BTE with anharmonic phonon interactions. First, computing IFCs using finite-displacement or MD-based approaches requires large supercells and a substantial number of first principles force calculations. Second, transforming real-space IFCs to the reciprocal space ph-ph matrix elements involves summations over atoms in the unit cell and neighboring cells, demanding a large amount of computing resources, especially for low-symmetry systems. Third, the collision integrals in Eqs. 1.6, 1.9 and 1.10 are evaluated at every time step. Accurately resolving long-range interactions, such as those present in polar materials, requires that these integrals be summed over dense momentum grids, which further increases computational cost. When higher-order phonon interactions are desired, the computational cost increases dramatically due to the increased number of scattering channels and the complexity of computing higher-order IFCs. Together, these challenges make it difficult to study complex materials

with large unit cells or to include higher-order phonon interactions, motivating the development of efficient algorithms to reduce the costs of computing IFCs, ph-ph matrix elements, and collision integrals in the rt-BTE framework.

1.4 Thesis outline

This thesis addresses the above challenges of simulating coupled electron-phonon dynamics with anharmonic phonon interactions. We not only develop and implement new algorithmic improvements across different aspects of the simulation, but also discuss the physical insights gained from these numerical techniques.

In Chapter 2, we develop an adaptive and multirate time integration scheme to more efficiently time-step the rt-BTE. Using graphene as a benchmark, we show that our new scheme significantly reduces the computational cost of evolving coupled dynamics while maintaining accuracy compared to single-rate, fixed-step methods. We then study dynamics in 3D materials, which were previously inaccessible due to high computational cost. We show that the rt-BTE predicts momentum-resolved thermal diffuse scattering signals, which are often experimentally measured to study lattice dynamics. Our framework also captures the importance of accounting anharmonicity to accurately describe hot-phonon effects in carrier relaxation. In the second half of the chapter, we introduce a fully-adaptive extension to the multirate integration scheme and discuss the microscopic insights provided by this class of numerical integration schemes.

In Chapter 3, we apply dynamic mode decomposition (DMD), a data-driven technique to accelerate the time evolution of the rt-BTE. For nonequilibrium dynamics under external fields, while resolving the transient dynamics remains necessary, the steady state is often of primary interest. In this case, avoiding explicit time evolution to steady state by interpolating long-time distributions based on existing simulation becomes especially efficient. We introduce the theory of DMD, discuss its implementation in the rt-BTE, and apply it to high-field charge transport, electron relaxation, and coupled electron-phonon dynamics. In all cases, we show that explicit time-stepping of a short window of the occupations provides sufficient data for DMD to extrapolate the dynamics to steady state. We further analyze the momentum-space modes extracted from DMD, which give insights to microscopic processes in the dynamics.

In Chapter 4, we demonstrate acceleration of the evaluation of collision integrals in the BTE using graphics processing units (GPUs), one step further from the

current MPI+OpenMP parallelization implemented in PERTURBO. We develop a GPU-optimized data structures and algorithm that reduce the overhead for data referencing, movement, and synchronization, which is applicable for both transport and nonequilibrium dynamics. Our approach is optimized for GPU implementation, and achieves significant speed-ups compared to CPU implementations. We also demonstrate the excellent strong scaling performance of this approach.

In Chapter 5, we focus on accelerating computation of anharmonic ph-ph matrix elements. We introduce a tensor decomposition approach to compress and efficiently evaluate collision integrals in the rt-BTE framework. Using tensor learning, we find low-rank approximations of the high-dimensional matrix elements, which reduces scaling of the summation in the ph-ph collision integral. We show that this approach can significantly reduce memory storage and computational cost for a diverse range of materials.

In Chapter 6, we attempt to expand the theoretical framework of the rt-BTE to bridge the gap between simulating coherent and incoherent regimes in ultrafast experiments. We present a theoretical framework for simulating coherently driven phonon dynamics using Lindblad master equations. And we also introduce an efficient approximation to model driven lattice dynamics within the rt-BTE framework.

Finally, in Chapter 7, we summarize the key findings and discuss future directions for advancing first-principles simulations of coupled electron-phonon dynamics.

References

[1] J. M. Ziman, *Electrons and Phonons: the Theory of Transport Phenomena in Solids* (Oxford university press, 2001).

[2] J. Bardeen and W. Shockley, Phys. Rev. **80**, 72 (1950).

[3] S. Poncé, E. R. Margine, and F. Giustino, Phys. Rev. B **97**, 121201 (2018).

[4] A. Verma, A. P. Kajdos, T. A. Cain, S. Stemmer, and D. Jena, Phys. Rev. Lett. **112**, 216601 (2014).

[5] M. Cardona and M. L. W. Thewalt, Rev. Mod. Phys. **77**, 1173 (2005).

[6] I. Meric, M. Y. Han, A. F. Young, B. Ozyilmaz, P. Kim, and K. L. Shepard, Nat. Nanotechnol. **3**, 654–659 (2008).

[7] I. Maliyov, J. Park, and M. Bernardi, Phys. Rev. B **104**, L100303 (2021).

[8] J. Bardeen, L. N. Cooper, and J. R. Schrieffer, Phys. Rev. **108**, 1175 (1957).

[9] T. Ritschel, J. Trinckauf, K. Koepernik, B. Büchner, M. v. Zimmermann, H. Berger, Y. I. Joe, P. Abbamonte, and J. Geck, Nat. Phys. **11**, 328 (2015).

[10] T. Holstein, Ann. Phys. **8**, 343 (1959).

[11] A. S. Disa, T. F. Nova, and A. Cavalleri, Nat. Phys. **17**, 1087 (2021).

[12] D. Basov, R. Averitt, and D. Hsieh, Nat. Mater. **16**, 1077 (2017).

[13] H. Li, Q. Wang, Y. Oteki, C. Ding, D. Liu, Y. Guo, Y. Li, Y. Wei, D. Wang, Y. Yang, T. Masuda, M. Chen, Z. Zhang, T. Sogabe, S. Hayase, Y. Okada, S. Iikubo, and Q. Shen, Adv. Mater. **35**, 2301834 (2023).

[14] D. Fausti, R. Tobey, N. Dean, S. Kaiser, A. Dienst, M. C. Hoffmann, S. Pyon, T. Takayama, H. Takagi, and A. Cavalleri, Science **331**, 189 (2011).

[15] M. Mitrano, A. Cantaluppi, D. Nicoletti, S. Kaiser, A. Perucchi, S. Lupi, P. Di Pietro, D. Pontiroli, M. Riccò, S. R. Clark, D. Jaksch, and A. Cavalleri, Nature **530**, 461 (2016).

[16] A. Cavalleri, T. Dekorsy, H. H. W. Chong, J. C. Kieffer, and R. W. Schoenlein, Phys. Rev. B **70**, 161102 (2004).

[17] S. Hellmann, T. Rohwer, M. Kalläne, K. Hanff, C. Sohrt, A. Stange, A. Carr, M. M. Murnane, H. C. Kapteyn, L. Kipp, M. Bauer, and K. Rossnagel, Nat. Commun. **3**, 1069 (2012).

[18] K. W. Kim, A. Pashkin, H. Schäfer, M. Beyer, M. Porer, T. Wolf, C. Bernhard, J. Demsar, R. Huber, and A. Leitenstorfer, Nat. Mater. **11**, 497 (2012).

[19] D. Afanasiev, J. R. Hortensius, B. A. Ivanov, A. Sasani, E. Bousquet, Y. M. Blanter, R. V. Mikhaylovskiy, A. V. Kimel, and A. D. Caviglia, Nat. Mater. **20**, 607 (2021).

[20] E. A. Mashkovich, K. A. Grishunin, R. M. Dubrovin, A. K. Zvezdin, R. V. Pisarev, and A. V. Kimel, Science **374**, 1608–1611 (2021).

[21] E. D. Murray, D. M. Fritz, J. K. Wahlstrand, S. Fahy, and D. A. Reis, Phys. Rev. B **72**, 060301 (2005).

[22] M. Först, C. Manzoni, S. Kaiser, Y. Tomioka, Y. Tokura, R. Merlin, and A. Cavalleri, Nat. Phys. **7**, 854 (2011).

[23] L. P. René de Cotret, J.-H. Pöhls, M. J. Stern, M. R. Otto, M. Sutton, and B. J. Siwick, Phys. Rev. B **100**, 214115 (2019).

[24] J. E. Katz, X. Zhang, K. Attenkofer, K. W. Chapman, C. Frandsen, P. Zarzycki, K. M. Rosso, R. W. Falcone, G. A. Waychunas, and B. Gilbert, Science **337**, 1200 (2012).

[25] M. Zürch, H.-T. Chang, L. J. Borja, P. M. Kraus, S. K. Cushing, A. Gandman, C. J. Kaplan, M. H. Oh, J. S. Prell, D. Prendergast, C. D. Pemmaraju, D. M. Neumark, and S. R. Leone, Nat. Commun. **8**, 15734 (2017).

[26] S. K. Cushing, M. Zürch, P. M. Kraus, L. M. Carneiro, A. Lee, H.-T. Chang, C. J. Kaplan, and S. R. Leone, Struct. Dyn. **5**, 054302 (2018).

[27] T. Valla, A. V. Fedorov, P. D. Johnson, and S. L. Hulbert, Phys. Rev. Lett. **83**, 2085 (1999).

[28] L. Perfetti, P. A. Loukakos, M. Lisowski, U. Bovensiepen, H. Eisaki, and M. Wolf, Phys. Rev. Lett. **99**, 197001 (2007).

[29] H. Tanimura, J. Kanasaki, K. Tanimura, J. Sjakste, N. Vast, M. Calandra, and F. Mauri, Phys. Rev. B **93**, 161203 (2016).

[30] M. X. Na, A. K. Mills, F. Boschini, M. Michiardi, B. Nosarzewski, R. P. Day, E. Razzoli, A. Sheyerman, M. Schneider, G. Levy, S. Zhdanovich, T. P. Devereaux, A. F. Kemper, D. J. Jones, and A. Damascelli, Science **366**, 1231 (2019).

[31] U. De Giovannini, H. Hübener, S. A. Sato, and A. Rubio, Phys. Rev. Lett. **125**, 136401 (2020).

[32] E. Baldini, A. Zong, D. Choi, C. Lee, M. H. Michael, L. Windgaetter, I. I. Mazin, S. Latini, D. Azoury, B. Lv, A. Kogar, Y. Su, Y. Wang, Y. Lu, T. Takayama, H. Takagi, A. J. Millis, A. Rubio, E. Demler, and N. Gedik, Proc. Natl. Acad. Sci. USA **120**, e2221688120 (2023).

[33] A. A. Maradudin and A. E. Fein, Phys. Rev. **128**, 2589 (1962).

[34] R. A. Cowley, Rep. Prog. Phys. **31**, 123 (1968).

[35] G. A. Slack, J. Phys. Chem. Solids **34**, 321 (1973).

[36] L. Lindsay, D. A. Broido, and T. L. Reinecke, Phys. Rev. Lett. **111**, 025901 (2013).

[37] R. Clady, M. J. Y. Tayebjee, P. Aliberti, D. König, N. J. Ekins-Daukes, G. J. Conibeer, T. W. Schmidt, and M. A. Green, Prog. Photovolt: Res. Appl. **20**, 82 (2012).

[38] C. W. Li, J. Ma, H. B. Cao, A. F. May, D. L. Abernathy, G. Ehlers, C. Hoffmann, X. Wang, T. Hong, A. Huq, O. Gourdon, and O. Delaire, Phys. Rev. B **90**, 214303 (2014).

[39] J. He and T. M. Tritt, Science **357**, eaak9997 (2017).

[40] J. Yang, X. Wen, H. Xia, R. Sheng, Q. Ma, J. Kim, P. Tapping, T. Harada, T. W. Kee, F. Huang, Y.-B. Cheng, M. Green, A. Ho-Baillie, S. Huang, S. Shrestha, R. Patterson, and G. Conibeer, Nat. Commun. **8**, 14120 (2017).

[41] S. Manzeli, D. Ovchinnikov, D. Pasquier, O. V. Yazyev, and A. Kis, Nat. Rev. Mater. **2**, 17033 (2017).

[42] U. D. Wdowik, K. Parlinski, S. Rols, and T. Chatterji, Phys. Rev. B **89**, 224306 (2014).

[43] X. Li, T. Qiu, J. Zhang, E. Baldini, J. Lu, A. M. Rappe, and K. A. Nelson, Science **364**, 1079 (2019).

[44] C. Ferrante, A. Virga, L. Benfatto, M. Martinati, D. De Fazio, U. Sassi, C. Fasolato, A. K. Ott, P. Postorino, D. Yoon, G. Cerullo, F. Mauri, A. C. Ferrari, and T. Scopigno, Nat. Commun. **9**, 308 (2018).

[45] M. J. Stern, L. P. René de Cotret, M. R. Otto, R. P. Chatelain, J.-P. Boisvert, M. Sutton, and B. J. Siwick, Phys. Rev. B **97**, 165416 (2018).

[46] A. A. Maznev, J. A. Johnson, and K. A. Nelson, Phys. Rev. B **84**, 195206 (2011).

[47] A. M. Lindenberg, I. Kang, S. L. Johnson, T. Missalla, P. A. Heimann, Z. Chang, J. Larsson, P. H. Bucksbaum, H. C. Kapteyn, H. A. Padmore, R. W. Lee, J. S. Wark, and R. W. Falcone, Phys. Rev. Lett. **84**, 111 (2000).

[48] M. Kozina, M. Fechner, P. Marsik, T. van Driel, J. M. Glownia, C. Bernhard, M. Radovic, D. Zhu, S. Bonetti, U. Staub, and M. C. Hoffmann, Nat. Phys. **15**, 387 (2019).

[49] M. Mohr, J. Maultzsch, E. Dobardžić, S. Reich, I. Milošević, M. Damnjanović, A. Bosak, M. Krisch, and C. Thomsen, Phys. Rev. B **76**, 035439 (2007).

[50] M. Trigo, M. Fuchs, J. Chen, M. P. Jiang, M. Cammarata, S. Fahy, D. M. Fritz, K. Gaffney, S. Ghimire, A. Higginbotham, S. L. Johnson, M. E. Kozina, J. Larsson, H. Lemke, A. M. Lindenberg, G. Ndabashimiye, F. Quirin, K. Sokolowski-Tinten, C. Uher, G. Wang, J. S. Wark, D. Zhu, and D. A. Reis, Nat. Phys. **9**, 790 (2013).

[51] J. Bardeen and W. Shockley, Phys. Rev. **80**, 72 (1950).

[52] H. Fröhlich, Adv. Phys. **3**, 325 (1954).

[53] G. D. Mahan, *Many-Particle Physics*, 3rd ed. (Springer, 2000).

[54] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods*, 2nd ed. (Cambridge University Press, 2020).

[55] S. Baroni, S. de Gironcoli, A. Dal Corso, and P. Giannozzi, Rev. Mod. Phys. **73**, 515 (2001).

[56] P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. B. Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni, N. Colonna, I. Carnimeo, A. D. Corso, S. de Gironcoli, P. Delugas, R. A. D. Jr, A. Ferretti, A. Floris, G. Fratesi, G. Fugallo, R. Gebauer, U. Gerstmann, F. Giustino, T. Gorni, J. Jia, M. Kawamura, H.-Y. Ko, A. Kokalj, E. Küçükbenli, M. Lazzeri, M. Marsili, N. Marzari, F. Mauri, N. L. Nguyen, H.-V. Nguyen, A. O. de-la Roza, L. Paulatto, S. Poncé, D. Rocca, R. Sabatini, B. Santra, M. Schlipf, A. P. Seitsonen, A. Smogunov, I. Timrov, T. Thonhauser, P. Umari, N. Vast, X. Wu, and S. Baroni, J. Phys. Condens. Matter **29**, 465901 (2017).

[57] D. Sangalli and A. Marini, Europhy. Lett. **110**, 47004 (2015).

[58] M. Bernardi, D. Vigil-Fowler, J. Lischner, J. B. Neaton, and S. G. Louie, Phys. Rev. Lett. **112**, 257402 (2014).

[59] M. A. Marques, N. T. Maitra, F. M. Nogueira, E. K. Gross, and A. Rubio, *Fundamentals of Time-Dependent Density Functional Theory*, Vol. 837 (Springer, 2012).

[60] C. Andrea Rozzi, S. Maria Falke, N. Spallanzani, A. Rubio, E. Molinari, D. Brida, M. Maiuri, G. Cerullo, H. Schramm, J. Christoffers, *et al.*, Nat. Commun. **4**, 1602 (2013).

[61] A. V. Akimov, A. J. Neukirch, and O. V. Prezhdo, Chem. Rev. **113**, 4496 (2013).

[62] Z. Zheng, Y. Shi, J.-J. Zhou, O. V. Prezhdo, Q. Zheng, and J. Zhao, Nat. Comput. Sci **3**, 532–541 (2023).

[63] Z. Wang, Z. Zheng, Q. Zheng, and J. Zhao, J. Phys. Chem. Lett. **15**, 3907–3913 (2024).

[64] E. Perfetto, Y. Pavlyukh, and G. Stefanucci, Phys. Rev. Lett. **128**, 016801 (2022).

[65] E. Perfetto and G. Stefanucci, Nano Lett. **23**, 7029 (2023).

[66] G. Stefanucci, R. van Leeuwen, and E. Perfetto, Phys. Rev. X **13**, 031026 (2023).

[67] G. Stefanucci and E. Perfetto, SciPost Phys. **16**, 073 (2024).

[68] S. I. Anisimov, B. L. Kapeliovich, and T. L. Perelman, Sov. Phys. JETP (1974).

[69] S. Poncé, E. Margine, C. Verdi, and F. Giustino, Comput. Phys. Commun. **209**, 116 (2016).

[70] G. K. Madsen and D. J. Singh, Comput. Phys. Commun. **175**, 67 (2006).

[71] J. Zhou, B. Liao, and G. Chen, Semicond. Sci. Technol. **31**, 043001 (2016).

[72] L. Paulatto, I. Errea, M. Calandra, and F. Mauri, Phys. Rev. B **91**, 054304 (2015).

[73] V. A. Jhalani, J.-J. Zhou, and M. Bernardi, Nano Lett. **17**, 5012 (2017).

[74] X. Tong and M. Bernardi, Phys. Rev. Res. **3**, 023072 (2021).

[75] F. Caruso, J. Phys. Chem. Lett. **12**, 1734 (2021).

[76] T. L. Britt, Q. Li, L. P. René de Cotret, N. Olsen, M. Otto, S. A. Hassan, M. Zacharias, F. Caruso, X. Zhu, and B. J. Siwick, Nano Lett. **22**, 4718 (2022).

[77] I.-T. Lu, J.-J. Zhou, and M. Bernardi, Phys. Rev. Mater. **3**, 033804 (2019).

[78] J. Park, J.-J. Zhou, and M. Bernardi, Phys. Rev. B **101**, 045202 (2020).

[79] H.-Y. Chen, D. Sangalli, and M. Bernardi, Phys. Rev. Res. **4**, 043203 (2022).

[80] K. B. Le, A. Esquembre-Kučukalić, H.-Y. Chen, I. Maliyov, Y. Luo, J.-J. Zhou, D. Sangalli, A. Molina-Sánchez, and M. Bernardi, Phys. Rev. B **112**, L180403 (2025).

[81] M. Simoncelli, N. Marzari, and F. Mauri, Phys. Rev. X **12**, 041011 (2022).

[82] F. Caruso and M. Zacharias, Phys. Rev. B **107**, 054102 (2023).

[83] S. Mocatti, G. Marini, G. Volpato, P. Cudazzo, and M. Calandra, Nonequilibrium photocarrier and phonon dynamics from first principles: a unified treatment of carrier-carrier, carrier-phonon, and phonon-phonon scattering (2025), arXiv:2512.08618 [cond-mat.mtrl-sci] .

[84] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[85] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).

[86] G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).

[87] A. A. Mostofi, J. R. Yates, G. Pizzi, Y.-S. Lee, I. Souza, D. Vanderbilt, and N. Marzari, Comput. Phys. Commun. **185**, 2309 (2014).

[88] J.-J. Zhou, J. Park, I.-T. Lu, I. Maliyov, X. Tong, and M. Bernardi, Comput. Phys. Commun. **264**, 107970 (2021).

[89] A. Debernardi, S. Baroni, and E. Molinari, Phys. Rev. Lett. **75**, 1819 (1995).

[90] T. Tadano, Y. Gohda, and S. Tsuneyuki, J. Phys. Condens. Matter **26**, 225402 (2014).

[91] O. Hellman and I. A. Abrikosov, Phys. Rev. B **88**, 144301 (2013).

[92] L. Monacelli, R. Bianco, M. Cherubini, M. Calandra, I. Errea, and F. Mauri, J. Phys. Condens. Matter **33**, 363001 (2021).

Chapter 2

COUPLED ELECTRON-PHONON DYNAMICS USING ADAPTIVE AND MULTIRATE TIME INTEGRATION

Part of this chapter is adapted from the published work:

J. Yao, I. Maliyov, D. J. Gardner, C. S. Woodward, and M. Bernardi, npj Comput. Mater. **11**, 256 (2025),
J.Y. participated in conception of the project, conducted computational research, and wrote the manuscript.

The remaining sections contain original, unpublished work by the author, currently in preparation for submission.

2.1 Introduction

Coherent ultrafast control of electronic and lattice degrees of freedom is a novel frontier in materials physics. From light-induced phase transitions [1] and carrier relaxation [2], to longer-lived phenomena, such as driven lattice response [3], spin-phonon dynamics [4], and lattice thermalization [5], experiments have revealed new pathways to manipulate material properties on ultrafast timescales.

Different first principles methods [6–8] have enabled quantitative simulations of nonequilibrium phenomena, addressing distinct regimes and timescales of nonequilibrium dynamics. This work focuses on the rt-BTE method, a set of coupled integro-differential equations for the time-dependent electron and phonon distributions that evolve under the influence of external fields and scattering processes, with e-ph and ph-ph interactions computed from first-principles. The rt-BTE framework has been successfully used to study a wide range of phenomena in the incoherent regime, such as carrier relaxation [9], high-field transport [10], ultrafast spectroscopies [11]. In particular, the ability to capture anharmonic ph-ph interactions in the rt-BTE [12–15] is particularly important for modeling nonequilibrium lattice dynamics at the ps timescale and in materials with strong anharmonicity.

However, two key challenges are present in numerically propagating the rt-BTE in time: the difference in timescales of electron and phonon dynamics and the computational cost of evaluating collision integrals, particularly for ph-ph interactions.

Typically, the rt-BTE is evolved with a fixed time step of a few fs to resolve the faster electron dynamics, while the simulation extends to hundreds of ps for the phonons to reach steady state or thermal equilibrium. At each time step, e-ph and ph-ph collision integrals are computed on dense momentum grids to ensure precise resolution of scattering processes and energy conservation, which are critical for reliable predictions. While in practice, only a subset of the electron momenta within a given energy range of the band edge or Fermi energy is included to reduce computational cost, the ph-ph collision integral is integrated over the entire Brillouin zone (BZ) for all phonon modes, because phonons across the BZ can participate in scattering processes. A 10-ps simulation for a 2D material with a few atoms in the unit cell can easily exceed thousands of CPU core hours on modern high-performance computers. In bulk systems, this requires summing over billions of scattering channels. As a result, only a handful of recent works have shown first-principles simulations of coupled electron and phonon dynamics [16–19]. Without a significant reduction in computational cost, simulations of coupled electron and phonon dynamics remain out of reach for bulk materials and all but the simplest 2D materials.

To address these challenges, we explore time integration methods from the SUNDIALS suite [20, 21], which offers an array of robust and efficient algorithms for integrating differential equations. The ARKODE package [22] within SUNDIALS contains adaptive step size Runge-Kutta (RK) [23] as well as multirate infinitesimal (MRI) methods [24]. Adaptive RK methods dynamically adjust the time step to obtain solutions within user-specified solution tolerances while maximizing efficiency, selecting step sizes that reflect the inherent dynamics of the system. MRI methods further improve efficiency by time-evolving different processes or solution components with different step sizes. They are particularly effective when the dynamics is governed by processes with well-separated timescales and the slower processes dominate the computational cost, making them an ideal fit for the rt-BTE. Methods in the MRI family have demonstrated improved efficiency in air pollution models [25] and reactive flow simulations [26] compared to single-rate integrators and multirate spectral deferred correction methods, respectively. Despite their success in these areas, their application to electron and lattice dynamics simulated from first principles remains unexplored.

In this chapter, we use adaptive RK and MRI time integration methods to simulate coupled electron and phonon dynamics in the rt-BTE framework. This is done by implementing an interface between the PERTURBO code [9] and the SUNDIALS

library [20, 21]. We achieve a significant reduction in computational cost while retaining the same accuracy by using different and adaptive time steps for e-ph and ph-ph interactions. We benchmark our approach on graphene, achieving a reduction in computational cost by orders of magnitude for any choice of solution tolerance. Finally, our adaptive and multirate time-stepping scheme allows us to solve the coupled rt-BTE for bulk materials with reasonable computational effort, a goal that has so far been considered out of reach. This is demonstrated by studying nonequilibrium lattice dynamics in silicon (Si) and gallium arsenide, and simulating the associated thermal diffuse scattering maps in silicon. Taken together, we develop efficient and accurate simulations of nonequilibrium physics in real materials, opening doors for future studies of ultrafast electronic and lattice dynamics in materials driven by optical or terahertz pulses.

2.2 Methods

Adaptive time integration (ERK)

Adaptive-step explicit Runge-Kutta (ERK) methods target ordinary differential equations (ODEs),

$$y' = f(t, y), \quad y(t_0) = y_0. \tag{2.1}$$

The solution vector y of Eq. 1.5 contains all carrier and phonon populations in the respective momentum grids:

$$y(t) = \begin{bmatrix} f_{n\mathbf{k}}(t) \\ N_{\nu\mathbf{q}}(t) \end{bmatrix}, \tag{2.2}$$

and the right-hand side (RHS) function in Eq. 2.1 is

$$f(t, y) = \begin{bmatrix} \mathcal{I}^{e\text{-ph}}(y) \\ \mathcal{I}^{\text{ph-}e}(y) + \mathcal{I}^{\text{ph-ph}}(y) \end{bmatrix}. \tag{2.3}$$

To advance Eq. 2.1 from t_{k-1} to $t_k = t_{k-1} + h_k$ with a method of order p, we use

$$z_i = y_{k-1} + h_k \sum_{j=1}^{i-1} A_{i,j} f(t_{k,j}, z_j),\ i = 1, \ldots, S, \tag{2.4a}$$

$$y_k = y_{k-1} + h_k \sum_{i=1}^{S} b_i f(t_{k,i}, z_i), \tag{2.4b}$$

$$\tilde{y}_k = y_{k-1} + h_k \sum_{i=1}^{S} \tilde{b}_i f(t_{k,i}, z_i), \tag{2.4c}$$

where there are S stages, z_i, at times $t_{k,j} = t_{k-1} + c_j h_k$, and the coefficients that define the method for obtaining the new solution, y_k, are given in the corresponding Butcher tableau with $A \in \mathbb{R}^{S\times S}$, $b \in \mathbb{R}^S$, and $c \in \mathbb{R}^S$. The coefficients $\tilde{b} \in \mathbb{R}^S$ are used to obtain an embedded solution, $\tilde{y}_k$, (typically of order $p-1$) and the difference between the solution and embedding, $y_k - \tilde{y}_k$, provides an estimate of the local truncation error (LTE) for adapting the step size [23]. The ERK results above utilize the default adaptive ERK method in ARKODE, which is a five-stage, fourth-order method from Zonneveld [27].

An attempted time step is accepted if it satisfies $\|\text{LTE}\|_{\text{WRMS}} \leq 1$ in the weighted root-mean-square (WRMS) norm,

$$\|v\|_{\text{WRMS}} = \left(\frac{1}{N_y}\sum_{i=1}^{N_y}(v_i w_i)^2\right)^{1/2}, \tag{2.5}$$

where $N_y = N_c + N_{ph}$ is the length of the vector v, and the weights w_i are defined by the most recent solution y_{n-1}, the relative tolerance (rtol) and, the vector of absolute tolerances (atol) as

$$w_i = \left(\text{rtol}\,|y_{n-1,i}| + \text{atol}_i\right)^{-1}. \tag{2.6}$$

The relative tolerance controls the number of digits of accuracy in the solution, while the absolute tolerance sets the level below which differences in small solution components are ignored. If a step attempt is rejected, a new step size is computed based on the LTE, and the step is repeated. After successfully completing a step, the LTE is similarly used to determine the step size for the next step attempt.

With an adaptive integration method, the time step is adjusted to satisfy user-defined error tolerances, the choice of which is critical to adaptive integrator performance. To solve the rt-BTE with adaptive time-stepping, we specify the relative and absolute tolerances for carriers and phonons in SUNDIALS.

We employ the default error controller in ARKODE, the PID controller [28–30], for step size selection,

$$h' = h_k \varepsilon_k^{-a_1/p_*}\, \varepsilon_{k-1}^{-a_2/p_*}\, \varepsilon_{k-2}^{-a_3/p_*}, \tag{2.7}$$

where h' is the new step size and ε_n, ε_{n-1}, and ε_{n-2} are the WRMS norms of the error estimates from the current and prior two time steps, respectively. The values of ε_{n-1} and ε_{n-2} are initialized to 1 and updated as steps are accepted. We use the default PID parameter values ($a_1 = 0.58$, $a_2 = 0.21$, and $a_3 = 0.1$) except for p_*, where we use the method order, p, rather than the embedding order. Additionally, we disable

the step size adjustment thresholds, allowing any step size change between each step rather than retaining the current step size if the step growth factor is small. These adjustments to the default values are shown to be more efficient and lead to fewer failed steps, larger step sizes, more frequent changes in step size, and smoother step size profiles compared to the current default settings.

Multirate time integration (MRI)

MRI methods target ODEs with the RHS function split into fast and slow parts,

$$y' = f^f(t, y) + f^s(t, y), \quad y(t_0) = y_0, \tag{2.8}$$

where e-ph interactions are considered the fast part and ph-ph interactions the slow component:

$$f^f(t, y) = \begin{bmatrix} \mathcal{I}^{e\text{-ph}}(y) \\ \mathcal{I}^{\text{ph-}e}(y) \end{bmatrix}, \quad f^s(t, y) = \begin{bmatrix} 0 \\ \mathcal{I}^{\text{ph-ph}}(y) \end{bmatrix}. \tag{2.9}$$

MRI methods advance the slow dynamics at a fixed time step, h_s, while the fast dynamics are evolved by solving an auxiliary ODE with an adaptive time step using a sufficiently accurate method, such as an adaptive ERK method.

Explicit MRI methods for Eq. 2.8 with $\hat{S}$ stages advance the solution from t_{k-1} to $t_k = t_{k-1} + h_s$ with the following algorithm:

1. Set $z_1 = y_{k-1}$ and $\hat{t}_{k,1} = t_{k-1}$
2. For $i = 2, \ldots, \hat{S}$
   a) Let $\hat{t}_{k,i} = t_{k-1} + \hat{c}_i h_s$ and $v(\hat{t}_{k,i-1}) = z_{i-1}$
   b) Solve $v'(t) = f_f(t, v) + r_i(t)$ on $t \in [\hat{t}_{k,i-1}, \hat{t}_{k,i}]$ where:

$$r_i(t) = \frac{1}{\Delta\hat{c}_i} \sum_{j=1}^{i-1} \gamma_{i,j}(\tau)\, f^s(\hat{t}_{k,j}, z_j),$$
$$\tau = (t - \hat{t}_{k,i-1})/(\Delta\hat{c}_i h_s),$$
$$\Delta\hat{c}_i = \hat{c}_i - \hat{c}_{i-1}$$

   c) Set $z_i = v(\hat{t}_{k,i})$
3. Set $y_k = z_{\hat{S}}$

The abscissae are sorted, $0 = \hat{c}_1 \leq \cdots \leq \hat{c}_{\hat{S}} = 1$, and they define the intervals over which the fast auxiliary ODE in step 2(b) is solved to compute the stage values z_i. If $\Delta\hat{c}_i = 0$, solving the fast auxiliary ODE in step 2(b) reduces to a standard ERK stage update as in Eq. 2.4. The coefficient function,

$$\gamma_{i,j}(\tau) = \sum_{k=1}^{K} \Omega_{i,j,k}\, \tau^{k-1}, \tag{2.10}$$

is a polynomial in time with coefficients $\Omega_{i,j,k} \in \mathbb{R}^{\hat{S}\times\hat{S}\times K}$ that define the coupling between slow and fast timescales.

The results above utilize the default MRI algorithm in ARKODE, a third-order multirate infinitesimal step (MIS) method [25, 31, 32] based on the explicit method in Ref. [33]. MIS methods are a subset of MRI methods where the coupling coefficients are uniquely defined from an ERK method with $S = \hat{S} - 1$ stages and sorted abscissae. For third-order accuracy, the ERK method must also satisfy an additional order condition [33]. In the MIS case, the forcing function $r_i(t)$ reduces to a constant value ($K = 1$) that depends on the stage index, i. The resulting MIS method has abscissae $\hat{c} = [c_1 \ \cdots \ c_S \ 1]^T$ and coupling coefficients:

$$\Omega_{i,j,1} = \begin{cases} 0, & \text{if } i = 1, \\ A_{i,j} - A_{i-1,j}, & \text{if } 2 \leq i \leq S, \\ b_j - A_{S-1,j}, & \text{if } i = S+1 = \hat{S} \end{cases}. \tag{2.11}$$

In theory, the auxiliary ODE in step 2(b) is solved exactly with an infinitesimally small step size; however, in practice, it is solved using any sufficiently accurate method. Typically, the time steps h_f that are used in evolving the "fast" dynamics in step 2(b) are significantly smaller than the steps h_s that evolve the "slow" dynamics, resulting in a multirate time step that evaluates f^s considerably less frequently than f^f. As a result, MRI methods may achieve significant speedup over traditional single-rate methods for problems in which f^s is much more costly to evaluate than f^f. We use the same ERK method applied in the single-rate adaptive integration tests as the fast timescale integrator.

Computational details

Choices of tolerances and reference solutions

Population error in this work is defined as $\|f_{n\mathbf{k}}(t) - \tilde{f}_{n\mathbf{k}}(t)\|/\|\tilde{f}_{n\mathbf{k}}(t)\|$ for carriers and $\|N_{\nu\mathbf{q}}(t) - \tilde{N}_{\nu\mathbf{q}}(t)\|/\|\tilde{N}_{\nu\mathbf{q}}(t)\|$ for phonons. Since there is no analytical solution

for the coupled rt-BTE, we define reference solutions for error analysis, $\tilde{f}_{n\mathbf{k}}(t)$ and $\tilde{N}_{\nu\mathbf{q}}(t)$, which are obtained using the MRI method with a very small slow time step, $h_s = 0.01$ fs, and tight tolerances in the fast timescale integration, rtol $= 10^{-10}$ and $\text{atol}_\text{c} = \text{atol}_\text{ph} = 10^{-15}$ for both carriers and phonons. Analyzing solution errors with different reference solutions (Supplementary Fig. 2.12) confirms that the MRI method converges significantly faster than RK4 across time step sizes.

For benchmark results in graphene, relative tolerance values for the simulations (excluding the reference solution) are tested in the range 10^{-4} to 10^{-11}, while the absolute tolerances for MRI are fixed as $\text{atol}_\text{c} = 10^{-9}$ for carriers and $\text{atol}_\text{ph} = 10^{-12}$ for phonons, and for ERK they are set as $\text{atol}_\text{c} = \text{atol}_\text{ph} = 10^{-15}$ to observe a wider range of errors. rtol $= 10^{-5}$ was used for graphene simulations unless specified. For the simulation of Si, the fast timescale is integrated with the adaptive ERK method using a relative tolerance rtol $= 10^{-5}$ and absolute tolerances $\text{atol}_c = 10^{-9}$ for carriers and $\text{atol}_\text{ph} = 10^{-11}$ for phonons. Convergence at these tolerance values is confirmed by comparing with tests using tighter tolerances and smaller h_s values. For GaAs, the fast timescale tolerances are set as rtol $= 10^{-7}$ and $\text{atol}_\text{c} = 10^{-10}$ for carriers and $\text{atol}_\text{ph} = 10^{-12}$ for phonons.

First-principles calculations

We carry out first-principles calculations of e-ph and ph-ph scattering in graphene, Si, and GaAs. We obtain the electronic ground state and band structure from density functional theory (DFT) [6] with the Quantum ESPRESSO [8] package, using the local density approximation and norm-conserving pseudopotentials [34]. The e-ph matrix elements, defined in Eq. 1.1, are first computed on a coarse momentum grid using DFPT. The e-ph matrix elements are then transformed to a maximally-localized Wannier basis (generated using WANNIER90 [35]) and interpolated to finer electron and phonon momentum grids. The second- and third-order inter-atomic force constants (IFCs) in Eq. 1.2 using the TDEP [15], which computes the atomic forces on structures with random thermal displacements distributed according to a canonical ensemble. These calculations are carried out at 300 K employing supercells with dimensions $12 \times 12 \times 1$ for graphene, $6 \times 6 \times 6$ for Si, and $8 \times 8 \times 8$ for GaAs. The second-order IFCs are used to compute the phonon frequencies and eigenvectors [36], and the third-order IFCs to compute the ph-ph matrix elements.

The nonequilibrium rt-BTE framework for coupled dynamics is described in detail in Chapter 1 in Eq. 1.5 – 1.12.

For graphene, the phonon scattering rates are converged for ph-ph matrix elements computed on a $\mathbf{q}$-grid with size $200 \times 200 \times 1$, while full convergence of e-ph scattering requires a $400 \times 400 \times 1$ $\mathbf{q}$-grid. The nonequilibrium dynamics calculation employs this dense $400 \times 400 \times 1$ grid, with ph-ph matrix elements interpolated on the fly. The efficiency analyses in and Supplementary Figs. 1–3 are obtained using $200 \times 200 \times 1$ $\mathbf{k}$- and $\mathbf{q}$-grids. Delta functions enforcing energy conservation in Eqs. 1.6, 1.9, and 1.10 are approximated by the Gaussian function $\delta(E) \approx \frac{\pi}{\sigma} \exp[-(E/\sigma)^2]$, where σ is a broadening parameter typically set to a few meV. A small phonon frequency cutoff of 1 meV is employed. We use a Gaussian broadening of 20 meV for the collision integrals $\mathcal{I}^{e\text{-ph}}$ and $\mathcal{I}^{\text{ph-}e}$, and 2 meV for $\mathcal{I}^{\text{ph-ph}}$.

For silicon, the ph-ph matrix elements are computed on a $40 \times 40 \times 40$ $\mathbf{q}$-grid with a phonon frequency cutoff of 0.2 meV. The e-ph matrix elements and the dynamics are computed on a finer $80 \times 80 \times 80$ $\mathbf{q}$-grid. During the dynamics calculations, the ph-ph matrix elements are interpolated onto this finer grid. The Gaussian broadening is 5 meV for $\mathcal{I}^{e\text{-ph}}$ and $\mathcal{I}^{\text{ph-}e}$ and 1.1 meV for $\mathcal{I}^{\text{ph-ph}}$. Both the scattering rates and the dynamics have been tested for convergence with respect to broadening and grid size parameters. To find optimal grids for the nonequilibrium lattice dynamics, we compute mode-resolved three-phonon scattering rates at 300 K with the same ph-ph matrix elements employed in the time-domain simulations (see Supplementary Fig. 4). The MRI method in ARKODE is used with a fixed slow time step $h_s = 50$ fs.

For GaAs, ph-ph matrix elements are computed on a $40 \times 40 \times 40$ $\mathbf{q}$-grid with a phonon frequency cutoff of 0.2 meV. The e-ph matrix elements and the dynamics are computed on a finer $120 \times 120 \times 120$ $\mathbf{q}$-grid. Gaussian broadening values of 8 meV and 0.56 meV are used for computing the $\mathcal{I}^{e\text{-ph}}$ and $\mathcal{I}^{\text{ph-ph}}$ collision integrals, respectively. The MRI method is used with $h_s = 50$ fs.

The Wannier interpolation, e-ph and ph-ph matrix computation, and ultrafast dynamics simulation described above are implemented in the PERTURBO package [9]. Both MPI and OpenMP parallelizations are employed to sum over scattering channels when computing e-ph and ph-ph matrices, and to sum over $\mathbf{k}$- and $\mathbf{q}$-points when computing the collision integrals.

The computational cost of evaluating e-ph and ph-ph collision integrals is determined by the number of possible scattering channels, as constrained by momentum and energy conservation. Once these channels are identified, at each time step, the cost depends only on summations over momenta and bands (for electrons) or modes (for phonons) of two particles participating in the collision, while the state of the

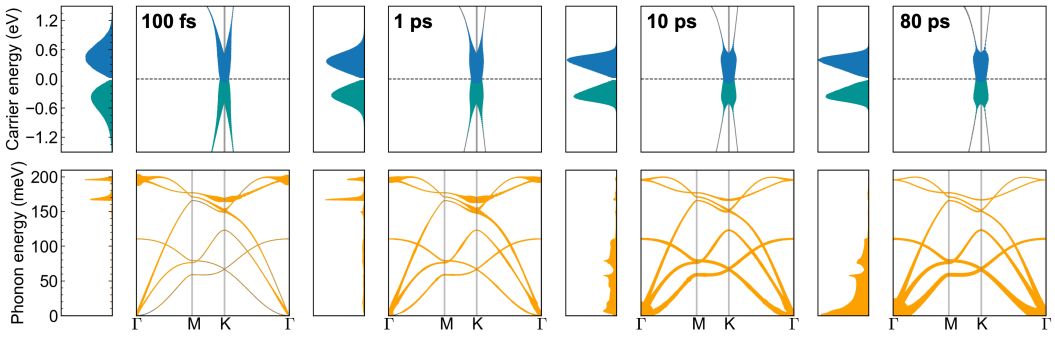

Figure 2.1: **Coupled electron and phonon dynamics in graphene**. Top row: electron populations, $f_{n\mathbf{k}}(t)$ (blue), and hole populations, $1 - f_{n\mathbf{k}}(t)$ (green), mapped onto the band structure at different time snapshots labeled in the panels, with point sizes proportional to the populations. To the left of each panel, we show the carrier populations as a function of energy by averaging over the BZ. Bottom row: change in phonon populations, $\Delta N_{\nu\mathbf{q}}(t)$, relative to the initial distribution, shown on the phonon dispersion. The BZ-averaged change in populations is shown to the left of each phonon dispersion plot. At time $t = 0$, the initial carriers are set to a hot Fermi-Dirac distribution at 4000 K, while the phonons follow a Bose-Einstein distribution at 300 K.

third particle is determined by conservation laws and does not affect the scaling complexity. Therefore, for e-ph interactions, the cost scales as $\mathcal{N}_\mathrm{c}\,\mathcal{N}_\mathrm{ph}$, where $\mathcal{N}_\mathrm{c}$ is the number of carrier momenta and band indices, and $\mathcal{N}_\mathrm{ph}$ is the number of phonon momenta and mode indices. In contrast, the computational cost for evaluating all ph-ph collision integrals scales as $\mathcal{N}_\mathrm{ph}^2$. In the low-excitation regime, the carriers can be sampled within a few-eV energy window near the band edge or Fermi energy, thus in practice $\mathcal{N}_\mathrm{ph} \gg \mathcal{N}_\mathrm{c}$. Given this, computing the ph-ph collision integral $\mathcal{I}^{\text{ph-ph}}$ and the phonon dynamics is significantly more expensive – in practical calculations by 1–2 orders of magnitude – than computing the dynamics governed by e-ph interactions.

Auger and phonon-induced recombination processes, which typically occur on longer timescales than those considered [37], are not expected to significantly impact the carrier dynamics and therefore omitted from the simulations. However, incorporating these processes into the rt-BTE remains a promising direction for future work.

2.3 Results

Electron and phonon dynamics in graphene

Leveraging adaptive and multirate time integration methods, we study the coupled dynamics of carriers and phonons in graphene, where evolving the system for long times, up to tens of ps to access lattice dynamics, is computationally expensive. The initial state of excited carriers is set to a hot Fermi-Dirac distribution at 4000 K [16] while the phonons are set to a Bose-Einstein distribution at 300 K.

Using a third-order MRI method, we are able to extend the simulations to 80 ps with a reasonable computational cost. Figure 2.1 shows the carrier and phonon populations along a high-symmetry path at various time snapshots, capturing the main trends of the coupled dynamics. In the first ps, electrons and holes undergo rapid cooling and emit A_1' and E_{2g} optical phonons with momenta near Γ and K, due to the strong e-ph couplings with these phonons [38]. From 1 to 10 ps, the excess optical phonons decay into acoustic phonons near Γ. The thermalization process for flexural phonons is slow and extends to more than 80 ps.

Efficiency of adaptive and multirate time integration methods

Figure 2.2 provides a detailed benchmark of the efficiency improvements obtained with the adaptive and multirate schemes for these graphene simulations. We compare results obtained using RK4 with a fixed time step, an adaptive 4th-order ERK method, and a third-order MRI method. The accuracy and computational cost depend strongly on the choice of the following key parameters: the fixed time step size h in RK4, the fixed slow time step size h_s in the MRI method, and the tolerances in the standalone ERK method (and also in the ERK method used for the MRI fast timescale). In Fig. 2.2, we vary these parameters for each method and plot the carrier and phonon population errors at $t = 0.5$ ps against the corresponding computational cost (runtime). The error is taken relative to a reference MRI solution with a small step size h_s.

With RK4, the solution error can be reduced systematically by using smaller time steps at the expense of taking longer computation time. When the time step is too long (here $h > 5$ fs), the populations diverge at long enough simulation times. Error-versus-cost results with the ERK method are obtained by tightening the relative tolerance (rtol) from 10^{-4} to 10^{-11} while keeping the absolute tolerance fixed. Results using the MRI method are obtained by varying h_s with fixed tolerance values for the adaptive ERK method used for the fast timescale.

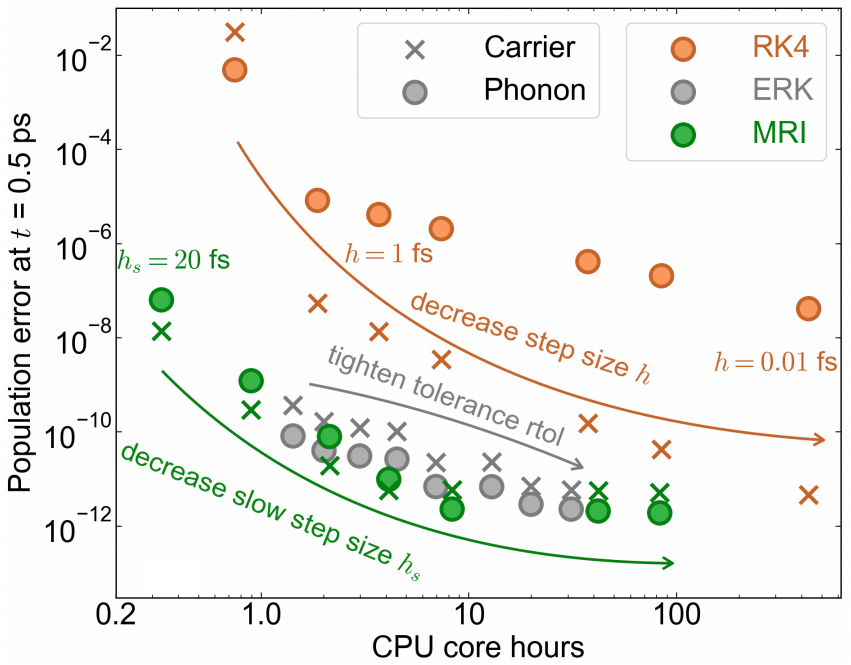

Figure 2.2: **Comparison of error and cost for solving the rt-BTE with different methods.** The carrier and phonon population errors at a fixed time (t = 0.5 ps) as a function of CPU core-hour cost, shown for RK4 (orange), ERK (gray), and MRI (green) time integration methods. Errors are computed as $\|f_{n\mathbf{k}}(t) - \tilde{f}_{n\mathbf{k}}(t)\|/\|\tilde{f}_{n\mathbf{k}}(t)\|$ for carriers and $\|N_{\nu\mathbf{q}}(t) - \tilde{N}_{\nu\mathbf{q}}(t)\|/\|\tilde{N}_{\nu\mathbf{q}}(t)\|$ for phonons, with the reference solutions $\tilde{f}_{n\mathbf{k}}(t)$ and $\tilde{N}_{\nu\mathbf{q}}(t)$ obtained using the MRI with h_s = 0.01 fs and tight tolerances. Each pair of cross and circle with the same color and CPU cost represents carrier and phonon population errors for the same simulation method and parameter settings. These parameter changes are indicated schematically with arrows for each method. For RK4, points from left to right represent results using a progressively smaller time step, h, from 5 fs to 0.01 fs. For MRI, results from left to right correspond to the slow time step, h_s, ranging from 20 fs to 0.05 fs. For ERK, the relative tolerance ranges from 10^{-4} to 10^{-11} from left to right with a fixed absolute tolerance.

Results for the ERK and MRI methods are in the lower-left corner of the error-versus-cost map in Fig. 2.2, which indicates their superior performance compared to RK4 with a fixed time step. These methods are more efficient over a wide range of tolerances (ERK) and slow time step values (MRI), and are particularly robust in reducing phonon population errors.

Detailed comparisons are helpful to quantify the performance improvement. The carrier population error using the MRI method with h_s = 20 fs is comparable to RK4 with a much shorter time step (h = 1 fs, a typical step size chosen for e-ph dynamics [16]), but the MRI method requires only ~10% of the computational effort

of RK4 for a 0.5-ps simulation. In addition, the phonon population error with the MRI method for the same settings is comparable to RK4 with an extremely short time step, $h = 0.01$ fs, but the computational cost for phonon dynamics is 1000x smaller relative to RK4. For the same computational cost as RK4 with $h = 1$ fs, the MRI method improves the carrier accuracy by 3 orders of magnitude and the phonon accuracy by 6 orders of magnitude. This dramatic improvement of efficiency and accuracy within the MRI method is game-changing—for example, it enables modeling nonequilibrium lattice dynamics up to long times (> 100 ps) and in bulk materials. Compared to RK4, the different error convergence behavior of MRI and ERK arises from the usage of adaptive time-stepping, which takes sufficiently small time steps in the early part of the dynamics, when the system is far from equilibrium, to limit error accumulation (see Supplementary Fig. 2.10d). In contrast, RK4 uses a fixed time step, leading to a larger integration error during this critical time window. Additional benchmarks for the ERK and MRI methods are shown in Supplementary Fig. 2.10-2.11. Overall, these results show that population errors are related to the adaptive step size and that both the absolute and relative tolerances contribute to controlling the errors by limiting h at early times. While adjusting the tolerance with adaptive methods influences the shape of the work-precision curves, it does not alter the overall conclusions for Fig. 2.2 in the main text.

The carrier and phonon population errors of the coupled dynamics simulated using different time integration methods and time-stepping parameters are shown in Fig. 2.3. In all simulations, we find that the errors approach constant values after 0.5 ps, showing that the error-cost results in Fig. 2.2 are representative of the entire dynamics of longer simulation times. Tolerances for both the ERK results and the fast timescale integration in the MRI method are set to rtol $= 10^{-5}$, $\text{atol}_\text{c} = 10^{-9}$, and $\text{atol}_\text{ph} = 10^{-12}$. With both the RK4 and MRI methods, the respective step sizes h and h_s determine how frequently the ph-ph collision integral $\mathcal{I}^{\text{ph-ph}}$ is evaluated, which is the main computational cost driver. To make a fair comparison, we set $h = h_s$ in both methods. As shown in Fig. 2.3, with this setup the MRI results are significantly more accurate than RK4, with accuracy greater by over 3 orders of magnitude for $h = 0.5$ fs and 8 orders of magnitude for $h = 5$ fs.

To highlight the superior performance of the MRI method, we also consider a first-order operator-splitting method where RK4 is employed with different fixed time steps for the fast (e-ph) and slow (ph-ph) components defined in Eq. 2.8. In this case, each simulation step performs RK4 advances of the fast and slow parts with

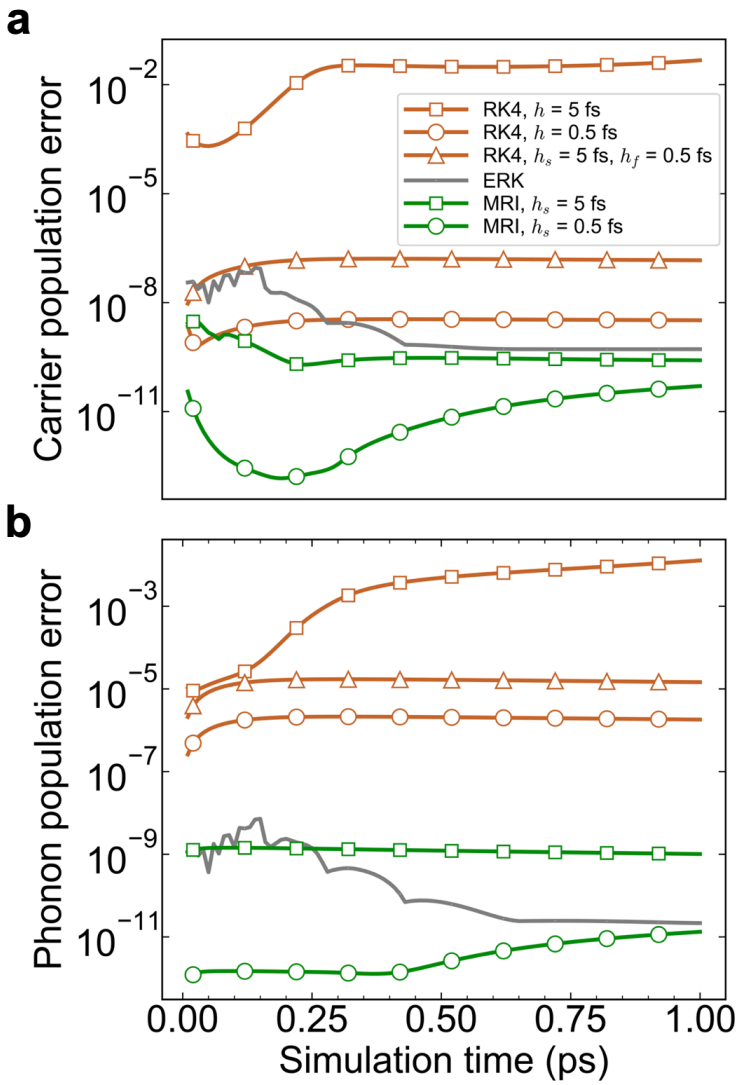

Figure 2.3: **Solution error for the coupled dynamics in graphene.** Population errors for **a** carrier and **b** phonon dynamics as functions of time using different time integration methods and parameters. The reference solution is the same as in Fig. 2.2. Orange curves show solution errors for RK4 with fixed time steps $h = 5$ fs (square) and 0.5 fs (circle), or an operator splitting approach with RK4 using separate time steps for ph-ph (5 fs) and e-ph/ph-e collisions (0.5 fs) (triangle). The green curves represent results for MRI with $h_s = 5$ fs (square) and 0.5 fs (circle). The gray curve shows results for ERK with relative tolerance 10^{-5}.

separate fixed time steps. The results of this method with a 0.5 fs time step for e-ph scattering and 5 fs for ph-ph scattering are plotted with orange triangles in Fig. 2.3. In terms of both speed and accuracy, this operator-splitting method falls between the RK4 results with $h = 0.5$ fs and $h = 5$ fs, without offering a clear improvement over RK4 with a single time step. This time-splitting approach is, therefore, significantly less efficient and accurate than the MRI and ERK methods. The same trends hold for other combinations of fixed fast and slow time steps when using the operator splitting method with RK4.

Our results for graphene demonstrate that the ERK and MRI methods can efficiently solve the coupled rt-BTE by taking advantage of the inherently different timescales of electron and phonon dynamics. More broadly, these findings suggest that adaptive time-stepping can improve the efficiency and accuracy of simulations involving coupled degrees of freedom in materials.

Silicon

We address the challenge of simulating lattice dynamics in bulk materials using first-principles e-ph and ph-ph scattering processes on dense momentum grids, choosing Si as a case study. This calculation is not feasible with fixed-step time integration methods such as RK4, even with the high-performance rt-BTE parallel implementation in PERTURBO [9]. As such, we employ a third-order MRI method paired with an adaptive ERK method as the fast timescale integrator for this simulation. At time zero, we initialize excited electrons in the conduction band of Si using a hot Fermi-Dirac distribution at 2000 K (with concentration 4.6×10^{20} cm^{-3}), while the phonons are initially in thermal equilibrium at 300 K.

The coupled electron and phonon dynamics in Si are analyzed in Fig. 2.4. Figure 2.4a illustrates the relaxation of excited electron populations in the X valleys of the lowest conduction band. The top two panels show the increase in electron populations $f_{n\mathbf{k}}(t)$ in the X valleys during the first picosecond. The bottom panel shows the change in populations from 1 ps to 20 ps, indicating a modest redistribution near the boundaries of the valleys and an even smaller change at their centers. The change in phonon populations for different phonon modes is visualized in Fig. 2.4b. In the first ps, most phonons are excited in the optical and longitudinal acoustic modes via e-ph interactions, mainly near the edge of the BZ or along the Γ-X high-symmetry line. Between 1 ps and 20 ps, these optically excited phonons decay into acoustic phonons, which progressively thermalize to long-wavelength lattice vibrations with

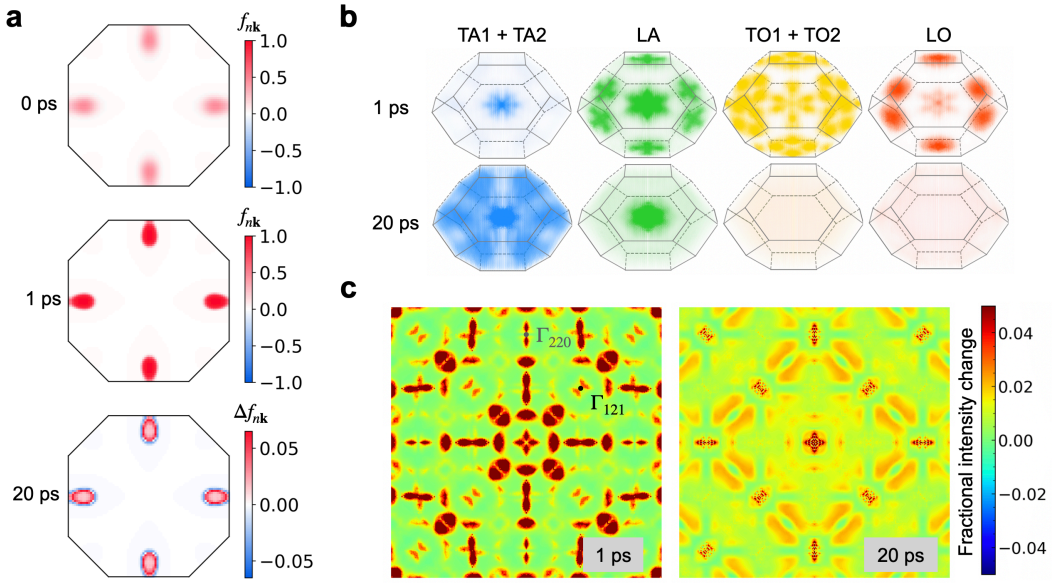

Figure 2.4: **Coupled dynamics of electrons and phonons in silicon.** **a** Electron populations, $f_{n\mathbf{k}}(t)$ in the lowest conduction band at 0 ps and 1 ps, and the change in the electron populations from 1 ps to 20 ps, $\Delta f_{n\mathbf{k}}$, shown in a 2D plane of the first BZ at $k_z = 0$. **b** Change in phonon populations, $\Delta N_{\nu\mathbf{q}}(t)$, in the first BZ at times 1 ps and 20 ps, plotted separately for transverse acoustic (TA1 + TA2), longitudinal acoustic (LA), transverse optical (TO1 + TO2), and longitudinal optical (LO) modes. **c** Fractional change in diffuse scattering intensity, $\Delta I(\mathbf{q}, t)$, for the (100) plane at 1 ps and 20 ps simulation times.

$q \to 0$ (Γ point in Fig. 2.4b).

Experiments have employed optical pump pulses to excite electrons in silicon and track their ultrafast dynamics [39–42]. In the absence of first-principles simulations, the two-temperature model [43] is commonly used to describe the energy transfer from excited carriers to the lattice during the relaxation. A three-temperature model (3TM) has also been used [41], which separately tracks the effective temperatures of electrons, acoustic phonons, and optical phonons. However, the 3TM fails to capture key aspects of the dynamics, such as the distinct relaxation times of different phonon modes and the long-time dynamics of acoustic phonons. This underscores the advantage of the rt-BTE as a first-principles, mode-resolved framework for simulating coupled carrier–lattice dynamics beyond the limitations of empirical models. Supplementary Fig. 5 shows the time evolution of effective phonon temperatures, highlighting the nonthermal behavior in the early stage of the dynamics and the different effective temperatures and dynamics of individual phonon modes.

Using the time-dependent phonon populations, we can connect our results with widely used experimental probes of time-domain lattice dynamics, particularly ultrafast diffraction techniques. Momentum-resolved thermal diffuse X-ray scattering (TDS) has been used extensively to determine phonon dispersions and study ultrafast phonon dynamics [44–46]. The transient TDS at scattering vector $\mathbf{q}$ at time t, $I(\mathbf{q}, t)$, is obtained from the time-dependent phonon populations, phonon frequencies, and structure factors $F_\nu(\mathbf{q}, t)$ [47] as

$$I(\mathbf{q}, t) \propto \sum_j \frac{1}{\omega_{\nu\mathbf{q}'}} \left[N_{\nu\mathbf{q}'}(t) + \frac{1}{2} \right] |F_\nu(\mathbf{q}, t)|^2, \tag{2.12}$$

where $\mathbf{q}' = \mathbf{q} - \mathbf{K}_\mathbf{q}$ is the momentum $\mathbf{q}$ folded to the first BZ, and $\mathbf{K}_\mathbf{q}$ is the nearest reciprocal lattice vector to $\mathbf{q}$. The one-phonon structure factor $F_\nu(\mathbf{q}, t)$ is defined as

$$F_\nu(\mathbf{q}, t) = \sum_\kappa \frac{f_\kappa(\mathbf{q})}{\sqrt{\mu_\kappa}} \exp\left[-g_\kappa(\mathbf{q})\right] (\mathbf{q} \cdot \mathbf{e}^\kappa_{\nu\mathbf{q}'}) \exp\left(-i\mathbf{K}_\mathbf{q} \cdot \mathbf{r}_\kappa\right) \tag{2.13}$$

where $f_\kappa(\mathbf{q})$ is the atomic scattering factor [48], $\mathbf{r}_\kappa$ is the equilibrium position of atom κ in the unit cell, and $g_\kappa(\mathbf{q})$ is the Debye-Waller factor.

Figure 2.4c shows the relative intensity change $\Delta I(\mathbf{q}, t) = \frac{I(\mathbf{q},t) - I(\mathbf{q},t=0)}{I(\mathbf{q},t=0)}$ at 1 ps and 20 ps on the (100) plane, where Γ_{000} is at the center of the plot, and Γ_{220} and Γ_{121} are labeled in the plot. At 1 ps, the TDS primarily shows an increase in optical phonons due to electron cooling. At 20 ps, the main contribution to the TDS is from acoustic phonons, with the LA modes yielding a signal close to Γ and the TA modes the "double-bar" patterns.

GaAs

To demonstrate the general applicability of our time-integration techniques beyond nonpolar materials, we simulate gallium arsenide (GaAs), a prototypical polar semiconductor where long-range e-ph interactions dominate the carrier cooling process [49–51]. In our simulation, the electrons are excited by a pump pulse centered at 1.88 eV, corresponding to an excitation of 0.45 eV above the band gap. The pulse amplitude is chosen to generate a carrier density of $1.1 \times 10^{19} \mathrm{cm}^{-3}$ at the end of the 50 fs pulse duration. We simulate the electron populations in the lowest conduction band and the excess phonon populations following the pulse and show them in Fig. 2.5. (Holes are omitted from the simulation due to their minor contribution to phonon dynamics.) The electrons exhibit slow relaxation in the time window from 0 to 20 ps, driven primarily by the accumulation of longitudinal optical phonons (LO) generated through the Fröhlich interaction [52]. The slow decay of these LO

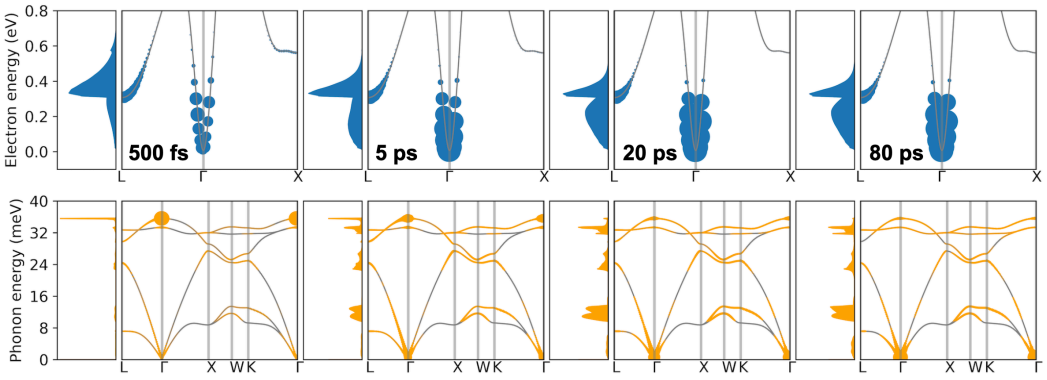

Figure 2.5: **Coupled electron and phonon dynamics in GaAs**. Top row: electron populations, $f_{n\mathbf{k}}(t)$, mapped onto the band structure at different time snapshots labeled in the panels, with point sizes proportional to the populations. Bottom row: excess phonon populations, $\Delta N_{\nu\mathbf{q}}(t)$, shown on the phonon dispersion. To the left of each panel: populations (electrons) or change of populations (phonons) as a function of energy averaged over the BZ. At time $t = 0$, the electrons are excited via a pump pulse 0.45 eV above the band gap, leading to a carrier concentration of $1.1 \times 10^{19}\text{cm}^{-3}$ at the end of the 50 fs pulse. The phonons are initially set at a Bose-Esinstein distribution at 300 K.

phonons creates a bottleneck that slows the electron cooling process [53]. This result is shown in Supplementary Fig. 6, where we analyze the contribution of nonequilibrium phonons to carrier cooling by simulating the same dynamics but including only a subset of interactions in the rt-BTE, and then comparing the corresponding time-dependent electronic temperatures.

Time step adaptivity is a key factor in improving the accuracy and efficiency of rt-BTE simulations. Therefore, it is instructive to analyze the time step sizes selected by the algorithm over the course of the simulation as the dynamics evolve. At early stages of the carrier dynamics in the graphene simulation, the adaptive step size in the ERK method is close to zero, and then it increases to about 5 fs after 400-fs simulation time (see Supplementary Fig. 1d). For both carrier and phonon populations, ERK and MRI methods exhibit distinctly different error convergence characteristics compared to RK4, when parameters are set to match the computational cost. This difference underscores the need to resolve early-time dynamics, when the system is far from equilibrium, to minimize error propagation to longer times. With reasonable tolerances and slow time-step size choices, the average adaptive step with the ERK method and the fast time step in the MRI method both reach a steady-state value of 5 fs, suggesting that adaptive methods can find the characteristic timescales

of physical interactions in the system.

These results show that adaptive and multirate time-stepping of the rt-BTEs enables first-principles simulations of phonon dynamics up to long timescales in bulk crystals, as well as modeling TDS experiments that are widely used to probe ultrafast lattice dynamics.

2.4 Extension

With the success achieved by the MRI method above, which uses a fixed slow time step for the ph-ph interactions, we further explore a fully adaptive ERK method to solve the coupled rt-BTEs. Unlike simpler subcycling methods that either alternate between evolving f^s and f^f or that freeze f^s throughout all fast steps, the fully adaptive MRI method allows high-order coupling between the processes, including methods with accuracy up to $O(h_s^5)$ and $O(h_s^6)$ [54, 55]. A considerably attractive feature of MRI methods is that the fast time scale is defined at the continuum level, allowing it to be solved using any desired algorithm that supports initial value problems (IVPs) of the form

$$v_i'(t) = f^f(v) + r_i(t), \quad t \in [t_{0,i}, t_{f,i}], \quad v(t_{n,i}) = v_{0,i}, \tag{2.14}$$

where the stage forcing function is a time-dependent linear combination of slow function evaluations,

$$r_i(t) = \sum_{j=0}^{i-1} \alpha_{i,j}(t) f^s(z_j), \tag{2.15}$$

and the updated stage is given by $z_i = v(t_{f,i})$. The stage time intervals $[t_{0,i}, t_{f,i}]$, stage initial condition $v_{0,i}$, and stage coefficients $\alpha_{i,j}$ are defined by the specific MRI method being used. The overall MRI time step solution is then given by the final stage, i.e., $y_{n+1} = z_s$. Note that this is different from the fixed-step MRI formulation in Eq. 2.8, where the fast time scale is only solved on non-overlapping intervals of the slow time step.

Results with fully adaptive MRI

We compare the performance of fully adaptive MRI methods against the fixed slow step MRI, which was used in previous sections. They hereon are referred to as ERK-MRI, as the fast scale is always solved with an adaptive ERK method of given order.

Figure 2.6 presents the error–cost curves of ERK-MRI methods with varying fast solver order and tolerance for simulations up to $t = 5$ ps. The default fast solver

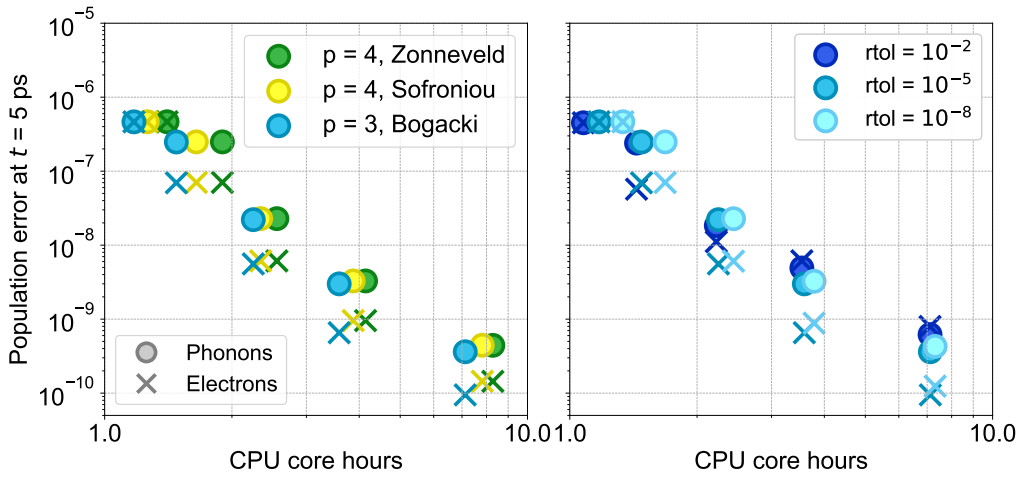

Figure 2.6: **Error-cost curves of ERK-MRI methods with varying order and tolerance.** Carrier and phonon population errors in graphene at $t = 5$ ps as a function of CPU core hours for MRI methods with fixed slow step h_s (denoted ERK-MRI), similar to Fig. 2.2. The left panel illustrates varying order p the fast solver. The right panel shows varying relative tolerance rtol. For each run, h_s takes values, from left to right of each panel, of 100, 50, 20, 10 and 5 fs. Circles and crosses denote phonon and electron population errors, respectively. The reference solution is obtained using ERK-MRI with $h_s = 0.1$ fs and tight tolerances.

used simulations in previous sections is ZONNEVELD_5_3_4 [27] a fourth-order method. In the left panel, the relative tolerance is fixed at rtol $= 10^{-5}$, while $\text{atol}_\text{c} = 10^{-9}$ and $\text{atol}_\text{ph} = 10^{-12}$, consistent with earlier simulations. An alternative fast solver of the 4th order, SOFRONIOU_SPALETTA_5_3_4 [56] (labeled as SOFRONIOU in the plot), achieves slightly improved accuracy as well as lower computational cost, yielding approximately 16% for a required error level of 10^{-6}, at $h_s = 100$ fs. Further reducing the fast solver order to $p = 3$ using BOGACKI_SHAMPINE_4_2_3 [57] leads to additional cost savings without losing accuracy. In the right panel of Fig. 2.6, the best-performing solver identified on the left, labeled as BOGACKI, is used to examine the effect of varying relative tolerance. Loosening rtol from 10^{-8} to 10^{-5} slightly improves computational cost without sacrificing accuracy. In contrast, further loosening rtol to 10^{-2} increases the error of carrier population by approximately an order of magnitude with negligible reduction in the computational time at $h_s = 5$ fs. A close to 10% reduction in computational cost is observed at looser rtol. Snapshots at $t = 0.5$ ps and $t = 40$ ps exhibit the same trend, which is not shown here for brevity. Overall, these results

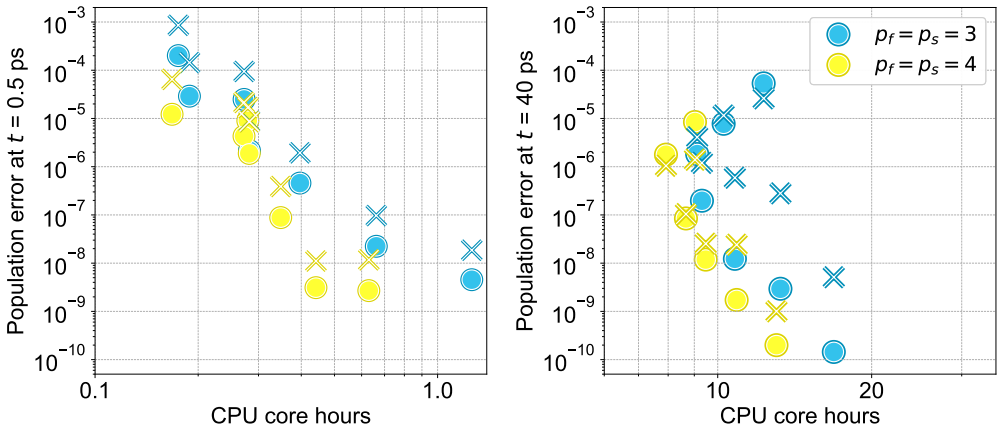

Figure 2.7: **Error–cost curves of GARK methods with varying order.** Population errors in graphene at $t = 0.5$ ps and $t = 40$ ps as a function of CPU core hours for MRI methods with adaptive slow steps (denoted GARK) with solver orders of either 3 or 4. For each panel, rtol takes values from left to right ranging from of 10^{-2} to 10^{-8} in increments of 10. 10^{-2} with fast and slow solver order $p_f = p_s = 4$ is omitted because the simulation failed to complete. The reference solution is obtained using ERK-MRI with $h_s = 0.01$ fs and tight tolerances for the left panel, and $h_s = 1$ fs for the right panel.

highlight that both the fast solver order and tolerance can be systematically tuned to optimize speedup while maintaining accurate carrier-population dynamics.

In the fully adaptive setting, performance is primarily governed by the adaptivity strategy, the type and order of solvers, and the chosen tolerances. Figure 2.7 presents the error–cost curves of fully adaptive MRI methods (denoted as GARK [54]) with varying rtol for two different fast solvers for simulations up to $t = 0.5$ ps and $t = 40$ ps. The case $p_f = p_s = 3$ corresponds to the same fast solver, BOGACKI, used in Fig. 2.6. The slow solver order [24] is chosen to match the fast solver. It can be shown that choosing $p_s = 4$ and $p_f = 2$ can further enhance performance, up to 30%, provided that the error controller can be optimized to more effectively identify and reject failed steps (see Fig. 2.16). For simplicity, tolerances are chosen to be the same for both slow and fast processes. The results are compared to using a higher order fast solver with $p_f = p_s = 4$ (ZONNEVELD in Fig. 2.6 for the fast solver). The simulation error decreases systematically as rtol is tightened from 10^{-2} to 10^{-8}, except for the case of rtol $= 10^{-2}$ using the fourth-order method, which is omitted due to solver failure to complete the simulation. Further analysis

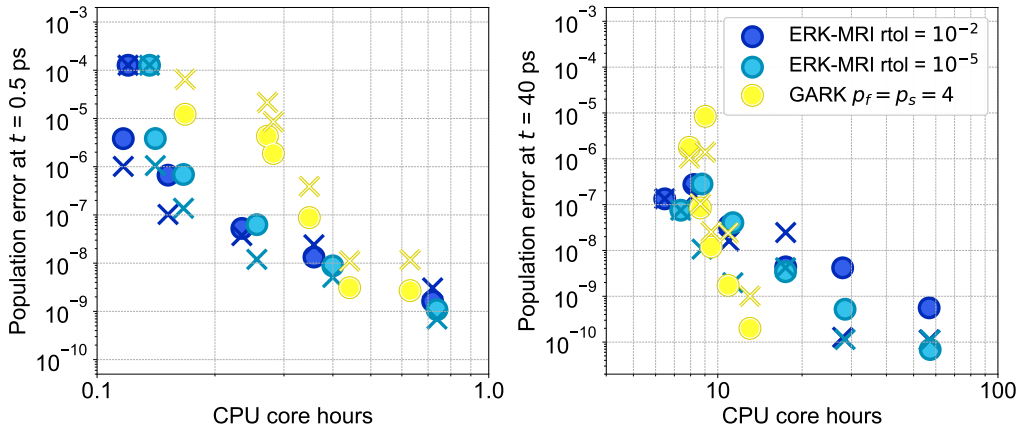

Figure 2.8: **Error–cost comparison of GARK and ERK-MRI methods.** Population errors in graphene at $t = 0.5$ ps and $t = 40$ ps as a function of CPU core hours for GARK and ERK-MRI. For each panel, rtol takes values from left to right ranging from 10^{-3} to 10^{-8} for the GARK method; h_s takes values of 200, 100, 50, 20, 10 and 5 fs for the ERK-MRI method.

of the time step evolution reveals that the solver is unable to identify a fast time step value that satisfies the error criteria at a specific point during the simulation. Tightening rtol for the slow process to 10^{-5}, while keeping it at 10^{-2} for the fast process, successfully resolves the issue.

Although for rtol ranging from 10^{-4} to 10^{-8}, the computational cost decreases monotonically with looser tolerances, for the fourth-order method with rtol = 10^{-2} at $t = 0.5$ ps, represented by the circle at phonon error of approximately 4×10^{-6}, the computational cost instead increases. Similarly, for rtol of 10^{-2} and 10^{-3} at $t = 40$ ps with $p_f = p_s = 3$ and 10^{-3} with $p_f = p_s = 4$, the cost is also higher than that obtained with the next tighter tolerance. This non-monotonic behavior can be attributed to the fact that loosening rtol leads to an increased number of failed steps, which require additional computational effort to reattempt and complete the simulation.

Overall, the higher-order GARK method shows consistently better performance across simulation times, especially for longer simulations. Comparing Fig. 2.7 to Fig. 2.6, we find that the choice of solver order plays a critical role. We also explored other fully adaptive methods, including SR and MERK schemes [54, 58]; however, both underperform GARK methods by an order of magnitude in computational

speed at comparable accuracy, regardless of solver order.

Figure 2.8 compares the performance of fully adaptive GARK methods against fixed slow step ERK-MRI methods for simulations up to $t = 0.5$ ps and $t = 40$ ps. The fixed step methods use the BOGACKI fast solver identified as optimal in Fig. 2.6, while the fully adaptive GARK methods uses ZONNEVELD as the fast and slow solvers, which was found to be optimal in Fig. 2.7. Two different choices of rtol for the fast scale are shown for ERK-MRI method, while GARK method varies rtol from 10^{-3} to 10^{-8}. Comparing the left and right panels, we observe a crossover for the efficiency of the two methods across simulation times. For a short simulation ($t = 0.5$ ps), the fixed step ERK-MRI method outperforms the fully adaptive GARK method across all tolerances considered. However, for a longer simulation ($t = 40$ ps), the fully adaptive GARK method becomes more efficient, especially at tighter tolerances. This crossover can be attributed to the ability of the fully adaptive GARK method to dynamically adjust the slow time step based on the evolving dynamics without an imposed upper limit. These results highlight the advantages of fully adaptive multirate methods for simulating coupled electron and phonon dynamics over extended timescales.

Finally, we analyze the time-step evolution of GARK methods over long simulation times in Fig. 2.9. As shown in Fig. 2.9a and d, for two choices of rtol, 10^{-3} and 10^{-5}, the slow step size h_s starts from a small value at the beginning of the simulation and gradually increases to the same steady-state value in both cases. A similar trend is observed for the fast step in GARK, as shown in Fig. 2.9b and e. This behavior of the fast step is consistent with that observed for the ERK method, as shown in Supplementary Fig. 2.10. Figure 2.9c and f show the simulation time advancement as a function of the number of fast steps taken, which more clearly illustrates the fast step evolution with the restarts induced by failed slow steps.

The choice in rtol primarily affects the number of failed steps taken during the simulation, as well as the elapsed simulation time before the steady-state step sizes are reached. For relatively tight tolerances, time step sizes for both slow and fast processes do not "overshoot" from steady-state value. For example, in the case of rtol $= 10^{-6}$ (not shown here for brevity), the slow time step size reaches steady state at around 75 ps simulation time without exceeding 2200 fs. Similarly, for the fast step, the simulation time elapsed before reaching steady state increases with tighter tolerance, while the steady-state value remains unchanged. This behavior aligns with the observation of ERK method shown in Supplementary Fig. 2.10.

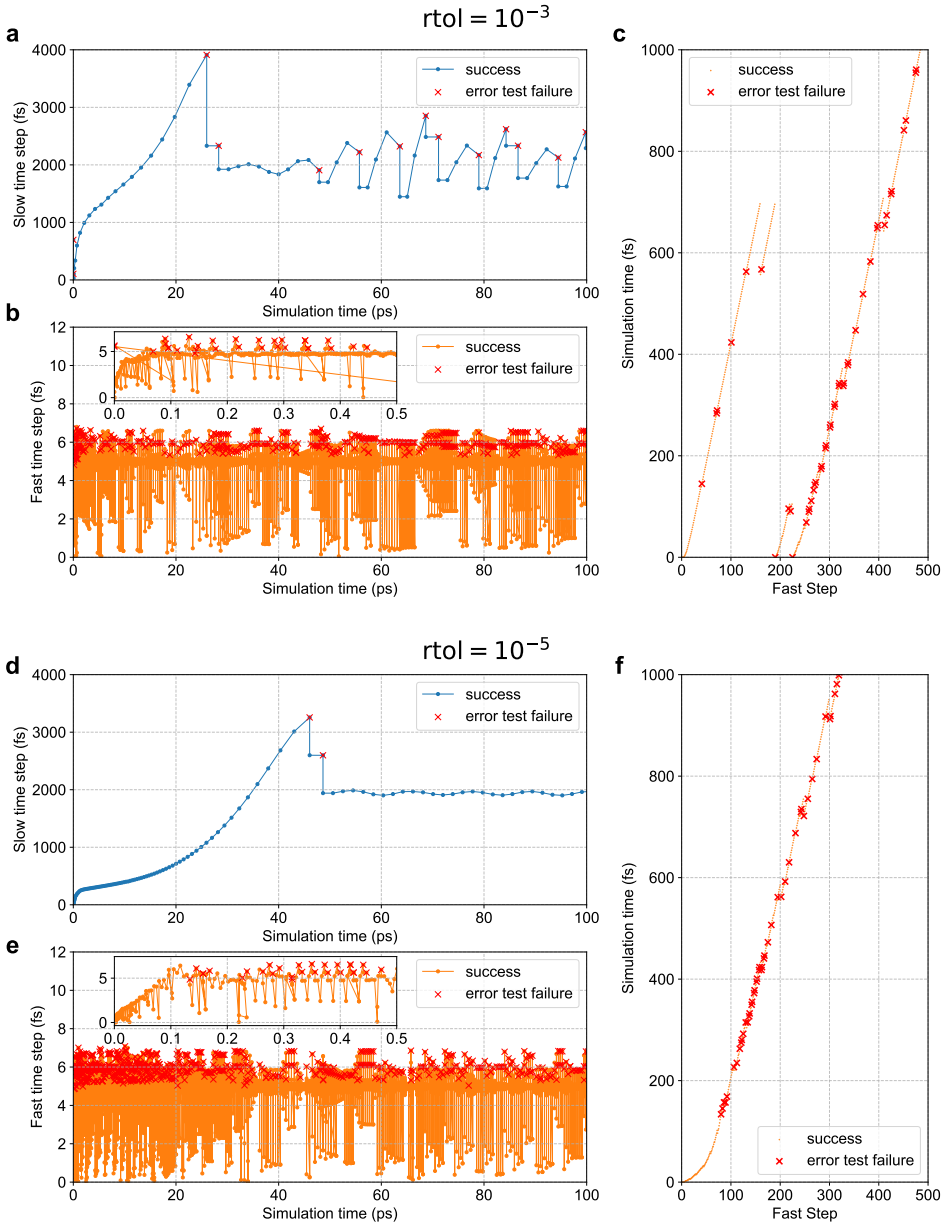

Figure 2.9: **Adaptive time step evolution for GARK methods in the graphene simulation. a** Slow step size h_s as a function of simulation time for GARK methods with rtol = 10^{-3}; red crosses represent failed steps. **b** Fast step size h_f as a function of simulation time. The inset resolves the first 0.5 ps of the simulation. Apparent reversals in time of lines connecting the fast steps correspond to restarts of the slow step after a failure. **c** Simulation time advancement in fs as a function of fast steps taken, which more clearly illustrates the fast step behaviors with slow step restarts. **d, e, f** Same as **a, b, c**, respectively, but for rtol = 10^{-5}.

With rtol = 10^{-5}, the GARK method can reach a steady-state h_s of approximately without large oscillations around 2000 fs, avoiding excessive failed steps and additional computational cost observed for 10^{-3}. Even without error analysis at longer simulation time ($\sim$ 100 ps), based on the results in Fig. 2.9 and Fig. 2.8, one can expect that the fully adaptive fourth-order GARK method with rtol = 10^{-5} will outperform the case with rtol = 10^{-4}, because the latter exhibits more failed steps beyond 40 ps.

From these results, we conclude that the optimal choice of tolerances exists and depends on the timescale of interest in the simulation. More importantly, the fully adaptive scheme provides further evidence that adaptive time-stepping can identify the intrinsic timescales of physical interactions in the system. Both the fast and slow time steps converge to steady-state values that reflect the characteristic relaxation times of e-ph and ph-ph interactions in graphene, respectively.

2.5 Discussion

In addition to improving efficiency, using adaptive methods eliminates the need to converge the solution with respect to the chosen time step for both ERK and fully adaptive MRI methods. Both schemes, as well as ERK-MRI methods, enhance efficiency compared to fixed time step methods. (The only point to note is that if the chosen slow time-step size is unnecessarily small, the solution will be more accurate than required, as the adaptive fast time step is constrained to be smaller than the slow time step, leading to a higher computation cost.)

The range of nonequilibrium dynamics that can be addressed with multirate methods is not limited to the coupled electron and phonon system studied here. Dynamics of excitons and other elementary excitations can also be studied from first principles, including their couplings with the lattice [59, 60]. Frameworks beyond the rt-BTE, including TDDFT, master-equation methods such as the Liouville or Lindblad equations, and NEGF could all greatly benefit from the multirate and adaptive time-stepping shown in this work. This advance provides a new tool for addressing a potentially wide range of problems in nonequilibrium materials physics, including modeling and interpreting time-domain spectroscopy and diffraction experiments.

2.6 Conclusion

In summary, we applied adaptive and multirate time-stepping methods to accelerate simulations of coupled electron and phonon dynamics using the rt-BTE. We achieve a 10- to 100-fold speedup relative to fixed time step methods with the same accuracy,

or orders of magnitude greater accuracy when setting simulation parameters to enforce equal computational cost. Beyond computational efficiency, our approach can automatically adapt the time integration to the intrinsic timescales of physical interactions in the system. This advance sets the stage for studying nonequilibrium dynamics at longer timescales and for a wider range of coupled degrees of freedom in materials, including electron, spin, phonon, and other excitations, opening the door to exploring new physical regimes. Future directions include adding explicit treatments of light pulses, coherent electron and phonon dynamics, and higher-order phonon interactions to explore novel quantum states and dynamical control of materials, such as phonon-driven Floquet engineering and time-domain tuning of physical properties, order parameters, spin texture, and crystal structure.

2.7 Supplementary information

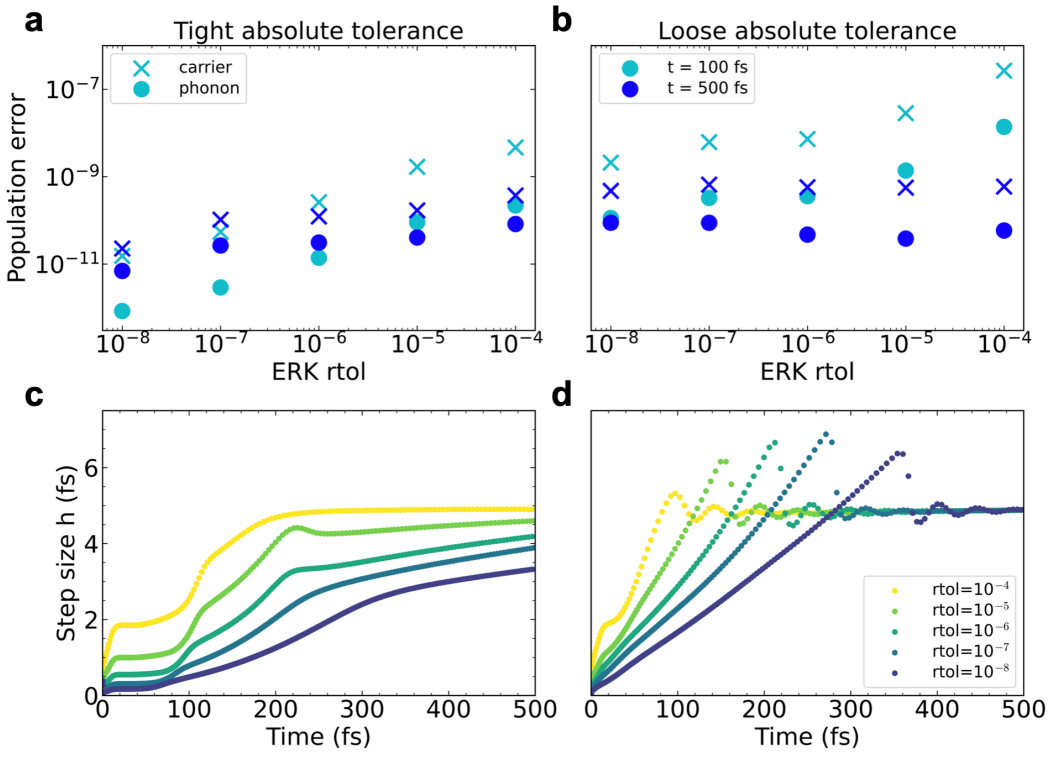

Figure 2.10: **Effect of tolerances with an adaptive ERK method. a.** Carrier and phonon population errors in graphene as a function of the relative tolerance, rtol, with the ERK method, obtained using $\text{atol}_c = \text{atol}_{ph} = 10^{-15}$, for simulation times 100 fs and 500 fs. **b.** Population errors as a function of rtol using the ERK method with $\text{atol}_c = 10^{-9}$ and $\text{atol}_{ph} = 10^{-12}$. **c.** Evolution of the ERK method adaptive time-step size, h, for different choices of rtol (see legend in panel **d**) using the same absolute tolerances as in **a**. **d.** Evolution of the ERK method adaptive time-step size, h, with absolute tolerance as in **b**, and rtol from 10^{-4} to 10^{-8}.

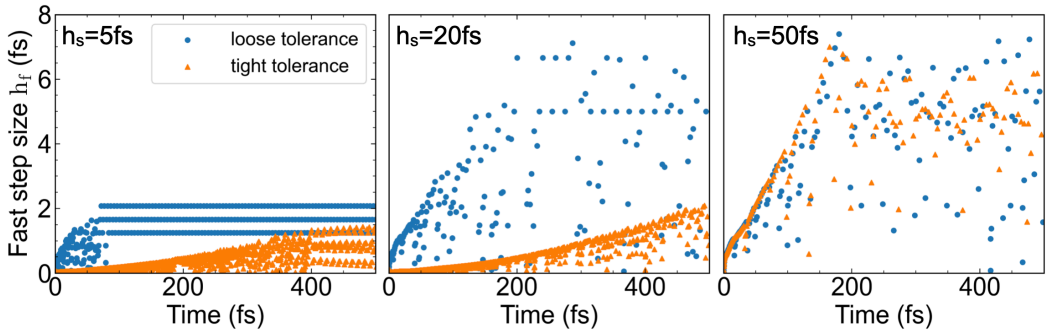

Figure 2.11: **Effect of tolerance and slow time-step size with an MRI method.** Evolution of the MRI method fast time-step size, h_f, for different choices of fixed slow time-step size, h_s. From left to right h_s = 5 fs, 20 fs, and 50 fs. The results are for ultrafast dynamics in graphene. Data in blue are obtained with loose tolerance values (rtol = 10^{-5}, atol_c= 10^{-9}, and atol_ph= 10^{-12}) for the adaptive ERK method used as the fast timescale integrator while data in orange is obtained with tight tolerances (rtol = 10^{-10} and atol_c = atol_ph = 10^{-15}). Small h_s values limit the maximum fast time-step size. For large enough h_s values, the average fast step size is determined by the underlying dynamics and tolerance values.

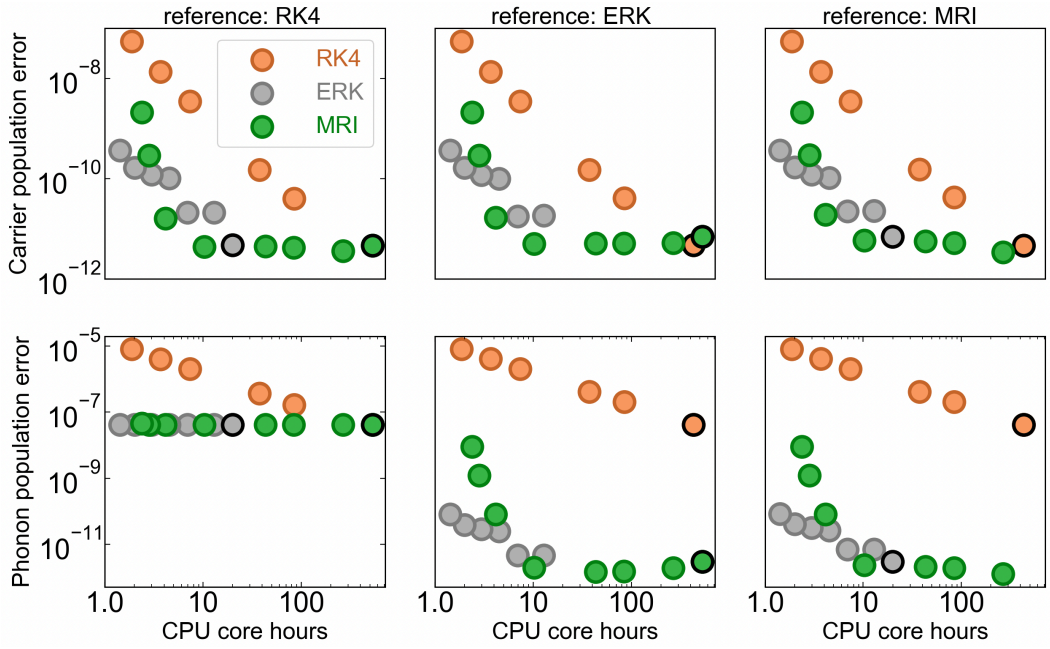

Figure 2.12: **Choice of reference solution for rt-BTE benchmarks.** To validate our use of an MRI method with $h_s = 0.01$ fs as the reference solution, we show the population errors for charge carriers (top) and phonons (bottom) for RK4 (orange), ERK (gray) and MRI (green) methods relative to different reference solutions. From left to right, errors are shown using as reference, respectively, RK4 with $h = 0.01$ fs, ERK with $\text{atol}_c = \text{atol}_{\text{ph}} = 10^{-15}$ and $\text{rtol} = 10^{-10}$, and MRI with $h_s = 0.01$ fs and the same tolerances for the fast timescale ERK method. The errors of candidate reference solutions are highlighted with black contours. Out of these three different reference solutions, only errors computed against accurate MRI simulations (right column) exhibit a progressive decrease with increasing accuracy – and thus increasing CPU core-hour computational cost – for all three methods, making accurate MRI calculations an ideal reference choice. (The increasing accuracy is achieved by decreasing step sizes or tightening tolerance levels in all three methods.)

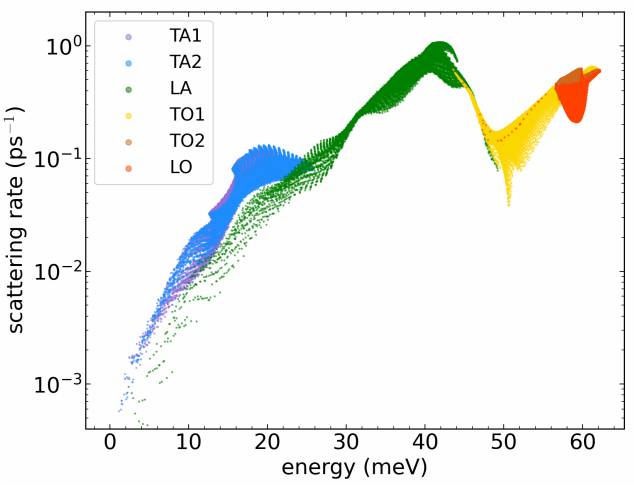

Figure 2.13: **Phonon scattering rates in silicon.** Mode-resolved phonon scattering rates in silicon from anharmonic ph-ph interactions, computed using an $80\times80\times80$ **q**-point grid and labeled according to the type of phonon mode – longitudinal (L) or transverse (T), and acoustic (A) or optical (O). Because of their lower scattering rates, TA phonons are expected to relax more slowly than other phonon modes, with a characteristic timescale of order 10–100 ps.

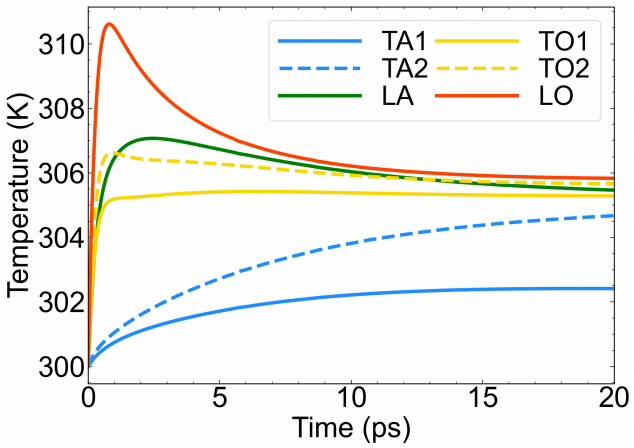

Figure 2.14: **Phonon effective temperatures in silicon.** Time-dependent effective temperatures for different phonon modes, obtained by analyzing the coupled electron and phonon dynamics in silicon. The effective temperature is computed as $\tilde{T}_\nu(t) = \frac{1}{N_\mathbf{q}} \sum_\mathbf{q} T_{\nu\mathbf{q}}(t)$, where $T_{\nu\mathbf{q}}(t)$ is obtained by inverting $N_{\nu\mathbf{q}}(t) = [e^{\hbar\omega_{\nu\mathbf{q}}/k_B T_{\nu\mathbf{q}}(t)} - 1]^{-1}$, k_B being the Boltzmann constant. The rapid increase of LO-mode temperature at short times is due to e-ph interactions, which generate a large excess of LO phonons on a 10–600 fs timescale. At longer times, different modes trade energy and thermalize through ph-ph interactions.

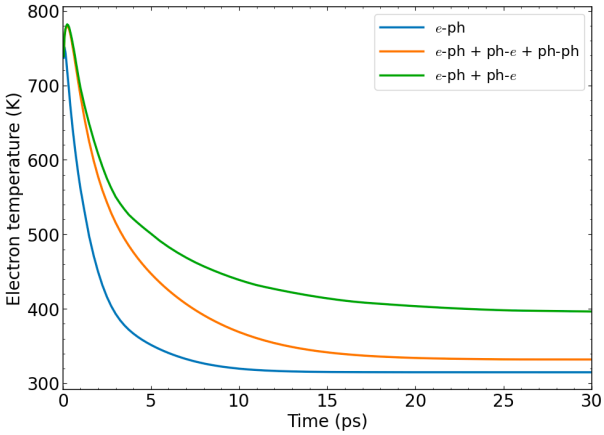

Figure 2.15: **Effective electron temperatures in GaAs.** Electron temperatures extracted from fitting the populations to a Fermi-Dirac distribution. The blue, green, and orange curves represent simulations including only electron-phonon (e–ph) scattering, both e–ph and phonon-electron (ph–e) scattering, and all three scattering processes: e–ph, ph–e, and ph–ph, respectively. The fastest electron cooling occurs when only e–ph scattering is included. A comparison between the orange and green curves shows that ph–ph scattering accelerates electron cooling at later times. This behavior reveals the hot phonon bottleneck effect, arising from an overpopulation of longitudinal optical (LO) phonons combined with weak ph-ph scattering that slows down energy dissipation in the electronic system.

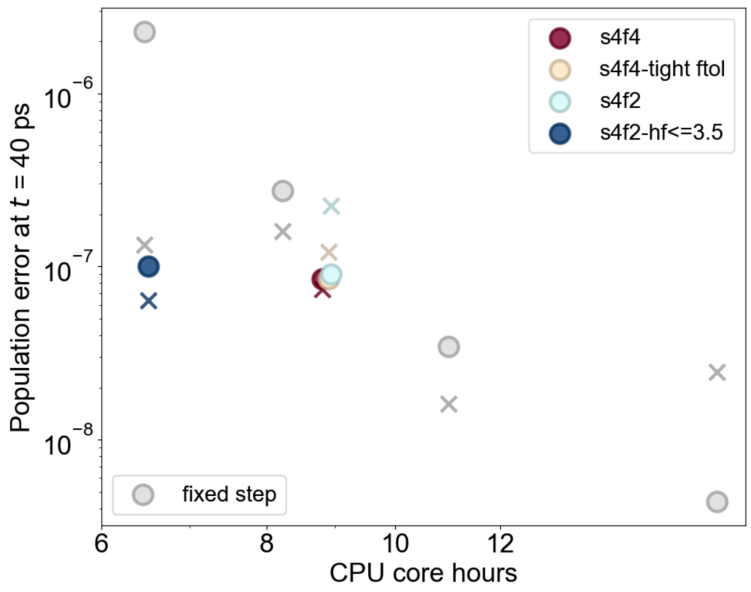

Figure 2.16: **Error-cost comparison of fine-tuning solver orders and step controller.** Population errors in graphene at $t = 40$ ps as a function of CPU core hours. The grey dots represent results using the ERK-MRI method with varying fixed slow step size. The label s4f4 represent using GARK with a slow solver of order $p_s = 4$ and a fast solver of order $p_f = 4$, which is carefully chosen as the optimal solver combination from results represented by Fig. 2.7. Tight ftol represent using an rtol for the fast process that is tighter than rtol for the slow process using the same solvers with GARK. We see that relaxing p_f to 2 does not change the simulation time but reduces accuracy of carrier population. However, capping the maximum allowed adaptive fast time step h_f to 3.5 fs significantly reduces computational time at similar accuracy. This is because the fast time step h_f can grow too large in some steps, leading to failed steps and extra computation time. By capping h_f, we avoid these failed steps and improve efficiency.

References

[1] M. Mitrano, A. Cantaluppi, D. Nicoletti, S. Kaiser, A. Perucchi, S. Lupi, P. Di Pietro, D. Pontiroli, M. Riccò, S. R. Clark, *et al.*, Nature **530**, 461 (2016).

[2] F. Adler, M. Geiger, A. Bauknecht, F. Scholz, H. Schweizer, M. H. Pilkuhn, B. Ohnesorge, and A. Forchel, J. Appl. Phys. **80**, 4019–4026 (1996).

[3] M. Först, C. Manzoni, S. Kaiser, Y. Tomioka, Y. Tokura, R. Merlin, and A. Cavalleri, Nat. Phys. **7**, 854 (2011).

[4] E. Mashkovich, R. V. Mikhaylovskiy, A. V. Kimel, R. V. Pisarev, A. M. Balbashov, A. K. Zvezdin, and D. A. Pisarev, Nat. Commun. **12**, 1 (2021).

[5] S. René de Cotret, N. Del Fatti, F. Vallée, P. A. Mante, and C. Thomsen, Phys. Rev. B **99**, 144305 (2019).

[6] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods*, 2nd ed. (Cambridge University Press, 2020).

[7] S. Baroni, S. de Gironcoli, A. Dal Corso, and P. Giannozzi, Rev. Mod. Phys. **73**, 515 (2001).

[8] P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. B. Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni, N. Colonna, I. Carnimeo, A. D. Corso, S. de Gironcoli, P. Delugas, R. A. D. Jr, A. Ferretti, A. Floris, G. Fratesi, G. Fugallo, R. Gebauer, U. Gerstmann, F. Giustino, T. Gorni, J. Jia, M. Kawamura, H.-Y. Ko, A. Kokalj, E. Küçükbenli, M. Lazzeri, M. Marsili, N. Marzari, F. Mauri, N. L. Nguyen, H.-V. Nguyen, A. O. de-la Roza, L. Paulatto, S. Poncé, D. Rocca, R. Sabatini, B. Santra, M. Schlipf, A. P. Seitsonen, A. Smogunov, I. Timrov, T. Thonhauser, P. Umari, N. Vast, X. Wu, and S. Baroni, J. Phys. Condens. Matter **29**, 465901 (2017).

[9] J.-J. Zhou, J. Park, I.-T. Lu, I. Maliyov, X. Tong, and M. Bernardi, Comput. Phys. Commun. **264**, 107970 (2021).

[10] I. Maliyov, J. Park, and M. Bernardi, Phys. Rev. B **104**, L100303 (2021).

[11] I. Maliyov, J. Yin, J. Yao, C. Yang, and M. Bernardi, Npj Comput. Mater. **10**, 123 (2024).

[12] W. Li, J. Carrete, N. A. Katcho, and N. Mingo, Comp. Phys. Commun. **185**, 1747–1758 (2014).

[13] A. Togo, L. Chaput, and I. Tanaka, Phys. Rev. B **91**, 094306 (2015).

[14] J. Carrete, B. Vermeersch, A. Katre, A. van Roekeghem, T. Wang, G. K. Madsen, and N. Mingo, Comput. Phys. Commun. **220**, 351 (2017).

[15] O. Hellman and I. A. Abrikosov, Phys. Rev. B **88**, 144301 (2013).

[16] X. Tong and M. Bernardi, Phys. Rev. Res. **3**, 023072 (2021).

[17] F. Caruso, J. Phys. Chem. Lett. **12**, 1734 (2021).

[18] T. L. Britt, Q. Li, L. P. René de Cotret, N. Olsen, M. Otto, S. A. Hassan, M. Zacharias, F. Caruso, X. Zhu, and B. J. Siwick, Nano Lett. **22**, 4718 (2022).

[19] Y. Pan and F. Caruso, Nano Lett. **23**, 7463 (2023).

[20] A. C. Hindmarsh, P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker, and C. S. Woodward, ACM Trans. Math. Softw. **31**, 363 (2005).

[21] D. J. Gardner, D. R. Reynolds, C. S. Woodward, and C. J. Balos, ACM Trans. Math. Softw. **48**, 1 (2022).

[22] D. R. Reynolds, D. J. Gardner, C. S. Woodward, and R. Chinomona, ACM Trans. Math. Softw. **49**, 1 (2023).

[23] E. Harier, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems*, 2nd ed., Springer Series in Computational Mathematics (Springer, 2008).

[24] A. Sandu, SIAM J. Numer. Anal. **57**, 2300 (2019).

[25] M. Schlegel, O. Knoth, M. Arnold, and R. Wolke, J. Comput. Appl. Math. **226**, 345 (2009).

[26] J. J. Loffeld, A. Nonaka, D. R. Reynolds, D. J. Gardner, and C. S. Woodward, Int. J. High Perform. Comput. Appl. **38**, 263 (2024).

[27] J. Zonneveld, *Automatic integration of ordinary differential equations*, Tech. Rep. R743 (Mathematisch Centrum, 1963).

[28] G. Söderlind, CWI Quarterly **11**, 55 (1998).

[29] G. Söderlind, ACM Trans. Math. Softw. **29**, 1 (2003).

[30] G. Söderlind, Appl. Numer. Math. **56**, 488 (2006).

[31] M. Schlegel, O. Knoth, M. Arnold, and R. Wolke, Geosci. Model Dev. **5**, 1395 (2012).

[32] M. Schlegel, O. Knoth, M. Arnold, and R. Wolke, Appl. Numer. Math. **62**, 1531 (2012).

[33] O. Knoth and R. Wolke, Appl. Numer. Math. **28**, 327 (1998).

[34] N. Troullier and J. L. Martins, Phys. Rev. B **43**, 1993 (1991).

[35] A. A. Mostofi, J. R. Yates, G. Pizzi, Y.-S. Lee, I. Souza, D. Vanderbilt, and N. Marzari, Comput. Phys. Commun. **185**, 2309 (2014).

[36] J.-J. Zhou, O. Hellman, and M. Bernardi, Phys. Rev. Lett. **121**, 226603 (2018).

[37] O. Madelung, U. Rössler, and M. Schulz, eds., *Group IV Elements, IV-IV and III-V Compounds. Part b - Electronic, Transport, Optical and Other Properties* (Springer-Verlag, 2002).

[38] M. X. Na, A. K. Mills, F. Boschini, M. Michiardi, B. Nosarzewski, R. P. Day, E. Razzoli, A. Sheyerman, M. Schneider, G. Levy, S. Zhdanovich, T. P. Devereaux, A. F. Kemper, D. J. Jones, and A. Damascelli, Science **366**, 1231 (2019).

[39] M. van Exter and D. Grischkowsky, Phys. Rev. B **41**, 12140–12149 (1990).

[40] A. J. Sabbah and D. M. Riffe, Phys. Rev. B **66**, 165217 (2002).

[41] S. K. Cushing, M. Zürch, P. M. Kraus, L. M. Carneiro, A. Lee, H.-T. Chang, C. J. Kaplan, and S. R. Leone, Struct. Dyn. **5**, 054302 (2018).

[42] M. Wörle, A. W. Holleitner, R. Kienberger, and H. Iglev, Phys. Rev. B **104**, L041201 (2021).

[43] R. B. Wilson, J. P. Feser, G. T. Hohensee, and D. G. Cahill, Phys. Rev. B **88**, 144305 (2013).

[44] M. Holt, Z. Wu, H. Hong, P. Zschack, P. Jemian, J. Tischler, H. Chen, and T.-C. Chiang, Phys. Rev. Lett. **83**, 3317 (1999).

[45] M. Trigo, M. Fuchs, J. Chen, M. P. Jiang, M. Cammarata, S. Fahy, D. M. Fritz, K. Gaffney, S. Ghimire, A. Higginbotham, S. L. Johnson, M. E. Kozina, J. Larsson, H. Lemke, A. M. Lindenberg, G. Ndabashimiye, F. Quirin, K. Sokolowski-Tinten, C. Uher, G. Wang, J. S. Wark, D. Zhu, and D. A. Reis, Nat. Phys. **9**, 790 (2013).

[46] D. Filippetto, P. Musumeci, R. K. Li, B. J. Siwick, M. R. Otto, M. Centurion, and J. P. F. Nunes, Rev. Mod. Phys. **94**, 045004 (2022).

[47] B. E. Warren, *X-Ray Diffraction* (Dover Publications, New York, 1990).

[48] S. Olukayode, C. Froese Fischer, and A. Volkov, Acta Crystallogr. Sect. A **79**, 59 (2023).

[49] J. Sjakste, N. Vast, M. Calandra, and F. Mauri, Phys. Rev. B **92**, 054307 (2015).

[50] J. Ma, A. S. Nissimagoudar, and W. Li, Phys. Rev. B **97**, 045201 (2018).

[51] N. H. Protik and D. A. Broido, Phys. Rev. B **101**, 075202 (2020).

[52] H. Fröhlich, Adv. Phys. **3**, 325 (1954).

[53] R. Clady, M. J. Y. Tayebjee, P. Aliberti, D. König, N. J. Ekins-Daukes, G. J. Conibeer, T. W. Schmidt, and M. A. Green, Prog. Photovolt.: Res. Appl. **20**, 82–92 (2011).

[54] V. T. Luan, R. Chinomona, and D. R. Reynolds, SIAM J. Sci. Comput. **42**, A1245 (2020).

[55] V. T. Luan, R. Chinomona, and D. R. Reynolds, SIAM J. Sci. Comput. **44**, A3265 (2022).

[56] M. Sofroniou and G. Spaletta, Math. Comput. Model. **40**, 1157 (2004).

[57] P. Bogacki and L. F. Shampine, Appl. Math. Lett. **2**, 321 (1989).

[58] A. C. Fish, D. R. Reynolds, and S. B. Roberts, J. Comput. Appl. Math. **438**, 115534 (2024).

[59] H.-Y. Chen, D. Sangalli, and M. Bernardi, Phys. Rev. Res. **4**, 043203 (2022).

[60] E. Perfetto, K. Wu, and G. Stefanucci, npj 2D Mater. Appl. **8**, 40 (2024).

Chapter 3

DYNAMIC MODE DECOMPOSITION OF NONEQUILIBRIUM ELECTRON-PHONON DYNAMICS

This chapter is adapted from the published work:

I. Maliyov, J. Yin, J. Yao, C. Yang, and M. Bernardi, npj Comput. Mater. **10**, 123 (2024),
J.Y. participated in the computational research and manuscript preparation.

3.1 Introduction

Data-driven techniques are increasingly employed in materials modeling, both for accelerating computational workflows and to gain physical insight using learning algorithms [1, 2]. In particular, dynamic mode decomposition (DMD), which was developed in the last decade to study fluid dynamics, is a valuable tool to linearize dynamical problems and reduce their dimensionality [3, 4]. In DMD, explicit simulation of a short initial time window allows one to learn the dominant modes governing the dynamics and extrapolate the simulation to future times at low computational cost. Recent work has employed DMD to study electron dynamics described by model Hamiltonians with purely electronic interactions [5, 6]. Yet, to date DMD has not been applied to more computationally intensive first-principles studies.

We combine DMD with first-principles calculations of nonequilibrium electron dynamics, using the framework of the rt-BTE in the presence of e-ph collisions and external fields. Following an initial excitation, the rt-BTE is propagated in time to reach thermal equilibrium or steady state in an external field. However, evaluating the scattering integral at each time step makes the rt-BTE approach computationally demanding even for materials with a handful of atoms in the unit cell.

We show that DMD provides an order-of-magnitude computational speed-up while retaining the full accuracy of the first-principles rt-BTE. In addition, DMD reveals key momentum-space temporal patterns and achieves a significant dimensionality reduction of the nonequilibrium physics. Our results include both high-field transport and transient excited-state dynamics, and are accompanied by a careful

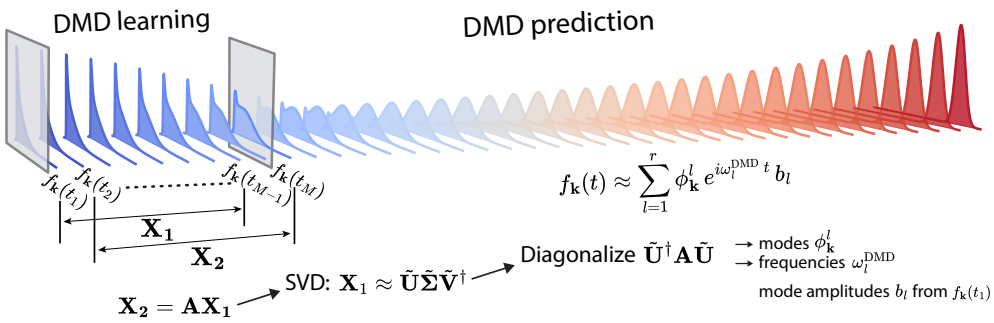

Figure 3.1: **Workflow of DMD plus rt-BTE calculations.** The first M steps of the dynamics, which make up the sampling window for DMD learning, are simulated by solving the rt-BTE. The resulting populations $f_{\mathbf{k}}(t)$ are stacked in the $\mathbf{X}_1$ and $\mathbf{X}_2$ matrices with a relative shift of one time step. The dynamics at later times $t > t_M$ is predicted with DMD using the r leading modes obtained by SVD of the matrix $\mathbf{X}_1$ and diagonalization of the matrix $\tilde{\mathbf{A}} = \tilde{\mathbf{U}}^{\dagger}\mathbf{A}\tilde{\mathbf{U}}$.

characterization of convergence with respect to the size of the sampling window during which DMD learns the dominant modes.

3.2 Methods

First-principles rt-BTE

We describe the electron distribution using the time-dependent populations $f_{n\mathbf{k}}(t)$, which quantify the occupation of each electronic state $|n\mathbf{k}\rangle$, where $\mathbf{k}$ is the electron crystal momentum and n is the band index (from now on we omit the band index to simplify the notation). Starting from an initial distribution at time zero, $f_{\mathbf{k}}(t = 0)$, in the rt-BTE the populations evolve according to [7]

$$\frac{\partial f_{\mathbf{k}}(t)}{\partial t} = -\frac{\mathbf{F}}{\hbar} \cdot \nabla_{\mathbf{k}} f_{\mathbf{k}}(t) + \mathcal{I}\left[f_{\mathbf{k}}(t)\right], \tag{3.1}$$

where $\mathcal{I}\left[f_{\mathbf{k}}(t)\right]$ is the collision integral accounting for e-ph scattering processes in momentum space and $\mathbf{F}$ includes any external fields applied to the system.

The rt-BTE simulations use dense momentum grids to accurately describe scattering between electronic states via absorption and emission of phonons. The required grid sizes are typically greater than 100×100×100 for both electron and phonon momenta. We time-step equation 3.1 using explicit solvers (Euler or 4th-order Runge-Kutta) or more advanced Strang splitting techniques [8].

DMD learning and prediction of the dynamics

We employ DMD in combination with rt-BTE simulations. The DMD approach linearizes the dynamics by relating the states of the system at times t and $t + \Delta t$ via a time-independent matrix $\mathbf{A}$ [9, 10]. Focusing on the e-ph dynamics, this amounts to advancing the electronic populations at time t using

$$f_{\mathbf{k}}(t + \Delta t) = \mathbf{A} f_{\mathbf{k}}(t), \tag{3.2}$$

where the populations $f_{\mathbf{k}}$ form a vector with size N equal to the number of $\mathbf{k}$-points in the electronic momentum grid (typically, $N \approx 10^5 - 10^6$). To obtain the matrix $\mathbf{A}$, we time-step the rt-BTE in a sampling window consisting of M time steps and then we form two matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ relating the populations at times t and $t + \Delta t$. The populations $f_{\mathbf{k}}(t)$ from t_1 to t_{M-1} are stacked column-wise in the matrix $\mathbf{X}_1$, with column i corresponding to time t_i and containing the populations $f_{\mathbf{k}}(t_i)$ for all $\mathbf{k}$-points and bands. The populations from t_2 to t_M are similarly stacked column-wise in the second matrix $\mathbf{X}_2$.

According to equation 3.2, these matrices are related by $\mathbf{X}_2 = \mathbf{A}\mathbf{X}_1$, but computing $\mathbf{A}$ naively from the pseudoinverse of $\mathbf{X}_1$ has a prohobitive cost due to the large size N of the $\mathbf{k}$-point grid. To circumvent this problem, in DMD one first performs a truncated singular value decomposition (SVD) [11, 12] of the $\mathbf{X}_1$ matrix:

$$\mathbf{X}_1 = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\dagger, \tag{3.3}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{N\times(M-1)}$ is a matrix with diagonal entries equal to the singular values σ_j arranged in decreasing order, while $\mathbf{U} \in \mathbb{C}^{N\times N}$ and $\mathbf{V} \in \mathbb{C}^{(M-1)\times(M-1)}$ are matrices collecting the mutually orthogonal singular vectors [13]. (Above, $\mathbf{V}^\dagger$ indicates the Hermitian conjugate of $\mathbf{V}$.)

Because $\mathbf{X}_1$ contains the time-dependent populations, this SVD procedure can single out the main patterns in the momentum-space dynamics. Here, we keep only the first r singular values (typically, $r \approx 10$) to restrict the solution space to the leading r momentum-space modes, and then project the matrix $\mathbf{A}$ onto this reduced r-dimensional space. This procedure provides the matrix $\tilde{\mathbf{A}}$, with reduced size $r \times r$, which can be diagonalized straightforwardly to obtain the dominant DMD modes. Using this procedure, the populations at future times $t > t_M$ are *predicted* – that is, obtained without explicit solution of the rt-BTE – using

$$f_{\mathbf{k}}(t > t_M) \approx \sum_{l=1}^{r} b_l \, \phi_{\mathbf{k}}^l \, e^{i\omega_l^{\mathrm{DMD}} t}, \tag{3.4}$$

where $\phi^l_{\mathbf{k}}$ are the momentum-space DMD modes obtained from the matrix $\tilde{\mathbf{A}}$, and ω^{DMD}_l and b_l are their frequencies and amplitudes.

We summarize the main steps of this DMD procedure, which are illustrated in Fig. 3.1:

1. Simulate the rt-BTE dynamics for the first M steps and construct the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$;
2. Perform SVD on $\mathbf{X}_1$ to find the matrix $\tilde{\mathbf{A}}$ in the reduced r-dimensional space;
3. Diagonalize $\tilde{\mathbf{A}}$ to find the DMD modes $\phi^l_{\mathbf{k}}$ and their frequencies ω^{DMD}_l, with $l = 1 \dots r$;
4. Obtain the mode amplitudes b_l from the initial condition $f_{\mathbf{k}}(t_1)$;
5. Predict the dynamics for $t > t_M$ using equation 3.4.

A key parameter is the duration of the sampling window (t_M) required for accurate DMD extrapolation of the dynamics beyond t_M. As the computational cost of DMD is negligible, the size of the sampling window, during which the rt-BTE is solved by explicit time-stepping, determines the computational cost of the entire workflow.

Computing DMD modes and frequencies

Let us describe in more detail the calculation of DMD modes and frequencies. We start from the snapshots $f_{\mathbf{k}}(t)$ evaluated explicitly with the rt-BTE in the sampling window $t_1 < t < t_M$, and then apply the SVD procedure to the matrix $\mathbf{X}_1$ (see Eq. 3.3). As shown in Fig. 3.2b, we find that the singular values σ_j decay rapidly. Keeping only the largest $r \approx 10$ singular values, we write the SVD of $\mathbf{X}_1$ as

$$\mathbf{X}_1 \approx \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^{\dagger}, \tag{3.5}$$

where we defined the economy-sized matrices in the r-dimensional subspace [13] as $\tilde{\mathbf{\Sigma}} = \mathbf{\Sigma}(1:r, 1:r)$, $\tilde{\mathbf{U}} = \mathbf{U}(1:N, 1:r)$, $\tilde{\mathbf{V}} = \mathbf{V}(1:M-1, 1:r)$. This way, the approximate pseudo-inverse of the matrix $\mathbf{X}_1$, denoted as $\mathbf{X}_1{}^{+}$, can be obtained with little effort as $\tilde{\mathbf{V}}\tilde{\mathbf{\Sigma}}^{-1}\tilde{\mathbf{U}}^{\dagger}$. Then the matrix $\mathbf{A}$ relating the snapshot matrices via $\mathbf{X}_2 = \mathbf{A}\mathbf{X}_1$ can be written as

$$\mathbf{A} = \mathbf{X}_2\mathbf{X}_1{}^{+} = \mathbf{X}_2\tilde{\mathbf{V}}\tilde{\mathbf{\Sigma}}^{-1}\tilde{\mathbf{U}}^{\dagger}. \tag{3.6}$$

Note that the matrix $\mathbf{A}$ depends on the sampling window. Due to its large $N \times N$ size (here, $N \approx 10^5$ is the number of $\mathbf{k}$-points), diagonalizing $\mathbf{A}$ is computationally expensive. In DMD, a key step is rewriting this matrix in the reduced r-dimensional space:

$$\tilde{\mathbf{A}} = \tilde{\mathbf{U}}^{\dagger}\mathbf{A}\tilde{\mathbf{U}} = \tilde{\mathbf{U}}^{\dagger}\mathbf{X}_2\tilde{\mathbf{V}}\tilde{\mathbf{\Sigma}}^{-1}, \tag{3.7}$$

allowing for straightforward eigenvalue decomposition:

$$\tilde{\mathbf{A}}\mathbf{W} = \mathbf{W}\mathbf{\Lambda}, \tag{3.8}$$

where the matrix $\mathbf{W}$ contains the eigenvectors of $\tilde{\mathbf{A}}$ and the eigenvalues $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_l\}$ are common to both matrices $\tilde{\mathbf{A}}$ and $\mathbf{A}$ [14]. The DMD modes, stacked column-wise in the matrix $\mathbf{\Phi} = \begin{pmatrix}\boldsymbol{\phi}^1 & \boldsymbol{\phi}^2 \cdots \boldsymbol{\phi}^r\end{pmatrix} \in \mathbb{C}^{N\times r}$, can be obtained using [14]

$$\mathbf{\Phi} = \mathbf{X}_2\tilde{\mathbf{V}}\tilde{\mathbf{\Sigma}}^{-1}\mathbf{W}. \tag{3.9}$$

The DMD frequency of mode l is obtained from the corresponding eigenvalue λ_l using equation 3.8,

$$\omega_l^{\mathrm{DMD}} = -i\frac{\ln\lambda_l}{\Delta t}, \tag{3.10}$$

where Δt is the simulation time step. To circumvent the potential addition of a $2\pi m\, i, m \in \mathbb{Z}$ term due to $\ln\lambda_l$ computation, we evaluate the logarithm in the following way: $\ln\lambda_l = \ln|\lambda_l| + i\arg(\lambda_l)$, where we take the principal value of a complex argument defined in $(-\pi, \pi]$.

The mode amplitudes $\mathbf{b} = \begin{pmatrix}b_1 & b_2 \cdots b_r\end{pmatrix} \in \mathbb{C}^r$ are obtained from the initial condition. Setting $t = 0$ in equation 3.4, we get

$$f_{\mathbf{k}}(0) = \mathbf{\Phi}\mathbf{b}, \tag{3.11}$$

and thus the mode amplitude vector $\mathbf{b}$ is obtained from the pseudo-inverse of the DMD mode matrix $\mathbf{\Phi}$:

$$\mathbf{b} = \mathbf{\Phi}^{+} f_{\mathbf{k}}(0). \tag{3.12}$$

The pseudo-inverse of the matrix $\mathbf{\Phi}$ is computed using truncated SVD and has a negligible computational cost compared to SVD of the $\mathbf{X}_1$ matrix due to (N, r) dimensions of the matrix $\mathbf{\Phi}$.

This approach provides the DMD modes $\phi_{\mathbf{k}}^l$, frequencies ω_l^{DMD}, and mode amplitudes b_l, and thus all the quantities needed for DMD prediction of the dynamics outside the sampling window ($t > t_M$) using equation 3.4.

Electron-phonon scattering from first principles

Our first-principles calculations of e-ph scattering employ an established workflow, which is summarized here and described in more detail in Ref. [7]. The electronic wave functions and band energies are obtained the same way as in Chapter 2. The electronic quasiparticle band structure is refined using GW calculations carried out with the YAMBO code [15]. This step improves the agreement with experiment of the electron effective masses and relative valley energies, which are essential for precise calculations of high-field dynamics [8] and excited electron relaxation [16]. The phonon dispersion, e-ph interactions and the Wannier-Fourier interpolation are computed the same way as in Chapter 2. Dirac delta functions expressing energy conservation are implemented as Gaussians with a small (~5 meV) broadening.

3.3 Results

High-field electron dynamics

We employ our DMD-based approach to simulate time-domain electron dynamics in an applied electric field in the presence of e-ph collisions. We recently demonstrated similar calculations using the rt-BTE without the aid of data-driven techniques [8]. Here we use this case study to explore the accuracy and efficiency of our rt-BTE plus DMD approach as well as find optimal values for the sampling window and analyze the momentum-space DMD modes. Our calculations focus on electrons in GaAs, where the conduction band has three sets of low-energy valleys, at Γ and near L and X in order of increasing energy [17] (see the inset in Fig. 3.2a). Upon applying an electric field, the electrons are accelerated to higher band energies while they also transfer part of that excess energy to the lattice via e-ph collisions. These competing mechanisms lead to a steady-state electronic distribution which is typically reached on a picosecond to nanosecond time scale.

Our simulations begin with electrons in thermal equilibrium with the lattice at 300 K. We apply a constant electric field $\mathbf{E}$ and time step the electron populations until they reach the steady state distribution, $f_{\mathbf{k}}^{\mathbf{E}}$, from which we compute the mean drift velocity, $v(\mathbf{E})$, a quantity routinely measured in experiments [18–20]. Repeating this procedure for multiple field values allows us to construct the full drift velocity versus electric field curve in a material, starting from linear response at low field to velocity saturation at high field [8].

Figure 3.2a shows the time-dependent populations in four regions of the Brillouin zone following the application of a high field (5 kVcm^{-1}). Electrons initially oc-

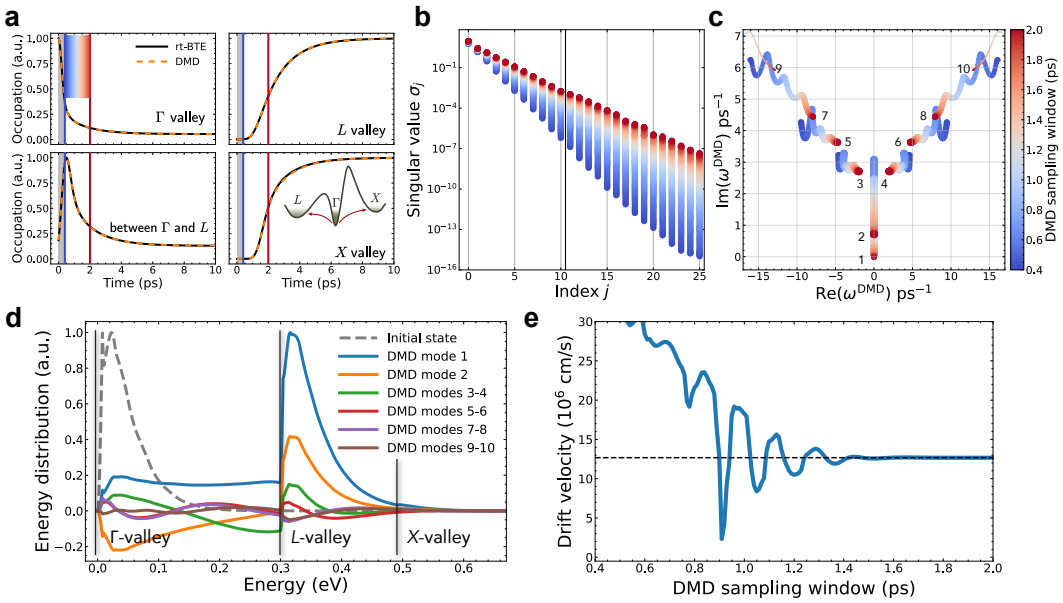

Figure 3.2: **DMD simulations of electrons in GaAs in an applied electric field**. **a**, Time-dependent electron populations $f_{\mathbf{k}}(t)$ for four electron momenta. The gray region indicates the shortest sampling window (0.4 ps) and the red vertical line the longest sampling window we tested (2 ps). Solid black lines show the rt-BTE results and orange dashed lines the DMD predictions obtained using the longest sampling window. The inset is a schematic of the low-energy band structure of GaAs showing the Γ- and higher energy L- and X-valleys. **b**, Singular values of the $\mathbf{X}_1$ matrix, with the ten largest singular values used in our DMD calculations separated by a vertical line. **c**, DMD frequencies plotted in the complex plane and shown as circles with radii proportional to the DMD mode amplitudes b_l. In panels **a**, **b**, and **c**, the colors indicate the duration of the DMD sampling window according to the legend given in (**c**). **d**, DMD momentum-space modes $\phi^l_{\mathbf{k}}$, multiplied by the corresponding amplitudes b_l, given as a function of energy. The initial state $f_{\mathbf{k}}(t_1)$ is shown with a dashed line. **e**, Convergence of the steady-state drift velocity with respect to duration of the DMD sampling window. The rt-BTE value is shown for reference as a dashed line.

cupying the Γ-valley scatter to the higher-energy L- and X-valleys. As a result, the electron populations in the Γ-valley decrease, with a corresponding increase in L- and X-valley populations. In regions of momentum space between the Γ- and L-valleys the populations peak at intermediate times and then relax to lower values.

This dynamics is nontrivial because the populations evolve differently in different momentum-space regions, making accurate predictions challenging. Our DMD approach can learn the dominant modes governing this intricate dynamics and extrapolate the time-dependent populations well beyond the sampling window. Re-

markably, we find that a short sampling window – 400 fs to 2 ps out of a total simulation time of 12.5 ps – is sufficient to extrapolate the dynamics all the way to steady state, with rt-BTE and DMD trajectories in nearly exact agreement outside the sampling window (Fig. 3.2a). This accuracy extends to the entire set of $\sim 10^5$ **k**-points considered in our simulations, providing carrier number conservation within 1% error.

The ability to learn key temporal momentum-space patterns is a consequence of the relatively rapid decay of the singular values of the $\mathbf{X}_1$ matrix used for learning the dynamics in the sampling window (Fig. 3.2b). This decay becomes slower as the sampling window increases, but it remains significant even for the longest sampling window of 2 ps used here (note the log scale in the plot). In turn, the singular value decay enables a striking dimensionality reduction, with DMD employing only $r \approx 10$ modes to solve the dynamics as opposed to 10^5 populations $f_{\mathbf{k}}$ and billions of e-ph scattering terms in the rt-BTE.

The choice of an ideal sampling window can rely on the appearance of specific DMD modes at steady state. Figure 3.2c shows the DMD mode frequencies ω^{DMD} in the complex plane, where the imaginary part of ω^{DMD} corresponds to the decay rate of a given mode and the real part gives its oscillation frequency. The populations $f_{\mathbf{k}}(t)$ are real-valued and are written as a summation of complex exponentials in equation 3.4. Therefore, physically meaningful results are possible only when $\mathrm{Re}(\omega^{\mathrm{DMD}}) = 0$ (modes 1, 2) or when ω^{DMD} appear as complex conjugate pairs (modes 3 – 10). Describing the steady state is particularly important in our simulations. In DMD, all modes with a non-zero imaginary frequency vanish in the long time limit, with only one mode surviving at steady state (mode 1 in Fig. 3.2c). As the sampling window increases, the imaginary frequency of this mode goes to zero, providing the correct steady state behavior. This analysis allows us to find the minimal sampling window required for accurate steady-state results by monitoring the zero-frequency mode.

The DMD eigenvector of the zero-frequency mode (mode 1 in Fig. 3.2d) determines the steady-state electron distribution $b_l \phi^1_{\mathbf{k}} = f_{\mathbf{k}}(t \to \infty)$, while the other modes control the transient dynamics. For example, mode 2 governs electron scattering from the Γ- to the L- and X-valleys, and higher modes appearing as conjugate pairs exhibit oscillating trends in energy (modes 3–10 in Fig. 3.2d). Converging the zero-frequency mode allows us to compute the steady-state drift velocity more efficiently. Figure 3.2e shows that a sampling window of 1.7 ps (170 snapshots) provides a drift

velocity nearly identical to the full rt-BTE calculation, which requires much longer simulation times of up to 12.5 ps (1250 snapshots). On this basis we conclude that DMD needs only ∼10% of the dynamics data for accurate steady-state predictions.

Velocity-field curves

We also employ DMD to accelerate calculations of entire velocity-field curves. This requires the drift velocity for a set of electric field values, and thus we adopt a modified workflow. Following our recent work [8], we gradually increase the electric field (black curve in Fig. 3.3a) and use the steady-state populations for a given field, $f^{\mathbf{E}}_{\mathbf{k}}$, as the initial condition for the next field value, $E + \Delta E$, where the field increment ΔE is typically $100 - 200$ Vcm^{-1}. As the applied field increases, the DMD frequencies and momentum-space modes change substantially. Therefore, for each new field value we repeat DMD learning in the initial stage of the simulation (see the DMD sampling regions shown as red rectangles in Fig. 3.3a). We then predict the steady-state populations using mode 1 from DMD, $f^{\mathbf{E}}_{\mathbf{k}} = b_1 \phi^1_{\mathbf{k}}$, and the drift velocity for that field value, and use $f^{\mathbf{E}}_{\mathbf{k}}$ as the initial condition for the next field value.

The velocity-field curves obtained with this approach are shown in Fig. 3.3b for GaAs and graphene and compared with rt-BTE results obtained without DMD. Using DMD lowers significantly the computational cost to obtain the full velocity-field curves, by a factor of 10.5 for GaAs and ∼16.5 for graphene, while fully preserving the accuracy. Because the drift velocity is computed as a weighted sum of $f^{\mathbf{E}}_{\mathbf{k}}$ [8], the nearly exact agreement between the DMD and full rt-BTE results demonstrates the accuracy of the DMD populations in momentum space. The DMD efficiency is a consequence of its ability to capture the dominant modes in the population dynamics using only a small number of snapshots, with a similar accuracy regardless of the electric field value. Our strategy of gradually increasing the electric field leads to an easier-to-extrapolate dynamics compared to the abrupt application of a strong field.

Excited electron relaxation

Next, we consider a different nonequilibrium dynamics where the material is initially prepared in an excited electronic state. This setting can be used, for example, to model the effect of an optical excitation with a laser pulse [21]. Different from the high-field dynamics, in this case the long-time limit is known and we are primarily interested in the *transient* dynamics. Following the initial excitation,

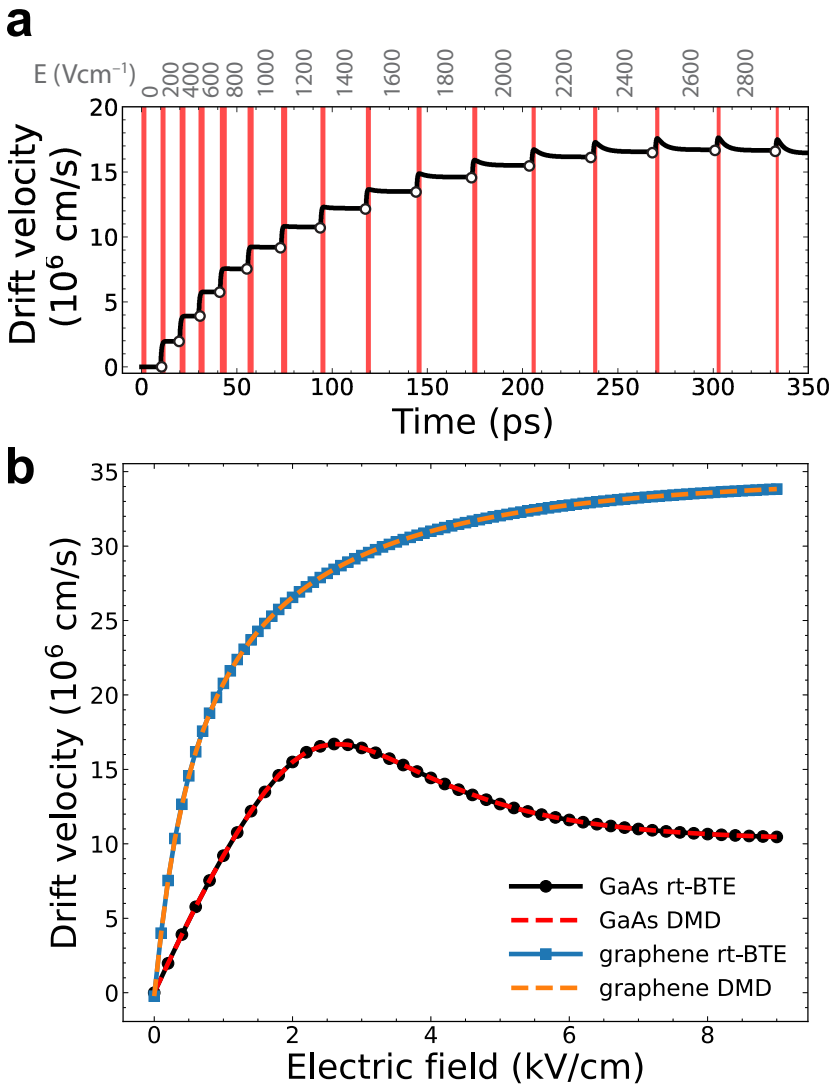

Figure 3.3: **Velocity-field curves from DMD. a**, Transient drift velocity in GaAs computed as a function of time (black curve). The external electric field is increased step-wise in the simulation (see the field values given above the plot). The DMD sampling window for each electric field is shown with a red rectangle, and the drift velocities outside this window are predicted with DMD. The steady-state drift velocities from DMD correspond to the plateaus for each field value, and agree with the reference drift velocities, obtained by explicitly time-stepping the rt-BTE until steady state, which are shown with white dots. **b**, Velocity-field curves in GaAs and graphene obtained from the rt-BTE (black and blue solid lines) and by combining the rt-BTE and DMD (red and orange dashed lines).

in the presence of e-ph interactions and without any external fields, the electrons relax to a thermal equilibrium Fermi-Dirac distribution [22], $f_{\mathbf{k}}^{\mathrm{FD}}$, typically on a sub-picosecond time scale. This ultrafast dynamics can be modeled by time-stepping the rt-BTE until reaching the equilibrium Fermi-Dirac distribution. Using this approach, our previous work has shown that electrons relax to the band edge significantly slower than holes in GaN semiconductor, with implications for optoelectronic devices [16].

Following that work, we model an excited state in GaN by placing the electrons ~1 eV above the conduction band edge, and then obtain the time-dependent electron populations by solving the rt-BTE (see Fig. 3.4a,b). We employ DMD to predict this transient dynamics, and find large errors when using a short time window of up to ~50 fs (solid red line in Fig. 3.4c). The correct steady state and transient dynamics are obtained by increasing the sampling window to 200 fs (dashed orange line in Fig. 3.4c). Our analysis of the DMD frequencies shows that the zero-frequency mode describing thermal equilibrium in the long-time limit appears when the sampling window reaches 100 fs (see the arrow in Fig. 3.4d) and fully converges for a ~200 fs sampling window. The need for such a long sampling window relative to the total duration of the dynamics (400 fs) makes DMD ineffective.

To address this issue and more efficiently study transient dynamics with DMD, we formulate a different learning procedure that incorporates knowledge of the equilibrium state. We focus on the difference between the transient and equilibrium populations, $\delta f_{\mathbf{k}}(t) = f_{\mathbf{k}}(t) - f_{\mathbf{k}}^{\mathrm{FD}}$, as opposed to just $f_{\mathbf{k}}(t)$ as we did in the high-field example. After predicting $\delta f_{\mathbf{k}}$(t) with DMD, we obtain the time-dependent populations $f_{\mathbf{k}}(t)$ by adding back the $f_{\mathbf{k}}^{\mathrm{FD}}$ term. As $\delta f_{\mathbf{k}}$ vanishes in the long-time limit (Fig. 3.4b), the zero-frequency DMD mode is missing when computing $\delta f_{\mathbf{k}}$ (Fig. 3.4e); all other DMD frequencies associated with $\delta f_{\mathbf{k}}$ are similar to those for $f_{\mathbf{k}}(t)$ (Fig. 3.4d-e). We find that the DMD method based on $\delta f_{\mathbf{k}}$ is far more effective and requires a significantly shorter sampling window for accurate DMD predictions – using a 50 fs sampling window, we achieve results similar to DMD for $f_{\mathbf{k}}(t)$ with a four times longer (200 fs) window (Fig. 3.4c).

With this improved DMD approach, using a sampling window of only ~12% of the total simulation time allows us to accurately predict the average electron relaxation rate in GaN, with a DMD computed value of 5.23 eVfs^{-1} in close agreement (within 0.8%) with the rt-BTE result. This result demonstrates that our DMD approach can predict excited electron relaxation with a high accuracy.

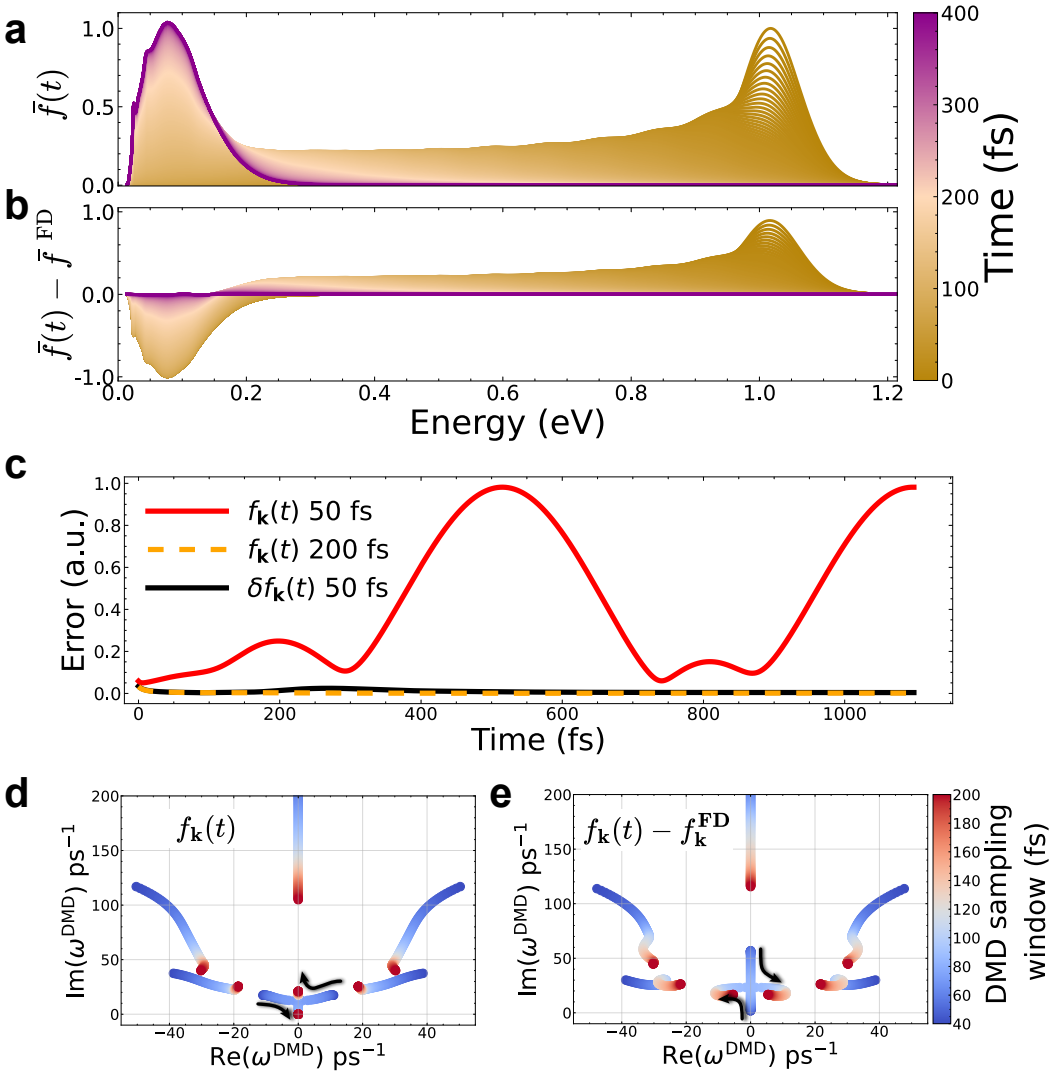

Figure 3.4: **Transient dynamics in GaN using DMD. a**, Energy dependence of the momentum-averaged electron populations, and **b**, difference between the time-dependent and equilibrium populations in GaN. In both panels, the simulation time is color-coded using sepia for the initial excited state and purple for the equilibrium state. The energy zero is set to the conduction band minimum. **c**, DMD error on the electron populations, computed as the root-mean-square difference between the reference values from rt-BTE and those obtained from DMD. Results are shown for different sampling windows, given in the legend, and for the two schemes where DMD is applied to $f_{\mathbf{k}}(t)$ or alternatively to $\delta f_{\mathbf{k}}(t) = f_{\mathbf{k}}(t) - f_{\mathbf{k}}^{\mathrm{FD}}$. **d**, **e**, DMD frequencies on the complex plane, respectively for $f_{\mathbf{k}}(t)$ and $f_{\mathbf{k}}(t) - f_{\mathbf{k}}^{\mathrm{FD}}$, with the duration of the sampling window color-coded.

3.4 Discussion

The DMD approach introduced here is very efficient: the rt-BTE is solved explicitly on a high-performance computer only for a small number of initial time steps, after which the entire dynamics can be computed straightforwardly with DMD, using only a laptop. The most demanding step is carrying out truncated SVD on the $\mathbf{X}_1$ matrix, but for comparison this step requires lower computational resources than even just a single rt-BTE time step.

This remarkable speed-up is achieved by reducing the dimensionality of the rt-BTE dynamics and is linked to the shape of the $\mathbf{X}_1$ matrix. The rt-BTE employs a large number of $\mathbf{k}$-points (about $10^5 - 10^6$), which equals the number of rows of the matrix $\mathbf{X}_1$, and a significantly smaller number of snapshots in the DMD sampling window, typically ~100 time steps, which sets the number of columns in $\mathbf{X}_1$. Following truncated SVD, the size of the problem is reduced to (at most) the number of snapshots and is typically of order 50−100, and thus smaller by orders of magnitude compared to the original rt-BTE. (Note that one could use the entire set of singular values, but here we prefer using only ~10 singular values to prevent numerical instabilities [13]).

This efficiency allows us to evaluate the accuracy of DMD on the fly, halting explicit time-stepping of the rt-BTE when the steady state or transient dynamics are fully converged. In addition, our approach addresses the key challenge of storing the rt-BTE populations, which are needed only in the sampling window in DMD, as opposed to the full dynamics. This is a critical improvement because in conventional rt-BTE simulations one needs to store the populations $f_{\mathbf{k}}(t)$ on dense momentum grids for thousands of time steps, resulting in terabytes of data. In contrast, after carrying out SVD in the sampling window, DMD stores only a handful of complex frequencies and momentum-space modes, using which the dynamics can be reconstructed for the entire simulation.

3.5 Conclusion

In summary, we have introduced a data-driven approach based on DMD to accelerate first-principles calculations of nonequilibrium electron dynamics in materials. Our method speeds-up the solution of the time-dependent Boltzmann equation with electron collisions computed from first principles. We have shown that DMD can capture dominant modes governing the microscopic dynamics, enabling accurate predictions of the steady-state properties such as the drift velocity as well as transient

processes such as electron relaxation and equilibration. In both steady-state and transient nonequilibrium calculations, DMD requires explicit time-stepping of the rt-BTE in a time window of only ∼10% of the full simulation, after which the dynamics is extrapolated from the DMD modes with negligible computational cost. This DMD workflow preserves the accuracy while requiring far more modest computational resources than full rt-BTE simulations.

These advances are broadly relevant to studying nonequilibrium quantum dynamics of elementary excitations. For example, in future work, our data-driven approach will be extended to study the coupled dynamics of electrons and phonons, which involves fast (electron) and slow (phonon) timescales. The current DMD approach is not designed to address such multiscale nonequilibrium dynamics, and extensions using multiresolution DMD will be explored.

References

[1] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, Npj Comput. Mater. **3** (2017).

[2] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, Npj Comput. Mater. **5** (2019).

[3] P. J. Schmid, J. Fluid Mech. **656**, 5 (2010).

[4] J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor, *Dynamic mode decomposition: data-driven modeling of complex systems* (SIAM, 2016).

[5] J. Yin, Y. hao Chan, F. H. da Jornada, D. Y. Qiu, C. Yang, and S. G. Louie, J. Comput. Phys. **477**, 111909 (2023).

[6] C. C. Reeves, J. Yin, Y. Zhu, K. Z. Ibrahim, C. Yang, and V. Vlček, Phys. Rev. B **107** (2023).

[7] J.-J. Zhou, J. Park, I.-T. Lu, I. Maliyov, X. Tong, and M. Bernardi, Comput. Phys. Commun. **264**, 107970 (2021).

[8] I. Maliyov, J. Park, and M. Bernardi, Phys. Rev. B **104**, L100303 (2021).

[9] B. O. Koopman, Proc. Natl. Acad. Sci. USA **17**, 315 (1931).

[10] C. W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson, J. Fluid Mech. **641**, 115 (2009).

[11] G. Golub and W. Kahan, SIAM J. Numer. Anal. **2**, 205 (1965).

[12] G. H. Golub and C. F. Van Loan, *Matrix Computations* (JHU press, 2013).

[13] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control* (Cambridge University Press, 2022).

[14] J. H. Tu, , C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. K. and, J. Comput. Dyn. **1**, 391 (2014).

[15] D. Sangalli, A. Ferretti, H. Miranda, C. Attaccalite, I. Marri, E. Cannuccia, P. Melo, M. Marsili, F. Paleari, A. Marrazzo, G. Prandini, P. Bonfà, M. O. Atambo, F. Affinito, M. Palummo, A. Molina-Sánchez, C. Hogan, M. Grüning, D. Varsano, and A. Marini, J. Phys.: Condens. Matter **31**, 325902 (2019).

[16] V. A. Jhalani, J.-J. Zhou, and M. Bernardi, Nano Lett. **17**, 5012 (2017).

[17] Y. Peter and M. Cardona, *Fundamentals of Semiconductors: Physics and Materials Properties* (Springer Science & Business Media, 2010).

[18] C. Canali, G. Ottaviani, and A. A. Quaranta, J. Phys. Chem. Solids **32**, 1707 (1971).

[19] K. Ashida, M. Inoue, J. Shirafuji, and Y. Inuishi, J. Phys. Soc. Japan **37**, 408 (1974).

[20] D. Ferry, Phys. Rev. B **12**, 2361 (1975).

[21] M. Bernardi, D. Vigil-Fowler, J. Lischner, J. B. Neaton, and S. G. Louie, Phys. Rev. Lett. **112**, 257402 (2014).

[22] D. K. Ferry, *Hot Carriers in Semiconductors*, 2053-2563 (IOP Publishing, 2021).

Chapter 4

EFFICIENT GPU PARALLELIZATION OF ELECTRONIC TRANSPORT AND NONEQUILIBRIUM DYNAMICS

This chapter is adapted from the following work:

S. Peng, D. Pinkston, J. Yao, S. Kliavinek, I. Maliyov, and M. Bernardi, Efficient GPU parallelization of electronic transport and nonequilibrium dynamics from electron-phonon interactions in the perturbo code (2025),
J.Y. contributed to algorithm design, simulation setup, and writing of the manuscript.

4.1 Introduction

In recent years, the combination of semiclassical Boltzmann transport equation (BTE) with first-principles e-ph interactions has enabled accurate predictions of electronic transport in metals [1, 2], inorganic and organic semiconductors [3–10], complex oxides [11, 12], and quantum materials [13–15].

For studies of ultrafast nonequilibrium dynamics, solving the rt-BTE provides a favorable balance between accuracy and computational cost [16, 17]. Despite the overall efficiency, the rt-BTE method still requires parallelization and extensive software optimization, particularly for simulations of materials with large unit cells, and/or using dense momentum grids and targeting long simulation times beyond the picosecond timescale. In the PERTURBO code, after identifying the relevant e-ph scattering processes, the collision integral – the key quantity in BTE calculations – is computed by looping over these scattering channels [18]. Even after selecting a relevant energy window and retaining only energy-conserving scattering channels, in a typical calculation, the total number of active scattering channels can still be as large as 10^8 or higher, which poses a major computational challenge for CPU hardware. Therefore, for both transport and time-domain dynamics, it is particularly important to design a data structure that addresses the high-dimensionality and sparsity of e-ph interactions and scattering processes [19].

Using graphic processing units (GPUs) [20], which are now prevalent and widely available, could be game changing for accelerating BTE calculations of transport and nonequilibrium dynamics, and potentially a broader range of e-ph physics. Unlike CPUs, which typically feature a limited number of high-performance cores

and are optimized to handle complex workloads, GPUs can process a large volume of lightweight tasks. Originally designed for processing images, in the last decade GPUs have greatly expanded their scope in scientific computing, becoming an increasingly popular option for high-performance tasks [21]. For example, on the Perlmutter cluster at the National Energy Research Scientific Computing Center (NERSC) [22], at present about 40% of the compute nodes are GPU nodes, each equipped with 28,000 CUDA cores. In contrast, each CPU node contains only 128 cores. Although each GPU core has lower computing power than a CPU core, GPUs can perform a large number of simple tasks with a high degree of parallelism.

However, two key considerations need to be addressed when designing algorithms for GPU execution. First, it is important to minimize data movement between the host and GPUs because it causes substantial overhead. Second, atomic operations must be optimized to avoid communication and synchronization between GPU cores, which causes significant performance loss [23]. This is particularly evident in a parallel CPU implementation of the BTE method, where different scattering channels contributing to the collision integral are typically handled by different processes or threads [18]. Therefore, we seek to design a data structure for e-ph scattering in the BTE that is optimal for use on GPUs.

Several programming frameworks are available for GPU algorithms, such as the widely used CUDA and OpenACC. Due to its low-level architecture, CUDA offers greater control and optimization, but at the cost of increased complexity for code implementation and maintenance. In contrast, OpenACC offers a directive-based approach that enhances code readability and maintainability, with only a slight performance loss. In addition, OpenACC is designed for portability across platforms and thus is not limited to GPUs from any specific vendor. These strengths led us to use OpenACC in this work.

Here, we design an efficient GPU data structure and algorithm for the BTE, and implement them with OpenACC in Perturbo, to accelerate calculations of transport and ultrafast dynamics using GPUs. Our new data structure optimizes data allocation and movement, as well as communication and synchronization between GPU cores. We show benchmarks for performance, memory consumption, and strong scaling in several materials. Our analysis shows a substantial performance improvement relative to the (already efficient) reference CPU implementation: we achieve a speed-up by ~40 times for BTE calculations of transport and nonequilibrium dynamics on GPUs, realizing nearly linear scaling up to 100 GPUs. This new implementation

was released in PERTURBO v3.0 in early 2025.

4.2 Methods

BTE, collision integrals, and scattering channels

The semiclassical BTE models the change in electronic occupations in response to external fields and collision processes. In a solid, these processes are naturally described in momentum space as in Eq. 3.1 in Chapter 3. Note that we assume slowly varying fields and homogeneous material and electronic occupations, which removes spatial derivatives. Under an external field $\boldsymbol{F}$ ($\boldsymbol{F} = -e\boldsymbol{E}$ for electrons in the presence of an electric field $\boldsymbol{E}$), the change in electron occupations is determined by the drift term (first term on the right-hand side) and the collision integral $\mathcal{I}$. The BTE is solved in the time domain for ultrafast dynamics and at steady state for transport (see Methods for details). Computing the collision integral is the main bottleneck in the algorithm for both ultrafast dynamics, where it is computed at each time step, and for transport calculations, where it is computed in each iteration step.

Using Fermi's golden rule, the collision integral $\mathcal{I}$ for nonequilibrium electron dynamics due to e-ph interactions is given by Eqs. 1.6, 1.7 and 1.8 in Chapter 1. Each absorption or emission process can be labeled using the notation $(\boldsymbol{k}, \boldsymbol{q}, n, m, \nu)$, here referred to as a scattering channel. The factors F_{em} and F_{abs} depend on the carrier occupations $f_{n\boldsymbol{k}}(t)$ and the phonon occupations $N_{\nu\boldsymbol{q}}$. In addition, $g_{mn\nu}(\boldsymbol{k}, \boldsymbol{q})$ are elements of the e-ph coupling matrix $\boldsymbol{g}(\boldsymbol{k}, \boldsymbol{q})$ computed from first principles, the δ functions enforce energy conservation, and crystal momentum conservation is satisfied by using commensurate electron and phonon grids (with $\mathcal{N}_{\mathbf{q}}$ grid points) and selecting appropriate $(\boldsymbol{k}, \boldsymbol{q})$ pairs for each scattering process. For convenience, in the following we denote the collision integral for state $|n, \boldsymbol{k}\rangle$ as $\mathcal{I}(n, \boldsymbol{k}) = \mathcal{I}^{e-\text{ph}}[f_{n\boldsymbol{k}}(t)]$. Although the expression for $\mathcal{I}(n, \boldsymbol{k})$ is different in transport calculations, the same data structure and algorithm apply to both transport and ultrafast dynamics, and are illustrated here for the ultrafast dynamics case.

Starting with an initial value of $f_{n\boldsymbol{k}}(t)$, the rt-BTE can be solved by explicit time-stepping, in our case using the fourth-order Runge-Kutta method. The collision integrals for each band and momentum are evaluated at each time step, with a typical simulation comprising 1,000−10,000 time steps. More advanced methods such as adaptive and multirate time integration have also recently been proposed for more efficient time stepping [19]. The choice of the initial electron occupation depends on the specific problem being studied, with common options including

Lorentzian, Fermi-Dirac, and Gaussian distributions in energy. We use Gaussian energy distributions to model the initial electronic occupations for ultrafast dynamics simulations of all materials studied in this work.

Electronic transport calculations

The BTE reaches a steady state when the time derivative $\partial f_{n\boldsymbol{k}}(t)/\partial t$ in Eq. 3.1 vanishes. At steady state, the drift term from the external field $\boldsymbol{E}$ is balanced by the e-ph collisions. For weak electric fields, the electron occupations $f_{n\boldsymbol{k}}$ can be expanded to first-order in the field as [18]

$$\begin{aligned} f_{n\boldsymbol{k}} &= f^0_{n\boldsymbol{k}} + f^1_{n\boldsymbol{k}} + O(E^2) \\ &= f^0_{n\boldsymbol{k}} + e\boldsymbol{E}\cdot\boldsymbol{F}_{n\boldsymbol{k}}\frac{\partial f^0_{n\boldsymbol{k}}}{\partial \epsilon_{n\boldsymbol{k}}} + O(E^2), \end{aligned} \tag{4.1}$$

where $f^0_{n\boldsymbol{k}}$ is the equilibrium Fermi-Dirac distribution and $\boldsymbol{F}_{n\boldsymbol{k}}$ characterizes the first-order deviation from the equilibrium distribution.

Substituting Eq. 4.1 into Eq. 3.1, we obtain an iterative approach to compute the deviation from equilibrium [18]:

$$\boldsymbol{F}^{i+1}_{n\boldsymbol{k}} = \boldsymbol{F}^0_{n\boldsymbol{k}} + \frac{\tau_{n\boldsymbol{k}}}{\mathcal{N}_{\boldsymbol{q}}}\sum_{m,\nu\boldsymbol{q}} \boldsymbol{F}^i_{m\boldsymbol{k}+\boldsymbol{q}} W^{\nu\boldsymbol{q}}_{n\boldsymbol{k},m\boldsymbol{k}+\boldsymbol{q}}, \tag{4.2}$$

from which the conductivity and other transport coefficients can be obtained. Above, the e-ph scattering rate is defined as

$$\begin{aligned} W^{\nu\boldsymbol{q}}_{n\boldsymbol{k},m\boldsymbol{k}+\boldsymbol{q}} &= \frac{2\pi}{\hbar}\left|g_{mn\nu}(\boldsymbol{k},\boldsymbol{q})\right|^2 \\ &\times \Big[\delta\Big(\epsilon_{n\boldsymbol{k}} - \hbar\omega_{\nu\boldsymbol{q}} - \epsilon_{m\boldsymbol{k}+\boldsymbol{q}}\Big)(1 + N_{\nu\boldsymbol{q}} - f_{m\boldsymbol{k}+\boldsymbol{q}}) \\ &+ \delta\Big(\epsilon_{n\boldsymbol{k}} + \hbar\omega_{\nu\boldsymbol{q}} - \epsilon_{m\boldsymbol{k}+\boldsymbol{q}}\Big)(N_{\nu\boldsymbol{q}} + f_{m\boldsymbol{k}+\boldsymbol{q}})\Big], \end{aligned} \tag{4.3}$$

and

$$\boldsymbol{F}^0_{n\boldsymbol{k}} = \tau_{n\boldsymbol{k}} v_{n\boldsymbol{k}} = \left(\frac{1}{N_{\boldsymbol{q}}}\sum_{m,\nu\boldsymbol{q}} W^{\nu\boldsymbol{q}}_{n\boldsymbol{k},m\boldsymbol{k}+\boldsymbol{q}}\right)^{-1} v_{n\boldsymbol{k}} \tag{4.4}$$

is the deviation from equilibrium in the relaxation time approximation, which is used as the initial guess in the iterative method.

The collision integral for transport calculations can be obtained from Eq. 4.2 as

$$\begin{aligned} \mathcal{I}(n,\boldsymbol{k},\alpha) = \frac{2\pi}{\hbar\mathcal{N}_{\boldsymbol{q}}}\tau_{n\boldsymbol{k}}\sum_{m\boldsymbol{q}\nu}|g_{mn\nu}(\boldsymbol{k},\boldsymbol{q})|^2 \times \boldsymbol{F}^i_{m\boldsymbol{k}+\boldsymbol{q},\alpha} \times \\ \Big[\delta\Big(\epsilon_{n\boldsymbol{k}} - \hbar\omega_{\nu\boldsymbol{q}} - \epsilon_{m\boldsymbol{k}+\boldsymbol{q}}\Big)(1 + N_{\nu\boldsymbol{q}} - f_{m\boldsymbol{k}+\boldsymbol{q}}) \\ + \delta\Big(\epsilon_{n\boldsymbol{k}} + \hbar\omega_{\nu\boldsymbol{q}} - \epsilon_{m\boldsymbol{k}+\boldsymbol{q}}\Big)(N_{\nu\boldsymbol{q}} + f_{m\boldsymbol{k}+\boldsymbol{q}})\Big], \end{aligned} \tag{4.5}$$

where α denotes the Cartesian directions of the applied electric field $\boldsymbol{E}$. For transport calculations using the iterative solution of the BTE, collision integrals are typically evaluated for 10−100 iterations to reach convergence, depending on the specific system and conditions.

Computing the collision integral $\mathcal{I}$ for all electronic states involves summing over all active scattering channels. The electronic structure and lattice dynamics are first obtained in the entire Brillouin zone. Then, we restrict the electronic states of interest to a given energy window, significantly reducing the total number of $(\boldsymbol{k}, \boldsymbol{q})$ pairs [18]. For each $(\boldsymbol{k}, \boldsymbol{q})$ pair, the nominal number of scattering channels prior to imposing any conservation constraints is $N_b^2 \times N_\nu$, where N_b is the number of included bands and N_ν is the number of phonon modes. By imposing an approximate energy conservation (with a Gaussian δ-function) and discarding channels with $|g_{mn\nu}(\boldsymbol{k}, \boldsymbol{q})|$ below a prescribed cutoff, the set of active scattering channels is further reduced to a small fraction of the total. This process makes $g_{nm\nu}(\boldsymbol{k}, \boldsymbol{q})$ highly sparse in the parameter space $(\boldsymbol{k}, \boldsymbol{q}, n, m, \nu)$ (see Supplementary Fig. 4.6), saving extensive memory and computational cost [24]. This scattering channel selection algorithm is implemented in PERTURBO v2.2.0 and earlier CPU-based versions [18], and achieves efficient performance and memory usage on CPUs.

However, the calculation of the collision integral $\mathcal{I}$ remains the most computationally demanding part of the BTE workflow and would greatly benefit from GPU acceleration. The CPU implementation is highly optimized using hybrid MPI and OpenMP parallelization but is not readily adapted to GPU parallelization for two main reasons. First, when computing the collision integral, the CPU implementation uses a large number of atomic operations to avoid competition (so-called "race conditions") between threads parallelized over scattering channels. Second, the use of numerous arrays of variable lengths to store information about the scattering channels $(\boldsymbol{k}, \boldsymbol{q}, n, m, \nu)$ requires referencing millions of individual heap allocations, introducing substantial overhead that limits GPU performance.

We propose a GPU-optimized data structure and algorithm that address these limitations and efficiently calculate the collision integral on GPUs. In the following, we describe this data structure and algorithm, and show benchmarks of performance, memory usage, and scaling behavior, using the CPU implementation of PERTURBO as a reference.

Benchmark setup and computational environment

We design benchmarks that use different momentum-grid sizes and computational resources. Performance tests are conducted on a single compute node and the strong-scaling analysis is performed on 4 to 64 nodes. The grid size for each material is chosen to fit on one node for performance tests and on four nodes for strong scaling tests. As the memory requirements for transport and ultrafast dynamics are different, we use different grid sizes for these two types of calculations.

For a fair comparison, all benchmark tests – including both CPU and GPU calculations – are performed on the heterogeneous GPU nodes of the Perlmutter cluster at NERSC. Each heterogeneous GPU node consists of one CPU (AMD EPYC 7763) with 64 cores and four GPUs (Nvidia A100 with 40GB) [22]. For transport and ultrafast dynamics simulations using the GPU-optimized code, only the collision integral, which is the most computationally expensive part in the solution of the BTE, is performed on GPUs, while the remaining part of the algorithm is executed on CPUs. Each CPU (GPU) calculation is repeated 10 (50) times independently, and we report the average wall-time to ensure statistical significance of the results. Finally, the accuracy of the GPU implementation was validated through integration tests, not discussed in this work, using the Python-based package PERTURBOPY [25].

First-principles calculations

The first-principles calculations and wannierization are performed following the same workflow as in previous chapters. The lattice dynamics and e-ph perturbation potentials are computed using DFPT on $8\times8\times8$, $18\times18\times1$, and $8\times8\times8$ $\boldsymbol{q}$-point grids for GaAs, graphene, and silicon, respectively. The momentum grids are chosen separately for different simulations and are given in the main text. The delta functions enforcing energy conservation during scattering processes are approximated using Gaussian functions with a 5 meV broadening. Relatively large energy windows – respectively, 0.7 eV for GaAs, 1.3 eV for graphene, and 0.55 eV for silicon – are used to generate grids that fill the computational capacity of one node in benchmark tests and 4 nodes in strong-scaling tests, as discussed above. For ultrafast dynamics simulations, the initial distributions of excited carriers are modeled as narrow Gaussian functions with broadening parameters of 20 meV for GaAs and graphene, and 10 meV for silicon. The fourth-order Runge-Kutta method is used to time-step the rt-BTE.

Data structure and implementation

Reference CPU algorithm

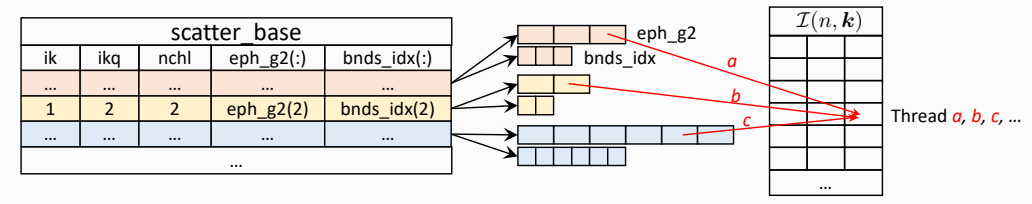

Figure 4.1: **Schematic of the `scatter_base` data structure.** The information for each $(\boldsymbol{k},\boldsymbol{q})$ pair is stored as a separate entry in `scatter_base`, as shown using different colors. The variables `ik`, `ikq`, `nchl` are indices of the $\boldsymbol{k}$ and $\boldsymbol{k}+\boldsymbol{q}$ points and the number of active scattering channels, respectively. The variables `eph_g2` and `bnds_idx` are arrays holding, respectively, the squared norm of the e-ph matrix elements and the joint indices of bands and phonon modes for each scattering channel. The relation of these variables to the collision integral $\mathcal{I}(n,\boldsymbol{k})$, whose components can be updated by multiple processes and threads simultaneously, is shown using red arrows.

As discussed above, the electronic bands, phonon modes, and momenta collectively label an active scattering channel between states $|n\boldsymbol{k}\rangle$ and $|m\boldsymbol{k}+\boldsymbol{q}\rangle$. The e-ph coupling matrix $\boldsymbol{g}$ is effectively sparse and irregular in this parameter space $(\boldsymbol{k},\boldsymbol{q},m,n,\nu)$. Consequently, using a single 5-dimensional array to store $g_{nm\nu}(\boldsymbol{k},\boldsymbol{q})$ leads to highly inefficient code as many of its entries are zero or very small (see Supplementary Fig. 4.6). In practice, we group together all active scattering channels involving the same $(\boldsymbol{k},\boldsymbol{q})$ pairs, and create an abstract type containing all relevant information for each pair. This design leverages benefits of object-oriented programming, such as flexibility and maintainability, while at the same time decoupling the $(\boldsymbol{k},\boldsymbol{q})$ pairs, which enables efficient distributed programming using MPI. Specifically, we define the object `scatter_base` to store all the relevant information for all the active scattering channels of each $(\boldsymbol{k},\boldsymbol{q})$ pair, thus filtering out all redundant e-ph and scattering channel data. A schematic of this data structure is shown in Fig. 4.1.

We use this implementation and data structure, available in PERTURBO v2.2.0 and earlier versions [18], as a reference or baseline for benchmarking code performance. This "Baseline-CPU" algorithm features hybrid MPI plus OpenMP CPU parallelization, where the $\boldsymbol{k}$ points are evenly distributed over different MPI processes. For ultrafast dynamics simulations, this code has two nested loops for each $\boldsymbol{k}$ point:

an outer loop over $(\boldsymbol{k}, \boldsymbol{q})$ pairs accelerated with OpenMP, and an inner loop over active scattering channels for the current $(\boldsymbol{k}, \boldsymbol{q})$ pair, executed sequentially by each OpenMP thread. For transport, there is an additional inner loop over Cartesian components of the external field. When the same collision integral $\mathcal{I}(n, \boldsymbol{k})$ is updated by multiple threads simultaneously, as in the red arrows in Fig. 4.1, the race condition between threads is avoided using OpenMP atomic operations.

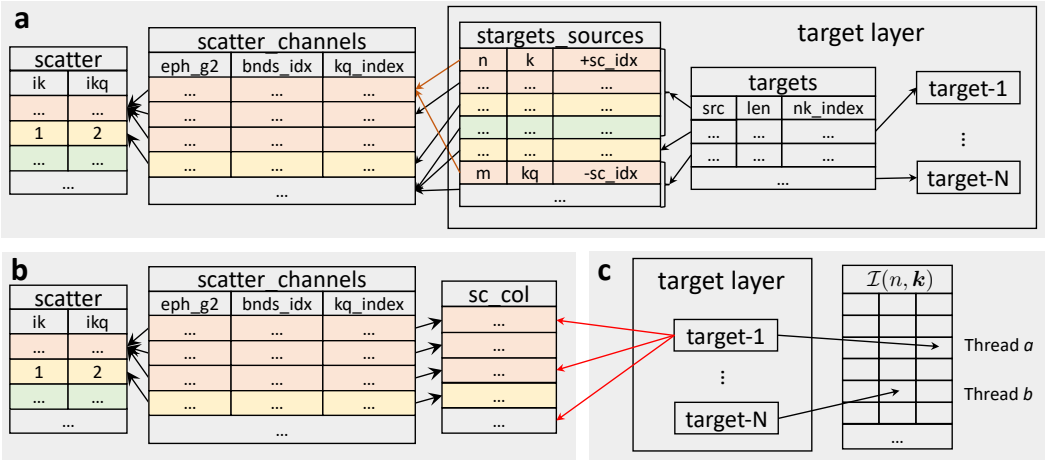

Figure 4.2: **Data structure optimized for GPUs.** **a** Setup of the scattering channels and target layer. Relevant quantities for the $(\boldsymbol{k}, \boldsymbol{q})$ pairs are stored into multiple arrays: `scatter`, which stores the indexes of the $\boldsymbol{k}$ and $\boldsymbol{k} + \boldsymbol{q}$ points, and `scatter_channels`, which stores information for all scattering channels, such as the square of the e-ph matrix elements (`eph_g2`), the joint indices of bands and phonon modes (`bnds_idx`), and the index of the $(\boldsymbol{k}, \boldsymbol{q})$ pair (`kq_index`). Scattering channels shown with the same color are associated with the same $(\boldsymbol{k}, \boldsymbol{q})$ pair. In addition, `stargets_sources` indexes the elements of the collision integral $\mathcal{I}(n, \boldsymbol{k})$ and the position of the scattering channels (`sc_idx`) in `scatter_channels`. The positive (negative) sign of `sc_idx` reflects how that entry contributes to the collision integral. The rows of `stargets_sources` are arranged in order, with rows sharing the same $(n, \boldsymbol{k})$ grouped together, as shown with curly braces. Each such group is called a target, and for each group, the position in `stargets_sources` (`src`), the length (`len`), and the combined $(n, \boldsymbol{k})$ index (`nk_index`) are stored in `targets`. Together, `stargets_sources` and `targets` constitute the target layer. **b** Calculation of the contribution to the collision integral from each scattering channel, defined in Eq. 1.6, which is computed and stored in `sc_col`. **c** Update of collision integrals $\mathcal{I}(n, \boldsymbol{k})$. Each element of $\mathcal{I}$ is updated by one target and one thread of execution. Using the target layer described in (**a**), each target is able to find the contribution of all the associated scattering channels in `sc_col`, as shown with red arrows.

Optimized GPU algorithm

A direct implementation of the Baseline-CPU algorithm on GPUs would be inherently inefficient as GPUs are optimized for executing many lightweight threads concurrently, and therefore perform poorly when frequent atomic updates are required. Moreover, OpenACC generates one data transfer per memory allocation, so allocating a large number of variable-length arrays incurs significant overhead. The combination of numerous heap allocations and frequent atomic operations across threads would severely limit the efficiency of a GPU version of the above algorithm.

To achieve GPU acceleration, we redesign the data structure and code implementation for the key step of the BTE algorithm, the calculation of the collision integral. In the optimized data structure, shown in Fig. 4.2, we allocate `scatter` and `scatter_channels`, which store the same information as `scatter_base` but using fixed-size buffers instead of variable-length arrays. This avoids dynamic GPU allocations and improves performance, while preserving the flexibility of object-oriented programming. In addition, to eliminate atomic updates on the GPU, we develop an algorithm that inverts the accumulation scheme: rather than assigning multiple threads to update the contribution of each scattering channel to one collision integral $\mathcal{I}(n, \boldsymbol{k})$, it distributes threads over $\mathcal{I}(n, \boldsymbol{k})$ itself. Each thread then identifies the scattering channels that contribute to its assigned $\mathcal{I}(n, \boldsymbol{k})$. The calculation of the collision integral with the optimized GPU algorithm and data structure consists of three steps:

1. *Create the target layer:*
   In the first step, the scattering channels are traversed to determine to which elements of the collision integral $\mathcal{I}(n, \boldsymbol{k})$ they contribute. Channels contributing to the same element of $\mathcal{I}$ are grouped together and form a `target`. This grouping is handled in the target layer (Fig. 4.2a), where each scattering channel of `scatter_channels` contributes to two different collision integrals, $\mathcal{I}(n, \boldsymbol{k})$ and $\mathcal{I}(m, \boldsymbol{k} + \boldsymbol{q})$. These contributions are equal in magnitude but opposite in sign: In Fig. 4.2a, the positive sign of `sc_idx` denotes the contribution to $\mathcal{I}(n, \boldsymbol{k})$ and the negative sign to $\mathcal{I}(m, \boldsymbol{k} + \boldsymbol{q})$. The rows of `stargets_sources` are sorted based on their contribution to the collision integral and grouped into a `target`. This step is not computationally intensive and is performed only once on CPUs. See Supplementary Note 1 for demo code related to this step.

2. *Compute the contribution from each scattering channel:*
   The second step evaluates the contribution of each scattering channel to the collision integral. As shown in Fig. 4.2b, the contribution to $\mathcal{I}(n, \boldsymbol{k})$ from each channel is stored in the variable `sc_col`, which has the same number of rows as `scatter_channels`. This computationally demanding step is accelerated on GPUs using heterogeneous programming with OpenACC. Example code is provided in Supplementary Note 2.

3. *Collect the contributions from all targets:*
   This step updates the collision integrals using contributions from all targets computed in Step 2. The routine loops over `targets` to update the elements of $\mathcal{I}$ by summing over contributions from all scattering channels in each target (Fig. 4.2c). This way, only one element of $\mathcal{I}$ will be updated for each thread, as shown with black arrows in Fig. 4.2c. Using `target`-1 as an example, the first row of `targets` records the position of this target in `stargets_sources`. By using the value of `sc_idx` for all relevant elements in `stargets_sources`, the code finds all the scattering channels in `sc_col` (see red arrows in Fig. 4.2c). Then those values are summed together to update the corresponding element of $\mathcal{I}$ in the current thread. This step is performed on GPUs for acceleration. Demo code for this step is provided in Supplementary Note 3.

The new data structure resolves the inefficiency of the Baseline-CPU method by minimizing host–device data transfers, reducing the number of atomic operations across threads, as well as eliminating memory padding through data alignment.

As our discussion has focused on ultrafast dynamics, we briefly mention the main differences in the implementation for transport calculations. For ultrafast dynamics, the contribution to the collision integrals $\mathcal{I}(n, \boldsymbol{k})$ and $\mathcal{I}(m, \boldsymbol{k} + \boldsymbol{q})$ from a single scattering channel is equal in magnitude and opposite in sign. This allows us to define only a 1D array for `sc_col`(:) and use the sign of `sc_idx` for bookkeeping. For transport, this is not possible, and thus `sc_col` needs an additional dimension. In practice, `sc_col` is defined as a 3D array to account for the external field, and the code has an additional loop over the directions of the field. These small differences do not affect the performance.

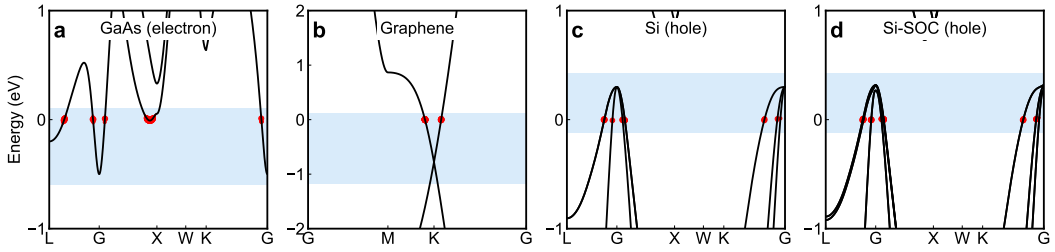

Figure 4.3: **Simulation setup for four systems. a** Electrons in GaAs, **b** electrons in graphene, **c** hole carriers in silicon, and **d** holes in silicon with SOC. Band structures are shown together with the selected energy windows (shaded regions) and the initial populations for the nonequilibrium dynamics simulations (red dots). Energies are shifted so that the Fermi energy is at 0 eV.

Results

Simulation setup

We benchmark the GPU implementation on four selected systems: electron carriers in gallium arsenide (GaAs), graphene, hole carriers in silicon (Si) modeled without spin-orbit coupling (SOC), and hole carriers in silicon with SOC (Si-SOC). These cases cover a range of scenarios, including metals and semiconductors, calculations with and without SOC, electron and hole carriers, and 2D and bulk materials. The band structures, energy windows for nonequilibrium dynamics, and the initial population for nonequilibrium simulation for all four systems are shown in Fig. 4.3.

As shown in Table 4.1, in the ultrafast dynamics simulation of electrons in GaAs, after imposing an energy window of 0.7 eV above the conduction band edge, the number of $\boldsymbol{k}$ and $\boldsymbol{q}$ points in GaAs are reduced from 135^3 to 56713 (2% of the original value) and 637412 (26% of the original value), respectively. Imposing energy conservation and a cutoff on $|\mathbf{g}(\boldsymbol{k},\boldsymbol{q})|$ further reduces the number of scattering channels, from a nominal value of 10^{11} to an actual value of 10^8. These values justify our approach of keeping only the active scattering channels (10^8 in the case of GaAs) and looping over these channels in the code.

Performance and memory usage

We first compare the wall-time of the optimized-GPU code with the Baseline-CPU algorithm to directly show the performance improvement. Figure 4.4a-d compare calculations carried out with the baseline CPU algorithm, used as a reference, and the optimized GPU algorithm, for both transport and ultrafast dynamics, for the four

Table 4.1: Summary of simulation parameters for the four systems studied, including the number of bands and phonon modes (n,ν), the number of $(\boldsymbol{k}, \boldsymbol{q})$ points, and the number of nominal and active scattering channels.

Systems	(n,ν)	(#$\boldsymbol{k}$, #$\boldsymbol{q}$)		# channels	
		Nominal	Energy window	Nominal	Active
GaAs	(1,6)	(135^3, 135^3)	(56713, 637412)	10^{11}	10^8
graphene	(1,6)	(1300^2, 1300^2)	(43206, 131605)	10^{11}	10^8
Si	(3,6)	(105^3, 105^3)	(44275, 232728)	10^{12}	10^8
Si-SOC	(6,6)	(95^3, 95^3)	(29317, 151560)	10^{12}	10^7

systems studied. For transport calculations, the wall time used in the plots is the average elapsed wall time of each iteration in the iterative BTE solution, and for ultrafast dynamics, the wall time is for one time step of the rt-BTE simulation. The speed up of the GPU implementation relative to the baseline CPU code is noteworthy. Our optimized GPU code achieves a speedup by a factor of 44 for transport and 35 for ultrafast dynamics. We find similar speed-ups for all systems in our test set, showing that the speed-up is an intrinsic feature of the GPU algorithm independent of the system studied. This order-of-magnitude speed up is the result of careful design and optimization of array structures, data transfer, and thread management in our GPU algorithm.

To demonstrate the importance of our novel data structure optimized for GPUs, we run the optimized GPU code with and without OpenACC directives in the same HPC settings. This test compares the performance of the same GPU-optimized code executed on GPU versus CPU hardware; this is a fair and established approach to compare CPU and GPU code. As shown in the Supplementary Fig. 4.7, the optimized-GPU code runs 25−50 times faster on GPUs than on CPU hardware, for both transport and ultrafast dynamics. This result demonstrates the considerable acceleration achieved on GPUs.

Next, we compare memory consumption in the baseline CPU and optimized GPU algorithms. In Fig. 4.4e-h, we show data on memory usage. For all calculations without SOC, we find that the memory allocation in the optimized GPU code is approximately 70% for ultrafast dynamics, and 200% for transport, relative to the memory used in the baseline CPU code. This difference arises from the need to store several intermediate variables in the GPU implementation of transport, in particular the direction of the applied electric field, which in the Baseline-CPU code is replaced by an iterative loop without loss of performance. Finally, the calculation

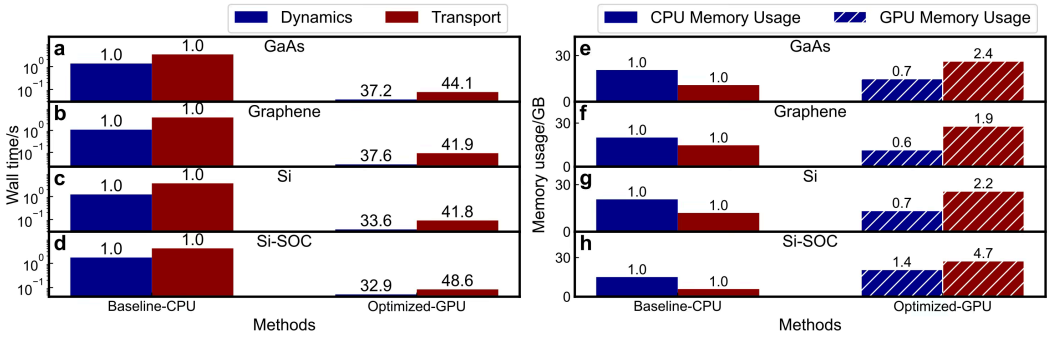

Figure 4.4: **Performance of the optimized GPU implementation. a-d**, performance, and **e-h**, memory usage, for the four systems studied here, respectively. The left panels, **a**–**d**, show the wall time (in seconds, on a logarithmic scale) for ultrafast dynamics (blue) and transport (red) calculations. The speedup values, obtained as the ratio of Baseline-CPU to optimized-GPU code wall times, are given above each bar in the optimized-GPU results. The right panels, **e-h**, give the memory usage (in GB) on CPU (solid colors) and GPU (striped bars) for the same systems. Memory usage values annotated in the plot are referenced to the baseline CPU results.

with SOC in Fig. 4.4h uses more memory than the cases without SOC. The reason is that the number of scattering channels in the arrays scales with the number of bands, resulting in a higher memory usage when SOC is included. Despite the higher memory usage, the wall-time speed-up is unchanged in the presence of SOC.

Strong scaling analysis

The strong scaling measures how the speedup scales with the number of computing nodes for a fixed simulation size. Therefore, strong-scaling benchmarks address a key question for computationally intensive simulations: What performance improvement can one obtain by increasing the computational resources? To avoid confusion with the speed-up of GPU versus CPU code discussed above, we define the strong-scaling speed-up as:

$$\text{Strong-scaling speedup}_N = \frac{T_{4-nodes}}{T_{\text{N}-nodes}} \tag{4.6}$$

where $T_{4-nodes}$ and $T_{\text{N}-nodes}$ are the wall times for simulations using 4 and N nodes, respectively.

We carry out strong-scaling benchmarks for the optimized GPU code, for both transport and ultrafast dynamics, using between 4 and 64 GPU nodes. Setup details

Table 4.2: Momentum grid size and computational resources for performance and strong scaling analysis on four systems. Note that the momentum grid is two-dimensional in graphene and three-dimensional for the other materials.

		GaAs	graphene	Si	Si-SOC
Performance	Transport	120^3	1200^2	95^3	75^3
(1 node, size fixed)	Dynamics	135^3	1300^2	105^3	90^3
Strong scaling	Transport	155^3	1700^2	125^3	95^3
(#nodes vary, size fixed)	Dynamics	195^3	2300^2	150^3	120^3

of these simulations, which employ very dense grids in momentum space, are provided in Table 4.2. Strong scaling results for the four systems studied here are shown in Fig. 4.5. Based on the definition of strong-scaling speedup given above, the ideal speedup is $\frac{N}{4}$, where N is the number of nodes. This ideal value is shown in Fig. 4.5 with a dashed line and used as a reference.

Our results in Fig. 4.5 show that in common scenarios for most users, consisting of calculations with up to 20 GPU nodes, our GPU code exhibits nearly ideal scaling performance, which extends up to 24 nodes or more nodes in most cases. The only case that deviates from this trend is transport calculations in Si with or without SOC (red lines in Fig. 4.5c-d). Due to the high memory demand for these calculations, the scaling remains nearly ideal up to 8 GPU nodes, but deviates increasingly beyond that. For all systems, the acceleration efficiency drops to around 40–60% of the ideal scaling at 64 nodes. This reduction is common in GPU codes and is mainly due

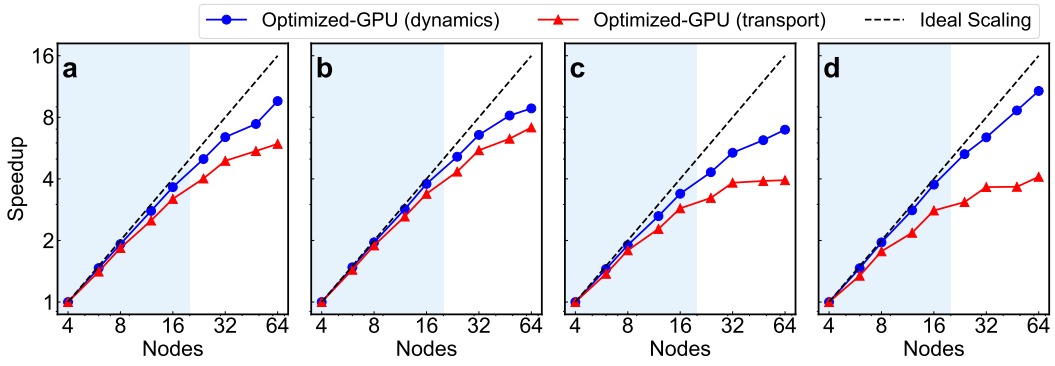

Figure 4.5: **Strong-scaling performance.** Speedup versus number of GPU nodes for **a** GaAs, **b** graphene, **c** Si, and **d** Si with SOC. Results for the optimized-GPU code are shown using solid lines with symbols for ultrafast dynamics (purple) and transport (red). The dashed line shows the ideal linear scaling. Common scenarios for most users ($\leq$ 20 GPU nodes) are indicated with shaded regions.

to insufficient workload per GPU when such a large number of nodes is employed. This result implies that allocating excessive GPU resources to a calculation that can be completed on a much smaller number of nodes is unnecessary and inefficient.

4.3 Conclusion

In this work, we have developed an efficient GPU algorithm and data structure to accelerate BTE calculations of electronic transport and ultrafast dynamics governed by e-ph interactions. The code is developed in OpenACC and is included in version 3.0 of the open-source code PERTURBO. Due to the directive-based approach used in OpenACC, the implementation is easy to maintain. With GPU acceleration, this new version of PERTURBO provides a highly efficient MPI+OpenMP+GPU parallelization that can fully take advantage of modern Exascale HPC computing environments with multi-core CPUs and GPU accelerators.

Through extensive benchmarks, we report speed-ups by a factor of ~40 for transport and ultrafast dynamics relative to the reference, state-of-the-art CPU implementation in the previous version of PERTURBO. This result is achieved by reformulating the data structure and algorithm to store e-ph interactions and carry out calculations of collision integrals. Our approach optimizes data allocation and movement between host and GPUs and synchronization between numerous GPU cores. We analyze the performance, memory consumption, and strong scaling for three materials. The strong scaling analysis shows nearly ideal scaling up to 16 GPU nodes (64 GPUs), with only a slight decrease in performance up to 24 GPU nodes (96 GPUs) for most cases.

While the optimized GPU data structure is discussed in the context of e-ph interactions, the same data structure can also be used for other scattering mechanisms. Note that we did not consider GPU acceleration for the interpolation of e-ph matrices, which has been studied in previous work [26, 27], mainly because recently developed compression algorithms [24] make the e-ph interpolation routines already highly efficient. Beyond e-ph interactions, the computation of ph-ph interactions is also restructured in data structure as part of this thesis work, which will be included for future releases of PERTURBO, following the same design principles discussed here. In a future release, we will extend the GPU acceleration feature to other modules of PERTURBO.

4.4 Supplementary information

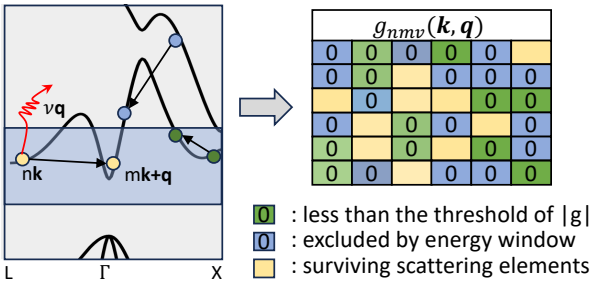

Figure 4.6: **Visualization of the sparsity of the g matrix.** The left panel shows schematically e-ph scattering processes mapped on the band structure, where electronic transitions are shown with black arrows and pictorial phonon lines using red arrows. The energy window where the real-time dynamics is simulated is shown as a shaded region. The right panel shows schematically the structure of the e-ph matrix, where scattering channels outside the energy window, or with coupling strength smaller than a cut-off value, are shown as purple and green zero elements, respectively.

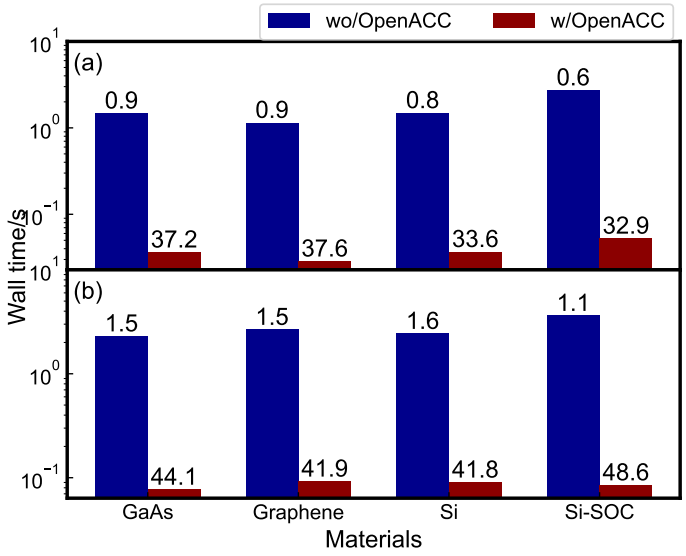

Figure 4.7: **Performance of optimized-GPU code.** Comparison of the performance of the GPU code without OpenACC directives (blue) and with OpenACC directives (red), for **a** ultrafast dynamics simulation and **b** transport calculations. Results are shown for the four systems discussed in main text. The speedup values given in the plot are relative to the Baseline-CPU code. These results show that the GPU code with OpenACC directives is 25−50 times faster than the GPU code built without OpenACC directives (and executed on CPUs).

```
      type :: scatter_type
         integer :: ik
         integer :: ikq
      end type scatter_type

      type :: scatter_channels_type
         integer :: eph_g2
         integer :: bnds_idx
         integer :: kq_index
      end type scatter_channels_type

      type :: stargets_sources_type
         integer :: n
         integer :: ik
         integer :: sc_idx
      end type stargets_sources_type

      type :: targets_type
         integer :: len
         integer :: src
         integer :: nk_index
      end type targets_type

      type(scatter_type) :: scatter(num_kq_pair)
      type(scatter_channels_type) :: scatter_channels(
         num_scatter_channel)
      type(stargets_sources_type) :: stargets_sources(2*
         num_scatter_channel)
      type(targets_type) :: targets(num_band*num_k)
```

Note 1: **Demo Fortran code for the key abstract types of the target layer in the GPU-optimized algorithm.**

```
      real(8) :: sc_col(num_scatter_channel)

      !$acc parallel loop private(bands) async
      do ns = 1, num_scatter_channel
         i = scatter_channels(ns)%kq_index
         ik = scatter(i)%ik
         ikq = scatter(i)%ikq
         bands = scatter_channels(ns)%bnds_idx

         fem = ...
         fabs = ...
         ...
         sc_col(ns) = - scatter_channels(ns)%eph_g2 * (fabs
            + fem)
      end do
```

Note 2: **Demo code for computing the contribution from each scattering channel using OpenACC in the GPU-optimized algorithm.** The `async` command is used to improve the performance by launching the GPU threads asynchronously. Note that this code block is for one MPI process, and thus `num_scatter_channel` is the total number of scattering channels for all $(\boldsymbol{k}, \boldsymbol{q})$ pairs in the current MPI process.

```
      !$acc parallel loop private(bands) async
      do nt = 1, num_targets
         total = 0
         start = targets(nt)%src
         stop = start + targets(nt)%len - 1
         !$acc loop reduction(+:total)
         do is = start, stop
            total = total + ...
         end do
         I(ik,n) = total
      end do
      !$acc end data
      !$acc wait
      call mp_sum(epcol, inter_pool_comm)
```

Note 3: **Demo code for collecting the contributions from all targets in the GPU-optimized algorithm.** The routine has two nested loops: the outer loop, executed asynchronously, iterates over targets, while the inner loop runs over the components of each target. A target and its components are computed by only one thread, so that each element of $\mathcal{I}$ is updated by only one thread of execution. The code block is for one MPI process. Before collecting information from all MPI processes, the `acc wait` command is used for synchronization across GPU threads.

References

[1] C.-H. Park, N. Bonini, T. Sohier, G. Samsonidze, B. Kozinsky, M. Calandra, F. Mauri, and N. Marzari, Nano Lett. **14**, 1113 (2014).

[2] J. I. Mustafa, M. Bernardi, J. B. Neaton, and S. G. Louie, Phys. Rev. B **94**, 155105 (2016).

[3] W. Li, Phys. Rev. B **92**, 075405 (2015).

[4] T.-H. Liu, J. Zhou, B. Liao, D. J. Singh, and G. Chen, Phys. Rev. B **95**, 075206 (2017).

[5] N.-E. Lee, J.-J. Zhou, L. A. Agapito, and M. Bernardi, Phys. Rev. B **97**, 115203 (2018).

[6] L. Cheng, C. Zhang, and Y. Liu, Phys. Rev. Lett. **125**, 177701 (2020).

[7] S. Poncé, W. Li, S. Reichardt, and F. Giustino, Rep. Prog. Phys. **83**, 036501 (2020).

[8] D. C. Desai, B. Zviazhynski, J.-J. Zhou, and M. Bernardi, Phys. Rev. B **103**, L161103 (2021).

[9] B. K. Chang, J.-J. Zhou, N.-E. Lee, and M. Bernardi, Npj Comput. Mater. **8**, 63 (2022).

[10] B. K. Chang and M. Bernardi, J. Phys.: Condens. Matter **37**, 095704 (2025).

[11] J.-J. Zhou, O. Hellman, and M. Bernardi, Phys. Rev. Lett. **121**, 226603 (2018).

[12] Y. Luo, J. Park, and M. Bernardi, Nat. Phys. **21**, 1275 (2025).

[13] D. J. Abramovitch, J. Mravlje, J.-J. Zhou, A. Georges, and M. Bernardi, Phys. Rev. Lett. **133**, 186501 (2024).

[14] S. Gao, J.-J. Zhou, Y. Luo, and M. Bernardi, Phys. Rev. Mater. **8**, L051001 (2024).

[15] D. C. Desai, J. Park, J.-J. Zhou, and M. Bernardi, Nano Lett. **23**, 3947 (2023).

[16] M. Bernardi, Nat. Comput. Sci **3**, 480 (2023).

[17] F. Caruso and D. Novko, Adv. Phys. X **7**, 2095925 (2022).

[18] J.-J. Zhou, J. Park, I.-T. Lu, I. Maliyov, X. Tong, and M. Bernardi, Comput. Phys. Commun. **264**, 107970 (2021).

[19] J. Yao, I. Maliyov, D. J. Gardner, C. S. Woodward, and M. Bernardi, Npj Comput. Mater. **11**, 256 (2025).

[20] M. Taher, in *2009 4th International Design and Test Workshop (IDT)* (IEEE, 2009) pp. 1–6.

[21] N. S. Abramov and S. M. Abramov, Supercomputing Frontiers and Innovations **10**, 4 (2023).

[22] NERSC, Perlmutter architecture (2025).

[23] D. Storti and M. Yurtoglu, *CUDA for engineers: an introduction to high-performance parallel computing* (Addison-Wesley Professional, 2015).

[24] Y. Luo, D. Desai, B. K. Chang, J. Park, and M. Bernardi, Phys. Rev. X **14**, 021023 (2024).

[25] Perturbopy: a suite of Python scripts for Perturbo testing and postprocessing., Official documentation (2025).

[26] A. Cepellotti, J. Coulter, A. Johansson, N. S. Fedorova, and B. Kozinsky, J. Phys. Mater. **5**, 035003 (2022).

[27] Z. Liu, B. Zhang, Z. Fan, and W. Li, arXiv 2306.16493 (2023).

Chapter 5

TENSOR LEARNING AND COMPRESSION OF N-PHONON INTERACTIONS

5.1 Introduction

Phonon-phonon (ph-ph) interactions are crucial to understanding thermal transport, lattice dynamics, and structural phase transitions in condensed matter [1]. They originate from the anharmonicity of the lattice potential and can be described by n-th order interatomic force constant (n-IFC) tensors, where $n \geq 3$. Density function theory (DFT) [2] calculations combined with fitting algorithms provide accurate n-IFC tensors [3–6], enabling quantitative predictions of thermal transport in materials [7–12]. However, the high-dimensionality of the IFC tensors obscures the underlying physics and poses significant computational challenges. The complexity of n-phonon (n-ph) interactions grows exponentially with order n, a clear example of the curse of dimensionality. For thermal conductivity calculations, 4-ph interactions require orders of magnitude more computational effort than 3-ph interactions, whereas 5-ph and higher-order ph-ph interactions remain inaccessible.

To overcome this complexity, discovering low-rank approximations of n-IFC could be game changing. There is growing interest in such dimensionality reduction approaches, including tensor network states for many-body wave functions [13, 14], tensor train for differential equations [15–17] and Feynman diagrams [16, 18], tensor hyper-contraction [19–21] and density fitting [22] for electron interactions in quantum chemistry, and feature optimization for atomic machine learning [23–26]. Recent work has employed singular value decomposition to compress electron-phonon (e-ph) interactions and greatly speed up first-principles e-ph calculations [27]. For ph-ph interactions, previous work has taken advantage of crystal symmetry to reduce the number of free parameters and compute the n-IFCs with fewer DFT force calculations [28, 29]. The compressed sensing technique has also been used to obtain n-IFCs using sparse solvers [4, 30]. However, in previous work the n-IFCs are still handled explicitly in full tensor form, Φ_{ijk} and Φ_{ijkl} for 3- and 4-IFCs respectively, and a low-rank representation of n-IFCs is still missing.

In this chapter, we introduce a low-rank tensor ansatz for n-ph interactions in momentum space based on the CANDECOMP/PARAFAC (CP) decomposition [31],

a compression method used in tensor learning [32–34]. The proposed ansatz is a permanent CP (PCP) decomposition, which generalizes the CP decomposition to enforce bosonic statistics. To find an optimal low-rank PCP decomposition, we formulate the tensor learning problem and solve it on GPU hardware. The optimized low-rank PCP tensors achieve large compression factors of 10^3–10^4 with minimal compression losses of only a few percent in all materials we study. This result reveals the inherent low-dimensionality of n-ph interactions and enables a speedup of nearly three orders of magnitude for calculations using n-IFCs, including phonon relaxation times and thermal conductivity. We also introduce constraints to treat ph-ph interactions for long-wavelength acoustic phonons, leading to nearly lossless predictions of thermal conductivity compared to calculations using full n-IFC tensors. Finally, we show that the PCP ansatz can uncover dominant modes in n-ph interactions, providing a valuable tool for understanding microscopic thermal transport mechanisms and formulating accurate minimal models. Beyond ph-ph interactions, the PCP ansatz and corresponding open-source routines developed here provide a blueprint for modeling momentum-dependent tensors, with broad applications in condensed matter physics.

5.2 Methods

We consider the n-ph interaction $V^{(n)}(\mathbf{Q}_1, ..., \mathbf{Q}_n)$, which describes the scattering amplitude of n phonon modes. Each mode is specified by a wave vector $\mathbf{q}_i$ and branch index ν_i, collectively denoted as $\mathbf{Q}_i = (\nu_i, \mathbf{q}_i)$ for $i = 1, \ldots, n$. For instance, the 3-ph interaction is given by

$$V^{(3)}(\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3) = \delta\left(\sum_{i=1}^{3} \boldsymbol{q}_i\right) \sum_{b_1,l_2b_2,l_3b_3} \sum_{\alpha_1\alpha_2\alpha_3} \times \Phi^{\alpha_1,\alpha_2,\alpha_3}_{0b_1,l_2b_2,l_3b_3} \frac{e^{\mathbf{Q}_1}_{\alpha_1 b_1} e^{\mathbf{Q}_2}_{\alpha_2 b_2} e^{\mathbf{Q}_3}_{\alpha_3 b_3}}{\sqrt{m_{b_1} m_{b_2} m_{b_3}}} e^{i\boldsymbol{q}_2 \boldsymbol{r}_{l_2} + i\boldsymbol{q}_3 \boldsymbol{r}_{l_3}}, \tag{5.1}$$

where each of the l, b, and α label the primitive cell, atom, and Cartesian coordinate, respectively; $e^{\mathbf{Q}}_{\alpha b}$ is the displacement eigenvector of phonon mode $\mathbf{Q}$, m_b is the mass of atom b, and $\boldsymbol{r}_l$ is the position of primitive cell l. Above, $\Phi^{\alpha,\alpha_1,\alpha_2}_{0b,l_1b_1,l_2b_2}$ is the 3-IFC tensor associated with three atomic displacements, the first displacement fixed at the cell origin. For general n-ph interactions, analogous n-IFC tensors $\mathbf{\Phi}^{(n)}$ can be defined.

To compress the n-ph interactions $V^{(n)}$, we propose the PCP decomposition as the

low-rank ansatz

$$\tilde{V}^{(n)}(\mathbf{Q}_1, ..., \mathbf{Q}_n) = \delta\left(\sum_{i=1}^{n} \boldsymbol{q}_i\right) \times \sum_{\xi=1}^{N_c^{(n)}} \frac{\lambda_\xi}{n!} \sum_{\sigma \in S_n} A_1^\xi(\mathbf{Q}_{\sigma_1}) A_2^\xi(\mathbf{Q}_{\sigma_2}) ... A_n^\xi(\mathbf{Q}_{\sigma_n}), \tag{5.2}$$

where $N_c^{(n)}$ is the PCP rank, λ_ξ the ξ-th PCP singular value, A_i^ξ are the corresponding PCP modes, and S_n is the symmetric group of order n containing all the permutations of $(1, ..., n)$. Note that Eq. (5.2) correctly preserves the bosonic exchange symmetry and momentum conservation of n-ph interactions.

The n-ph interactions in compressed form involve n momenta $(\mathbf{Q}_1, \mathbf{Q}_2, ..., \mathbf{Q}_n)$, each associated with $N_c^{(n)}$ PCP modes $A_\xi(\mathbf{Q})$, which are functions defined on a grid of size $N_\mathbf{Q}$. This gives a storage requirement of $O(nN_c^{(n)}N_\mathbf{Q})$ for the n-ph interactions in compressed form. For example, storing 3-ph interactions $V^{(3)}$ conventionally requires $O(N_\mathbf{Q}^2)$ resources versus $O(3N_c^{(3)}N_\mathbf{Q})$ for its PCP compressed counterpart $\tilde{V}^{(3)}$. This yields a large storage reduction, by a factor of $N_\mathbf{Q}/(3N_c^{(3)}) \approx 10^3$–$10^4$ in a typical calculation.

Since $V^{(n)}$ is obtained by transforming the n-IFC $\mathbf{\Phi}^{(n)}$ to momentum space, we derive the corresponding low-rank ansatz for 3- and 4-IFCs in real space:

$$\tilde{\mathbf{\Phi}}^{\alpha_1,\alpha_2,\alpha_3}_{l_1 b_1, l_2 b_2, l_3 b_3} = \sum_{\xi=1}^{N_c^{(3)}} \frac{\lambda_\xi}{3!} \sum_{\sigma \in S_3} \sum_{l} \prod_{i=1}^{3} A^\xi_{\sigma_i}(l_i - l, b_i \alpha_i), \tag{5.3}$$

$$\tilde{\mathbf{\Phi}}^{\alpha_1,\alpha_2,\alpha_3,\alpha_4}_{l_1 b_1, l_2 b_2, l_3 b_3, l_4 b_4} = \sum_{\xi=1}^{N_c^{(4)}} \frac{\lambda_\xi}{4!} \sum_{\sigma \in S_4} \sum_{l} \prod_{i=1}^{4} B^\xi_{\sigma_i}(l_i - l, b_i \alpha_i),$$

where $A_i^\xi(l, b\alpha)$ and $B_i^\xi(l, b\alpha)$ are real-space representations of $A_i^\xi(\mathbf{Q})$ in Eq. 5.2 for 3-ph and 4-ph interactions respectively. To enforce the acoustic sum rule (ASR) [29] on the compressed n-ph interactions $\tilde{V}^{(n)}$, we impose $\sum_{l,b} A_i^\xi(l, b\alpha) = 0$, which preserves the ASR in the PCP compression. Derivations of n-ph interactions and their PCP decomposition are provided in the Supplementary information [35]. The computational cost of PCP-compressed $\tilde{\mathbf{\Phi}}^{(n)}$ and $\tilde{V}^{(n)}$ scales linearly with system size, making it promising for calculations in complex materials with large unit cells.

To find optimal PCP modes A_i^ξ (or B_i^ξ) that best approximate the original IFC tensors $\mathbf{\Phi}^{(n)}$ computed with DFT, we minimize the loss function:

$$L = \|\tilde{\mathbf{\Phi}}^{(n)}[\boldsymbol{A}] - \mathbf{\Phi}^{(n)}\|^2, \tag{5.4}$$

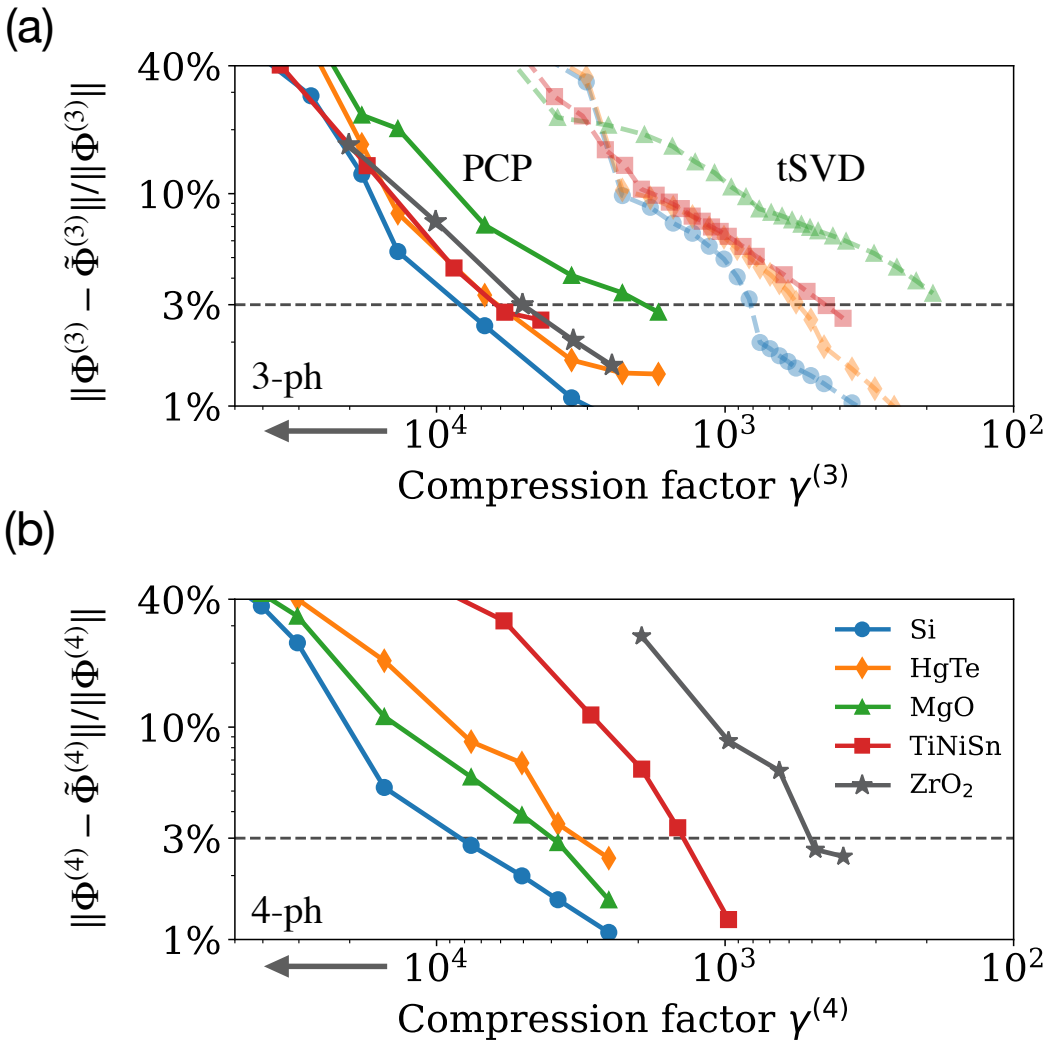

Figure 5.1: Relative compression loss vs. compression factor $\gamma^{(n)}$. (a) Results for 3-IFC tensors comparing PCP (solid lines) and tSVD (dashed lines). (b) Results for 4-IFC tensors. The compression factor $\gamma^{(n)}$ is defined in Eq. 5.5.

where $\tilde{\mathbf{\Phi}}^{(n)}$ is the low-rank IFC tensor in Eq. (5.3), and the norm is defined as $\|\boldsymbol{x}\| = \sqrt{\sum_I x_I^2}$ [1]. This is a typical tensor learning problem that lacks closed-form solutions for high-order tensors. In addition, existing optimization algorithms are computationally demanding for large tensors. We address this challenge by developing a computational toolkit, called PHONON-PCP, implemented using PYTORCH [37]. This optimization routine follows a standard "neural network"-like training loop,

[1] In the optimization, $A_i^{\xi}(l, b\alpha)$ is truncated in real space up to a cutoff r_c large enough for the low-rank IFC tensor $\tilde{\mathbf{\Phi}}^{(n)}$ to span all nonzero elements in the full IFC tensor $\mathbf{\Phi}^{(n)}$.

where $\tilde{\mathbf{\Phi}}^{(n)}$ is evaluated in a forward pass, and A_i^{ξ} is the learnable weight trained using automatic differentiation. A typical optimization involves 10^3 – 10^4 trainable parameters, depending on the material. The algorithm runs efficiently on GPUs – the compression process converges in a few minutes (Si and other simple systems) to one hour (ZrO_2) on a single NVIDIA A100 chip.

We apply our approach to five materials: Si, HgTe, MgO, TiNiSn, and monoclinic ZrO_2. These materials span a wide range of properties, including nonpolar and polar (or ionic) bonds, high and low crystal symmetry, and different levels of anharmonicity. For each material, we generate force-displacement datasets from DFT force calculations on supercells and then employ ALAMODE [3] to extract the 3- and 4-IFCs [2]. The symmetry of the n-IFCs is taken into account in the force constant fitting process [3]. We calculate DFT forces using VASP [39, 40] with PBEsol functional [41]. For polar materials, we calculate Born effective charges using density functional perturbation theory in VASP, and use the Ewald summation to compute the nonanalytic part of the dynamical matrix [42]. Additional computational details, are provided in the Supplementary Information [35].

To assess the accuracy of PCP for compressing n-ph interactions, we calculate the error relative to uncompressed n-IFCs, expressed as the compression loss $\|\tilde{\mathbf{\Phi}}^{(n)} - \mathbf{\Phi}^{(n)}\|/\|\mathbf{\Phi}^{(n)}\|$, as a function of the compression factor $\gamma^{(n)}$, which quantifies parameter reduction in the n-IFCs:

$$\gamma^{(n)} = \frac{\#\text{ of nonzero entries of } \mathbf{\Phi}^{(n)}}{N_c^{(n)}}. \tag{5.5}$$

This error analysis is shown for the 3-ph interactions in Fig. 5.1(a). We achieve compression factors of 10^3–10^4 at the 3% compression error threshold. For comparison, a truncated singular value decomposition (tSVD) of 3-ph interactions, inspired by our previous work on e-ph coupling [27], can only reach compression factors of a few hundred for the same 3% compression loss. Therefore, the PCP approach provides a 10-times more favorable dimensionality reduction than SVD for all materials we study. We attribute the superior performance of PCP to its flexible ansatz combined with the correct permutation symmetry and the explicit parameter optimization during tensor learning, which enable effective low-rank representation of n-ph interactions. We achieve similarly large compression factors for 4-ph interactions,

[2] We include 3-IFCs up to the fifth-nearest neighbors and 4-IFCs up to the second-nearest neighbors for Si, HgTe and MgO. For TiNiSn, the 3-IFCs are truncated to 6.4 Å and the 4-IFCs to 3.8 Å. For ZrO_2, the 3-IFCs are truncated to 5.3 Å and the 4-IFCs to 2.7 Å.

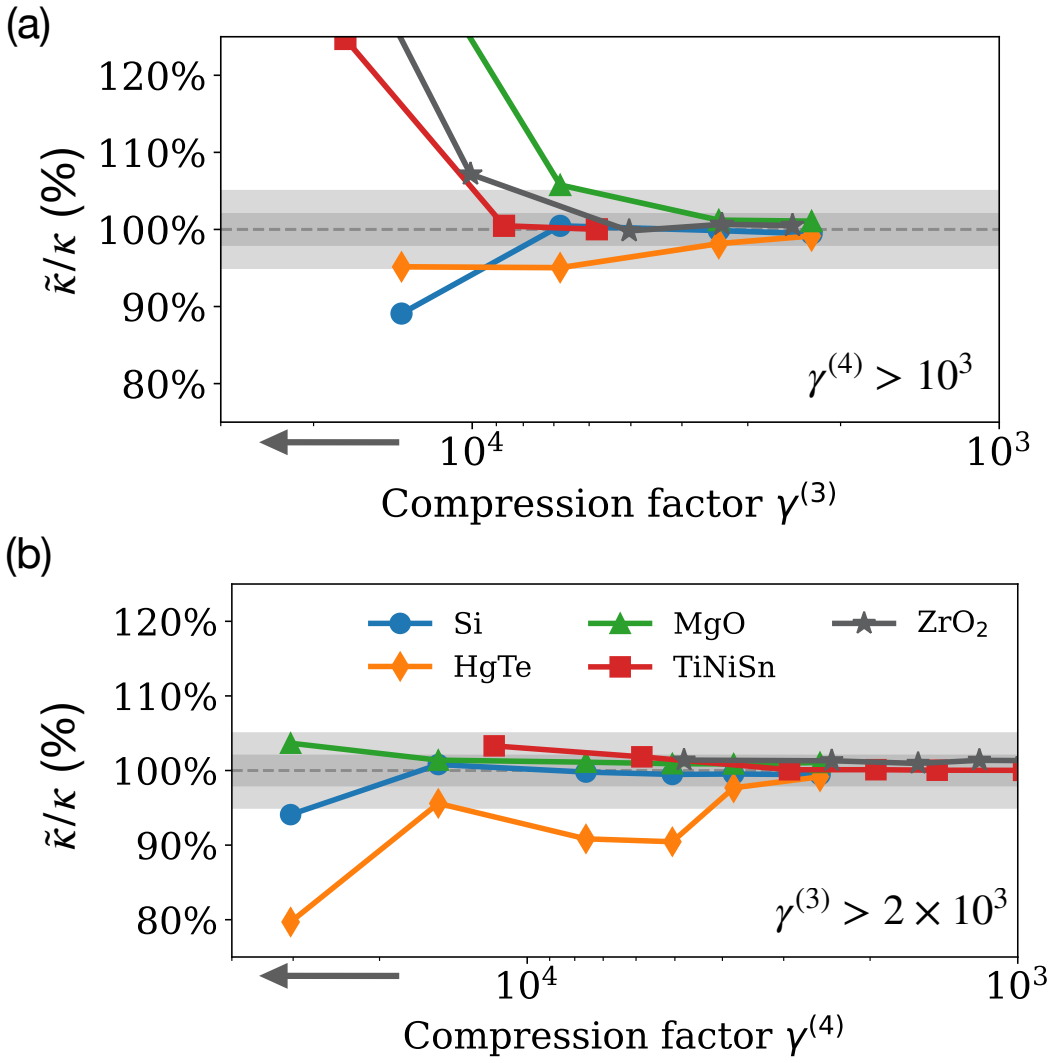

Figure 5.2: Ratio of the thermal conductivity $\tilde{\kappa}$, computed using compressed 3- and 4-ph interactions, to the thermal conductivity κ computed with uncompressed tensors at 300 K. Results are shown as a function of: (a) 3-ph compression factor, $\gamma^{(3)}$, for fixed $\gamma^{(4)} > 10^3$ ($N_c^{(4)} = 48$), and (b) 4-ph compression factor, $\gamma^{(4)}$, for fixed $\gamma^{(3)} > 2 \times 10^3$ ($N_c^{(3)} = 24$ for all materials except ZrO_2, and $N_c^{(3)} = 144$ for ZrO_2). The light (dark) shaded regions show the 95% (98%) accuracy windows.

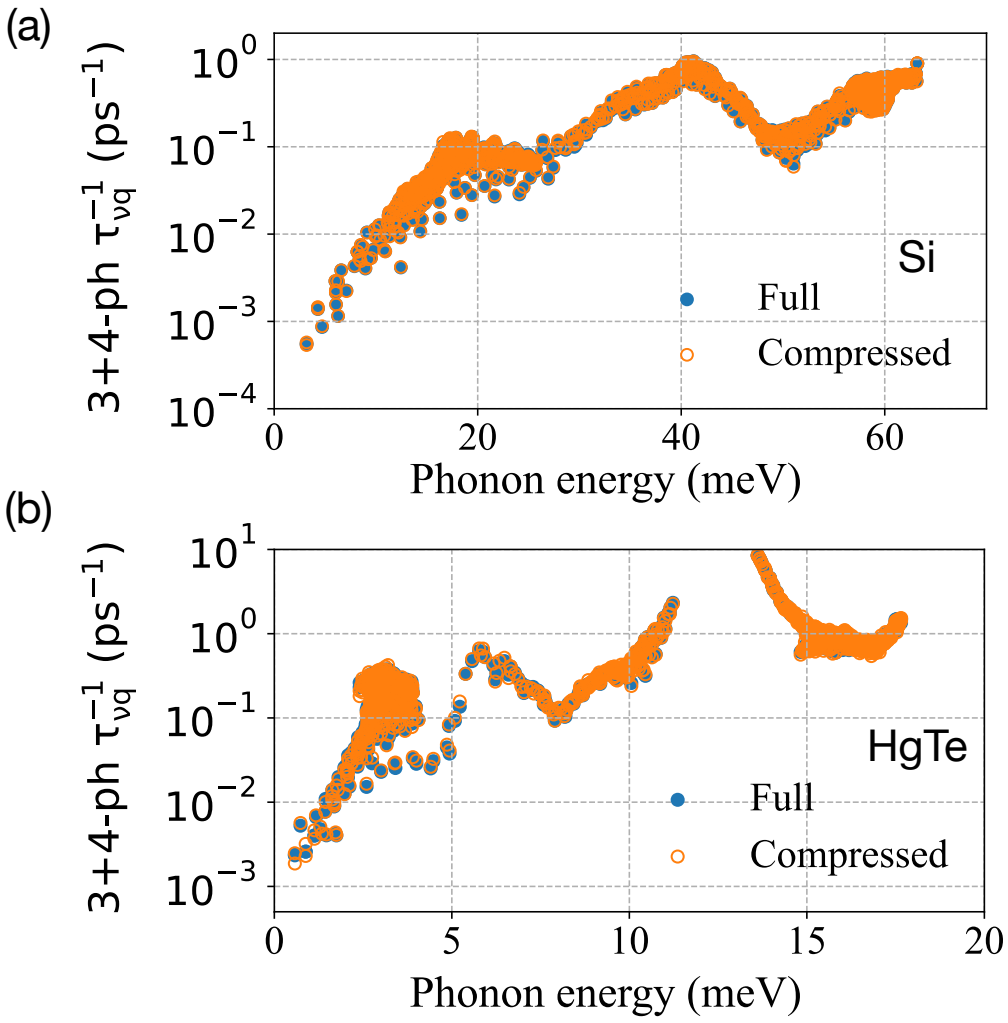

Figure 5.3: Phonon scattering rates combining 3- and 4-ph interactions, shown as a function of phonon energy for (a) Si and (b) HgTe, comparing results for full IFC tensors (blue dots) and compressed IFC tensors (orange circles). In both materials, we show results for compression factors $\gamma^{(3)} = 2200$ for 3-ph, and $\gamma^{(4)} = 2500$ for 4-ph interactions.

as shown in Fig. 5.1(b). These results highlight the inherent low-dimensionality of n-ph interactions and show their accurate low-rank approximation using tensor learning.

Note that crystal symmetry can also reduce the size of the n-IFCs, but it cannot be used to speed up calculations of phonon relaxation times and thermal conductivity, which require the full n-IFCs. In addition, the size reduction from symmetry is significantly smaller than the PCP compression factor shown here [35].

Using compressed n-ph interactions, we compute the thermal conductivity at 300 K for the four materials. Our calculations use the single-mode relaxation time approximation [7, 43, 44] and include both 3- and 4-ph interactions [45–48] with a converged number of scattering processes for 3- and 4-ph scattering rates [49] (see details in Supplementary Information [35]).

The thermal conductivity $\tilde{\kappa}$, computed with compressed n-IFC tensors, is compared with the reference value κ from full (uncompressed) tensors in Fig. 5.2(a) and (b). We find that $\tilde{\kappa}$ converges to the reference value for large compression factors, $\gamma^{(3)} \approx 7 \times 10^3$ and $\gamma^{(4)} \approx 10^4$ for Si, MgO, TiNiSn, and ZrO_2, and slightly smaller $\gamma^{(4)} = 3 \times 10^3$ for HgTe, where the 4-ph contribution is more important [9]. Even for these very large compression factors, the approximate thermal conductivity is still within 2% of the reference value obtained with full IFCs [35]. This high accuracy for thermal conductivity is a result of the low compression losses for n-IFCs.

We also examine the accuracy of compressed IFC tensors for calculating the microscopic scattering rate for each phonon mode, $\tau^{-1}_{\nu\boldsymbol{q}}$. In Figs. 5.3(a) and (b), the phonon scattering rates from compressed IFC tensors are nearly identical to those from full IFC tensors for each single phonon mode, even in cases where $\tau^{-1}_{\nu\boldsymbol{q}}$ varies by up to four orders of magnitude over the phonon spectrum. We calculate the coefficient of determination R^2 [50] between approximate and reference $\tau^{-1}_{\nu\boldsymbol{q}}$, and find values of $R^2 = 0.9996$ for Si and $R^2 = 0.9986$ for HgTe, further confirming the accuracy of the compressed n-ph interactions. The approximate scattering rates achieve a similar accuracy in TiNiSn and MgO [35].

The dimensionality reduction provided by PCP leads to massive cost savings for calculations involving n-ph interactions, with speed-up proportional to the compression factor. In Fig. 5.4, we compare CPU wall times for calculations of 3-ph and 4-ph scattering rates using full and compressed n-IFCs with large compression factors ($\gamma^{(3)} > 2200$ and $\gamma^{(4)} > 1000$). For all materials studied here, the use of compressed IFC tensors enables a speed-up of 260–7200 in calculations of thermal

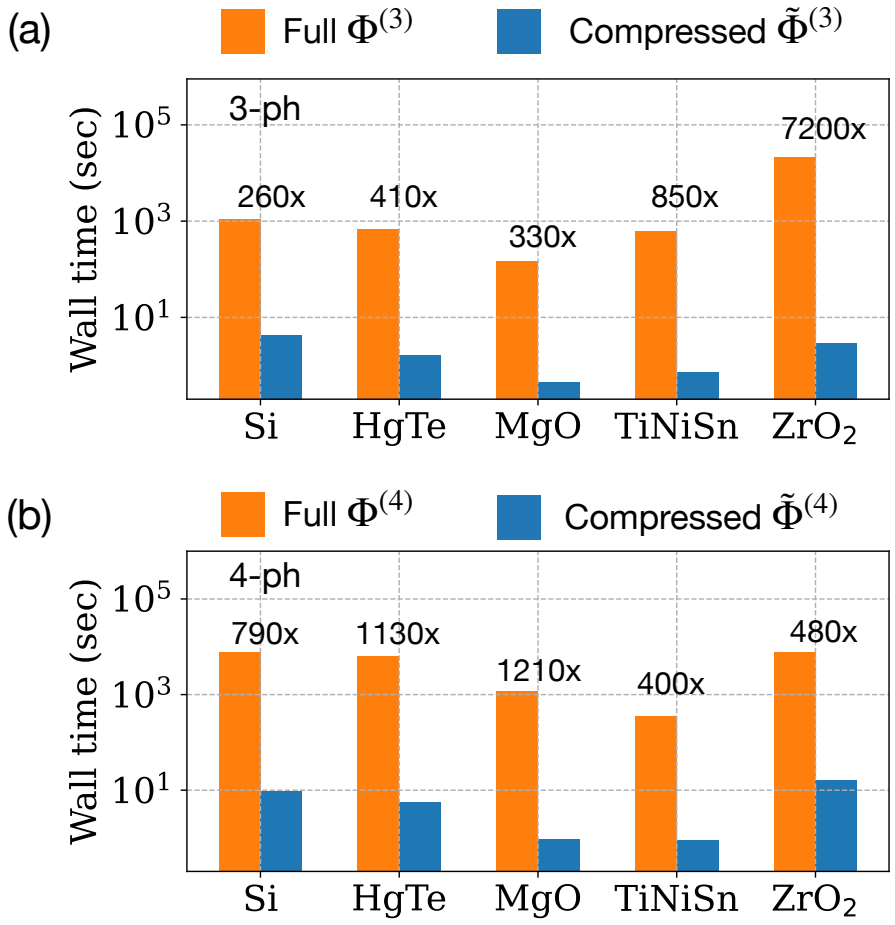

Figure 5.4: Comparison of computational cost (CPU wall time) for calculations of phonon scattering rates using compressed and full IFC tensors. Results are shown for (a) 3-ph scattering rates with compression factors $\gamma^{(3)} > 2200$ and (b) 4-ph scattering rates with compression factors $\gamma^{(4)} > 1000$ in all materials. The speedup achieved with compressed tensors is given in figure for each material.

conductivity. Calculations of $\tau_{\nu\boldsymbol{q}}^{-1}$ using compressed IFC tensors are so efficient that computing the thermal conductivity with 3- and 4-ph interactions takes less than a minute for each material studied here, including monoclinic ZrO_2 with a 12-atom unit cell. Using compressed IFC tensors, we are able to accurately extrapolate the thermal conductivity to the thermodynamic limit (TDL) by employing progressively denser momentum grids. The importance of extrapolating to the TDL is clear from the case of HgTe, where the TDL-extrapolated κ is 47% larger than the value obtained with reasonable grid sizes of 16^3 [35]. The calculation for the largest grid, 250^3, requires only 2.5 CPU node-hours. We attribute the slow convergence of the thermal conductivity in HgTe to the important role of 4-ph processes (see analysis

in Supplementary Information [35]).

In Fig. 5.4(a), the speed-up increases for increasing unit cell sizes. For large unit cells, such as in ZrO_2, the speed-up becomes comparable to the compression factor because the cost of computing the n-ph interactions dominates the total CPU wall time [35]. The significant speed-up achieved by PCP can be understood by comparing the n-ph interactions and their compressed counterparts. The size of the n-IFC tensor in real space, $\Phi^{(n)}$, scales as $N_R^{n-1}(3N_a)^n$, where N_R is the number of Wigner-Seitz cells, which is the same as the size of the coarse $\mathbf{q}$-point grid, and N_a is the number of atoms in the unit cell, so that $3N_a$ is the number of phonon modes. The summation over these variables in the conventional n-ph interactions, exemplified by the 3-ph case in Eq. 5.1, is replaced by a summation over $N_c^{(n)}$ PCP modes for the compressed n-ph interactions in Eq. 5.2. This reduces the computational cost of $V^{(n)}(\mathbf{Q}_1, \ldots, \mathbf{Q}_n)$ by a factor of $O\left(3N_a^n N_R^{n-1}/N_c^{(n)}\right)$ (see analysis in Supplementary Information [35]).

The PCP decomposition also offers interesting ways to analyze n-ph interactions. We express the n-ph anharmonic energy $E^{(n)}(\boldsymbol{u})$ in terms of PCP modes [35],

$$E^{(n)}(\boldsymbol{u}) = \sum_l \sum_\xi \frac{\lambda_\xi}{n!} \prod_{i=1}^{n} \phi_i^\xi(l, \boldsymbol{u}), \tag{5.6}$$

$$\begin{aligned} \phi_i^\xi(l, \boldsymbol{u}) &= \langle \boldsymbol{A}_i^\xi(l), \boldsymbol{u} \rangle \\ &= \sum_{l' b\alpha} A_{\sigma_i}^\xi(l' - l, b\alpha) u(l' b\alpha), \end{aligned} \tag{5.7}$$

where $u(lb\alpha)$ is the displacement of atom b in primitive cell l along the α direction. Here, the PCP modes $[\boldsymbol{A}_i^\xi(l)]_{l' b\alpha} = A_{\sigma_i}^\xi(l' - l, b\alpha)$ are viewed as local vibrational modes, centered in the l-th primitive cell, which give dominant contributions the n-th order anharmonic energy $E^{(n)}(\boldsymbol{u})$. Therefore, the projection $\phi_i^\xi(l, \boldsymbol{u})$ defined in Eq. 5.6 is a descriptor of the atomic environment that quantifies the similarity between the local atomic displacement $\boldsymbol{u}$ and the PCP modes $\boldsymbol{A}_i^\xi(l)$. (Interestingly, the anharmonic energy for PCP decomposition in Eq. (5.6) has a structure similar to the slave mode [51, 52] and the atomic cluster expansion models [53].)

These PCP modes can be viewed as generalized eigenvectors of the n-ph interaction tensor, and provide direct physical information about the dominant vibrational patterns. In Fig. 5.5, we visualize the three modes associated with the largest PCP singular value for 3-ph interactions in Si. The PCP-mode triplet $\{A_1^{\xi=1}, A_2^{\xi=1}, A_3^{\xi=1}\}$ provides the best rank-1 approximation of 3-IFC tensors in Si and is obtained by

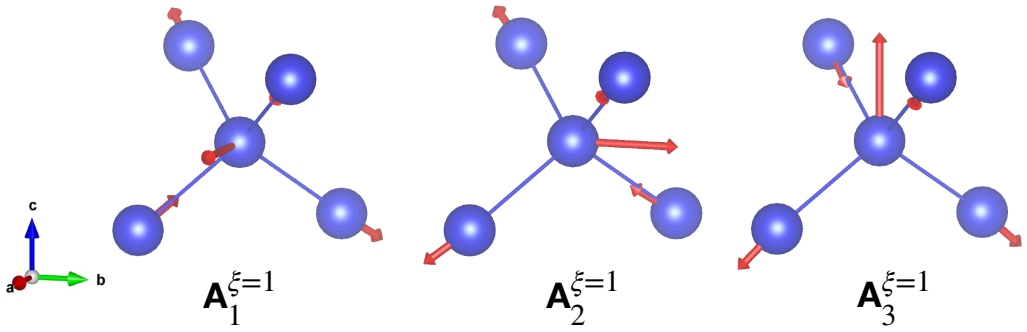

Figure 5.5: The PCP mode triplet $A^{\xi=1}_{i=1,2,3}$ with the largest 3-ph coupling strength in Si. The three modes are related by the C_3 rotation symmetry along the [111] direction in Si.

setting the PCP rank to $N_c^{(3)} = 1$ in the training process. These three modes are related by a C_3 rotation along the [111] crystal direction:

$$\hat{C}_3 A_1^{\xi=1} = A_2^{\xi=1},\ \hat{C}_3 A_2^{\xi=1} = A_3^{\xi=1},\ \hat{C}_3 A_3^{\xi=1} = A_1^{\xi=1}.$$

Even though we do not explicitly preserve the space-group symmetry when generating the PCP modes, the rank-1 PCP ansatz is capable of learning the C_3 symmetry of Si encoded in the uncompressed 3-IFC tensors during the optimization process. In general, the crystal symmetry is preserved in the compressed IFCs, with only a small symmetry loss resulting from compression (see Supplementary Information [35]).

5.3 Conclusion

In summary, we propose a tensor decomposition of n-ph interactions and show compression of n-IFC tensors using GPU-accelerated tensor learning. Our PCP decomposition reveals the inherent low-dimensionality of 3- and 4-ph interactions in crystals with generic symmetry and unit cell size, enabling compression factors greater than 10^3 for minimal compression errors of less than 3%. We achieve corresponding cost savings for calculations of phonon scattering rates and thermal conductivity. The large speed-up makes our method suitable for high-throughput screening of thermal transport in materials and offers a promising pathway to go beyond 3- and 4-ph interactions. The PCP-compressed n-ph interactions can be combined with CPU and GPU parallelization to accelerate modeling of ph-ph interactions. The PCP decomposition can also uncover dominant atomic environment descriptors, providing valuable information for formulating machine learning atomic force fields. Future work will explore 5-ph and higher n-ph interactions in strongly

anharmonic materials, use PCP to obtain simplified coupled-mode equations for nonlinear phonon processes, and accelerate simulations of ultrafast phonon dynamics.

5.4 Supplementary information

Derivation of n-ph interactions and their PCP decomposition

The n-IFC tensor in PCP format is

$$\tilde{\Phi}^{\alpha_1,\ldots,\alpha_n}_{l_1 b_1,\ldots,l_n b_n} = \sum_{\xi=1}^{N_c^{(n)}} \frac{\lambda_\xi}{n!} \sum_{\sigma\in S_n} \sum_{l} \prod_{i=1}^{n} A^{\xi}_{\sigma_i}(l_i - l, b_i\alpha_i), \tag{5.8}$$

where l, b, and α index the primitive cell, basis atom, and Cartesian coordinate, respectively. In the following, we show that this expression is consistent with the PCP decomposition of n-ph interactions, $\tilde{V}^{(n)}$, in Eq. (2) of main text.

The compressed n-ph interaction in momentum space, $\tilde{V}^{(n)}$, is obtained from a basis transformation of $\tilde{\Phi}$:

$$\tilde{V}^{(n)}(\mathbf{Q}_1,\ldots,\mathbf{Q}_n) = \frac{1}{N} \sum_{l_1 b_1,\alpha_1} \frac{e^{\mathbf{Q}_1}_{\alpha_1 b_1}}{\sqrt{m_{b_1}}} e^{i\boldsymbol{q}_1\boldsymbol{r}_{l_1}} \ldots \sum_{l_n b_n,\alpha_n} \frac{e^{\mathbf{Q}_n}_{\alpha_n b_n}}{\sqrt{m_{b_n}}} e^{i\boldsymbol{q}_n\boldsymbol{r}_{l_n}} \tilde{\Phi}^{\alpha_1,\ldots,\alpha_n}_{l_1 b_1,\ldots,l_n b_n}, \tag{5.9}$$

where N is the number of primitive cells in a Born–von Kármán (BvK) supercell. Substituting Eq. 5.8 into Eq. 5.9, and exchanging the order of $\prod$ and $\sum$, we get

$$\tilde{V}^{(n)}(\mathbf{Q}_1,\ldots,\mathbf{Q}_n) = \sum_{\xi=1}^{N_c^{(n)}} \frac{\lambda_\xi}{n!} \sum_{\sigma\in S_n} \frac{1}{N} \sum_{l} \prod_{i=1}^{n} \left(\sum_{l_i b_i,\alpha_i} \frac{e^{\mathbf{Q}_1}_{\alpha_i b_i}}{\sqrt{m_{b_i}}} e^{i\boldsymbol{q}_i\boldsymbol{r}_{l_i}} A_{\sigma_i}(l_i - l, b_i\alpha_i) \right). \tag{5.10}$$

Let us define the PCP modes in momentum space,

$$A^{\xi}_{\sigma_i}(\mathbf{Q}) = \sum_{l'_i b_i,\alpha_i} \frac{e^{\mathbf{Q}_1}_{\alpha_i b_i}}{\sqrt{m_{b_1}}} e^{i\boldsymbol{q}_i\boldsymbol{r}_{l'_i}} A_{\sigma_i}(l'_i, b_i\alpha_i).$$

After the change of variable $l'_i = l_i - l$, we obtain

$$\tilde{V}^{(n)}(\mathbf{Q}_1,\ldots,\mathbf{Q}_n) = \sum_{\xi=1}^{N_c^{(n)}} \frac{\lambda_\xi}{n!} \sum_{\sigma\in S_n} \frac{1}{N} \sum_{l} \prod_{i=1}^{n} \left(e^{i\boldsymbol{q}_i\boldsymbol{r}_l} A^{\xi}_{\sigma_i}(\mathbf{Q}_i) \right) \tag{5.11}$$

$$= \sum_{\xi=1}^{N_c^{(n)}} \frac{\lambda_\xi}{n!} \sum_{\sigma\in S_n} \frac{1}{N} \sum_{l} e^{i\sum_{i=1}^{n}\boldsymbol{q}_i\boldsymbol{r}_l} \prod_{i=1}^{n} A^{\xi}_{\sigma_i}(\mathbf{Q}_i) \tag{5.12}$$

$$= \sum_{\xi=1}^{N_c^{(n)}} \frac{\lambda_\xi}{n!} \sum_{\sigma\in S_n} \delta\left(\sum_{i=1}^{n} \boldsymbol{q}_i \right) \prod_{i=1}^{n} A^{\xi}_{\sigma_i}(\mathbf{Q}_i). \tag{5.13}$$

Using the property that a permanent is invariant under matrix transpose, we exchange i and σ_i in $A^{\xi}_{\sigma_i}(\mathbf{Q}_i)$, and obtain the n-ph interactions in momentum space in PCP format:

$$\tilde{V}^{(n)}(\mathbf{Q}_1,...,\mathbf{Q}_n) = \delta\left(\sum_{i=1}^{n}\boldsymbol{q}_i\right)\sum_{\xi=1}^{N_c^{(n)}}\frac{\lambda_\xi}{n!}\sum_{\sigma\in S_n}\prod_{i=1}^{n}A_i^{\xi}(\mathbf{Q}_{\sigma_i}), \tag{5.14}$$

which is Eq. (2) in the main text.

Anharmonic energy $E^{(n)}$ in PCP decomposition

Using the definition of the n-IFCs, the nth-order anharmonic energy $E^{(n)}(\boldsymbol{u})$ is:

$$E^{(n)}(\boldsymbol{u}) = \frac{1}{n!}\sum_{l_1b_1\alpha_1}...\sum_{l_nb_n\alpha_n}\tilde{\Phi}^{\alpha_1,...,\alpha_n}_{l_1b_1,...,l_nb_n}u(l_1b_1\alpha_1)u(l_2b_2\alpha_2)...u(l_nb_n\alpha_n). \tag{5.15}$$

Substituting Eq. 5.8 into the equation above, we obtain

$$E^{(n)}(\boldsymbol{u}) = \frac{1}{n!}\sum_{l_1b_1\alpha_1}u(l_1b_1\alpha_1)...\sum_{l_nb_n\alpha_n}u(l_nb_n\alpha_n)\sum_{\xi=1}^{N_c^{(n)}}\frac{\lambda_\xi}{n!}\sum_{\sigma\in S_n}\sum_{l}\prod_{i=1}^{n}A^{\xi}_{\sigma_i}(l_i-l,b_i\alpha_i). \tag{5.16}$$

Exchanging the order of $\prod$ and $\sum$ in Eq. (5.10), we rewrite this equation as

$$E^{(n)}(\boldsymbol{u}) = \sum_{l}\sum_{\xi=1}^{N_c^{(n)}}\frac{\lambda_\xi}{n!}\frac{1}{n!}\sum_{\sigma\in S_n}\prod_{i=1}^{n}\left(\sum_{l_ib_i\alpha_i}A^{\xi}_{\sigma_i}(l_i-l,b_i\alpha_i)u(l_ib_i\alpha_i)\right). \tag{5.17}$$

To simplify this expression, we define the projections appearing in Eq. (7) of main text:

$$\phi_i^{\xi}(l,\boldsymbol{u}) = \langle \boldsymbol{A}_i^{\xi}(l),\boldsymbol{u}\rangle = \sum_{l'b\alpha}A^{\xi}_{\sigma_i}(l'-l,b\alpha)u(l'b\alpha). \tag{5.18}$$

In terms of these quantities, the nth-order anharmonic energy $E^{(n)}(\boldsymbol{u})$ becomes

$$E^{(n)}(\boldsymbol{u}) = \sum_{l}\sum_{\xi=1}^{N_c^{(n)}}\frac{\lambda_\xi}{n!}\frac{1}{n!}\sum_{\sigma\in S_n}\prod_{i=1}^{n}\phi^{\xi}_{\sigma_i}(l,\boldsymbol{u}). \tag{5.19}$$

Since $\prod_{i=1}^{n}\phi^{\xi}_{\sigma_i}(l,\boldsymbol{u})$ is independent of σ, we can further simplify this expression to

$$E^{(n)}(\boldsymbol{u}) = \sum_{l}\sum_{\xi=1}^{N_c^{(n)}}\frac{\lambda_\xi}{n!}\prod_{i=1}^{n}\phi_i^{\xi}(l,\boldsymbol{u}), \tag{5.20}$$

which is Eq. (6) in the main text.

Computational cost analysis

From a computational viewpoint, Eq. (2) in main text can greatly accelerate calculations of ph-ph interactions. The main bottleneck in such calculations is computing the n-ph coupling $V^{(n)}(\mathbf{Q}_1, \mathbf{Q}_2, ..., \mathbf{Q}_n)$ for many scattering channels starting from the n-IFC, $\Phi^{(n)}$. When using the uncompressed n-ph interactions in Eq. (1), the computational cost scales as $O(N_\Phi N_{\text{channel}})$, where N_Φ is the number of non-zero entries in $\Phi^{(n)}$ and N_{channel} is the number of active scattering channels in $V^{(n)}(\mathbf{Q}_1, \mathbf{Q}_2, ..., \mathbf{Q}_n)$. In contrast, utilizing the compressed n-ph interaction in Eq. (2), the same calculation has a cost that scales as $O(N_c^{(n)} N_{\text{channel}})$, where $N_c^{(n)}$ is the PCP rank. Therefore, the computational cost savings from using Eq. (1) instead of Eq. (2) is proportional to the compression factor, $\gamma^{(n)} = N_\Phi / N_c^{(n)}$. In Fig. 4 of main text, we compare calculations using full (uncompressed) and compressed n-ph interactions. The two sets of calculations are identical except for the method used to calculate $V^{(n)}(\mathbf{Q}_1, \mathbf{Q}_2, ..., \mathbf{Q}_n)$. Therefore, the computational speed-up in Fig. 4 is entirely due to the more efficient computation of $V^{(n)}(\mathbf{Q}_1, \mathbf{Q}_2, ..., \mathbf{Q}_n)$ when using compressed n-ph interactions.

In thermal conductivity calculations, the speed-up analysis is more complex. The total CPU time for computing phonon scattering rates (and thus the thermal conductivity) can be split into two contributions, with respective computational cost:

T_1 = cost of evaluating the n-phonon interactions $V^{(n)}(\mathbf{Q}_1, \mathbf{Q}_2, ..., \mathbf{Q}_n)$,

T_2 = cost of computing the scattering rates from those interaction tensors.

Our PCP method only accelerates T_1, reducing it to $\frac{T_1}{k\gamma_n}$, where γ_n is the compression factor and k is a small pre-factor (in practice $k \approx 2$). The cost T_2 is left unchanged. Therefore, the speedup for the phonon scattering rate calculation, defined as the ratio of CPU wall times, is

$$\text{speedup} \;=\; \frac{T_1 + T_2}{T_1/(k\,\gamma_n) + T_2}.$$

In conventional calculations of phonon scattering rates, evaluating n-phonon interactions is the bottleneck and T_2 is negligible. However, since our PCP approach reduces T_1 very efficiently, T_2 eventually becomes non-negligible. Consequently, there is an upper bound for the speedup:

$$\text{speedup} \;<\; \frac{T_1 + T_2}{T_2} \;\approx\; \frac{T_1}{T_2}.$$

For example, in the case of Si, $T_1/(k\,\gamma_n)$ is smaller than T_2, so the observed speedup is dominated by the fixed cost T_2 and thus it is close to the upper bound:

$$\text{speedup} \;=\; \frac{T_1 + T_2}{T_1/(k\,\gamma_n) + T_2} \;\approx\; \frac{T_1 + T_2}{T_2} \;\approx\; 260.$$

For all cubic materials with small unit cells discussed in the manuscript, the wall time of the calculation is dominated by the fixed cost T_2, and thus it is close to the upper bound T_1/T_2. For ZrO_2, which has 12 atoms in the unit cell, we use a much larger PCP rank ($N_c = 144$), so that $T_1/(k\,\gamma_n) > T_2$ and the observed speedup approaches a different limit where the fixed overhead cost T_2 becomes irrelevant:

$$\text{speedup} \;=\; \frac{T_1 + T_2}{T_1/(k\,\gamma_n) + T_2} \;\approx\; k\,\gamma_3 \;\approx\; 7200.$$

This value reflects the true speedup of our PCP approach, which is close to twice the compression ratio: $k \approx 2$, $\gamma_3 \approx 4000$, and thus a speedup close to 8000.

DFT calculations and force constants generation

We generate displacement-force datasets using the ALAMODE package [3]. For all materials studied here, we sample 150 random configurations at 300 K using the harmonic phonon dispersion, and calculate forces for these randomly sampled configurations with the PBEsol functional [41] using VASP [39, 40] with a 2×2×2 **k**-point grid for all supercells. To extract the IFCs, we employ the adaptive LASSO solver implemented in ALAMODE [3], with hyperparameters optimized by the cross-validation method. For Si, we use a 5×5×5 supercell containing 250 atoms, with a unit cell lattice constant of 5.43 Å and a kinetic energy cutoff of 500 eV. For HgTe, we use a 4×4×4 supercell containing 128 atoms, with a unit cell lattice constant of 6.52 Å and a kinetic energy cutoff of 300 eV. For MgO, we use a 3×3×3 supercell containing 250 atoms, with a unit cell lattice constant of 4.25 Å and a kinetic energy cutoff of 400 eV. For TiNiSn, we use a 2×2×2 supercell containing 96 atoms, with a unit cell lattice constant of 5.87 Å and a kinetic energy cutoff of 500 eV. For ZrO_2, we use a 2×2×2 supercell containing 96 atoms, with a relaxed unit cell lattice constant of $a = 4.94$Å, $b = 5.16$Å, $c = 5.08$Å and a kinetic energy cutoff of 350 eV.

Phonon scattering rate calculations

We compute the phonon scattering rate as the sum of 3-ph and 4-ph scattering rates [58]. The 3-ph scattering rate $\tau^{-1}_{3ph,\mathbf{Q}}$, for a phonon mode $\mathbf{Q} = \nu\boldsymbol{q}$, reads

$$\tau^{-1}_{3ph,\mathbf{Q}} = \frac{\pi\hbar}{4N_{\boldsymbol{q}}} \sum_{\mathbf{Q}_1} \sum_{\mathbf{Q}_2} \left|V^{(3)}(\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3)\right|^2 \frac{1}{\omega_{\mathbf{Q}}\omega_{\mathbf{Q}_1}\omega_{\mathbf{Q}_2}} \tag{5.21}$$
$$\times \left[\frac{1}{2}(1 + n_{\mathbf{Q}_1} + n_{\mathbf{Q}_2})\delta(\omega_{\mathbf{Q}} - \omega_{\mathbf{Q}_1} - \omega_{\mathbf{Q}_2}) + (n_{\mathbf{Q}_1} - n_{\mathbf{Q}_2})\delta(\omega_{\mathbf{Q}} + \omega_{\mathbf{Q}_1} - \omega_{\mathbf{Q}_2})\right].$$

The 4-ph scattering rate $\tau^{-1}_{4ph,\mathbf{Q}}$, for a phonon mode $\mathbf{Q} = \nu\boldsymbol{q}$, reads

$$\tau^{-1}_{4ph,\mathbf{Q}} = \frac{\pi\hbar}{4N_{\boldsymbol{q}}} \frac{\hbar}{2N_{\boldsymbol{q}}} \sum_{\mathbf{Q}_1} \sum_{\mathbf{Q}_2} \sum_{\mathbf{Q}_3} \left|V^{(4)}(\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3, \mathbf{Q}_4)\right|^2 \frac{1}{\omega_{\mathbf{Q}}\omega_{\mathbf{Q}_1}\omega_{\mathbf{Q}_2}\omega_{\mathbf{Q}_3}} \tag{5.22}$$
$$\times \left[\frac{1}{6}\frac{n_{\mathbf{Q}_1} n_{\mathbf{Q}_2} n_{\mathbf{Q}_3}}{n_{\mathbf{Q}}}\delta(\omega_{\mathbf{Q}} - \omega_{\mathbf{Q}_1} - \omega_{\mathbf{Q}_2} - \omega_{\mathbf{Q}_3})\right.$$
$$+ \frac{1}{2}\frac{(1 + n_{\mathbf{Q}_1}) n_{\mathbf{Q}_2} n_{\mathbf{Q}_3}}{n_{\mathbf{Q}}}\delta(\omega_{\mathbf{Q}} + \omega_{\mathbf{Q}_1} - \omega_{\mathbf{Q}_2} - \omega_{\mathbf{Q}_3})$$
$$\left. + \frac{1}{2}\frac{(1 + n_{\mathbf{Q}_1})(1 + n_{\mathbf{Q}_2}) n_{\mathbf{Q}_3}}{n_{\mathbf{Q}}}\delta(\omega_{\mathbf{Q}} + \omega_{\mathbf{Q}_1} + \omega_{\mathbf{Q}_2} - \omega_{\mathbf{Q}_3})\right],$$

where $N_{\boldsymbol{q}}$ is the number of $\mathbf{q}$-points in the phonon momentum grid, $\omega_{\mathbf{Q}}$ is the energy of the phonon mode $\mathbf{Q}$ and $n_{\mathbf{Q}}$ the corresponding Bose-Einstein thermal occupation. The delta function $\delta(x)$ is approximated by a Gaussian function $\mathrm{e}^{-(x/\epsilon)^2}$, where ϵ is the Gaussian smearing parameter. We set ϵ = 0.5 meV, 0.2 meV, 0.2 meV, 0.2 meV, 1.0 meV for Si, HgTe, MgO, TiNiSn and ZrO_2, respectively. We employ the random sampling technique in Ref. [49] to speed up the integration. For Si, HgTe, MgO and TiNiSn, we sample $2{\times}10^5$ scattering processes for both 3- and 4-ph scattering rates for each $\tau^{-1}_{\mathbf{Q}}$, which ensures convergence of the phonon scattering rates. For ZrO_2, we sample $4{\times}10^4$ scattering processes for 3-ph interactions and $4{\times}10^5$ scattering processes for 4-ph interactions.

Constrained optimization

For acoustic phonons, the ph-ph interactions vanish in the long-wavelength limit due to the acoustic sum rule (ASR). However, the thermal occupation of acoustic phonons can be large at finite temperature and diverge in the long-wavelength limit. This interplay results in acoustic phonon scattering being non-negligible, even though their interaction matrix elements are vanishingly small. We address this problem using a constrained optimization approach inspired by a method we previously developed for compressing e-ph interactions using SVD [27]. The ASR

for the 3-IFCs gives the constraint

$$\sum_{l_1 b_1} \Phi^{\alpha_1,\alpha_2,\alpha_3}_{l_1 b_1, l_2 b_2, l_3 b_3} = 0. \tag{5.23}$$

To enforce the acoustic sum rule [29], we impose $\sum_{l,b} A_i^{\xi}(l, b\alpha) = 0$. This ASR leads to the following expansion in the long-wavelength limit ($\boldsymbol{q} \to 0$),

$$\sum_{l_1 b_1} e^{i\boldsymbol{q}_1 \boldsymbol{r}_{l_1}} \Phi^{\alpha_1,\alpha_2,\alpha_3}_{l_1 b_1, l_2 b_2, l_3 b_3} \approx i\boldsymbol{q}_1 \cdot \sum_{l_1 b_1} \Phi^{\alpha_1,\alpha_2,\alpha_3}_{l_1 b_1, l_2 b_2, l_3 b_3} \boldsymbol{r}_{l_1} \approx i\boldsymbol{q}_1 \cdot \boldsymbol{F}^{\alpha_1,\alpha_2,\alpha_3}_{l_2 b_2, l_3 b_3}[\Phi], \tag{5.24}$$

where

$$\boldsymbol{F}^{\alpha_1,\alpha_2,\alpha_3}_{l_2 b_2, l_3 b_3}[\Phi] = \sum_{l_1 b_1} \Phi^{\alpha_1,\alpha_2,\alpha_3}_{l_1 b_1, l_2 b_2, l_3 b_3} \boldsymbol{r}_{l_1} \tag{5.25}$$

is the ph-ph deformation potential, which is analogous to the e-ph deformation potentials discussed in Refs. [54, 55].

To preserve the ph-ph deformation potential $\boldsymbol{F}$, we introduce a Lagrange multiplier in the loss function, which becomes:

$$L = \|\tilde{\boldsymbol{\Phi}}[\boldsymbol{A}] - \boldsymbol{\Phi}\|^2 + \lambda_F \|\boldsymbol{F}[\tilde{\boldsymbol{\Phi}}[\boldsymbol{A}]] - \boldsymbol{F}[\boldsymbol{\Phi}]\|^2. \tag{5.26}$$

In our calculations, we set λ_F to a large value to numerically enforce the ph-ph deformation potential.

Phonon scattering rates for MgO and TiNiSn

Thermal conductivity results

We calculate the thermal conductivity κ at 300 K using momentum grids of 40^3, 30^3, 20^3, 20^3 and 10^3 for Si, HgTe, MgO, TiNiSn, and ZrO_2 respectively. In the following, we show the convergence of κ with respect to the PCP decomposition rank for the 3-ph and 4-ph interactions, respectively, in Tables 5.1 and 5.2. For $N_c^{(3)} = 24$ and $N_c^{(4)} = 48$, the thermal conductivity predicted using compressed IFC tensors is within 98% of the value obtained using full 3- and 4-IFC tensors for all materials studied here.

Extrapolation to the thermodynamic limit

Origin of the slow convergence of thermal conductivity for HgTe

In HgTe, the 4-ph interactions give an important contribution. Without 4-ph scattering, the thermal conductivity is overestimated by around 400%, as has been reported

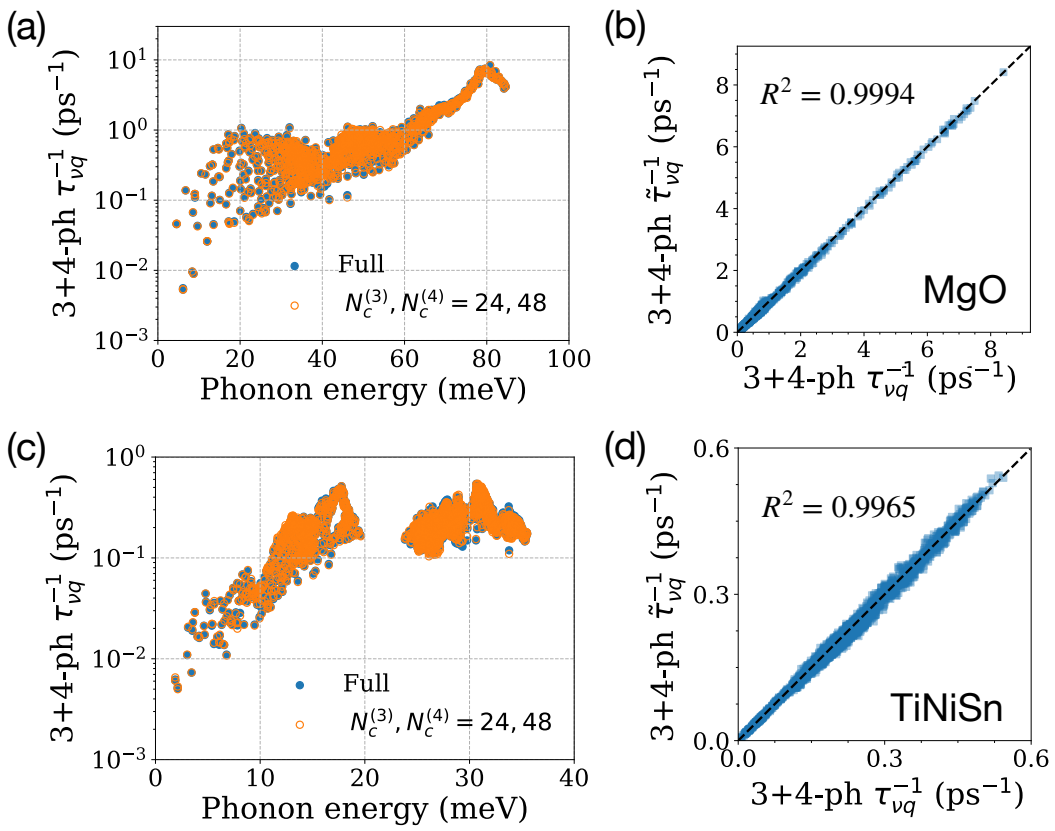

Figure 5.6: Phonon scattering rates as a function of phonon energy for (a) MgO and (c) TiNiSn. The plots compare results computed with full IFC tensors and compressed IFC tensors (we use $N_c^{(3)} = 24$ for MgO and $N_c^{(4)} = 48$ for TiNiSn). Corresponding parity plots comparing results from compressed vs. full IFC tensors for 3+4-phonon scattering rates are given the same materials in panels (b) and (d).

previously [9]. In figure. 5.8, we calculate the thermal conductivity as a function of grid size for two values of the smearing parameter, 0.2 meV and 0.8 meV, with and without 4-ph interaction. We find that the thermal conductivity without 4-ph interactions is greatly overestimated, as is seen by comparing the y-axis values in the two figure panels. Figure. 5.8(a) shows the calculation without 4-ph interactions, where the thermal conductivity is nearly converged for a grid of 20^3 and is close to the results extrapolated to the thermodynamic limit. This confirms that the 3-ph interactions converge rapidly with the grid size. In Fig. 5.8(b), in the calculation with 4-ph interactions, the convergence with grid size is slow and persists up to much larger smearing values of 0.8 meV. Therefore, we attribute the slow convergence of the thermal conductivity in HgTe to the importance of the 4-ph processes, which are

Table 5.1: Calculated thermal conductivity, in units of W/(mK), for Si, HgTe, MgO, TiNiSn. We compare reference results from full (uncompressed) n-IFCs (rightmost column) with calculations using compressed 3-ph and 4-ph interactions with different choices of PCP rank for 3-ph interactions. The PCP rank for 4-ph interactions is fixed to $N_c^{(4)} = 48$. Thermal conductivities that are within 98% of the uncompressed reference result are shown with bold font.

Crystals	$N_c^{(3)} = 4$	$N_c^{(3)} = 8$	$N_c^{(3)} = 16$	$N_c^{(3)} = 24$	Full $\mathbf{\Phi}^{(3)}$, $\mathbf{\Phi}^{(4)}$
Si	116.6	**131.6**	**130.7**	**130.2**	**130.9**
HgTe	2.13	2.13	**2.20**	**2.22**	**2.24**
MgO	59.0	44.7	**42.8**	**42.7**	**42.3**
TiNiSn	27.7	18.7	**15.1**	**15.0**	**15.0**

Table 5.2: Calculated thermal conductivity as in Table S1 above but with results given for different choices of PCP rank for 4-ph interactions while fixing the PCP rank for 3-ph interactions to $N_c^{(3)} = 24$. Bold font indicates thermal conductivities within 98% of the respective reference result obtained from full n-IFC tensors (rightmost column).

Crystals	$N_c^{(4)} = 4$	$N_c^{(4)} = 8$	$N_c^{(4)} = 16$	$N_c^{(4)} = 24$	$N_c^{(4)} = 32$	$N_c^{(4)} = 48$	Full $\mathbf{\Phi}^{(3)}$, $\mathbf{\Phi}^{(4)}$
Si	123.2	132.0	**130.7**	**130.2**	**130.3**	**130.2**	**130.9**
HgTe	1.79	2.14	2.04	2.03	2.19	**2.22**	**2.24**
MgO	43.8	**42.9**	**42.8**	**42.7**	**42.7**	**42.7**	**42.3**
TiNiSn	15.5	15.3	**15.0**	**15.0**	**15.0**	**15.0**	**15.0**

expected to be less smooth than the 3-ph scattering processes in the Brillouin zone and thus require denser grids to reach convergence.

Number of force constants for 3rd order force constants

In this section, we compare the compression factor for our PCP technique, $\gamma^{(3)}$ defined in the manuscript, with the size reduction (or "compression") resulting from symmetry, defined as

$$\gamma_{\text{sym}}^{(3)} = \frac{\|\Phi_{\text{full}}^{(3)}\|_0}{\|\Phi_{\text{irr}}^{(3)}\|_0}. \tag{5.27}$$

where $\|\boldsymbol{x}\|_0$ indicates the L_0 norm, which measures the number of nonzero elements, and $\Phi_{\text{irr}}^{(3)}$ is the irreducible IFC tensor, which accounts for the reduction of independent entries in the IFC tensor due to symmetry. In Table (5.3), we compare the PCP and symmetry-derived compression factors for all materials studied in this work, including ZrO_2. The size reduction deriving from symmetry is of order 30–300 for all materials, and thus much smaller than the PCP compression factors, which are

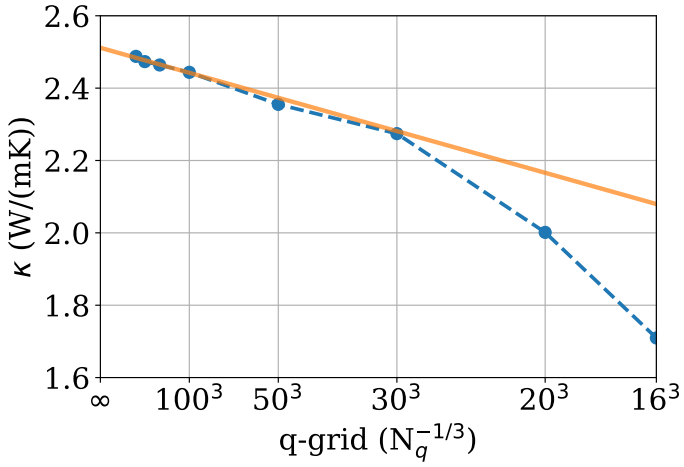

Figure 5.7: Extrapolation of the thermal conductivity for HgTe at 300 K to the thermodynamic limit (TDL). The largest grid size we computed is 250^3, which corresponds to the leftmost data point. We extrapolate to the TDL by fitting the thermal conductivity values for the four largest grid sizes using $\kappa(N_q) = \kappa(\infty) - bN_q^{-1/3}$, where N_q is the number of $\mathbf{q}$-points in the grid, and b is the slope. Using this procedure, we obtain an extrapolated value of $\kappa(\infty) = 2.5$ W/(mK), which is 47% larger than the value using a grid size of 16^3.

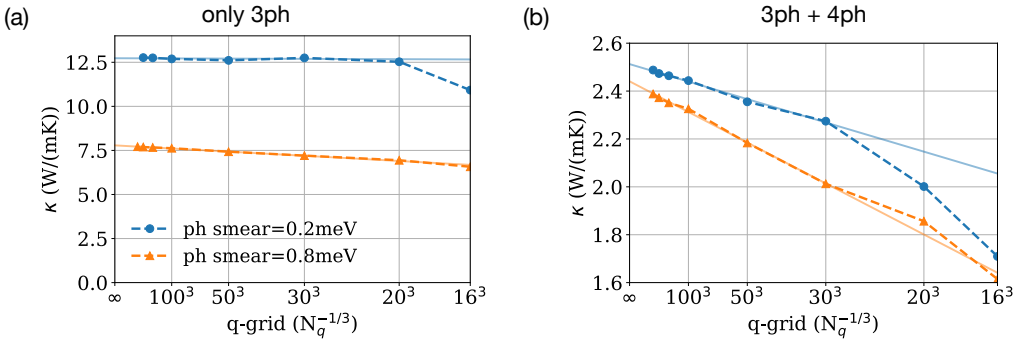

Figure 5.8: Convergence of the thermal conductivity for HgTe at 300 K with (a) 3-ph and (b) 3+4-ph interactions, using two smearing parameters. Blue and orange markers correspond to smearing values of 0.2 meV and 0.8 meV, respectively.

in the range of 2000–6000. In the case of monoclinic ZrO_2, the symmetry-derived compression factor is relatively small ($\gamma^{(3)}_{\text{sym}} = 36$) due to the low symmetry of ZrO_2, while the PCP compression is nearly 100 times greater (to be precise, $93 \approx 3358/36$ times greater). This means that our tensor learning approach becomes even more favorable for low-symmetry crystals.

Table 5.3: Summary of the compression factors of PCP methods and symmetrization for all the studied materials.

Crystals	bravis lattice	$\|\Phi^{(3)}_{\text{irr}}\|_0$	$\|\Phi^{(3)}_{\text{full}}\|_0$	PCP rank $N_c^{(3)}$	$\gamma^{(3)}$	$\gamma^{(3)}_{\text{sym}}$
Si	Cubic	174	54480	24	2270	313
HgTe	Cubic	383	54480	24	2270	142
MgO	Cubic	219	54480	24	2270	249
TiNiSn	Cubic	515	139230	24	5801	270
ZrO_2	Monoclinic	13354	483600	144	**3358**	**36**

Symmetry loss from compression

We quantify the symmetry loss resulting from compression for the i-th symmetry operation, $\hat{S}_i$, using

$$\epsilon_i = \frac{\sqrt{\sum_{\mathbf{Q}_1,\mathbf{Q}_2,\mathbf{Q}_3} \left|\tilde{V}^{(3)}(\mathbf{Q}_1,\mathbf{Q}_2,\mathbf{Q}_3) - \hat{S}_i[\tilde{V}^{(3)}](\mathbf{Q}_1,\mathbf{Q}_2,\mathbf{Q}_3)\right|^2}}{\sqrt{\sum_{\mathbf{Q}_1,\mathbf{Q}_2,\mathbf{Q}_3} \left|\tilde{V}^{(3)}(\mathbf{Q}_1,\mathbf{Q}_2,\mathbf{Q}_3)\right|^2}}, \tag{5.28}$$

where $\hat{S}_i[\tilde{V}^{(3)}]$ is the compressed 3-ph interaction after applying the symmetry operation $\hat{S}_i$. The symmetry transformation follows Ref. [56]. In Fig. 5.9, we show ϵ_i for each symmetry operation in the point group of Si. In all cases, the symmetry loss resulting from compression is less than 1%.

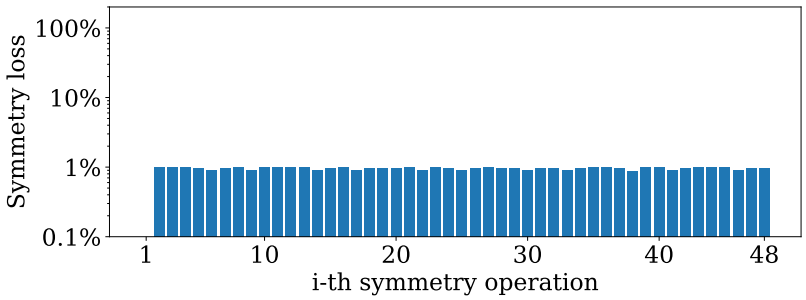

Figure 5.9: Relative symmetry loss for the compressed 3-ph interactions in Si, using a compression factor $\gamma^{(3)} = 2200$ ($N_c^{(3)} = 24$), given for each symmetry operations in the point group. In all cases, the symmetry loss, ε_i defined above, is less than 1%. The first symmetry operation is the identity, which is always satisfied exactly. We use SPGLIB [57] to find all the point-group symmetry operations.

References

[1] M. Born and K. Huang, *Dynamical Theory of Crystal Lattices* (Oxford University Press, 1996).

[2] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods* (Cambridge University Press, 2004).

[3] T. Tadano, Y. Gohda, and S. Tsuneyuki, J. Condens. Matter Phys. **26**, 225402 (2014).

[4] F. Zhou, W. Nielson, Y. Xia, and V. Ozoliņš, Phys. Rev. Lett. **113**, 185501 (2014).

[5] O. Hellman and I. A. Abrikosov, Phys. Rev. B **88**, 144301 (2013).

[6] A. H. Romero, E. K. U. Gross, M. J. Verstraete, and O. Hellman, Phys. Rev. B **91**, 214310 (2015).

[7] D. A. Broido, M. Malorny, G. Birner, N. Mingo, and D. A. Stewart, Appl. Phys. Lett. **91**, 231922 (2007).

[8] N. K. Ravichandran and D. Broido, Phys. Rev. X **10**, 021063 (2020).

[9] Y. Xia, V. I. Hegde, K. Pal, X. Hua, D. Gaines, S. Patel, J. He, M. Aykol, and C. Wolverton, Phys. Rev. X **10**, 041029 (2020).

[10] J. Carrete, W. Li, N. Mingo, S. Wang, and S. Curtarolo, Phys. Rev. X **4**, 011019 (2014).

[11] W. Li, J. Carrete, N. A. Katcho, and N. Mingo, Comp. Phys. Commun. **185**, 1747–1758 (2014).

[12] G. Barbalinardo, Z. Chen, N. W. Lundgren, and D. Donadio, J. Appl. Phys. **128**, 135104 (2020).

[13] J. I. Cirac, D. Pérez-García, N. Schuch, and F. Verstraete, Rev. Mod. Phys. **93**, 045003 (2021).

[14] R. Orús, Ann. Phys. **349**, 117 (2014).

[15] I. V. Oseledets, SIAM J. Sci. Comput. **33**, 2295 (2011).

[16] Y. Núñez Fernández, M. Jeannin, P. T. Dumitrescu, T. Kloss, J. Kaye, O. Parcollet, and X. Waintal, Phys. Rev. X **12**, 041018 (2022).

[17] R. D. Peddinti, S. Pisoni, A. Marini, P. Lott, H. Argentieri, E. Tiunov, and L. Aolita, Commun. Phys. **7**, 135 (2024).

[18] H. Shinaoka, M. Wallerberger, Y. Murakami, K. Nogaki, R. Sakurai, P. Werner, and A. Kauch, Phys. Rev. X **13**, 021015 (2023).

[19] E. G. Hohenstein, R. M. Parrish, and T. J. Martínez, J. Chem. Phys. **137**, 044103 (2012).

[20] R. M. Parrish, E. G. Hohenstein, T. J. Martínez, and C. D. Sherrill, J. Chem. Phys. **137**, 224106 (2012).

[21] E. G. Hohenstein, R. M. Parrish, C. D. Sherrill, and T. J. Martínez, J. Chem. Phys. **137**, 221101 (2012).

[22] B. I. Dunlap, Phys. Chem. Chem. Phys. **2**, 2113 (2000).

[23] M. J. Willatt, F. Musil, and M. Ceriotti, Phys. Chem. Chem. Phys. **20**, 29661 (2018).

[24] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, J. Chem. Phys. **148**, 241730 (2018).

[25] R. K. Cersonsky, B. A. Helfrecht, E. A. Engel, S. Kliavinek, and M. Ceriotti, Mach. Learn.: Sci. Technol. **2**, 035038 (2021).

[26] J. P. Darby, J. R. Kermode, and G. Csányi, Npj Comput. Mater. **8**, 166 (2022).

[27] Y. Luo, D. Desai, B. K. Chang, J. Park, and M. Bernardi, Phys. Rev. X **14**, 021023 (2024).

[28] L. Fu, M. Kornbluth, Z. Cheng, and C. A. Marianetti, Phys. Rev. B **100**, 014303 (2019).

[29] K. Esfarjani and H. T. Stokes, Phys. Rev. B **77**, 144112 (2008).

[30] F. Zhou, W. Nielson, Y. Xia, and V. Ozoliņš, Phys. Rev. B **100**, 184308 (2019).

[31] T. G. Kolda and B. W. Bader, SIAM Rev. **51**, 455 (2009).

[32] V. Lebedev, Y. Ganin, M. Rakhuba, I. Oseledets, and V. Lempitsky, arXiv:1412.6553 [cs.CV] (2015).

[33] T. Lacroix, N. Usunier, and G. Obozinski, in *Proceedings of the 35th International Conference on Machine Learning*, Proc. Mach. Learn. Res., Vol. 80, edited by J. Dy and A. Krause (PMLR, 2018) pp. 2863–2872.

[34] K. Pierce and M. Morales, arXiv:2502.17683 [physics.chem-ph] (2025).

[35] See Supplemental Material for: derivations of n-ph interactions and their PCP decomposition, derivation of n-ph anharmonic energy in Eq. (6), computational cost analysis, DFT calculations and IFC generation, comparisons with compression from symmetrization, phonon scattering rate calculations, constrained optimization for acoustic modes, phonon scattering rates for MgO and TiNiSn, thermal conductivity for all studied materials, extrapolation of κ to the TDL in HgTe, origin of the slow convergence of thermal conductivity for HgTe, number of force constants for 3rd order force constants, and symmetry loss from compression in Si.

[36] In the optimization, $A_i^{\xi}(l, b\alpha)$ is truncated in real space up to a cutoff r_c large enough for the low-rank IFC tensor $\tilde{\mathbf{\Phi}}^{(n)}$ to span all nonzero elements in the full IFC tensor $\mathbf{\Phi}^{(n)}$.

[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, Pytorch: an imperative style, high-performance deep learning library (Curran Associates Inc., Red Hook, NY, USA, 2019).

[38] We include 3-IFCs up to the fifth-nearest neighbors and 4-IFCs up to the second-nearest neighbors for Si, HgTe and MgO. For TiNiSn, the 3-IFCs are truncated to 6.4 Å and the 4-IFCs to 3.8 Å. For ZrO_2, the 3-IFCs are truncated to 5.3 Å and the 4-IFCs to 2.7 Å.

[39] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).

[40] G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).

[41] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, Phys. Rev. Lett. **100**, 136406 (2008).

[42] X. Gonze and C. Lee, Phys. Rev. B **55**, 10355 (1997).

[43] R. Peierls, *Quantum Theory of Solids* (Clarendon Press, Oxford, 1955) p. 40.

[44] J. M. Ziman, *Electrons and Phonons* (Oxford University Press, London, 1960) p. 298.

[45] A. A. Maradudin and A. E. Fein, Phys. Rev. **128**, 2589 (1962).

[46] R. S. Tripathi and K. N. Pathak, Il Nuovo Cimento B **21**, 289 (1974).

[47] M. Balkanski, R. F. Wallis, and E. Haro, Phys. Rev. B **28**, 1928 (1983).

[48] T. Feng and X. Ruan, Phys. Rev. B **93**, 045202 (2016).

[49] Z. Guo, Z. Han, D. Feng, G. Lin, and X. Ruan, Npj Comput. Mater. **10**, 31 (2024).

[50] Coefficient of determination, in *The Concise Encyclopedia of Statistics* (Springer New York, New York, NY, 2008) pp. 88–91.

[51] X. Ai, Y. Chen, and C. A. Marianetti, Phys. Rev. B **90**, 014308 (2014).

[52] Y. Chen, X. Ai, and C. A. Marianetti, Phys. Rev. Lett. **113**, 105501 (2014).

[53] R. Drautz, Phys. Rev. B **99**, 014104 (2019).

[54] J. Bardeen and W. Shockley, Phys. Rev. **80**, 72 (1950).

[55] C. Herring and E. Vogt, Phys. Rev. **101**, 944 (1956).

[56] A. A. Maraduin and S. H. Vosko, Rev. Mod. Phys. **40**, 1 (1968).

[57] K. S. Atsushi Togo and I. Tanaka, Sci. Technol. Adv. Mater., Meth. **4**, 2384822 (2024).

[58] T. Feng, L. Lindsay, and X. Ruan, Phys. Rev. B **96**, 161201 (2017).

Chapter 6

COHERENT PHONON DYNAMICS

6.1 Introduction

Nonlinear phononics, where one vibrational mode of a crystal drives or mixes with another, has become a central route to steering solids far from equilibrium in a symmetry-selective way. Intense mid-infrared or THz pulses can resonantly excite an infrared-active phonon and, through anharmonic couplings, transiently reshape the lattice potential experienced by other modes. This mechanism underlies a growing class of light-induced phenomena: structural chirality control [1], transient paraelectric-ferroelectric switching [2], magnetic phase control [3], insulator-metal transitions[4], and reports of superconducting-like signatures at temperatures far above equilibrium critical values[5]. The common feature across these examples is that the experimentally relevant control is not merely a change in phonon populations, but a coherent, phase-defined lattice displacement that can selectively break or restore symmetries on ultrafast timescales.

The rt-BTE describes electrons and phonons through occupations. However, a purely population-based description cannot represent coherent lattice motion: a coherent phonon corresponds to a nonzero expectation value of a displacement operator, which is invisible to phonon population alone. In pump-probe, time-resolved diffraction, and nonlinear optical probes, the measured signal is often directly proportional to lattice displacements, making it essential to compute the coherent coordinate itself.

This chapter develops a microscopic and phase-consistent framework for coupled coherent and incoherent electron-phonon dynamics. The guiding goal is to connect three ingredients within a single set of equations of motion (EOMs): electronic populations, phonon populations, and phonon coherences. The central object that accounts for coherence is the phonon displacement coordinate $\langle Q_{\mathbf{q}\nu}(t)\rangle$, derived from creation/annihilation operators. In the incoherent limit, this quantity vanishes, and the formalism reduces to the standard rt-BTE. Under coherent driving, however, $\langle Q_{\mathbf{q}\nu}(t)\rangle$ represents a macroscopic lattice displacement with a definite phase.

Microscopic NEGF frameworks provide a rigorous treatment of e-ph interactions and phonon coherence in nonequilibrium [6]. However, the work does not rely on

a NEGF formulation. Instead, this chapter derives the coupled dynamics directly from the Heisenberg EOMs for electron and phonon operators, allowing phonon coherence to be retained explicitly at the level of operator expectation values. This approach provides a transparent microscopic route to equations of motion for electronic populations, phonon populations, and phonon coherences, while avoiding the technical overhead associated with two-time Green's functions.

Within this Heisenberg formulation, e-ph scattering is treated in a phase-consistent Markov approximation. We show that, when phonon coherence is neglected, the resulting equations reduce to the standard rt-BTE with Fermi's golden rule scattering rates. When phonon coherence is retained, additional phase factors appear in the Markov reduction, and careful accounting of these phases is required to obtain physically meaningful resonance conditions. With the phase-consistent prescription developed here, the electron and phonon population equations recover the correct physical limits and agree with results obtained from the mirrored generalized Kadanoff-Baym ansatz (MGKBA) [6]. In this sense, the present approach reproduces the established e-ph and ph-e kinetics while making the role of phonon coherence explicit and transparent.

Ph-ph interactions are not included at the same microscopic level within the Heisenberg scattering formalism presented here. Their fully phase-resolved incorporation remains an open problem and is deferred to future work. To capture the leading effects of anharmonicity on coherent lattice dynamics, we also explore a complement treatment, a Liouville-von Neumann formulation of phonon dynamics. At lowest order, this provides a consistent description of ph-ph scattering, phonon dephasing, and energy relaxation.

Finally, to enable practical simulations of nonlinear-phononics experiments, we introduce an efficient computational framework using a controlled separation between coherent and thermal phonons. Coherent, optically driven modes are described by a classical EOM for the phonon displacement, while all other phonons are treated within the rt-BTE. Under this approximation, coherent lattice dynamics, e-ph scattering, and ph-ph relaxation are coupled in a way that preserves the correct limiting behavior and global energy conservation. Together, these elements provide a unified and computationally tractable framework for studying coherent phonon dynamics in driven solids, bridging microscopic scattering theory and experimentally relevant observables.

6.2 Electron and phonon dynamics from Heisenberg equations of motion

Electron, phonon and Hamiltonian and EOMs

The noninteracting electronic and phononic Hamiltonians are, respectively [7],

$$H_e = \sum_{kn} \epsilon_{kn} c^\dagger_{kn} c_{kn}, \tag{6.1}$$

and

$$H_p = \sum_{q\nu} \hbar\omega_{q\nu} b^\dagger_{q\nu} b_{q\nu}. \tag{6.2}$$

The electron-phonon interaction Hamiltonian is [7]

$$H_{e-p} = \sum_{k,m,n,q,\nu} g_{mn\nu}(k,q) c^\dagger_{k+qm} c_{kn} (b_{q\nu} + b^\dagger_{-q\nu}) \tag{6.3}$$

Here, the **k** and **q** are electron and phonon wave vectors, respectively; n and ν are band indices for electrons and phonons, respectively; $c^\dagger_{kn}(c_{kn})$ is the creation (annihilation) operator for an electron in state $|kn\rangle$; $b^\dagger_{q\nu}(b_{q\nu})$ is the creation (annihilation) operator for a phonon in mode $|q\nu\rangle$; ϵ_{kn} is the energy of the electronic state $|kn\rangle$; and $\omega_{q\nu}$ is the frequency of the phonon mode $|q\nu\rangle$.

We define the electron density matrix as

$$\rho_{nm}(k;Q) \equiv \langle c^\dagger_{k+Q,n} c_{k,m} \rangle, \tag{6.4}$$

and occupation as

$$f_{nk} \equiv \rho_{nn}(k;0) = \langle c^\dagger_{nk} c_{nk} \rangle. \tag{6.5}$$

And the phonon occupation is defined as

$$n_{q\nu} \equiv \langle b^\dagger_{q\nu} b_{q\nu} \rangle, \tag{6.6}$$

and phonon coherence as

$$\Theta^{\nu'\nu}_{q'q}(t) = \langle b_{q'\nu'} b_{q\nu} \rangle - \langle b_{q'\nu'} \rangle \langle b_{q\nu} \rangle. \tag{6.7}$$

In this work, we focus on the diagonal part of the electron density matrix, i.e., $k' = k + q'$ and $\alpha = \beta$, and the phonon coherence at the same mode, i.e., $q' = -q$ and $\nu' = \nu$. For brevity, the phonon coherence at the same mode is written as

$$\Theta_{q\nu} = \langle b_{q\nu} b_{-q\nu} \rangle - \langle b_{q\nu} \rangle \langle b_{-q\nu} \rangle. \tag{6.8}$$

The EOM of the electronic density matrix under the Heisenberg picture is

$$\begin{aligned}\frac{d\langle c^\dagger_{k'+q'\beta}c_{k'\alpha}\rangle}{dt} = -\frac{i}{\hbar}\Big[\sum_{n,q,\nu} g_{\alpha n\nu}(k'-q,q)(\langle c^\dagger_{k'+q'\beta}c_{k'-qn}b_{q\nu}\rangle \\ + \langle c^\dagger_{k'+q'\beta}c_{k'-qn}b^\dagger_{-q\nu}\rangle) - g_{n\beta\nu}(k'+q',q)(\langle c^\dagger_{k'+q'+qn}c_{k'\alpha}b_{q\nu}\rangle \\ + \langle c^\dagger_{k'+q'+qn}c_{k'\alpha}b^\dagger_{-q\nu}\rangle)\Big].\end{aligned} \tag{6.9}$$

Define $s^{q\nu}_{mk',nk} \equiv \langle b_{q\nu}c^\dagger_{k'm}c_{kn}\rangle$, and $\delta s^{q\nu}_{mk',nk} = s^{q\nu}_{mk',nk} - \langle b_{q\nu}\rangle\langle c^\dagger_{k'm}c_{kn}\rangle$. We obtain the EOM for $\delta s^{q'\nu'}_{k''m',k'n'}$. (See Appendix A for detailed derivations.) Disgarding higher-order terms in the form of $\delta s\langle b\rangle$ that describe energy renormalization, we have

$$\begin{aligned}\frac{d\,\delta s^{q'\nu'}_{k''m',k'n'}}{dt} = -\frac{i}{\hbar}\Big\{&\delta s^{q'\nu'}_{k''m',k'n'}(\epsilon_{k'n'} - \epsilon_{k''m'} + \hbar\omega_{q'\nu'}) \\ &- \sum_{k,m,n} g_{mn\nu'}(k,-q')\langle c^\dagger_{k-q'm}c_{k'n'}\rangle\langle c^\dagger_{k''m'}c_{kn}\rangle \\ &+ \sum_{n,q,\nu} g_{n'n\nu}(k'-q,q)(\langle b_{q'\nu'}b_{q\nu}\rangle - \langle b_{q'\nu'}\rangle\langle b_{q\nu}\rangle)\langle c^\dagger_{k''m'}c_{k'-qn}\rangle \\ &+ \sum_{n} g_{n'n\nu'}(k'+q',-q')(n_{\nu'q'}+1)\langle c^\dagger_{k''m'}c_{k'+q'n}\rangle \\ &- \sum_{n,q,\nu} g_{nm'\nu}(k'',q)(\langle b_{q'\nu'}b_{q\nu}\rangle - \langle b_{q'\nu'}\rangle\langle b_{q\nu}\rangle)\langle c^\dagger_{k''+qn}c_{k'n'}\rangle \\ &- \sum_{n} g_{nm'\nu'}(k'',-q')n_{\nu'q'}\langle c^\dagger_{k''-q'n}c_{k'n'}\rangle\Big\}.\end{aligned} \tag{6.10}$$

Phase-consistent Markov approximation of phonon-coherence contribution

To obtain closed equations of motion from the Heisenberg formulation, we invoke a Markovian approximation when formally integrating Eq. 6.10. The imaginary (on-shell) part of the resulting memory integral yields dissipative scattering terms, corresponding to the collision integral in a Boltzmann description, while the real (off-shell) part produces energy renormalization effects, which are neglected here for simplicity. The Markov kernel is defined as [7]

$$\mathcal{K}(\Delta) \equiv \mathcal{P}\frac{1}{\Delta} - i\pi\,\delta(\Delta), \tag{6.11}$$

where $\mathcal{P}$ denotes the Cauchy principal value.

In practice, the application of the Markov kernel depends on how two-time quantities are reduced to a single-time description. In the absence of phonon coherence, the

oscillatory phases associated with the two-time structure cancel under the Markov approximation, leading to the usual Fermi's golden rule expressions for e-ph scattering. When phonon coherence is retained, however, additional phase factors appear and must be accounted for explicitly in order to obtain physically meaningful results. This issue is closely related to the ambiguity encountered in single-time closures of the Kadanoff-Baym equations using the generalized Kadanoff-Baym ansatz (GKBA) [6, 8].

At the level of the Heisenberg equations, the Markov reduction is controlled by the net oscillatory phase appearing in the formal solution of mixed e-ph correlators. Denoting such a correlator schematically by δs, its formal solution may be written as

$$\delta s(t) = \int_0^\infty d\tau \; e^{-i\Omega\tau} \, \mathcal{S}(t-\tau), \tag{6.12}$$

where Ω collects the phase generated by the homogeneous part of the equation of motion, and $\mathcal{S}$ is a source term.

For correlators containing a single phonon annihilation operator $b_{q'\nu'}$, the homogeneous evolution contributes a frequency $+\omega_{q'\nu'}$, so that

$$\Omega = \frac{\epsilon_{kn} - \epsilon_{k+q'm}}{\hbar} + \omega_{q'\nu'}. \tag{6.13}$$

If the source $\mathcal{S}$ does not contain additional coherent phases, the Markov approximation yields the on-shell condition $\delta(\Omega)$.

When the source contains a two-annihilator phonon coherence, $\langle b_{q'\nu'} b_{q\nu} \rangle$, its free evolution carries an intrinsic oscillatory phase. Retaining this phase prior to applying the Markov kernel shifts the effective detuning in Eq. 6.12, yielding an on-shell contribution of the form

$$\delta\big[\Omega - (\omega_{q'\nu'} + \omega_{q\nu})\big]. \tag{6.14}$$

As a result, the phonon frequency entering the resonance condition differs from that obtained for population-driven processes.

In the physically relevant same-mode limit, $q' = q$ and $\nu' = \nu$, the coherence contributes a phase $2\omega_{q\nu}$. With Ω defined by Eq. 6.13, the on-shell condition reduces to

$$\delta\left(\frac{\epsilon_{kn} - \epsilon_{k+qm}}{\hbar} - \omega_{q\nu}\right), \tag{6.15}$$

demonstrating that a term whose homogeneous evolution contains $+\omega_{q\nu}$ yields a resonance condition involving $-\omega_{q\nu}$ once the phonon-coherence phase is properly retained.

Practical rule. *In the Markov limit, anomalous phonon coherences* $\langle bb\rangle$ *and* $\langle b^\dagger b^\dagger\rangle$ *shift the effective detuning by* $\pm(\omega_{q'\nu'}+\omega_{q\nu})$*, which in the same-mode limit induces an explicit swap* $\delta(\Delta+\hbar\omega)\leftrightarrow\delta(\Delta-\hbar\omega)$ *in the coherence-driven channels, with the conjugate coherence producing the opposite swap so that the total contribution remains Hermitian.*

The EOMs with phonon coherence

The intermediate results of deriving these EOMs can be found in Appendix A. Here we only present the final results.

Electron occupation EOM

Omitting electron coherences, and only considering phonon coherences at the same mode, the EOM of the electronic occupation $f_{nk}=\rho_{nn}(k;0)$ is given by

$$\begin{aligned}\frac{df_{nk}}{dt}=&\frac{2\pi}{\hbar}\sum_{q,\nu,n'}|g_{nn'\nu}(k,q)|^2\Big\{\Big[(n_{q\nu}+1)f_{n',k+q}(1-f_{nk})-n_{q\nu}f_{nk}(1-f_{n',k+q})\\&+\mathrm{Re}\big[\Theta_{q\nu}\big](f_{n'k+q}-f_{nk})\Big]\delta(\epsilon_{n',k+q}-\epsilon_{nk}-\hbar\omega_{q\nu})\\&\Big[n_{q\nu}f_{n',k+q}(1-f_{nk})-(n_{q\nu}+1)f_{nk}(1-f_{n',k+q})\\&+\mathrm{Re}\big[\Theta_{q\nu}\big](f_{n',k+q}-f_{nk})\Big]\delta(\epsilon_{nk}-\epsilon_{n',k+q}-\hbar\omega_{q\nu})\Big]\Big\},\end{aligned}\tag{6.16}$$

where $\Theta_{q\nu}=\langle b_{q\nu}b_{-q\nu}\rangle-\langle b_{q\nu}\rangle\langle b_{-q\nu}\rangle$. Notice that the phase change associated with $\mathrm{Re}[\Theta_{q\nu}]$ discussed in the previous section does not affect the result for electron occupation EOM.

Phonon occupation EOM

Similarly the phonon occupation EOM is

$$\frac{d\langle b^\dagger_{q'\nu'}b_{q'\nu'}\rangle}{dt}|=\frac{d\langle b^\dagger_{q'\nu'}b_{q'\nu'}\rangle}{dt}|_{1st}+\frac{d\langle b^\dagger_{q'\nu'}b_{q'\nu'}\rangle}{dt}|_{2nd},\tag{6.17}$$

where

$$\begin{aligned}\frac{d\langle b^\dagger_{q'\nu'}b_{q'\nu'}\rangle}{dt}|_{1st}=-\frac{i}{\hbar}\sum_{k,m,n}\Big[&g_{mn\nu'}(k,-q')\langle c^\dagger_{k-q'm}c_{kn}\rangle\langle b^\dagger_{-q'\nu'}\rangle\\&-g_{mn\nu'}(k,q')\langle c^\dagger_{k+q'm}c_{kn}\rangle\langle b_{q\nu'}\rangle\Big]\end{aligned}\tag{6.18}$$

is the Ehrenfest contribution, which vanishes if there is no phonon displacement. And the second order contribution that comes from e-ph scattering is

$$\begin{aligned}\frac{d\langle b^{\dagger}_{q'\nu'}b_{q'\nu'}\rangle}{dt}|_{2nd} = -\frac{2\pi}{\hbar}\sum_{k',m',n'}|g_{m'n'\nu'}(k',q')|^2\Big\{&\delta(\epsilon_{k'n'}-\epsilon_{k'+q'm'}+\hbar\omega_{q'\nu'})\\ &\times[f_{k'+q'm'}f_{k'n'}-(n_{\nu'q'}+1)f_{k'n'}+n_{\nu'q'}f_{k'+q'm'}]\\ &+\delta(\epsilon_{k'+q'm'}-\epsilon_{k'n'}+\hbar\omega_{q'\nu'})\mathrm{Re}\big[\Theta_{q'\nu'}\big](f_{k'+q'm'}-f_{k'n'})\Big\}.\end{aligned} \tag{6.19}$$

Here the Markov approximation with phase correction discussed in the previous section has been applied, leading to the same result as using the MGKBA [6]. A conventional treatment of the Markov approximation agrees with the result from GKBA.

Phonon coherence EOM

The EOM of the phonon coherence at the same mode, $\Theta_{q\nu}$, is given by

$$\frac{d\Theta_{q\nu}}{dt}|_{1st} = -\frac{i}{\hbar}(\hbar\omega_{q\nu}+\hbar\omega_{-q\nu})\Theta_{q\nu}, \tag{6.20}$$

and $\omega_{q\nu} = \omega_{-q\nu}$ for time-reversal invariant (non-magnetic) crystals.

The second-order contribution from e-ph scattering is

$$\begin{aligned}\frac{d\Theta_{q'\nu'}}{dt}|_{2nd} = \frac{\pi}{\hbar}\sum_{k',m',n'}&|g_{m'n'\nu'}(k',q')|^2\delta(\epsilon_{k'+q'm'}-\epsilon_{k'n'}+\hbar\omega_{-q'\nu'})\\ &\times\Big\{f_{k'n'}f_{k'+q'm'}-(n_{\nu'-q'}+1)f_{k'n'}+n_{\nu'-q'}f_{k'+q'm'}+\Theta_{q'\nu'}(f_{k'n'}-f_{k'+q'm'})\Big\}\\ &+\delta(\epsilon_{k'n'}-\epsilon_{k'+q'm'}+\hbar\omega_{q'\nu'})\\ &\times\Big\{f_{k'n'}f_{k'+q'm'}-(n_{\nu'q'}+1)f_{k'+q'm'}+n_{\nu'q'}f_{k'n'}+\Theta_{q'\nu'}(f_{k'+q'm'}-f_{k'n'})\Big\}.\end{aligned} \tag{6.21}$$

Notice that the difference in the sign in front of the coherence term for the emission and absorption processes now produce the correct physical limits.

Outlook: Phonon occupation and coherence EOMs with ph-ph scattering

The immediate next step, which is under development, is to include three-phonon scattering into the phonon occupation and coherence EOMs and applying the correct phase corrections. This will allow us to study the coherent phonon dynamics including both e-ph and ph-ph scattering with phonon coherence effects.

6.3 Liouville-von Neumann formalism of coherent phonon dynamics

Liouville-von Neumann equation

With the adiabatic approximation, the Liouville-von Neumann equation dictates the evolution of the density operator in the Schrödinger picture using the Lindbladian super-operator:

$$\frac{d\rho}{dt} = \mathcal{L}(\rho) + C. \tag{6.22}$$

The super-operator $\mathcal{L}$:

$$\mathcal{L}(\rho) = \frac{1}{i\hbar}[H_0, \rho] - \frac{1}{\hbar}[H', [K, \rho]], \tag{6.23}$$

and K is defined as:

$$K = \frac{1}{\hbar}\int_{t_0}^{t} dt'\, U_0(t-t')H'U_0^\dagger(t-t'). \tag{6.24}$$

For a single-particle density matrix $\rho^{\mathrm{sp}}_{\alpha_1\alpha_2} = \langle b^\dagger_{\alpha_1} b_{\alpha_2}\rangle$, its EOM reads:

$$\begin{aligned}\frac{d}{dt}\langle b^\dagger_{\alpha_1} b_{\alpha_2}\rangle = &-i(\omega_{\alpha_2} - \omega_{\alpha_1})\langle b^\dagger_{\alpha_1} b_{\alpha_2}\rangle + \frac{1}{i\hbar}\mathrm{Tr}\{[b^\dagger_{\alpha_1} b_{\alpha_2}, H]\rho^{\mathrm{int}}\} \\ &- \frac{1}{\hbar}\langle[[b^\dagger_{\alpha_1} b_{\alpha_2}, H'], K]\rangle.\end{aligned} \tag{6.25}$$

Electron-phonon scattering

The derivations are shown in Appendix A. Without $\langle b_{\boldsymbol{q}_1\nu_1} b_{\boldsymbol{q}_2\nu_2}\rangle$ terms, assume $\boldsymbol{q}_1, \boldsymbol{q}_2 \neq 0$, and ignore electronic coherences (only taking populations) and taking $\boldsymbol{q}_1\nu_1 = \boldsymbol{q}_2\nu_2 = \boldsymbol{q}\nu$:

$$\begin{aligned}\frac{d}{dt}n_{\boldsymbol{q}\nu} = \frac{2\pi}{\hbar}\sum_{\boldsymbol{k}nn',\nu'} &\delta(\epsilon_{\boldsymbol{k}+\boldsymbol{q}n'} - \epsilon_{\boldsymbol{k}n} - \hbar\omega_{\boldsymbol{q}\nu'})\times \\ &\mathrm{Re}\Big\{g_{n'n\nu'}(\boldsymbol{k}+\boldsymbol{q}, -\boldsymbol{q})g_{nn'\nu}(\boldsymbol{k}, \boldsymbol{q})\left[(\delta_{\nu\nu'} + \rho_{\boldsymbol{q}\nu\nu'})f_{\boldsymbol{k}+\boldsymbol{q}n'} - \rho_{\boldsymbol{q}\nu\nu'}f_{\boldsymbol{k}n}\right] \\ &-|g_{nn'\nu}(\boldsymbol{k}, \boldsymbol{q})|^2 f_{\boldsymbol{k}n}f_{\boldsymbol{k}+\boldsymbol{q}n'}\Big\}.\end{aligned} \tag{6.26}$$

The part proportional to population, taking $\nu' = \nu$:

$$\begin{aligned}\frac{d}{dt}n_{\boldsymbol{q}\nu}\Big|_{\mathrm{pop.}} = \frac{2\pi}{\hbar}\sum_{\boldsymbol{k}nn'} &\delta(\epsilon_{\boldsymbol{k}+\boldsymbol{q}n'} - \epsilon_{\boldsymbol{k}n} - \hbar\omega_{\boldsymbol{q}\nu})\times \\ &|g_{nn'\nu}(\boldsymbol{k}, \boldsymbol{q})|^2\Big\{(n_{\boldsymbol{q}\nu}+1)f_{\boldsymbol{k}+\boldsymbol{q}n'}f_{\boldsymbol{k}n} - n_{\boldsymbol{q}\nu}f_{\boldsymbol{k}n}(1 - f_{\boldsymbol{k}+\boldsymbol{q}n'})\Big\}.\end{aligned} \tag{6.27}$$

This is the same result from the Heisenberg formalism, resembling the classical collision integral.

The terms proportional to coherence:

$$\frac{d}{dt}n_{\boldsymbol{q}\nu}\Big|_{\text{coh}} = \frac{2\pi}{\hbar}\sum_{\boldsymbol{k}nn'}\sum_{\nu'\neq\nu}\delta(\epsilon_{\boldsymbol{k}+\boldsymbol{q}n'}-\epsilon_{\boldsymbol{k}n}-\hbar\omega_{\boldsymbol{q}\nu'}) \times \text{Re}\Big\{g_{n'n\nu'}(\boldsymbol{k}+\boldsymbol{q},-\boldsymbol{q})g_{nn'\nu}(\boldsymbol{k},\boldsymbol{q})\big[f_{\boldsymbol{k}+\boldsymbol{q}n'}(1-f_{\boldsymbol{k}n})-f_{\boldsymbol{k}n}(1-f_{\boldsymbol{k}+\boldsymbol{q}n'})\big]\rho_{\boldsymbol{q}\nu\nu'}\Big\}. \tag{6.28}$$

This implies that the emission or absorption of phonon $(\boldsymbol{q}', \nu')$ will stimulate the emission or absorption of other phonons $(\boldsymbol{q}, \nu)$, with the stimulated population proportional to the real part of the corresponding phonon coherence $\rho_{\boldsymbol{q}\nu;\boldsymbol{q}'\nu'}$. However, only using electron populations and ignoring electron coherences means that only $\boldsymbol{q}' = \boldsymbol{q}$ shows up in the equations. For $\boldsymbol{q}' \neq \boldsymbol{q}$, it seems to be a "second-order" process that involves both electron and phonon coherences.

Electron dephasing

The diagonal dephasing terms are

$$\frac{d}{dt}\rho_{\boldsymbol{q}_1\nu_1;\boldsymbol{q}_2\nu_2}\Big|_{\text{e-ph, diag}} = \frac{1}{2}(\Gamma_{\text{in}}-\Gamma_{\text{out}})\rho_{\boldsymbol{q}_1\nu_1;\boldsymbol{q}_2\nu_2}, \tag{6.29}$$

where

$$\Gamma_{\text{in}} = \frac{2\pi}{\hbar}\sum_{\boldsymbol{k}nn'}\Big\{\delta(\epsilon_{\boldsymbol{k}+\boldsymbol{q}_1n'}-\epsilon_{\boldsymbol{k}n}-\hbar\omega_{\boldsymbol{q}_1\nu_1})|g_{nn'\nu_1}(\boldsymbol{k},\boldsymbol{q}_1)|^2 f_{\boldsymbol{k}+\boldsymbol{q}_1n'}(1-f_{\boldsymbol{k}n}) +\delta(\epsilon_{\boldsymbol{k}+\boldsymbol{q}_2n'}-\epsilon_{\boldsymbol{k}n}-\hbar\omega_{\boldsymbol{q}_2\nu_2})|g_{nn'\nu_2}(\boldsymbol{k},\boldsymbol{q}_2)|^2 f_{\boldsymbol{k}+\boldsymbol{q}_2n'}(1-f_{\boldsymbol{k}n})\Big\}, \tag{6.30}$$

and

$$\Gamma_{\text{out}} = \frac{2\pi}{\hbar}\sum_{\boldsymbol{k}nn'}\Big\{\delta(\epsilon_{\boldsymbol{k}+\boldsymbol{q}_1n'}-\epsilon_{\boldsymbol{k}n}-\hbar\omega_{\boldsymbol{q}_1\nu_1})|g_{nn'\nu_1}(\boldsymbol{k},\boldsymbol{q}_1)|^2 f_{\boldsymbol{k}n}(1-f_{\boldsymbol{k}+\boldsymbol{q}_1n'}) +\delta(\epsilon_{\boldsymbol{k}+\boldsymbol{q}_2n'}-\epsilon_{\boldsymbol{k}n}-\hbar\omega_{\boldsymbol{q}_2\nu_2})|g_{nn'\nu_2}(\boldsymbol{k},\boldsymbol{q}_2)|^2 f_{\boldsymbol{k}n}(1-f_{\boldsymbol{k}+\boldsymbol{q}_2n'})\Big\}. \tag{6.31}$$

Phonon-phonon scattering

The form taken for K, using the fact that permutations of the matrix element indices are a symmetry:

$$\begin{aligned} K_{\text{ph-ph}} = \frac{\pi}{3!}\sum_{\boldsymbol{q}_1\nu_1}\sum_{\boldsymbol{q}_2\nu_2}\sum_{\boldsymbol{q}_3\nu_3}\Psi(\boldsymbol{q}_1\nu_1;\boldsymbol{q}_2\nu_2;\boldsymbol{q}_3\nu_3) \\ \Big\{\delta(\omega_1+\omega_2+\omega_3)b_{\boldsymbol{q}_1\nu_1}b_{\boldsymbol{q}_2\nu_2}b_{\boldsymbol{q}_3\nu_3} \\ +3\delta(\omega_1-\omega_2-\omega_3)b^\dagger_{-\boldsymbol{q}_1\nu_1}b_{\boldsymbol{q}_2\nu_2}b_{\boldsymbol{q}_3\nu_3} \\ +3\delta(\omega_1+\omega_2-\omega_3)b^\dagger_{-\boldsymbol{q}_1\nu_1}b^\dagger_{-\boldsymbol{q}_2\nu_2}b_{\boldsymbol{q}_3\nu_3} \\ +\delta(\omega_1+\omega_2+\omega_3)b^\dagger_{-\boldsymbol{q}_1\nu_1}b^\dagger_{-\boldsymbol{q}_2\nu_2}b^\dagger_{-\boldsymbol{q}_3\nu_3}\Big\}. \end{aligned} \tag{6.32}$$

Taking a thermal average, using the cluster expansion, with zeroth and first order terms with respect to off-diagonal terms $f_{\mu\nu}$ with $\mu \neq \nu$, we arrive at EOMs for phonon populations and coherences:

$$\frac{d}{dt}n_{\boldsymbol{q}\nu} = -\frac{1}{2\hbar}\sum_{\alpha\beta}\Big\{\Psi(\alpha;\beta;-\boldsymbol{q}\nu)\Big(\langle[b^\dagger_{\boldsymbol{q}\nu}b_\alpha b_\beta, K_{\text{ph-ph}}]\rangle - 2\langle[b^\dagger_\beta b_{-\alpha}b_{\boldsymbol{q}\nu}, K^\dagger_{\text{ph-ph}}]\rangle^*\Big) + \Psi(\alpha;\beta;\boldsymbol{q}\nu)\Big(\langle[b^\dagger_{\boldsymbol{q}\nu}b_{-\alpha}b_{-\beta}, K^\dagger_{\text{ph-ph}}]\rangle^* - 2\langle[b^\dagger_{-\beta}b_\alpha b_{\boldsymbol{q}\nu}, K_{\text{ph-ph}}]\rangle\Big)\Big\}. \tag{6.33}$$

The semiclassical limit is recovered with the zeroth order:

$$\begin{aligned}\frac{d}{dt}n_{\boldsymbol{q}\nu}\Big|_0 =& \frac{\pi}{\hbar}\sum_{\alpha\beta}|\Psi(\alpha;\beta;\boldsymbol{q}\nu)|^2 \\ &\Big\{\delta(\omega_{\boldsymbol{q}\nu}-\omega_\alpha-\omega_\beta)[n_\alpha n_\beta(n_{\boldsymbol{q}\nu}+1) - n_{\boldsymbol{q}\nu}(n_\alpha+1)(n_\beta+1)] \\ &-2\delta(\omega_\beta-\omega_{\boldsymbol{q}\nu}-\omega_\alpha)[n_{\boldsymbol{q}\nu}n_\alpha(n_\beta+1) - n_\beta(n_{\boldsymbol{q}\nu}+1)(n_\alpha+1)]\Big\}.\end{aligned} \tag{6.34}$$

The EOM of the first order for $\boldsymbol{q} \neq 0$ is in Appendix A.

Phonon dephasing

With

$$\begin{aligned}\frac{d}{dt}\langle b_1^\dagger b_2\rangle =& -i(\omega_2-\omega_1)\langle b_1^\dagger b_2\rangle - \frac{1}{2\hbar}\sum_{\alpha,\beta}\Big\{ \\ &+\Psi(\alpha;\beta;-2)\Big(\big\langle[b^\dagger_\mu b_\alpha b_\beta, K]\big\rangle - 2\big\langle[b^\dagger_\beta b_{-\alpha}b_\mu, K^\dagger]\big\rangle^*\Big) \\ &+\Psi(\alpha;\beta;1)\Big(\big\langle[b^\dagger_\nu b_{-\alpha}\, b_{-\beta}, K^\dagger]\big\rangle^* - 2\big\langle[b^\dagger_{-\beta}b_\alpha b_\nu, K]\big\rangle\Big)\Big\}.\end{aligned} \tag{6.35}$$

we can compute the zeroth and first order terms,

$$\frac{d}{dt}\langle b_1^\dagger b_2\rangle\Big| = \frac{d}{dt}\langle b_1^\dagger b_2\rangle\Big|_0 + \frac{d}{dt}\langle b_1^\dagger b_2\rangle\Big|_1, \tag{6.36}$$

and the zeroth-order term is shown in Appendix A, which vanishes for a thermal distribution, as expected.

Diagonal dephasing operator is given by:

$$\frac{d}{dt}\langle b_1^\dagger b_2\rangle\Big|_{\text{scat, diag}} = -\frac{1}{2}\Gamma\langle b_1^\dagger b_2\rangle, \tag{6.37}$$

where

$$\begin{aligned}\Gamma = \frac{\pi}{\hbar}\Bigg\{ &\sum_{\alpha}\Big[2\delta(\omega_\alpha - \omega_1 - \omega_2)|\Psi(\alpha;\mu;\nu)|^2[(n_1 - n_\alpha) + (n_2 - n_\alpha)]\Big] \\ &+ \sum_{\alpha\beta}\Big[\delta(\omega_1 - \omega_\alpha - \omega_\beta)|\Psi(\alpha;\beta;1)|^2(n_\alpha + n_\beta + 1) \\ &\quad + \delta(\omega_2 - \omega_\alpha - \omega_\beta)|\Psi(\alpha;\beta;2)|^2(n_\alpha + n_\beta + 1) \\ &\quad + 2\delta(\omega_\beta - \omega_\alpha - \omega_1)|\Psi(\alpha;\beta;1)|^2(n_\alpha - n_\beta) \\ &\quad + 2\delta(\omega_\beta - \omega_\alpha - \omega_2)|\Psi(\alpha;\beta;2)|^2(n_\alpha - n_\beta)\Big]\Bigg\}.\end{aligned} \tag{6.38}$$

When separated into "in-" and "out-" scattering, Eq. 6.3 becomes:

$$\frac{d}{dt}\langle b_1^\dagger b_2\rangle = \frac{1}{2}(\Gamma_{\text{in}} - \Gamma_{\text{out}})\langle b_1^\dagger b_2\rangle, \tag{6.39}$$

where

$$\begin{aligned}\Gamma_{\text{in}} = \frac{\pi}{\hbar}\Bigg\{ &\sum_{\alpha}\Big[2\delta(\omega_\alpha - \omega_1 - \omega_2)|\Psi(\alpha;\mu;\nu)|^2[n_\alpha(n_2 + 1) + n_\alpha(n_1 + 1)]\Big] \\ &+ \sum_{\alpha\beta}\Big[\delta(\omega_1 - \omega_\alpha - \omega_\beta)|\Psi(\alpha;\beta;1)|^2 n_\alpha n_\beta \\ &\quad + \delta(\omega_2 - \omega_\alpha - \omega_\beta)|\Psi(\alpha;\beta;2)|^2 n_\alpha n_\beta \\ &\quad + 2\delta(\omega_\beta - \omega_\alpha - \omega_1)|\Psi(\alpha;\beta;1)|^2 n_\beta(n_\alpha + 1) \\ &\quad + 2\delta(\omega_\beta - \omega_\alpha - \omega_2)|\Psi(\alpha;\beta;2)|^2 n_\beta(n_\alpha + 1)\Big]\Bigg\},\end{aligned} \tag{6.40}$$

and

$$\begin{aligned}\Gamma_{\text{out}} = \frac{\pi}{\hbar}\Bigg\{ &\sum_{\alpha}\Big[2\delta(\omega_\alpha - \omega_1 - \omega_2)|\Psi(\alpha;\mu;\nu)|^2[n_2(n_\alpha + 1) + n_1(n_\alpha + 1)]\Big] \\ &+ \sum_{\alpha\beta}\Big[\delta(\omega_1 - \omega_\alpha - \omega_\beta)|\Psi(\alpha;\beta;1)|^2(n_\alpha + 1)(n_\beta + 1) \\ &\quad + \delta(\omega_2 - \omega_\alpha - \omega_\beta)|\Psi(\alpha;\beta;2)|^2(n_\alpha + 1)(n_\beta + 1) \\ &\quad + 2\delta(\omega_\beta - \omega_\alpha - \omega_1)|\Psi(\alpha;\beta;1)|^2 n_\alpha(n_\beta + 1) \\ &\quad + 2\delta(\omega_\beta - \omega_\alpha - \omega_2)|\Psi(\alpha;\beta;2)|^2 n_\alpha(n_\beta + 1)\Big]\Bigg\}.\end{aligned} \tag{6.41}$$

Computational implementation

To obtain analytical insight, we formulate a simplified model for the lattice dynamics. The ph-ph interaction is approximated by a constant effective coupling parameter

M, with a low-frequency cutoff imposed to regularize long-wavelength modes. The phonon coherence coupling matrix $f_{\mathbf{q}\nu;\mathbf{q}'\nu'}$ is assumed to be diagonal in crystal momentum, such that only terms with $\mathbf{q} = \mathbf{q}'$ are retained. Under this approximation, the coupling reduces to $f_{\mathbf{q}\nu\nu'} = f^*_{\mathbf{q}\nu'\nu}$, describing coherence transfer between phonon branches at the same momentum. This assumption is expected to be valid for non-degenerate, well-separated phonon branches. Finally, off-diagonal dephasing terms in the phonon coherence EOM are neglected, allowing us to focus on the coherent dynamics. The EOMs obtained from these approximations are summarized in Appendix A.

6.4 Approximate description of coupled coherent electron-phonon dynamics

Within the BTE framework, the system is described by electron and phonon occupation factors $f_{\mathbf{k}n}$ and $n_{\mathbf{q}\nu}$. In order to capture coherent lattice motion in addition to population dynamics, we explicitly introduce the phonon displacement operator

$$\hat{Q}_{\mathbf{q}\nu} = \hat{b}_{\mathbf{q}\nu} + \hat{b}^\dagger_{-\mathbf{q}\nu}, \tag{6.42}$$

defined in terms of phonon creation and annihilation operators. Its expectation value,

$$Q_{\mathbf{q}\nu}(t) = \langle \hat{Q}_{\mathbf{q}\nu}(t) \rangle, \tag{6.43}$$

serves as a direct measure of phonon coherence: in the incoherent regime $Q_{\mathbf{q}\nu} = 0$, whereas optically driven coherent phonons correspond to finite macroscopic displacements.

Beyond its formal role in the EOM, the coherent displacement $Q_{\mathbf{q}\nu}(t)$ is directly relevant to experiments. In pump-probe, second-harmonic generation, and time-resolved spectroscopy measurements, the observed signals are often proportional to lattice displacements or their time derivatives rather than to phonon populations. As a result, computing $Q_{\mathbf{q}\nu}(t)$ provides direct access to the amplitude, phase, and symmetry of coherent phonon motion, enabling quantitative comparison with experiments and a clear separation of coherent lattice dynamics from purely thermal effects [9–11].

EOM for the coherent displacement

Starting from the Heisenberg EOM for $\hat{Q}_{\mathbf{q}\nu}$ and including harmonic phonon dynamics together with ph-ph, e-ph, and infrared-field interactions, one obtains a second-order EOM for the coherent displacement,

$$\partial_t^2 Q_{\mathbf{q}\nu}(t) + \omega^2_{\mathbf{q}\nu} Q_{\mathbf{q}\nu}(t) = \mathcal{D}_{\mathbf{q}\nu}(t), \tag{6.44}$$

which has the form of a driven harmonic oscillator. The total driving force $\mathcal{D}_{\mathbf{q}\nu}(t)$ decomposes into contributions from ph-ph interactions, e-ph coupling, and direct coupling to an external infrared field, each derived microscopically from the corresponding interaction Hamiltonian.

Central approximation: separation of coherent and thermal phonons

To efficiently couple the coherent phonon EOM to the rt-BTE, we decompose the phonon displacement operator into coherent and thermal components,

$$\hat{Q}_{\mathbf{q}\nu} = \hat{Q}^{\mathrm{coh}}_{\mathbf{q}\nu} + \hat{Q}^{\mathrm{th}}_{\mathbf{q}\nu}. \tag{6.45}$$

The coherent component is treated at the classical mean-field level,

$$\langle \hat{Q}^{\mathrm{coh}}_{\mathbf{q}\nu} \rangle = Q^{\mathrm{coh}}_{\mathbf{q}\nu}, \tag{6.46}$$

while the thermal component satisfies $\langle \hat{Q}^{\mathrm{th}}_{\mathbf{q}\nu} \rangle = 0$ and is described statistically through the phonon occupations obtained from the rt-BTE.

Under this separation, the displacement-displacement correlator entering the ph-ph driving force is approximated as

$$\begin{aligned} \langle \hat{Q}_{\mathbf{q}'\nu'} \hat{Q}_{\mathbf{q}''\nu''} \rangle = {} & Q^{\mathrm{coh}}_{\mathbf{q}'\nu'} Q^{\mathrm{coh}}_{\mathbf{q}''\nu''} \\ & + \left[2n^{\mathrm{th}}_{\mathbf{q}'\nu'}(t) + 1 \right] \delta_{-\mathbf{q}'\mathbf{q}''} \delta_{\nu'\nu''}. \end{aligned} \tag{6.47}$$

As a result, the ph-ph contribution to $\mathcal{D}_{\mathbf{q}\nu}(t)$ separates into a coherent-coherent term that drives nonlinear coherent oscillations and a thermal term determined by the time-dependent phonon populations.

Assumptions and regime of validity

The present formulation relies on the following assumptions:

1. Coherent phonons are macroscopically occupied and can be treated at the mean-field level.
2. Coherent-thermal phonon correlations are neglected, implying the absence of phase locking between coherent and thermal modes.
3. Thermal phonons obey Gaussian statistics and are fully characterized by $n_{\mathbf{q}\nu}(t)$.
4. Coherent dynamics is restricted to optically active zone-center phonons, while all finite-momentum phonons are treated within the rt-BTE.

These assumptions define the regime in which the hybrid EOM-rt-BTE description is both valid and computationally efficient.

Damping and energy conservation

Damping of the coherent phonon EOM is derived from first-principles mode-resolved three-phonon ph-ph scattering processes. The same anharmonic interaction matrix elements enter both the decay rates of coherent phonons and the ph-ph collision integrals in the phonon rt-BTE. Consequently, energy exchange between coherent phonons and thermal phonon populations is treated consistently, ensuring global energy conservation and avoiding unphysical energy drift during time propagation.

6.5 Conclusion

This chapter established a Heisenberg-equation framework for coupled coherent and incoherent electron-phonon dynamics, recovering the correct limiting behavior for e-ph and ph-e scattering while providing direct access to coherent lattice displacements. A complementary Liouville-von Neumann formulation was introduced to capture ph-ph scattering at lowest order, and an efficient hybrid description was proposed to enable practical simulations of nonlinear-phononics experiments. An important direction for future work is the extension of the Heisenberg formalism to include ph-ph interactions in a fully phase-resolved manner, allowing higher-order anharmonic processes to be treated on the same microscopic footing as e-ph scattering. Further developments will focus on systematically connecting the approximate hybrid framework to the underlying Heisenberg EOMs, assessing its regime of validity, and implementing the resulting equations suitable for first-principles simulations of driven solids.

References

[1] Z. Zeng, M. Först, M. Fechner, M. Buzzi, E. Amuah, C. Putzke, P. Moll, D. Prabhakaran, P. Radaelli, and A. Cavalleri, Science **387**, 431 (2025).

[2] T. Nova, A. Disa, M. Fechner, and A. Cavalleri, Science **364**, 1075 (2019).

[3] A. Disa, J. Curtis, M. Fechner, A. Liu, A. Von Hoegen, M. Först, T. Nova, P. Narang, A. Maljuk, A. Boris, *et al.*, Nature **617**, 73 (2023).

[4] M. Rini, R. Tobey, N. Dean, J. Itatani, Y. Tomioka, Y. Tokura, R. W. Schoenlein, and A. Cavalleri, Nature **449**, 72 (2007).

[5] M. Mitrano, A. Cantaluppi, D. Nicoletti, S. Kaiser, A. Perucchi, S. Lupi, P. Di Pietro, D. Pontiroli, M. Riccò, S. R. Clark, *et al.*, Nature **530**, 461 (2016).

[6] G. Stefanucci and E. Perfetto, SciPost Physics **16**, 073 (2024).

[7] F. Rossi and T. Kuhn, Rev. Mod. Phys. **74**, 895 (2002).

[8] P. Lipavský, V. Špička, and B. Velický, Phys. Rev. B **34**, 6933 (1986).

[9] F. Caruso and M. Zacharias, Phys. Rev. B **107**, 054102 (2023).

[10] C. Emeis, S. Jauernik, S. Dahiya, Y. Pan, C. E. Jensen, P. Hein, M. Bauer, and F. Caruso, Phys. Rev. X **15**, 021039 (2025).

[11] Y. Pan, C. Emeis, S. Jauernik, M. Bauer, and F. Caruso, (2025), arXiv:2502.01529 [cond-mat.str-el] .

Chapter 7

SUMMARY AND FUTURE DIRECTIONS

This thesis advances the first-principles framework for simulating nonequilibrium coupled electron–phonon dynamics that explicitly incorporates e-ph and anharmonic ph-ph interactions using the real-time Boltzmann transport equation (rt-BTE). The demonstrated framework addresses the high computational cost associated with long simulation times, dense Brillouin-zone sampling, and anharmonic lattice dynamics, with improvements that enable simulations across a wide range of time scales and physical systems.

In Chapter 2, adaptive and multirate time-integration methods were developed to efficiently propagate the coupled rt-BTEs. By exploiting the intrinsic separation of electronic and phononic timescales, these methods substantially reduce computational cost while maintaining desired accuracy. The results demonstrate that the optimal choice of numerical parameters, such as tolerance and solver order, depends sensitively on the simulation timescale of interest, and that fully adaptive multirate schemes provide a robust framework for long-time nonequilibrium simulations without the need to converge the results time step sizes. The methods also reveal insights of the underlying timescales of dynamics. These developments significantly accelerate the simulation by 1-2 orders of magnitude, extending the predictive reach of rt-BTE calculations to regimes that were previously computationally inaccessible.

Chapter 3 introduced dynamic mode decomposition (DMD) as a data-driven acceleration technique for nonequilibrium transport simulations. By reconstructing long-time dynamics and steady states from short-time trajectories, DMD enables efficient access to steady-state regimes without explicit long-time time stepping. Beyond its computational advantages, the extracted DMD modes provide insight into dominant relaxation pathways and collective behaviors in momentum space, offering a complementary perspective on nonequilibrium e-ph dynamics.

In Chapter 4, the computational efficiency of rt-BTE simulations was further enhanced through GPU parallelization of collision-integral evaluations. A GPU-optimized data layout and algorithm were developed to reduce memory overhead and synchronization costs existed in computing collision integrals of e-ph and ph-ph interactions, yielding substantial speedups and favorable strong-scaling behavior.

This work demonstrates the potential to leverage modern high-performance computing architectures and efficient data structure and algorithms that enables the study of complex materials and extended Brillouin-zone sampling with reduced computational resources.

Chapter 5 focused on the efficient treatment of anharmonic ph-ph interactions through tensor learning and compression. By constructing low-rank representations of high-dimensional phonon scattering tensors, the computational costs associated with anharmonicity were significantly reduced both in computing time and memory. This approach enables the inclusion of higher-order phonon interactions and the treatment of materials with large unit cells, paving the way toward systematically improving the accuracy of lattice dynamics.

In Chapter 6, the scope of the thesis was extended beyond incoherent population dynamics to address coherently driven lattice phenomena. Starting from operator-level equations of motion, a theoretical framework was constructed to describe electron and phonon coherences induced by external driving and e-ph coupling. A computational framework that only allows Γ-point phonon coherences are explored, where finite-momentum phonons are treated as incoherent populations subject to anharmonic scattering. Anharmonic coupling between coherent and incoherent modes would then provide a microscopic mechanism for decoherence and energy redistribution. While the treatment remains exploratory and relies on controlled approximations, it establishes a foundation for bridging the gap between fully quantum-coherent descriptions and semiclassical transport theories.

While the methods developed in this thesis significantly advance simulations of incoherent nonequilibrium dynamics, the treatment of coherent phonon phenomena remains an open and important frontier. Many modern ultrafast experiments probe regimes in which phonon coherence, nonlinear lattice motion, and strong external driving play a central role, and these effects are not fully captured by population-based transport equations. The framework introduced in Chapter 6 represents a first step toward addressing this gap, but substantial theoretical and computational developments remain to be pursued.

Therefore, one central future direction is the analytical development of equations of motion for coherent phonons that explicitly include anharmonic ph-ph interactions. In the present work, phonon coherences were primarily derived within the harmonic approximation, with dissipation treated phenomenologically or through simplified coupling to incoherent populations. Extending this formulation to include cubic

and quartic phonon interactions at the operator or density-matrix level is essential for capturing dephasing, frequency renormalization, and nonlinear mode coupling observed in driven lattice experiments.

From a numerical perspective, an important near-term objective is the implementation of coherent phonon dynamics using controlled approximations that reduce computational complexity while retaining essential physics. Limiting coherence to a small subset of phonon modes that are IR or Raman active dramatically reduces dimensionality and allows coherent dynamics to be coupled efficiently to existing rt-BTE simulations of incoherent populations.

In addition, future work should aim to combine adaptive multirate time integration, DMD-based acceleration, GPU parallelization, and tensor-compressed phonon interactions into a unified framework. Coherent lattice oscillations often coexist with slow incoherent thermalization processes, making them well suited for multirate strategies. DMD may further enable efficient analysis of long-time behavior or extraction of dominant collective modes in driven systems. GPU acceleration and tensor compression will be essential for extending coherent simulations to realistic materials with strong anharmonicity and complex unit cells. The implementation of these techniques will be integrated into future releases of PERTURBO [1], making them broadly accessible to the materials science community.

With a robust theoretical and computational framework in place, a wide range of material systems become accessible. Of particular interest are materials in which driven phonons play an active role in modifying electronic, structural, or collective properties. $SrTiO_3$ [2] is a prototypical example, exhibiting soft optical phonons, strong anharmonicity, and rich coupling between lattice and electronic degrees of freedom.

A specific example of application of the coherent phonon framework is the photo-induced high-temperature ferromagnetic state recently observed in the Mott-insulating titanate $YTiO_3$ [3]. In this experiment, a resonant THz excitation of selected infrared-active $B_{2\mu}$ phonons was used to control the magnetic order of the material, where the strongest response was observed with a driving frequency of 9 THz. A transient ferromagnetic state was induced at a temperature greater than 80 K, nearly three times the equilibrium Curie temperature. The response was highly mode selective, where a 4 THz pump frequency, for example, was observed to suppress ferromagnetism. The magnetic state was induced by the pump on the time scale of 10 to 40 ps and persisted for at least several nanoseconds.

The central puzzle that makes the coherent phonon framework preferable is the lack of detailed nonlinear effects in a simple coherent frozen-phonon DFT calculation, where the predicted exchange change had the opposite sign from the observed pump-induced magnetic response, independent of the Hubbard U parameter used in the calculation [3]. The authors therefore argued that physics beyond a naïve adiabatic spin-phonon coupling is needed. They proposed that driven phonons modify the orbital wavefunction and orbital gap, thereby moving $YTiO_3$ from the phase boundary between ferromagnetic and competing antiferromagnetic spin-orbital states. This interpretation is physically compelling, but it remains incomplete because it treats the lattice primarily through the coherent zone-center coordinate and does not calculate the time-dependent, mode-resolved phonon population generated by anharmonic decay of the driven phonon.

This distinction is important because the experimental magnetic response coincides with the coherent phonon lifetime. This suggests that the coherent phonons may gradually transfer energy into selected finite-q lattice fluctuations, which in turn modify orbital polarization, exchange competition, and spin relaxation pathways. In this picture, the induced magnetization does not follow only $Q^{coh}_{\mathbf{q}\nu}(t)$, but also the nonequilibrium phonon population $N_{\mathbf{q}\nu}(t)$.

A calculation to clarify the origin of the opposite sign of the magnetic response at 9 THz pump frequency is therefore to simulate ultrafast dynamics using the framework in Chapter 6. Both third-order and fourth-order phonon interactions can be included, as the $B_{2\mu}$ modes can couple to a broad set of lattice fluctuations due to its symmetry. This term is especially crucial if the leading cubic decay channels are restricted by symmetry, weak, or energetically off resonance. The tensor-compression approach developed in Chapter 5 is relevant, because four-phonon interactions are otherwise extremely expensive to evaluate on dense momentum grids.

With the excited mode labeled as λ, we can compute the time-dependent, mode-resolved phonon population dynamics:

$$\frac{dN_{\mathbf{q}\nu}}{dt} = S^{(3)}_{\lambda\mathbf{q}\nu} + S^{(4)}_{\lambda\mathbf{q}\nu} + I^{\text{ph}-\text{ph}(3)}[N_{\mathbf{q}\nu}] + I^{\text{ph}-\text{ph}(4)}[N_{\mathbf{q}\nu}], \tag{7.1}$$

where

$$S^{(3)}_{\lambda\mathbf{q}\nu} \propto \sum_{\nu'} \left|\Phi^{(3)}_{\lambda,\mathbf{q}\nu,-\mathbf{q}\nu'}\right|^2 |Q_\lambda|^2 \, \delta\left(\omega_\lambda - \omega_{\mathbf{q}\nu} - \omega_{-\mathbf{q}\nu'}\right), \tag{7.2}$$

and

$$S^{(4)}_{\lambda\mathbf{q}\nu} \propto \sum_{\nu'} \left|\Phi^{(4)}_{\lambda,\lambda,\mathbf{q}\nu,-\mathbf{q}\nu'}\right|^2 |Q_\lambda|^4 \, \delta\left(2\omega_\lambda - \omega_{\mathbf{q}\nu} - \omega_{-\mathbf{q}\nu'}\right). \tag{7.3}$$

This calculation will elucidate the energy relaxation pathway during the lifetime of the coherent phonon with a mode-specific analysis for different pump frequencies.

With the phonon populations and coherent displacements, one can connect them to magnetic and orbital degrees of freedom. In the existing interpretation, the relevant low-energy electronic physics is governed by the t_{2g} orbital of Ti. The proposed calculation would generalize the frozen-phonon analysis by computing orbital and magnetic quantities to leading order if spin-phonon interactions are incorporated in the calculation. By sampling distorted structures from the nonequilibrium phonon populations produced with the simulations, the orbital gap for those sampled structures can be computed using DFT+U and fit the result to a low-order response model. Although this calculation would not by itself predict the full time-dependent magnetization, it would establish whether the nonequilibrium phonon distribution produces the experimentally observed enhancement or suppression of ferromagnetism.

Other materials of interests include alkali-doped fullerides such as K_3C_{60} [4] represent another promising class, where coupling between molecular vibrations and electronic correlations has been implicated in light-induced superconducting-like states. Similarly, complex oxide systems such as $LaAlO_3$-based thin films or heterostructures [5] host lattice and electronic phenomena that are highly sensitive to lattice distortions and phonon excitation. Applying the coherent phonon framework developed here to these materials would enable direct comparison with ultrafast experiments and provide microscopic insight into how driven lattice dynamics influence emergent phases. More broadly, such studies could guide the design of materials and driving protocols for phonon-engineered functionalities.

References

[1] J.-J. Zhou, J. Park, I.-T. Lu, I. Maliyov, X. Tong, and M. Bernardi, Comput. Phys. Commun. **264**, 107970 (2021).

[2] T. Nova, A. Disa, M. Fechner, and A. Cavalleri, Science **364**, 1075 (2019).

[3] A. S. Disa, J. Curtis, M. Fechner, A. Liu, A. von Hoegen, M. Först, T. F. Nova, P. Narang, A. Maljuk, A. V. Boris, B. Keimer, and A. Cavalleri, Nature **617**, 73 (2023).

[4] M. Mitrano, A. Cantaluppi, D. Nicoletti, S. Kaiser, A. Perucchi, S. Lupi, P. Di Pietro, D. Pontiroli, M. Riccò, S. R. Clark, D. Jaksch, and A. Cavalleri, Nature **530**, 461 (2016).

[5] J. Gollwitzer, J. Z. Kaaret, Y. E. Suyolcu, G. Khalsa, R. C. Fernandes, O. Gorobtsov, S. Buchenau, C. You, J. Higgins, R. S. Russell, Z. Shao, Y. A. Birkhölzer, T. Sato, M. Chollet, G. Coslovich, M. Brützam, C. Guguschev, J. W. Harter, A. S. Disa, D. G. Schlom, N. A. Benedek, and A. Singer, Phys. Rev. Lett. **135**, 116906 (2025).

Appendix A

DERIVATION OF EQUATIONS OF MOTION FOR ELECTRON-PHONON COHERENCE

Correction to Markov approximation for phonon coherences

This appendix provides the detailed derivation underlying Sec. 6.2. We focus on the phase structure of phonon-coherence contributions and its impact on the Markovian reduction of the Heisenberg equations of motion.

The dissipative part of the Markov approximation follows from

$$\int_0^\infty d\tau\; e^{-i\Delta\tau} = \pi\,\delta(\Delta) - i\,\mathcal{P}\frac{1}{\Delta}, \tag{A.1}$$

so that the argument of the delta function is fixed by the net oscillatory phase in the memory integral.

Consider the formal solution of a mixed electron–phonon correlator,

$$\delta s(t) = \int_0^\infty d\tau\; e^{-i\Omega\tau}\,\mathcal{S}(t-\tau), \tag{A.2}$$

where Ω arises from the homogeneous evolution and $\mathcal{S}$ denotes the source term. For a correlator containing a single phonon annihilation operator $b_{q'\nu'}$, one has

$$\Omega = \frac{\epsilon_{kn} - \epsilon_{k+q'm}}{\hbar} + \omega_{q'\nu'}. \tag{A.3}$$

When the source contains a two-annihilator phonon coherence,

$$\Theta^{\nu'\nu}_{q'q}(t) = \langle b_{q'\nu'}(t)\, b_{q\nu}(t)\rangle - \langle b_{q'\nu'}(t)\rangle\langle b_{q\nu}(t)\rangle, \tag{A.4}$$

its free evolution yields

$$\Theta^{\nu'\nu}_{q'q}(t-\tau) \simeq \Theta^{\nu'\nu}_{q'q}(t)\, e^{+i(\omega_{q'\nu'}+\omega_{q\nu})\tau}. \tag{A.5}$$

Substitution into Eq. A.2 gives

$$\delta s^{(bb)}(t) \propto \int_0^\infty d\tau\; e^{-i(\Omega-(\omega_{q'\nu'}+\omega_{q\nu}))\tau}, \tag{A.6}$$

and hence the on-shell contribution

$$\pi\,\delta\big[\Omega - (\omega_{q'\nu'} + \omega_{q\nu})\big]. \tag{A.7}$$

With Ω defined in Eq. A.3, this yields

$$\delta\left(\epsilon_{kn} - \epsilon_{k+q'm} - \hbar\omega_{q\nu}\right). \tag{A.8}$$

In the same-mode limit $q' = q$, $\nu' = \nu$, the coherence phase reduces to $2\omega_{q\nu}$, and the effective detuning becomes $\Omega - 2\omega_{q\nu}$, leading to the same resonance condition as obtained in the main text.

Heisenberg EOM for electron-phonon interactions

The EOM of s by computing its with Hamiltonians H_e, H_p and H_{e-p} is given by,

$$\begin{aligned}
\frac{ds^{q'\nu'}_{k''m',k'n'}}{dt} &= -\frac{i}{\hbar}\Big\{ s^{q'\nu'}_{k''m',k'n'}(\epsilon_{k'n'} - \epsilon_{k''m'} + \hbar\omega_{q'\nu'}) \\
&+ \sum_{n,q,\nu}\Big[g_{n'n\nu}(k'-q,q)(\langle c^\dagger_{k''m'} c_{k'-qn} b_{q'\nu'} b_{q\nu}\rangle + \langle c^\dagger_{k''m'} c_{k'-qn} b_{q'\nu'} b^\dagger_{-q\nu}\rangle) \\
&- g_{nm'\nu}(k'',q)(\langle c^\dagger_{k''+qn} c_{k'n'} b_{q'\nu'} b_{q\nu}\rangle + \langle c^\dagger_{k''+qn} c_{k'n'} b^\dagger_{-q\nu} b_{q'\nu'}\rangle)\Big] \\
&+ \sum_{k,m,n} g_{mn\nu'}(k,-q')(\langle c^\dagger_{k-q'm} c_{kn}\rangle\langle c^\dagger_{k''m'} c_{k'n'}\rangle - \langle c^\dagger_{k-q'm} c_{k'n'}\rangle\langle c^\dagger_{k''m'} c_{kn}\rangle)\Big\}.
\end{aligned} \tag{A.9}$$

where

$$\begin{aligned}
\langle c^\dagger_{k''m'} c_{k'-qn} b_{q'\nu'} b_{q\nu}\rangle = (\langle b^\dagger_{q'\nu'} b_{q\nu}\rangle - \langle b^\dagger_{q'\nu'}\rangle\langle b_{q\nu}\rangle)\langle c^\dagger_{k''m'} c_{k'-qn}\rangle \\
+\langle b_{q'\nu'}\rangle s^{q\nu}_{k''m',k'-qn} + \langle b_{q\nu}\rangle \delta s^{q'\nu'}_{k''m',k'-qn},
\end{aligned} \tag{A.10}$$

Subtracting the EOM of $s^{q'\nu'}_{0k''m',k'n'}$ from that of $s^{q'\nu'}_{k''m',k'n'}$, we get the EOM of $\delta s^{q'\nu'}_{k''m',k'n'}$ as stated in Eq. 6.10.

We can simplify by only using the delta function part, which describes energy conservation in scattering processes. We can collect the terms in Eq. 6.2 to obtain the final EOM for the electron density matrix including coherences under the influence of electron-phonon interactions:

$$\begin{aligned}
\frac{d}{dt}\rho_{m'n'}(k';q') = -\frac{2}{\hbar}\sum_{q,\nu,n'} \\
\mathrm{Re}\Big\{ g_{n'n'\nu}(k'-q,q)\,\mathcal{K}(\epsilon_{k'-q,n'} - \epsilon_{k'+q',m'} + \hbar\omega_{q\nu})\,\mathcal{S}_1 \\
- g_{n'm'\nu}(k'+q',q)\,\mathcal{K}(\epsilon_{k',n'} - \epsilon_{k'+q'+q,n'} + \hbar\omega_{q\nu})\,\mathcal{S}_2\Big\},
\end{aligned} \tag{A.11}$$

where Re{} takes the real part of the expression inside the braces.

The source terms are given by:

$$\begin{aligned}\mathcal{S}_1 = &\sum_n g_{n'n\nu}(k',-q)\left[(n_{\nu q}+1) - f_{k'-q,n'}\right]\rho_{m'n}(k';q') \\ &- \sum_n g_{nm'\nu}(k'+q',-q)\, n_{\nu q}\, \rho_{nn'}(k'-q;q'),\end{aligned} \tag{A.12}$$

$$\begin{aligned}\mathcal{S}_2 = &\sum_n g_{n'n\nu}(k'+q',-q)\left[(n_{\nu q}+1) - f_{k',n'}\right]\rho_{n'n}(k'+q';q) \\ &- \sum_n g_{n'n\nu}(k'+q'+q,-q)\, n_{\nu q}\, \rho_{nn'}(k';q'+q),\end{aligned} \tag{A.13}$$

$$\begin{aligned}\mathcal{S}_1^{(\text{ph-coh})} = &\sum_n \sum_{Q,\mu} g_{n'n\mu}(k'-q-Q,Q)\ \Theta_{q\nu,Q\mu}\ \rho_{m'n}(k'-q-Q;\, q'+q+Q) \\ &- \sum_n \sum_{Q,\mu} g_{nm'\mu}(k'+q',Q)\ \Theta_{q\nu,Q\mu}\ \rho_{nn'}(k'-q;\, q'+q+Q),\end{aligned} \tag{A.14}$$

and

$$\begin{aligned}\mathcal{S}_2^{(\text{ph-coh})} = &\sum_n \sum_{Q,\mu} g_{n'n\mu}(k'-Q,Q)\ \Theta_{q\nu,Q\mu}\ \rho_{n'n}(k'-Q;\, q'+q+Q) \\ &- \sum_n \sum_{Q,\mu} g_{nn'\mu}(k'+q'+q,Q)\ \Theta_{q\nu,Q\mu}\ \rho_{nn'}(k';\, q'+q+Q),\end{aligned} \tag{A.15}$$

where $\Theta_{q\nu,Q\mu}$ are phonon coherences.

To compute the part with phonon coherence, we need to compute the EOM of $\langle b^\dagger_{q''\nu''} b_{q'\nu'}\rangle$. Splitting it into first and second order parts, we have, the first order as

$$\begin{aligned}\frac{d\langle b^\dagger_{q''\nu''} b_{q'\nu'}\rangle}{dt}|_{1st} = -\frac{i}{\hbar}\sum_{k,m,n}\Big[&g_{mn\nu'}(k,-q')\langle c^\dagger_{k-q'm}c_{kn}\rangle\langle b^\dagger_{-q''\nu''}\rangle - g_{mn\nu''}(k,q'') \\ &\langle c^\dagger_{k+q''m}c_{kn}\rangle\langle b_{q\nu'}\rangle\Big] - \frac{i}{\hbar}(\hbar\omega_{q'\nu'} - \hbar\omega_{q''\nu''})\langle b^\dagger_{q''\nu''}b_{q'\nu'}\rangle,\end{aligned} \tag{A.16}$$

and the second order as

$$\begin{aligned}\frac{d\langle b^\dagger_{q''\nu''} b_{q'\nu'}\rangle}{dt}|_{2nd} = -\frac{i}{\hbar}\sum_{k',m',n'}\Big[&g_{m'n'\nu'}(k',-q')\delta s^{*q''\nu''}_{k'n',k'-q'm'} \\ &- g_{m'n'\nu''}(k',q'')\delta s^{q'\nu'}_{k'+q''m',kn'}\Big]\end{aligned}. \tag{A.17}$$

For the coherence $\langle b_{q''\nu''} b_{q'\nu'} \rangle$, the first order part is

$$\begin{aligned} \frac{d\langle b_{q''\nu''} b_{q'\nu'} \rangle}{dt}|_{1st} = -\frac{i}{\hbar} \sum_{k,m,n} \Big[& g_{mn\nu'}(k,-q')\langle c^\dagger_{k-q'm} c_{kn} \rangle \langle b_{q''\nu''} \rangle \\ & + g_{mn\nu''}(k,-q'')\langle c^\dagger_{k-q''m} c_{kn} \rangle \langle b_{q'\nu'} \rangle \Big] \\ & - \frac{i}{\hbar}(\hbar\omega_{q'\nu'} + \hbar\omega_{q''\nu''})\langle b_{q''\nu''} b_{q'\nu'} \rangle, \end{aligned} \tag{A.18}$$

and the second order part is

$$\begin{aligned} \frac{d\langle b_{q''\nu''} b_{q'\nu'} \rangle}{dt}|_{2nd} = -\frac{i}{\hbar} \sum_{k,m,n} \Big[& g_{mn\nu'}(k,-q')\delta s^{q''\nu''}_{k-q'm,kn} \\ & + g_{mn\nu''}(k,-q'')\delta s^{q'\nu'}_{k-q''m,kn} \Big] \end{aligned} \tag{A.19}$$

EOM from Liouville-von Neumann equations

Equation 6.3 can be written in terms of matrix elements, including a generalised scattering term:

$$\frac{d\rho_{\lambda_1\lambda_2}}{dt} = \frac{E_{\lambda_1} - E_{\lambda_2}}{i\hbar}\rho_{\lambda_1\lambda_2} + C_{\lambda_1\lambda_2} + \sum_{\lambda'_1\lambda'_2} \Gamma_{\lambda_1\lambda_2,\lambda'_1\lambda'_2}\rho_{\lambda'_1\lambda'_2}, \tag{A.20}$$

The latter sum can be separated into "in-" and "out-" scattering terms. The Markov limit is intrinsically assumed here. C is a time-dependent operator describing quantum correlation effects propagating from time t_0 to t:

$$C(t) = \frac{1}{i\hbar}[H, U_0(t-t_0)\rho(t_0)U_0^\dagger(t-t_0)]. \tag{A.21}$$

In the semiclassical limit and ignoring energy re-normalizations for a one-body perturbation:

$$K_{\lambda\lambda'} = \pi H'_{\lambda\lambda'}\delta(E_\lambda - E_{\lambda'}). \tag{A.22}$$

This evolution is non-Lindblad-like, and producing a closed equation of motion requires that the equations above only apply to a subsystem, with ρ being a reduced density matrix.

For e-ph interactions, we have

$$\begin{aligned} [b^\dagger_{\boldsymbol{q}_1\nu_1} b_{\boldsymbol{q}_2\nu_2}, H'_{\text{e-ph}}] = \sum_{\boldsymbol{k}nn'} \Big[& g_{nn'\nu_2}(\boldsymbol{k},-\boldsymbol{q}_2) b^\dagger_{\boldsymbol{q}_1\nu_1} c^\dagger_{\boldsymbol{k}-\boldsymbol{q}_2,n'} c_{\boldsymbol{k}n} \\ & - g_{nn'\nu_1}(\boldsymbol{k},\boldsymbol{q}_1) b_{\boldsymbol{q}_2\nu_2} c^\dagger_{\boldsymbol{k}+\boldsymbol{q}_1,n'} c_{\boldsymbol{k}n} \Big], \end{aligned} \tag{A.23}$$

and

$$K'_{\text{e-ph}} = \pi \sum_{\boldsymbol{k}nn',\boldsymbol{q}\nu} g^{\nu}_{nn'}(\boldsymbol{k},\boldsymbol{q}) \Big\{ \delta(\epsilon_{\boldsymbol{k}+\boldsymbol{q},n'} - \epsilon_{\boldsymbol{k}n} - \hbar\omega_{\boldsymbol{q}\nu}) c^{\dagger}_{\boldsymbol{k}+\boldsymbol{q},n'} c_{\boldsymbol{k}n} b_{\boldsymbol{q}\nu} + \delta(\epsilon_{\boldsymbol{k}+\boldsymbol{q},n'} - \epsilon_{\boldsymbol{k}n} + \hbar\omega_{\boldsymbol{q}\nu}) c^{\dagger}_{\boldsymbol{k}+\boldsymbol{q},n'} c_{\boldsymbol{k}n} b^{\dagger}_{-\boldsymbol{q}\nu} \Big\}. \tag{A.24}$$

Computational implementation of EOMs

The equations of motion for phonon populations and coherences including ph-ph interactions can be approximated and simplified as follows:

$$\frac{d}{dt} n_{\boldsymbol{q}\nu}\Big|_0 = \frac{\pi}{\hbar}|M|^2 \sum_{\alpha\beta} \Big\{ \delta(\omega_{\boldsymbol{q}\nu} - \omega_\alpha - \omega_\beta)[n_\alpha n_\beta - n_{\boldsymbol{q}\nu}(n_\alpha + n_\beta + 1)] - 2\delta(\omega_\beta - \omega_{\boldsymbol{q}\nu} - \omega_\alpha)[n_{\boldsymbol{q}\nu} n_\alpha - n_\beta(n_{\boldsymbol{q}\nu} + n_\alpha + 1)] \Big\} \tag{A.25}$$

$$\begin{aligned} \frac{d}{dt} n_{\boldsymbol{q}\nu}\Big|_1 = \frac{\pi}{\hbar}|M|^2 \sum_{\alpha\beta} \Big\{ & \sum_{\nu'\neq\nu} \big[-\delta(\omega_{\boldsymbol{q}\nu'} - \omega_\alpha - \omega_\beta)(n_\alpha + n_\beta + 1) \\ & + 2\delta(\omega_\beta - \omega_\alpha - \omega_{\boldsymbol{q}\nu'})(n_\beta - n_\alpha) \big] \mathrm{Re}\{f_{\boldsymbol{q}\nu\nu'}\} \\ & + \sum_{\nu'\neq\nu_\alpha} \big[2\delta(\omega_{\boldsymbol{q}\nu} - \omega_\beta - \omega_{\boldsymbol{q}_\alpha\nu'})(n_{\boldsymbol{q}\nu} - n_\beta) \\ & - 2\delta(\omega_{\boldsymbol{q}_\alpha\nu'} - \omega_\beta - \omega_{\boldsymbol{q}\nu})(n_{\boldsymbol{q}\nu} + n_\beta + 1) \\ & + 2\delta(\omega_\beta - \omega_{\boldsymbol{q}\nu} - \omega_{\boldsymbol{q}_\alpha\nu'})(n_\alpha - n_\beta) \big] \mathrm{Re}\{f_{\boldsymbol{q}_\alpha\nu_\alpha\nu'}\} \Big\} \end{aligned} \tag{A.26}$$

$$\begin{aligned} \frac{d}{dt} f_{\boldsymbol{q}\nu_1\nu_2}\Big|_0 = \frac{\pi}{2\hbar}|M|^2 \sum_{\alpha\beta} \Big\{ & \delta(\omega_{\boldsymbol{q}\nu_1} - \omega_\alpha - \omega_\beta)[n_\alpha n_\beta - n_{\boldsymbol{q}\nu_1}(n_\alpha + n_\beta + 1)] \\ & - 2\delta(\omega_\beta - \omega_{\boldsymbol{q}\nu_1} - \omega_\alpha)[n_\alpha n_{\boldsymbol{q}\nu_1} - n_\beta(n_\alpha + n_{\boldsymbol{q}\nu_1} + 1)] \\ & + \delta(\omega_{\boldsymbol{q}\nu_2} - \omega_\alpha - \omega_\beta)[n_\alpha n_\beta - n_{\boldsymbol{q}\nu_2}(n_\alpha + n_\beta + 1)] \\ & - 2\delta(\omega_\beta - \omega_{\boldsymbol{q}\nu_2} - \omega_\alpha)[n_\alpha n_{\boldsymbol{q}\nu_2} - n_\beta(n_\alpha + n_{\boldsymbol{q}\nu_2} + 1)] \Big\} \end{aligned} \tag{A.27}$$

$$\frac{d}{dt} f_{\boldsymbol{q}\nu_1\nu_2} = [i(\omega_{\boldsymbol{q}\nu_1} - \omega_{\boldsymbol{q}\nu_2}) - \Gamma] f_{\boldsymbol{q}\nu_1\nu_2} \tag{A.28}$$

$$\begin{aligned}\Gamma = \frac{\pi}{2\hbar}|M|^2 \sum_{\alpha\beta} \Big\{ & \delta(\omega_{\boldsymbol{q}\nu_1} - \omega_\alpha - \omega_\beta)(n_\alpha + n_\beta + 1) \\ & +2\delta(\omega_\beta - \omega_\alpha - \omega_{\boldsymbol{q}\nu_1})(n_\alpha - n_\beta) \\ & +\delta(\omega_{\boldsymbol{q}\nu_2} - \omega_\alpha - \omega_\beta)(n_\alpha + n_\beta + 1) \\ & +2\delta(\omega_\beta - \omega_\alpha - \omega_{\boldsymbol{q}\nu_2})(n_\alpha - n_\beta) \Big\}\end{aligned} \tag{A.29}$$

Higher order terms in phonon population and coherence EOMs

$$\begin{aligned}\frac{d}{dt} n_{\boldsymbol{q}\nu}\Big|_1 = -\frac{\pi}{\hbar} \sum_{\alpha\beta} \Big\{ & \sum_{\boldsymbol{q}'\nu' \neq \boldsymbol{q}\nu} \Big[\delta(\omega' - \omega_\alpha - \omega_\beta) \mathrm{Re}\{\Psi(\alpha;\beta;\boldsymbol{q}'\nu') \\ & \Psi(-\alpha;-\beta;-\boldsymbol{q}\nu) f_{\boldsymbol{q}\nu;\boldsymbol{q}'\nu'}\}(n_\alpha + n_\beta + 1) \\ & -2\delta(\omega_\beta - \omega_\alpha - \omega') \mathrm{Re}\{\Psi(\alpha;\beta;\boldsymbol{q}'\nu') \\ & \Psi(-\alpha;-\beta;-\boldsymbol{q}\nu) f_{\boldsymbol{q}\nu;\boldsymbol{q}'\nu'}\}(n_\beta - n_\alpha) \Big] \\ & + \sum_{\boldsymbol{q}'\nu' \neq \alpha} \Big[2\delta(\omega_{\boldsymbol{q}\nu} - \omega_\beta - \omega') \mathrm{Re}\{\Psi(-\boldsymbol{q}'\nu';\beta;\boldsymbol{q}\nu) \\ & \Psi(\alpha;-\beta;-\boldsymbol{q}\nu) f_{\boldsymbol{q}'\nu';\alpha}\}(n_{\boldsymbol{q}\nu} - n_\beta) \\ & -2\delta(\omega' - \omega_\beta - \omega_{\boldsymbol{q}\nu}) \mathrm{Re}\{\Psi(-\boldsymbol{q}'\nu';\beta;\boldsymbol{q}\nu) \\ & \Psi(\alpha;-\beta;-\boldsymbol{q}\nu) f_{\boldsymbol{q}'\nu';\alpha}\}(n_{\boldsymbol{q}\nu} + n_\beta + 1) \\ & -2\delta(\omega_\beta - \omega_{\boldsymbol{q}\nu} - \omega') \mathrm{Re}\{\Psi(\boldsymbol{q}'\nu';\beta;\boldsymbol{q}\nu) \\ & \Psi(-\alpha;-\beta;-\boldsymbol{q}\nu) f_{\alpha;\boldsymbol{q}'\nu'}\}(n_\beta - n_\alpha) \Big] \Big\},\end{aligned} \tag{A.30}$$

which represents all other collisions $(\boldsymbol{q}'\nu', \alpha, \beta)$ influencing the population of $(\boldsymbol{q}\nu)$ through coherences.

For phonon dephasing,

$$\begin{aligned}\frac{d}{dt}\langle b_1^\dagger b_2\rangle\Big|_0 = \frac{\pi}{2\hbar} \sum_{\alpha\beta} & \Psi(\alpha;\beta;1)\Psi(-\alpha;-\beta;-2) \\ & \Big\{ \delta(\omega_1 - \omega_\alpha - \omega_\beta)[n_\alpha n_\beta(n_1+1) - n_1(n_\alpha+1)(n_\beta+1)] \\ & -2\delta(\omega_\beta - \omega_\alpha - \omega_1)[n_\alpha n_1(n_\beta+1) - n_\beta(n_\alpha+1)(n_1+1)] \\ & +\delta(\omega_2 - \omega_\alpha - \omega_\beta)[n_\alpha n_\beta(n_2+1) - n_2(n_\alpha+1)(n_\beta+1)] \\ & -2\delta(\omega_\beta - \omega_\alpha - \omega_2)[n_\alpha n_2(n_\beta+1) - n_\beta(n_\alpha+1)(n_2+1)] \Big\}.\end{aligned} \tag{A.31}$$

Appendix B

ELECTRON AND PHONON SCATTERING RATES

This appendix is intended as a reference connecting the nonequilibrium BTE formalism used throughout this thesis to the more familiar equilibrium lifetime expressions arising from e-ph and ph-ph interactions. These quantities are closely related to the collision integrals introduced in Chapter 1, and provide useful equilibrium intuition for the characteristic timescales governing nonequilibrium dynamics.

The electron scattering rate due to e-ph interactions can be derived either from Fermi's golden rule or within a Green's-function formalism. The corresponding lowest-order electron self-energy is given by

$$\begin{aligned}\Sigma^{e-\mathrm{ph}}_{n\mathbf{k}} =& \frac{2\pi}{\hbar}\frac{1}{\mathcal{N}_{\mathbf{q}}}\sum_{\nu\mathbf{q}}|g_{mn\nu}(\mathbf{k},\mathbf{q})|^2\Big[(N_{\nu\mathbf{q}}+1-f_{m\mathbf{k}+\mathbf{q}})\delta(\epsilon_{n\mathbf{k}}-\epsilon_{m\mathbf{k}+\mathbf{q}}-\hbar\omega_{\nu\mathbf{q}}) \\ &+(N_{\nu\mathbf{q}}+f_{m\mathbf{k}+\mathbf{q}})\delta(\epsilon_{n\mathbf{k}}-\epsilon_{m\mathbf{k}+\mathbf{q}}+\hbar\omega_{\nu\mathbf{q}})\Big],\end{aligned} \tag{B.1}$$

where $f_{n\mathbf{k}}$ and $N_{\nu\mathbf{q}}$ reduce to the Fermi-Dirac and Bose-Einstein distributions at given temperature, respectively.

The phonon self-energy due to ph-ph scattering can be computed as [33]:

$$\begin{aligned}\Sigma^{\text{ph-ph}}_{\nu\mathbf{q}} =& \frac{18\pi}{\hbar}\frac{1}{\mathcal{N}_{\mathbf{q}}}\sum_{\nu'\nu''}\sum_{\mathbf{q}'}|\Psi_{\nu\nu'\nu''}(\mathbf{q},\mathbf{q}',\mathbf{q}'')|^2 \\ &\times\Big[(N_{\nu'\mathbf{q}'}+N_{\nu''\mathbf{q}''}+1)\delta(\omega_{\nu\mathbf{q}}-\omega_{\nu'\mathbf{q}'}-\omega_{\nu''\mathbf{q}''}) \\ &+2(N_{\nu'\mathbf{q}'}-N_{\nu''\mathbf{q}''})\delta(\omega_{\nu\mathbf{q}}+\omega_{\nu'\mathbf{q}'}-\omega_{\nu''\mathbf{q}''})\Big],\end{aligned} \tag{B.2}$$

where $\mathbf{q}'' = \mathbf{q} - \mathbf{q}'$ is implied by momentum conservation.

The scattering rate (or inverse lifetime) associated with either electrons or phonons is obtained from the imaginary part of the corresponding self-energy as

$$\Gamma = \tau^{-1} = \frac{2}{\hbar}\left|\mathrm{Im}\,\Sigma\right|. \tag{B.3}$$

While equilibrium lifetimes do not capture the full time-dependent evolution of carrier and phonon populations, they provide valuable intuition for the characteristic timescales of scattering processes and serve as a practical guide for convergence tests, such as momentum-grid resolution, prior to performing full nonequilibrium rt-BTE simulations.