MULTISCALE MODEL. SIMUL.
Vol. 15, No. 1, pp. 29–55

ADAPTIVELY COMPRESSED POLARIZABILITY OPERATOR FOR ACCELERATING LARGE SCALE *AB INITIO* PHONON CALCULATIONS[*]

LIN LIN[†], ZE XU[‡], AND LEXING YING[§]

Abstract. Phonon calculations based on first principle electronic structure theory, such as the Kohn–Sham density functional theory, have wide applications in physics, chemistry, and material science. The computational cost of first principle phonon calculations typically scales steeply as $\mathcal{O}(N_e^4)$, where N_e is the number of electrons in the system. In this work, we develop a new method for reducing the computational complexity of computing the full dynamical matrix, and hence the phonon spectrum, to $\mathcal{O}(N_e^3)$. The key concept for achieving this is to compress the polarizability operator adaptively with respect to the perturbation of the potential due to the change of the atomic configuration. Such an adaptively compressed polarizability operator allows accurate computation of the phonon spectrum. The reduction of complexity only weakly depends on the size of the band gap, and our method is applicable to insulators as well as semiconductors with small band gaps. We demonstrate the effectiveness of our method using one-dimensional and two-dimensional model problems.

Key words. density functional perturbation theory, phonon calculations, adaptive compression, polarizability operator, Sternheimer equation, Dyson equation

AMS subject classifications. 65F10, 65F30, 65Z05

DOI. 10.1137/16M1077325

1. Introduction.

Kohn–Sham density functional theory (KSDFT) [24, 27] is the most widely used electronic structure theory for molecules and systems in condensed phase. In principle, KSDFT provides an exact description of ground state properties of a many-body quantum system, such as electron density, energy, and atomic forces. Once the electronic ground state is obtained, many physical and chemical properties of the system can be described by studying the *response* of the quantum system under small perturbation. The theory for describing such response behavior is called the density functional perturbation theory (DFPT) [4, 19, 3].

One important application of DFPT is the description of lattice vibrations. In the Born–Oppenheimer approximation, lattice vibrations can be described by the dynamical matrix, which is related to the Hessian matrix of the ground state energy

[*]Received by the editors May 27, 2016; accepted for publication (in revised form) October 27, 2016; published electronically January 5, 2017.
http://www.siam.org/journals/mms/15-1/M107732.html

Funding: The work of the first and second authors was partially supported by Laboratory Directed Research and Development (LDRD) funding from Berkeley Lab, provided by the Director, Office of Science, of the U.S. Department of Energy (DOE) under contract DE-AC02-05CH11231, the first author by the Scientific Discovery through Advanced Computing (SciDAC) program and the Center for Applied Mathematics for Energy Research Applications (CAMERA) funded by U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research and Basic Energy Sciences, and by the Alfred P. Sloan fellowship. The third author was partially supported by the National Science Foundation under grant DMS-0846501 and the DOE's Advanced Scientific Computing Research program under grant DEFC02-13ER26134/DESC0009409.

[†]Department of Mathematics, University of California, Berkeley, Berkeley, CA 94720, and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 (linlin@math.berkeley.edu).

[‡]Department of Mathematics, University of California, Berkeley, CA 94720 (zexu@math.berkeley.edu).

[§]Department of Mathematics and Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305 (lexing@math.stanford.edu).

with respect to the atomic positions. The eigenfunctions of the dynamical matrix give the phonon modes, and the eigenvalues give the phonon frequencies. A large variety of physical properties of solids depends on such phonon calculations. A few examples include infrared spectroscopy, elastic neutron scattering, specific heat, heat conduction, and electron-phonon interaction related behaviors such as superconductivity [19, 3]. Furthermore, the computational procedure of phonon calculations are largely transferable to the calculation of other types of response behavior, such as response to homogeneous electric fields, piezoelectric properties, magnons, and many-body perturbation theory for the description of electrons at excited states such as the GW theory [23, 35].

The term "phonon calculation" usually describes the calculation of vibrational properties of condensed matter systems. In this paper, we slightly abuse this term to refer to calculations of vibration properties of general systems, including condensed matter systems as well as isolated molecule clusters, since such calculations share the same mathematical structure. Mathematically, the procedure for phonon calculations can be straightforward. When atoms are at their equilibrium positions, the atomic forces (i.e., first order derivatives of the energy with respect to the atomic position) are zero for all atoms. To compute the Hessian matrix, one can move one atom at a time slightly away from its equilibrium position and compute the corresponding atomic forces. This amounts to the finite difference (FD) approximation of the Hessian matrix and is referred to as the "frozen phonon" [49, 5] approach in physics. The FD approach is simple to implement and can be used to obtain a phonon spectrum quickly for systems of small sizes. However, FD requires in total $d \times N_A \sim \mathcal{O}(N_e)$ KSDFT calculations, where d is the spatial dimension (usually $d = 3$), and N_A is the number of atoms. The computational complexity of a single KSDFT calculation typically scales as $\mathcal{O}(N_e^3)$, where N_e is the number of electrons in the system. Since $N_A \sim \mathcal{O}(N_e)$, the total cost of the FD approximation is $\mathcal{O}(N_e^4)$. This is prohibitively expensive for systems of large sizes. Furthermore, the accuracy of the FD approximation is limited by the size of the perturbation, which cannot be too small due to the numerical noise in the evaluation of the atomic forces in KSDFT calculations (usually the accuracy of forces is set to be $10^{-4} \sim 10^{-3}$ Hartree/Bohr). Such numerical noise also makes it difficult to compute nonlinear response properties, which can require even higher order derivatives of the energy.

DFPT, on the other hand, can be viewed as the "proper" way for computing derivative quantities in the context of KSDFT. The central quantity in DFPT is the polarizability operator, which characterizes the linear response of the electron density with respect to the perturbation of the external potential. More specifically, phonon calculations require applying the polarizability operator to $d \times N_A$ perturbation vectors induced by the change of the atomic configuration. The polarizability operator can be obtained by solving a Dyson equation iteratively [3], and each iteration step requires the solutions to $\mathcal{O}(N_e^2)$ Sternheimer equations. In general the complexity of DFPT is still $\mathcal{O}(N_e^4)$. So the main advantage of DFPT is that it gives accurate linear response properties. Furthermore, the same framework can be used to compute nonlinear response properties [20, 18, 3]. The mathematical aspect of DFPT for reduced Hartree–Fock model systems was recently been analyzed [10]. It is also possible to reduce the computational complexity of phonon calculations by "linear scaling methods" [17, 7]. Such methods can be successful in reducing the computational cost for large-sized systems with substantial band gaps, but this can be challenging for medium-sized systems with relatively small band gaps.

The main computational bottleneck of DFPT is the solution of the $\mathcal{O}(N_e^2)$ Sternheimer equations [3]. KSDFT can be defined as a nonlinear eigenvalue problem with $\mathcal{O}(N_e)$ eigenfunctions. Each of the $\mathcal{O}(N_e^2)$ equations in DFPT represents the response of an eigenfunction to a different external perturbation. Hence at first sight it is not possible to reduce the number of equations. However, as N_e becomes large, there will be asymptotically more equations to solve than the size of the matrix. Hence there is potential room to obtain a set of "compressed perturbations," which leads to methods for solving DFPT with lower complexity.

At this point it might be enticing to compress the $\mathcal{O}(N_e^2)$ equations using standard compression schemes such as singular value decomposition (SVD). However, there is some immediate difficulty associated with SVD-type compression schemes: The matrix to be compressed is of size $\mathcal{O}(N_e^2) \times \mathcal{O}(N_e)$ and of approximate rank $\mathcal{O}(N_e)$. The associated cost of the SVD type of compression is $\mathcal{O}(N_e^4)$, and hence there are no savings in asymptotic complexity. Furthermore, the $\mathcal{O}(N_e^2)$ equations need to be solved self-consistently according to the Dyson equation. Hence the initially compressed vectors might not be applicable anymore as the iteration proceeds towards the converged solution. This leads to inaccurate phonon calculations.

In this paper, we develop a new method called the adaptively compressed polarizability operator (ACP) to overcome the above difficulties. ACP reduces the complexity for applying the polarizability operator to $\mathcal{O}(N_e)$ vectors as follows. (1) ACP compresses the $\mathcal{O}(N_e^2)$ right-hand side vectors of the Sternheimer equations into $\mathcal{O}(N_e)$ vectors, using a recently developed interpolative separable density fitting method [32]. Together with a Chebyshev interpolation procedure to disentangle the energy dependence from the right-hand side vectors, ACP reduces the number of equations from $\mathcal{O}(N_e^2)$ to only $\mathcal{O}(N_e)$. (2) ACP reformulates the Dyson equation into an equivalent fixed point problem, where the compression of the Sternheimer equations depends adaptively on the unknown solutions. Using such adaptive compression procedure, we demonstrate that the self-consistent solution to the Dyson equation no longer hinders the accuracy of the compressed polarizability operator. Such an adaptive compression strategy is similar in spirit to the recently developed adaptively compressed exchange operator (ACE) for accelerating KSDFT calculations with hybrid exchange-correlation functionals [29]. We demonstrate that the overall computational complexity for phonon calculations can be reduced to $\mathcal{O}(N_e^3)$, and the cost depends only weakly on the band gap of the system. Hence the method can be applied to both insulators and semiconductors with small gaps. To the extent of our knowledge, this is the first result of this type in literature. We demonstrate the numerical performance of the ACP formulation for accelerating phonon calculations using model systems for one-dimensional (1D) and two-dimensional (2D) systems, both for periodic lattices and for systems with defects and random perturbations. Our numerical results confirm the low complexity of the ACP formulation for computing the full dynamical matrix and hence the phonon spectrum.

The rest of the paper is organized as follows. Section 2 introduces the basic formulation of KSDFT and DFPT. Section 3 describes the ACP formulation. Numerical results are presented in section 4, followed by the conclusion and discussion in section 5.

2. Preliminary. For completeness we first provide a brief introduction to the Kohn–Sham density functional theory (KSDFT) and the density functional perturbation theory (DFPT) in the context of phonon calculations. To simplify our discussion, we neglect the spin degeneracy, temperature dependence, as well as the usage of nonlocal pseudopotential. We assume all orbitals $\{\psi_i(\mathbf{r})\}$ are real. The spatial dimension

$d=3$ is assumed in the treatment of, for example, Coulomb interaction unless otherwise specified. We remark that such simplified treatment does not reduce the core difficulty of the problem.

2.1. Kohn–Sham density functional theory. Consider a system consisting of N_A nuclei and N_e electrons. In the Born–Oppenheimer approximation, for each set of nuclear positions $\{\mathbf{R}_I\}_{I=1}^{N_A}$, the electrons are relaxed to their ground state. The ground state total energy is denoted by $E_{\mathrm{tot}}(\{\mathbf{R}_I\}_{I=1}^{N_A})$ and can be computed in KSDFT [24, 27] according to the minimization of the following Kohn–Sham energy functional:

$$\begin{aligned} E_{\mathrm{KS}}(\{\psi_i\};\{\mathbf{R}_I\}) =& \frac{1}{2}\sum_{i=1}^{N_e}\int |\nabla\psi_i(\mathbf{r})|^2 \,\mathrm{d}\mathbf{r} + \int V_{\mathrm{ion}}(\mathbf{r};\{\mathbf{R}_I\})\rho(\mathbf{r})\,\mathrm{d}\mathbf{r} \\ &+ \frac{1}{2}\iint v_c(\mathbf{r},\mathbf{r}')\rho(\mathbf{r})\rho(\mathbf{r}')\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}' + E_{\mathrm{xc}}[\rho] + E_{\mathrm{II}}(\{\mathbf{R}_I\}). \end{aligned} \tag{2.1}$$

Here the minimization is with respect to the Kohn–Sham orbitals $\{\psi_i\}_{i=1}^{N_e}$ satisfying the orthonormality condition

$$\int \psi_i^*(\mathbf{r})\psi_j(\mathbf{r})\,\mathrm{d}\mathbf{r} = \delta_{ij}, \quad i,j=1,\ldots,N_e.$$

In (2.1), $\rho(\mathbf{r}) = \sum_{i=1}^{N_e} |\psi_i(\mathbf{r})|^2$ defines the electron density. In the discussion below we will omit the range of indices I, i unless otherwise specified. In (2.1), $v_c(\mathbf{r},\mathbf{r}') = \frac{1}{|\mathbf{r}-\mathbf{r}'|}$ defines the kernel for Coulomb interaction in $\mathbb{R}^3$. V_{ion} is a local potential characterizing the electron-ion interaction in all-electron calculations and is independent of the electronic states $\{\psi_i\}$. More specifically, V_{ion} is the summation of local potentials from each atom I:

$$V_{\mathrm{ion}}(\mathbf{r};\{\mathbf{R}_I\}) = \sum_I V_I(\mathbf{r}-\mathbf{R}_I). \tag{2.2}$$

In a pseudopotential approximation, $V_I(\mathbf{r}-\mathbf{R}_I)$ is defined as

$$V_I(\mathbf{r}-\mathbf{R}_I) := \int v_c(\mathbf{r},\mathbf{r}')m_I(\mathbf{r}'-\mathbf{R}_I)\,\mathrm{d}\mathbf{r}', \tag{2.3}$$

where m_I is a localized function in the real space and is called a pseudocharge [33, 37]. The normalization condition for each pseudocharge is $\int m_I(\mathbf{r})\,\mathrm{d}\mathbf{r} = -Z_I$, and Z_I is the atomic charge for the Ith atom. The total pseudocharge is defined as $m(\mathbf{r}) = \sum_I m_I(\mathbf{r}-\mathbf{R}_I)$. We assume the system is charge neutral, i.e.,

$$\int m(\mathbf{r})\,\mathrm{d}\mathbf{r} = -\sum_I Z_I = -N_e.$$

E_{xc} is the exchange-correlation energy, and here we assume semilocal functionals such as local density approximation (LDA) [11, 39] and generalized gradient approximation (GGA) functionals [6, 28, 38] are used. The last term in (2.1) is the ion-ion Coulomb interaction energy. For isolated clusters in three dimensions,

$$E_{\mathrm{II}}(\{\mathbf{R}_I\}) = \frac{1}{2}\sum_{I\neq J}\frac{Z_I Z_J}{|\mathbf{R}_I-\mathbf{R}_J|}. \tag{2.4}$$

We also note that for extended systems, modeled as infinite periodic structures, all the terms with Coulomb kernel require special treatment in order to avoid divergence due to the long-range $1/r$ nature of the Coulomb interaction [33].

The Euler–Lagrange equation associated with the Kohn–Sham energy functional gives rise to the Kohn–Sham equations as

$$H[\rho]\psi_i = \left(-\frac{1}{2}\Delta + \mathcal{V}[\rho]\right)\psi_i = \varepsilon_i \psi_i, \tag{2.5}$$

$$\int \psi_i^*(\mathbf{r})\psi_j(\mathbf{r})\,\mathrm{d}\mathbf{r} = \delta_{ij}, \quad \rho(\mathbf{r}) = \sum_{i=1}^{N_e} |\psi_i(\mathbf{r})|^2 . \tag{2.6}$$

Here the eigenvalues $\{\varepsilon_i\}$ are ordered nondecreasingly. $\psi_1, \ldots, \psi_{N_e}$ are called the occupied orbitals, while $\psi_{N_e+1}, \ldots$ are called the unoccupied orbitals. ψ_{N_e} is often referred to as the highest occupied molecular orbital (HOMO), and ψ_{N_e+1} the lowest unoccupied molecular orbital (LUMO). The difference of the corresponding eigenvalues $\varepsilon_g = \varepsilon_{N_e+1} - \varepsilon_{N_e}$ defines the HOMO-LUMO gap. Here we are interested in insulating and semiconducting systems with positive energy gap ε_g.

For a given electron density ρ, the effective potential $\mathcal{V}[\rho]$ is

$$\begin{aligned}\mathcal{V}[\rho](\mathbf{r}) &= V_{\mathrm{ion}}(\mathbf{r}; \{\mathbf{R}_I\}) + \int v_c(\mathbf{r}, \mathbf{r}')\rho(\mathbf{r}')\,\mathrm{d}\mathbf{r}' + V_{\mathrm{xc}}[\rho](\mathbf{r}) \\ &= \int v_c(\mathbf{r}, \mathbf{r}')(\rho(\mathbf{r}') + m(\mathbf{r}'))\,\mathrm{d}\mathbf{r}' + V_{\mathrm{xc}}[\rho](\mathbf{r}).\end{aligned} \tag{2.7}$$

Here $V_{\mathrm{xc}}[\rho](\mathbf{r}) = \frac{\delta E_{\mathrm{xc}}}{\delta \rho(\mathbf{r})}$ is the exchange-correlation potential, which is the functional derivative of the exchange-correlation energy with respect to the electron density. The Kohn–Sham Hamiltonian depends nonlinearly on the electron density ρ, and the electron density should be solved self-consistently. When the Kohn–Sham energy functional E_{KS} achieves its minimum, the self-consistency of the electron density is simultaneously achieved. Then the total energy can be equivalently computed as [33]

$$\begin{aligned}E_{\mathrm{tot}} = &\sum_{i=1}^{N_e} \varepsilon_i - \frac{1}{2}\iint v_c(\mathbf{r}, \mathbf{r}')\rho(\mathbf{r})\rho(\mathbf{r}')\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}' \\ &- \int V_{\mathrm{xc}}[\rho](\mathbf{r})\rho(\mathbf{r})\,\mathrm{d}\mathbf{r} + E_{\mathrm{xc}}[\rho] + E_{\mathrm{II}}(\{\mathbf{R}_I\}).\end{aligned} \tag{2.8}$$

Here $E_{\mathrm{band}} = \sum_{i=1}^{N_e} \varepsilon_i$ is referred to as the band energy. Using the Hellmann–Feynman theorem [33], the atomic force can be computed as

$$\begin{aligned}\mathbf{F}_I &= -\frac{\partial E_{\mathrm{tot}}(\{\mathbf{R}_I\})}{\partial \mathbf{R}_I} = -\int \frac{\partial V_{\mathrm{ion}}}{\partial \mathbf{R}_I}(\mathbf{r}; \{\mathbf{R}_I\})\rho(\mathbf{r})\,\mathrm{d}\mathbf{r} - \frac{\partial E_{\mathrm{II}}(\{\mathbf{R}_I\})}{\partial \mathbf{R}_I} \\ &= -\int \frac{\partial V_I}{\partial \mathbf{R}_I}(\mathbf{r} - \mathbf{R}_I)\rho(\mathbf{r})\,\mathrm{d}\mathbf{r} - \frac{\partial E_{\mathrm{II}}(\{\mathbf{R}_I\})}{\partial \mathbf{R}_I}.\end{aligned} \tag{2.9}$$

The atomic force allows the performance of structural relaxation of the atomic configuration by minimizing the total energy E_{tot} with respect to the atomic positions $\{\mathbf{R}_I\}$. When the atoms are at their equilibrium positions, all atomic forces should be 0.

2.2. Density functional perturbation theory. In DFPT, we assume that the self-consistent ground state electron density ρ has been computed. In this paper, we focus on phonon calculations using DFPT. Assume the system deviates from its equilibrium position $\{\mathbf{R}_I\}$ by some small magnitude; then the changes of the total energy are dominated by the Hessian matrix with respect to the atomic positions. The dynamical matrix D consists of $d \times d$ blocks in the form

$$D_{I,J} = \frac{1}{\sqrt{M_I M_J}} \frac{\partial^2 E_{\mathrm{tot}}(\{\mathbf{R}_I\})}{\partial \mathbf{R}_I \partial \mathbf{R}_J},$$

where M_I is the mass of the Ith nuclei. The dimension of the dynamical matrix is $d \times N_A$. The equilibrium atomic configuration is at least at the local minimum of the total energy, and all the eigenvalues of D are real and nonnegative. Hence the eigendecomposition of D is

$$Du_k = \omega_k^2 u_k,$$

where u_k is called the kth phonon mode, and ω_k is called the kth phonon frequency. The phonon spectrum is defined as the distribution of the eigenvalues $\{\omega_k\}$, i.e.,

$$\varrho_D(\omega) = \frac{1}{dN_A} \sum_k \delta(\omega - \omega_k). \tag{2.10}$$

Here δ is the Dirac-δ distribution. ϱ_D is also referred to as the density of states of D [33, 30].

In order to compute the Hessian matrix, we obtain from (2.9) that

$$\begin{aligned} &\frac{\partial^2 E_{\mathrm{tot}}(\{\mathbf{R}_I\})}{\partial \mathbf{R_I} \partial \mathbf{R_J}} \\ &= \int \frac{\partial V_I}{\partial \mathbf{R}_I}(\mathbf{r} - \mathbf{R}_I) \frac{\delta \rho(\mathbf{r})}{\delta \mathbf{R}_J} \,\mathrm{d}\mathbf{r} + \delta_{I,J} \int \rho(\mathbf{r}) \frac{\partial^2 V_I}{\partial \mathbf{R}_I^2}(\mathbf{r} - \mathbf{R}_I) \,\mathrm{d}\mathbf{r} + \frac{\partial^2 E_{\mathrm{II}}(\{\mathbf{R}_I\})}{\partial \mathbf{R}_I \partial \mathbf{R}_J}. \end{aligned} \tag{2.11}$$

In (2.11), the second term can be readily computed with numerical integration, and the third term involves only ion-ion interaction that is independent of the electronic states. Hence the first term is the most challenging one due to the response of the electron density with respect to the perturbation of atomic positions. Applying the chain rule, we have

$$\int \frac{\partial V_I}{\partial \mathbf{R}_I}(\mathbf{r} - \mathbf{R}_I) \frac{\delta \rho(\mathbf{r})}{\delta \mathbf{R}_J} \,\mathrm{d}\mathbf{r} = \iint \frac{\partial V_I}{\partial \mathbf{R}_I}(\mathbf{r} - \mathbf{R}_I) \frac{\delta \rho(\mathbf{r})}{\delta V_{\mathrm{ion}}(\mathbf{r}')} \frac{\partial V_J}{\partial \mathbf{R}_J}(\mathbf{r}' - \mathbf{R}_J) \,\mathrm{d}\mathbf{r} \,\mathrm{d}\mathbf{r}'. \tag{2.12}$$

Here the Fréchet derivative $\chi(\mathbf{r}, \mathbf{r}') = \frac{\delta \rho(\mathbf{r})}{\delta V_{\mathrm{ion}}(\mathbf{r}')}$ is referred to as the reducible polarizability operator [35], which characterizes the *self-consistent* linear response of the electron density at $\mathbf{r}$ with respect to an external perturbation of V_{ion} at $\mathbf{r}'$. However, since the Kohn–Sham equations (2.5) are a set of nonlinear equations with respect to ρ, the self-consistent response is still difficult to compute. Instead, the computation of χ must be obtained through a simpler quantity,

$$\chi_0(\mathbf{r}, \mathbf{r}') = \frac{\delta \rho(\mathbf{r})}{\delta \mathcal{V}(\mathbf{r}')},$$

which is called the irreducible polarizability operator (a.k.a., the independent particle polarizability operator) [35].

For simplicity in the discussion below, we will not distinguish the continuous and discretized representations of various quantities. In the case when a discretized representation is needed, we assume that the computational domain is uniformly discretized into a number of grid points $\{\mathbf{r}_\alpha\}_{\alpha=1}^{N_g}$. We may refer to quantities such as $U(\mathbf{r}) \equiv [u_1(\mathbf{r}), \dots, u_{N_e}(\mathbf{r})]$ as a matrix of dimension $N_g \times N_e$. In particular, all indices in the subscript are interpreted as the column indices of the matrix, and row indices are given in parentheses if necessary. For example, u_i refers to the ith column of U, and $u_i(\mathbf{r}_\alpha)$ refers to the matrix element with row index $\mathbf{r}_\alpha$. We denote by M_{ij} a vector with a stacked column index ij, which refers to the $(i+(j-1)N_1)$th column of the matrix M. Here the index i ranges from 1 to N_1, and j from 1 to N_2, respectively. We also employ the following linear algebra notation. The matrix-vector product Av should be interpreted as

$$(Av)(\mathbf{r}) = \int A(\mathbf{r}, \mathbf{r}')v(\mathbf{r}')\,\mathrm{d}\mathbf{r}'.$$

Similarly the matrix-matrix product AB should be interpreted as

$$(AB)(\mathbf{r}, \mathbf{r}') = \int A(\mathbf{r}, \mathbf{r}'')B(\mathbf{r}'', \mathbf{r}')\,\mathrm{d}\mathbf{r}''.$$

The Hadamard product of two vectors $u \odot v$ should be interpreted as

$$(u \odot v)(\mathbf{r}) = u(\mathbf{r})v(\mathbf{r}).$$

Using the chain rule and the definition of $\mathcal{V}$ in (2.7), we have

$$\chi = \frac{\delta\rho}{\delta\mathcal{V}}\left(I + \frac{\delta\mathcal{V}}{\delta\rho}\frac{\delta\rho}{\delta V_{\mathrm{ion}}}\right) = \chi_0(I + v_{\mathrm{hxc}}\chi), \tag{2.13}$$

and thus

$$\chi = (I - \chi_0 v_{\mathrm{hxc}})^{-1}\chi_0 := \varepsilon^{-1}\chi_0. \tag{2.14}$$

Here I is the identity operator, and

$$v_{\mathrm{hxc}} = \frac{\delta\mathcal{V}}{\delta\rho} = v_c + \frac{\delta V_{\mathrm{xc}}}{\delta\rho} := v_c + f_{\mathrm{xc}}.$$

Here f_{xc} is called the exchange-correlation kernel and is a diagonal matrix in the LDA and GGA formulations of the exchange-correlation functionals. Equation (2.14) also defines the operator $\varepsilon = I - \chi_0 v_{\mathrm{hxc}}$, which is called the dielectric operator. Using linear algebra notation, (2.12) requires the computation of $g_{I,a}^T \chi g_{J,b}$. Here $g_{I,a} = \frac{\partial V_I}{\partial \mathbf{R}_{I,a}}(\mathbf{r} - \mathbf{R}_I)$ is the derivative of local pseudopotential V_I with respect to atomic position R_I along the ath direction ($a = 1, \dots, d$). Hence for phonon calculations, we need to compute χg for $d \times N_A$ vectors in the set of $\{g_{I,a}\}$. In the discussion below, we may simply use g to reflect any vector from the set $\{g_{I,a}\}$ unless otherwise specified.

In order to compute $u = \chi g$, we apply both sides of (2.13) to g and obtain

$$u = \chi_0 g + \chi_0 v_{\mathrm{hxc}} u. \tag{2.15}$$

Equation (2.15) is a fixed point problem for u and is often referred to as the Dyson equation in the physics literature. The simplest way to solve (2.15) is to use a fixed point iteration

$$u^{k+1} = \chi_0 g + \chi_0 v_{\mathrm{hxc}} u^k, \tag{2.16}$$

where u^k is the approximate solution to u at the kth iteration. The fixed point iteration (2.16) corresponds to the Neumann expansion of the matrix inverse in (2.14), and hence only converges when the spectral radius of $\chi_0 v_{\mathrm{hxc}}$ is small enough (usually smaller than 1). In order to improve the convergence behavior, more advanced numerical schemes such as Anderson mixing [2] can be used, similarly to the situation in the self-consistent field iteration for KSDFT calculations. We refer readers to [31] for more details on solving fixed point problems in the context of electronic structure calculations.

In order to solve the Dyson equation (2.15), we need to apply χ_0 to vectors of the form g or $v_{\mathrm{hxc}}u$. For systems with a finite band gap at the zero temperature, $\chi_0(\mathbf{r},\mathbf{r}')$ can be computed using the Adler–Wiser formula [1, 47],

$$\chi_0(\mathbf{r},\mathbf{r}') = 2\sum_{i=1}^{N_e}\sum_{j=N_e+1}^{\infty} \frac{\psi_i(\mathbf{r})\psi_j(\mathbf{r})\psi_i(\mathbf{r}')\psi_j(\mathbf{r}')}{\varepsilon_i-\varepsilon_j}, \tag{2.17}$$

where (ε_i,ψ_i), $i=1,2,\ldots,$ are the eigenpairs in (2.5). Note that χ_0 is a Hermitian operator, and is negative semidefinite since $\varepsilon_i<\varepsilon_j$. In linear algebra notation, χ_0 can be written as

$$\chi_0 = 2\sum_{i=1}^{N_e}\sum_{j=N_e+1}^{\infty}\frac{1}{\varepsilon_i-\varepsilon_j}(\psi_i\odot\psi_j)(\psi_i\odot\psi_j)^T.$$

However, the Adler–Wiser formula in its original form (2.17) can be very expensive, since it requires in principle *all* the unoccupied orbitals $\{\psi_j\}_{j=N_e+1}^{\infty}$. These unoccupied orbitals are in general not available in KSDFT calculations and are costly to compute and even to store in memory. However, from the Adler–Wiser formula, we can compute the multiplication of χ_0 with any given vector g, without computing nonoccupied orbitals, by solving the so-called Sternheimer equation [19]. This strategy has also been employed by several recent works related to the polarizability operator [45, 16, 34] in the context of the many-body perturbation theory. Introducing the projection operator to the unoccupied space

$$Q = I - \sum_{i=1}^{N_e}\psi_i\psi_i^T,$$

we can then compute $\chi_0 g$ as

$$\begin{aligned}\chi_0 g &= 2\sum_{i=1}^{N_e}\sum_{j=N_e+1}^{\infty}\frac{1}{\varepsilon_i-\varepsilon_j}(\psi_i\odot\psi_j)(\psi_j\odot\psi_i)^T g\\ &= 2\sum_{i=1}^{N_e}\psi_i\odot\left[\sum_{j=N_e+1}^{\infty}\psi_j(\varepsilon_i-\varepsilon_j)^{-1}\psi_j^T(\psi_i\odot g)\right]\\ &= 2\sum_{i=1}^{N_e}\psi_i\odot\left[Q(\varepsilon_i-H)^{-1}Q(\psi_i\odot g)\right].\end{aligned} \tag{2.18}$$

Here we have used that (ε_j,ψ_j) is an eigenpair of the Hamiltonian operator H. In principle, the last equation of (2.18) requires only one Q operator to be present. However, we choose the form $Q(\varepsilon_i-H)^{-1}Q$ to emphasize that this operator is symmetric. Equation (2.18) provides a practical numerical scheme for evaluating $\chi_0 g$. Let

$\zeta_i := Q(\varepsilon_i - H)^{-1}Q(\psi_i \odot g)$. The matrix inverse in (2.18) can be avoided by solving the Sternheimer equation [35]

$$Q(\varepsilon_i - H)Q\zeta_i = Q(\psi_i \odot g),$$

using standard direct or iterative linear solvers. The choice of the solver can depend on practical matters such as the discretization scheme and the availability of preconditioners. In practice, for planewave discretization, we find that the use of the minimal residual method (MINRES) [36] gives the best numerical performance. Algorithm 1 summarizes the algorithm for computing $\chi_0 g$, without explicitly computing the unoccupied orbitals.

Algorithm 1 Compute $\chi_0 g$.

```
Input:  Vector $g$, occupied orbitals $\{\psi_i\}_{i=1}^{N_e}$ and eigenvalues $\{\varepsilon_i\}_{i=1}^{N_e}$.
Output: $u = \chi_0 g$.
1. Initialize $u \leftarrow 0$.
2. For $i = 1, \ldots, N_e$
     (a) Solve the Sternheimer equation $Q(\varepsilon_i - H)Q\zeta_i = Q(\psi_i \odot g)$;
     (b) $u \leftarrow u + 2\psi_i \odot \zeta_i$;
   end for
```

The use of Algorithm 1 together with a proper method for solving the Dyson equation (2.15) gives rise to the basic formulation for phonon calculations in DFPT. This method will be simply referred to as DFPT for the discussion below. Assuming Dyson equations always converge in a constant number of iterations that is independent of the system size N_e, then the main cost of DFPT is associated with the application of χ_0 to $\mathcal{O}(N_e)$ vectors. Each application requires solving N_e equations, and hence there are $\mathcal{O}(N_e^2)$ equations to solve. The computational cost of applying the projection operator Q to a vector is $\mathcal{O}(N_e^2)$, and hence the overall complexity is $\mathcal{O}(N_e^4)$ [3]. Compared to the "frozen phonon" approach discussed in section 1, DFPT provides an accurate description of the linear response properties of materials.

3. Adaptively compressed polarizability operator.

In this section, we develop a new method for reducing the computational complexity of DFPT from $\mathcal{O}(N_e^4)$ to $\mathcal{O}(N_e^3)$. The reduction of the computational complexity is achieved by means of reducing the $\mathcal{O}(N_e^2)$ equations in DFPT to $\mathcal{O}(N_e)$ equations with systematic control of the accuracy. In particular, our method does not employ the "nearsightedness" property of electrons for insulating systems with substantial band gaps as in linear scaling methods [26]. Hence our method can be applied to insulators as well as semiconductors with small band gaps. In section 2, we have reduced the problem of computing the dynamic matrix to the computation of $\chi g_{I,a}$, where $\{g_{I,a}\}$ is a set of fixed vectors given by the derivative of the local pseudopotential with respect to the atomic positions. Let us stack the indices I, a into a single index j and denote by

$$G := [g_1, \ldots, g_j, \ldots, g_{d\times N_A}] \tag{3.1}$$

the matrix collecting all these vectors. More generally, G can be any fixed matrix with $\mathcal{O}(N_e)$ columns as required in different applications of DFPT. Then our method consists of two main steps: (1) Find a compressed representation of χ_0, which allows the computation of $\chi_0 G$ by solving only $\mathcal{O}(N_e)$ linear equations. (2) Update the compressed representation of χ_0, which allows the accurate computation of χG without significant increase of the computational cost. In particular, step 2 requires the compression strategy of χ_0 to be *adaptive* to the solution of the Dyson equation (2.15). Hence we refer to our representation of χ_0 as the adaptively compressed polarizability operator (ACP). Steps 1 and 2 of the ACP formulation are given in sections 3.1 and 3.2, respectively.

3.1. Compression of χ_0.

Consider first the computation of $\chi_0 G$ as required in the initial step in (2.15). In general, the singular values of χ_0 decay slowly, and a forcefully applied low rank decomposition of χ_0, such as those based on SVD, will lead to inaccurate results. Nonetheless, it is possible to find a compressed representation of χ_0 when we only need to evaluate $\chi_0 G$ for a fixed matrix G.

According to Algorithm 1, computing $\chi_0 G$ involves solving the following $\mathcal{O}(N_e^2)$ Sternheimer equations:

$$Q(\varepsilon_i - H)Q\zeta_{ij} = Q(\psi_i \odot g_j), \quad i = 1, \dots, N_e, \quad j = 1, \dots, d \times N_A. \tag{3.2}$$

As N_e becomes large, asymptotically there can be many more equations to solve than the dimension of the matrix $N_g \sim \mathcal{O}(N_e)$, and hence it should be possible to compress the redundant information in the right-hand side vectors. In fact this observation has been used in various contexts in computational chemistry for compressing the Hadamard product of occupied and unoccupied orbitals, which is called "density fitting" or "resolution of identity" techniques to compress $\mathcal{O}(N_e^2)$ vectors into $\mathcal{O}(N_e)$ vectors with a relatively small preconstant [46, 40].

It should be noted that density fitting techniques alone do not reduce the number of equations to solve. The reason is that equations (3.2) have the dependence on the shift ε_i on the left-hand side. Hence even if the number of right-hand side vectors is reduced to $\mathcal{O}(N_e)$, multiplied with the N_e shifts, we still have $\mathcal{O}(N_e^2)$ equations to solve! Therefore, in order to reduce the complexity for computing $\chi_0 G$, we must disentangle the right-hand side vectors and the shifts. Note that all $\{\varepsilon_i\}$ are eigenvalues corresponding to occupied orbitals, and are typically contained in a relatively small interval (in the order of eV), at least in the pseudopotential framework.

More specifically, consider the parameterized equation

$$Q(\varepsilon - H)Q\zeta = \xi, \tag{3.3}$$

where ξ is any vector in the range of Q. Since $\varepsilon \in \mathcal{I} \equiv [\varepsilon_1, \varepsilon_{N_e}]$, we can systematically obtain the solution to the parameterized equation by evaluating on a few sampled points in $\mathcal{I}$. In this work, we choose the Chebyshev nodes $\{\widetilde{\varepsilon}_c\}_{c=1}^{N_c}$, which are obtained by a linear map of the Chebyshev nodes in the reference interval $[-1, 1]$ to $\mathcal{I}$, i.e.,

$$\widetilde{\varepsilon}_c = \frac{\varepsilon_1 + \varepsilon_{N_e}}{2} + \frac{\varepsilon_1 - \varepsilon_{N_e}}{2} \cos\theta_c, \quad \theta_c = \frac{\pi(c - \frac{1}{2})}{N_c}, \quad c = 1, \dots, N_c.$$

Typically it is sufficient to choose the number of Chebyshev nodes N_c to be $10 \sim 40$. Denote by $\widetilde{\zeta}_c$ the solution to (3.3) corresponding to $\varepsilon = \widetilde{\varepsilon}_c$, $c = 1, \dots, N_c$; then any solution ζ with $\varepsilon \in \mathcal{I}$ can be obtained by a Lagrange interpolation procedure as

$$\zeta = \sum_{c=1}^{N_c} \widetilde{\zeta}_c \prod_{c' \neq c} \frac{\varepsilon - \widetilde{\varepsilon}_{c'}}{\widetilde{\varepsilon}_c - \widetilde{\varepsilon}_{c'}}. \tag{3.4}$$

Asymptotically, the number of Chebyshev points needed to reach a given error tolerance is $\mathcal{O}(\sqrt{\varepsilon_g/|\mathcal{I}|})$, where $|\mathcal{I}|$ is the width of the interval $\mathcal{I}$. To understand this, let us consider the error of the Chebyshev interpolation for the following scalar function:

$$f(z) = \frac{|\mathcal{I}|}{(z-1)\frac{|\mathcal{I}|}{2} - \varepsilon_g}, \qquad z \in [-1,1]. \tag{3.5}$$

Note that

$$f(-1) = \frac{|\mathcal{I}|}{-|\mathcal{I}| - \varepsilon_g} = \frac{|\mathcal{I}|}{\varepsilon_1 - \varepsilon_{N_e+1}}, \quad f(1) = \frac{|\mathcal{I}|}{-\varepsilon_g} = \frac{|\mathcal{I}|}{\varepsilon_{N_e} - \varepsilon_{N_e+1}}.$$

Therefore $f(z)$ reflects the worst-case behavior of the operator $Q(\varepsilon - H)^{-1}Q$. It is known that (see, e.g., [43]) the L^∞ error of Chebyshev interpolation with N_c points on the interval $[-1,1]$, denoted by E_{N_c}, should satisfy

$$E_{N_c} \le \frac{2M}{(\alpha-1)\alpha^{N_c}}.$$

Here the function $f(z)$ is analytic in the region bounded by the ellipse with foci ±1 and major and minor semiaxis lengths summing to $\alpha > 1$, and $|f(z)| < M$ is bounded in this region. For the specific function $f(z)$ in (3.5) with the presence of the band gap ε_g, the major semiaxis could be chosen to be $1 + \varepsilon_g/|\mathcal{I}|$, which means that $\alpha \approx 1 + \sqrt{2\varepsilon_g/|\mathcal{I}|} + \varepsilon_g/|\mathcal{I}| \approx 1 + \sqrt{2\varepsilon_g/|\mathcal{I}|}$. Within this ellipse, we obtain the bound $M = 2|\mathcal{I}|/\varepsilon_g$. This gives the error bound of Chebyshev interpolation as

$$E_{N_c} \le C \left(\frac{\mathcal{I}}{\varepsilon_g}\right)^{3/2} e^{-N_c\sqrt{2\varepsilon_g/|\mathcal{I}|}}. \tag{3.6}$$

When $\varepsilon_g/|\mathcal{I}|$ is sufficiently small, C is a constant that is independent of $\varepsilon_g/|\mathcal{I}|$. Therefore N_c should scale as $\sqrt{|\mathcal{I}|/\varepsilon_g}\log(|\mathcal{I}|/\varepsilon_g)$. Our numerical results in section 4 indicate that $N_c = 10 \sim 30$ often yields sufficiently accurate results. We also remark that if $\varepsilon_g/|\mathcal{I}|$ is very small, one can further reduce the number of interpolation points using contour integral techniques [14], where $N_c = \mathcal{O}(\log(|\mathcal{I}|/\varepsilon_g))$ is sufficient. However, we observe that the preconstant of such a technique tends to be larger than that of the Chebyshev interpolation. Hence we choose to present the method using Chebyshev interpolation in this paper.

Using Chebyshev interpolation (3.4), we need to solve (3.2) with ε_i replaced by $\widetilde{\varepsilon}_c$. At first sight, the number of equations does not decrease but actually increases by a factor of N_c compared to the original formulation (3.2). However, Chebyshev interpolation disentangles the index i that appears both in the shift and on the right-hand side. Since N_c is a constant that is independent of the system size, if we can find a compressed representation of the right-hand side vectors using $\mathcal{O}(N_e)$ vectors, we reduce to $\mathcal{O}(N_e)$ the overall number of equations we need to solve.

Let us denote by M the collection of right-hand side vectors in (3.2) without the Q factor. More specifically,

$$M_{ij} = \psi_i \odot g_j, \quad \text{or} \quad M_{ij}(\mathbf{r}) = \psi_i(\mathbf{r})g_j(\mathbf{r}).$$

Here we have used ij as a stacked column index for the matrix M. The dimension of M is $N_g \times \mathcal{O}(N_e^2)$. Typically, the computational complexity for the compression for such a dense matrix M with approximate rank $\mathcal{O}(N_e)$ is $\mathcal{O}(N_e^4)$, even with the help of the recently developed randomized algorithms (see [22] for a review). Nonetheless, note that for a fixed row index $\mathbf{r}_\alpha$, the row vector given by $\{M_{ij}(\mathbf{r}_\alpha)\}_{i,j=1}^{i=N_e,j=d\times N_A}$ is the Kronecker product between the row vector given by $\{g_j(\mathbf{r}_\alpha)\}_{j=1}^{d\times N_A}$ and that given by $\{\psi_i(\mathbf{r}_\alpha)\}_{i=1}^{N_e}$. As will be seen below, this structure allows the computational complexity of the compression of M to be reduced to $\mathcal{O}(N_e^3)$.

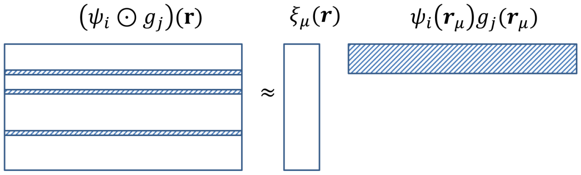

$(\psi_i \odot g_j)(\mathbf{r})$
$\approx$
$\xi_\mu(\mathbf{r})$
$\psi_i(\mathbf{r}_\mu)g_j(\mathbf{r}_\mu)$

FIG. 1. *Interpolative decomposition of* $M_{ij}(\mathbf{r})$.

To this end, we seek the following interpolative decomposition type of compression [13] for the matrix M, i.e.,

$$M_{ij}(\mathbf{r}) \approx \sum_{\mu=1}^{N_\mu} \xi_\mu(\mathbf{r}) M_{ij}(\mathbf{r}_\mu) \equiv \sum_{\mu=1}^{N_\mu} \xi_\mu(\mathbf{r})\psi_i(\mathbf{r}_\mu)g_j(\mathbf{r}_\mu). \tag{3.7}$$

Here $\{\mathbf{r}_\mu\}_{\mu=1}^{N_\mu}$ denotes a collection of selected row indices (see Figure 1 for an illustration). Mathematically, the meaning of the indices $\{\mathbf{r}_\mu\}$ is clear: (3.7) simply states that all rows $M_{:}(\mathbf{r})$ can be approximately expressed as the linear combination of the selected rows $\{M_{:}(\mathbf{r}_\mu)\}$. However, we are not aware of any direct physical interpretation of such selected indices. Since $N_g \sim N_e$, as N_e increases, the column dimension of M (which is $\mathcal{O}(N_e^2)$) can be larger than its row dimension (which is N_g), and we can expect that the vectors $\{\psi_i \odot g_j\}$ are approximately linearly dependent. Such an observation has been observed in the electronic structure community [46, 41, 15, 44, 40], and the numerical rank of the matrix M after truncation can be only $\mathcal{O}(N_e)$ with a relatively small preconstant. In the context of the interpolative decomposition, our numerical results also indicate that it is sufficient to choose $N_\mu \sim \mathcal{O}(N_e)$, and the preconstant is small.

One possible way of finding interpolative decomposition is to use a pivoted QR factorization [12, 21]. However, the cost of the pivoted QR factorization applied to the matrix M is $\mathcal{O}(N_g N_e^2 N_\mu) \sim \mathcal{O}(N_e^4)$ and is therefore not desirable. The interpolative separable density fitting method [32] employs a two-step procedure to reduce this cost (see Figure 2 for an illustration). The first step is to use a fast preprocessing procedure, such as a subsampled random Fourier transform (SRFT) [48], to transform the matrix M into a matrix $\widetilde{M}$ of smaller dimension $N_g \times rN_e$, with r a relatively small constant so that rN_e is slightly larger than N_μ. The second step is to apply the pivoted QR decomposition to $\widetilde{M}$,

$$\widetilde{M}^T \widetilde{\Pi} = \widetilde{Q}\widetilde{R}, \tag{3.8}$$

where $\widetilde{\Pi}$ is a permutation matrix and encodes the choice of the row indices $\{\mathbf{r}_\mu\}$ from $\widetilde{M}$. The interpolation vectors $\{\xi_\mu\}$ in (3.7) can also be computed from this pivoted QR decomposition. It should be noted that the preprocessing procedure does not affect the quality of the interpolative decomposition, while the cost of the pivoted QR factorization in (3.8) is now reduced to $\mathcal{O}(N_g N_\mu^2) \sim \mathcal{O}(N_e^3)$. We summarize the

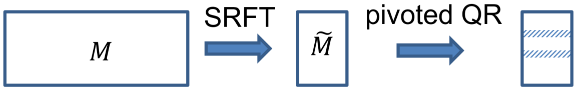


FIG. 2. *The two-step procedure of the interpolative separable density fitting method.*

Algorithm 2 Interpolative decomposition for M using an interpolative separable density fitting method [32].

```
Input: Matrix $M$. Threshold tolerance $\epsilon$.
Output: Selected row indices $\{\mathbf{r}_\mu\}$ and interpolation vectors $\{\xi_\mu\}$.
1. Subsampled random Fourier transform of $M$:
    (a) Compute for $\nu = 1, \ldots, N_e \times d \times N_A$ the discrete Fourier transform
        $$\hat{M}_\nu(\mathbf{r}) = \sum_{I=1}^{N_e \times d \times N_A} e^{-2\pi\imath I\nu/(N_e \times d \times N_A)} \eta_I M_I(\mathbf{r}),$$
        where $\eta_I$ is a random complex number with unit modulus for each $I$.
    (b) Choose a submatrix $\widetilde{M}$ of matrix $\hat{M}$ by randomly choosing $rN_e$ columns. In practice, $r = 8$ and $r = 16$ are used in our implementation for 1D and 2D numerical examples, respectively.
2. Compute the pivoted QR decomposition of the $rN_e \times N_g$ matrix $\widetilde{M}^T$ : $\widetilde{M}^T\widetilde{\Pi} = \widetilde{Q}\widetilde{R}$, where the absolute values of the diagonal entries of $\widetilde{R}$ are ordered nonincreasingly.
3. Determine the number of selected columns $N_\mu$, such that $|\widetilde{R}_{N_\mu+1,N_\mu+1}| < \epsilon|\widetilde{R}_{1,1}| \leq |\widetilde{R}_{N_\mu,N_\mu}|$. Form $\{\mathbf{r}_\mu\}$, $\mu = 1, \ldots, N_\mu$, such that the $\mathbf{r}_\mu$ column of $\widetilde{M}^T$ corresponds to the $\mu$th column of $\widetilde{M}^T\widetilde{\Pi}$.
4. Denote by $\widetilde{R}_{1:N_\mu,1:N_\mu}$ the submatrix of $\widetilde{R}$ consisting of its first $N_\mu \times N_\mu$ entries, and $\widetilde{R}_{1:N_\mu,:}$ the submatrix consisting of the first $N_\mu$ rows of $\widetilde{R}$. Compute
        $$\Xi^T = \widetilde{R}^{-1}_{1:N_\mu,1:N_\mu} \widetilde{R}_{1:N_\mu,:} \widetilde{\Pi}^{-1}.$$
   Then the $\mu$th column of the $N_g \times N_\mu$ matrix $\Xi$ gives the interpolation vector $\xi_\mu$.
```

procedure for compressing M in Algorithm 2 and refer readers to [32] for more details of the algorithm.

We remark that in Algorithm 2, step 1(a), it is possible to avoid the explicit construction of the matrix M. Instead of performing SRFT on the entire matrix M, we could apply SRFT only to the matrix G and select r columns as a matrix $\widetilde{G}$. Then for a fixed row index $\mathbf{r}_\alpha$, the Kronecker product between the rows of subsampled matrix $\widetilde{G}$, $\{\widetilde{g}_i(\mathbf{r}_\alpha)\}_{i=1}^{r}$, and $\{\psi_i(\mathbf{r}_\alpha)\}_{i=1}^{N_e}$ gives one row for $\widetilde{M}$. In practice we find that this heuristic procedure also works well for compressing the matrix M in phonon calculations.

Once the compressed representation (3.7) is obtained, we solve the following set of modified Sternheimer equations:

$$Q(\widetilde{\varepsilon}_c - H)Q\widetilde{\zeta}_{c\mu} = Q\xi_\mu, \quad c = 1, \ldots, N_c, \quad \mu = 1, \ldots, N_\mu. \tag{3.9}$$

Here $c\mu$ is the stacked column index for $\widetilde{\zeta}$. The number of equations is hence reduced to $N_cN_\mu \sim \mathcal{O}(N_e)$. Using (3.4), we construct the quantity $W = \left[W_1, \ldots, W_{N_\mu}\right]$. Each column of W is defined by

$$W_\mu = 2\sum_{i=1}^{N_e}\psi_i \odot \left(\sum_{c=1}^{N_c}\widetilde{\zeta}_{c\mu}\prod_{c'\neq c}\frac{\varepsilon_i-\widetilde{\varepsilon}_{c'}}{\widetilde{\varepsilon}_c-\widetilde{\varepsilon}_{c'}}\right)\psi_i(\mathbf{r}_\mu). \tag{3.10}$$

Combining (3.10) with (2.18), we directly obtain $\chi_0 g_j$ as

$$\chi_0 g_j \approx \sum_{\mu=1}^{N_\mu} W_\mu g_j(\mathbf{r}_\mu). \tag{3.11}$$

It should be noted that in (3.11), we have avoided the explicit reconstruction of the solution vectors ζ_{ij} as in equations (3.2), of which the computational cost is again $\mathcal{O}(N_e^4)$.

Formally, (3.11) can further be simplified by defining a matrix Π with N_μ columns, which consists of selected columns of a permutation matrix, i.e., $\Pi = \widetilde{\Pi}_{:,1:N_\mu}$ as the first N_μ columns of the permutation matrix obtained from pivoted QR decomposition. More specifically, $\Pi_\mu = e_{\mathbf{r}_\mu}$ and $e_{\mathbf{r}_\mu}$ is a unit vector with only one nonzero entry at $\mathbf{r}_\mu$ such that $e_{\mathbf{r}_\mu}^T g_j = g_j(\mathbf{r}_\mu)$. Then

$$\chi_0 g_j \approx W\Pi^T g_j := \widetilde{\chi}_0 g_j. \tag{3.12}$$

Note that the compressed polarizability operator $\widetilde{\chi}_0 = W\Pi^T$ is formally independent of the right-hand side vector $\{g_j\}$, and the rank of $\widetilde{\chi}_0$ is only N_μ, while the singular values of χ_0 have a much slower decay rate. This is because $\widetilde{\chi}_0$ only agrees with χ_0 when applied to vectors g_j. In other words, the difference between $\widetilde{\chi}_0$ and χ_0 is not controlled in the space orthogonal to that spanned by G. Algorithm 3 summarizes the algorithm for computing the compressed polarizability operator $\widetilde{\chi}_0$.

Algorithm 3 Computing compressed polarizability operator $\widetilde{\chi}_0$.

```
Input:  Vectors $\{g_j\}$. Hamiltonian matrix $H$.
        Eigenpairs corresponding to occupied orbitals $\{\psi_i, \varepsilon_i\}$
Output: $\widetilde{\chi}_0 = W\Pi^T$.
1. Use Algorithm 2 to obtain $\{\mathbf{r}_\mu\}$, $\Pi$, and hence the compressed representation of $M$.
2. Solve compressed equations (3.9).
3. Compute $W$ using (3.10).
```

The computational complexity of Algorithm 3 can be analyzed as follows. For simplicity we neglect all possible logarithmic factors in the complexity analysis, The cost for constructing the compressed representation of M is $\mathcal{O}(N_e^3)$. Equations (3.9) require solving $N_c N_\mu \sim \mathcal{O}(N_e)$ equations. Assuming the computational cost for applying H to a vector is $\mathcal{O}(N_g)$, and assuming that the number of iterations using an iterative solver to solve equations (3.9) is bounded by a constant, then the cost for solving all equations is dominated by the computation of $\{Q\xi_\mu\}$, which is $\mathcal{O}(N_e^3)$. In order to construct W, for each μ and i, we can first compute the term in parentheses on the right-hand side of (3.10). Then the computational complexity for constructing W is again $\mathcal{O}(N_e^3)$. Therefore, the overall asymptotic computational cost for constructing the compressed polarizability operator $\widetilde{\chi}_0$ is $\mathcal{O}(N_e^3)$. In practice, we find that the computational cost is dominated by solving the $\mathcal{O}(N_e)$ linear equations in step 2 of Algorithm 3.

3.2. Compression of χ.

According to (2.15), $\chi_0 G$ is the leading order approximation to $U = \chi G$, and this approximation can be inaccurate if $\chi_0 v_{\mathrm{hxc}}$ is not small. From the perspective of section 3.1, the self-consistent solution to the Dyson equation (2.15) introduces two additional difficulties: (1) we need to find compressed representation $\widetilde{\chi}_0$ that agrees with χ_0 when applied to both G and $v_{\mathrm{hxc}} U$; (2) U is not known a priori. Hence if we apply Algorithm 3 directly, we may need to increase the rank of $\widetilde{\chi}_0$ to $2N_\mu$ or higher to maintain the accuracy. Below we introduce the ACP method that simultaneously addresses these two difficulties.

We assume that v_{hxc} is invertible, and $v_{\mathrm{hxc}}^{-1} g$ for a vector g can be computed easily. This is the case in the absence of the exchange-correlation kernel f_{xc}, and $v_{\mathrm{hxc}}^{-1} g = v_c^{-1} g$ can simply be obtained by applying the Laplacian operator to g. This approximation is referred to as the "random phase approximation" in the physics literature [35]. In the presence of f_{xc} in the LDA and GGA formulations, f_{xc} is a diagonal matrix, and $v_{\mathrm{hxc}}^{-1} g = v_c^{-1} g$ can be solved using iterative methods.

We introduce the change of variables

$$U = \widetilde{U} - B, \quad B = v_{\mathrm{hxc}}^{-1} G, \tag{3.13}$$

and the Dyson equation (2.15) becomes

$$\widetilde{U} = \chi_0 v_{\mathrm{hxc}} \widetilde{U} + B. \tag{3.14}$$

The advantage of using (3.14) over (2.15) is that formally we only need to find $\widetilde{\chi}_0$ that is accurate when applied to $v_{\mathrm{hxc}} \widetilde{U}$. In an iterative algorithm, for a given matrix $\widetilde{U}$, we can use Algorithm 3 to construct $\widetilde{\chi}_0[\widetilde{U}]$ by replacing G with $v_{\mathrm{hxc}} \widetilde{U}$, with $\widetilde{U}$ in brackets to highlight the $\widetilde{U}$-dependence of the compression scheme, i.e.,

$$\widetilde{U} = \widetilde{\chi}_0[\widetilde{U}] v_{\mathrm{hxc}} \widetilde{U} + B. \tag{3.15}$$

We note that when self-consistency is reached for (3.15), with the self-consistent solution denoted by $\widetilde{U}^*$, $\widetilde{\chi}_0[\widetilde{U}^*] v_{\mathrm{hxc}} \widetilde{U}^*$ remains a good approximation to $\chi_0 v_{\mathrm{hxc}} \widetilde{U}^*$, even if $\widetilde{U}^*$ deviates away from the initial guess. In each step, the approximate rank of $\widetilde{\chi}_0[\widetilde{U}]$ remains N_μ. Hence $\widetilde{\chi}_0[\widetilde{U}]$ is adaptive to the solution $\widetilde{U}$ and hence is called the adaptively compressed polarizability operator (ACP). This concept of adaptively constructing a low rank matrix is similar in spirit to the recently developed adaptively compressed exchange operator (ACE) for the efficient solution of Hartree–Fock-like calculations [29].

Equation (3.15) can be solved using the fixed point iteration or more advanced methods for solving fixed point problems, similarly to that in (2.15) in DFPT. However, thanks to the low rank structure of $\widetilde{\chi}_0$ in (3.12), we can significantly accelerate the convergence. Let us denote the value of $\widetilde{U}$ at the kth iteration as $\widetilde{U}^k$, which gives rise to the ACP $\widetilde{\chi}_0[\widetilde{U}^k] = W^k (\Pi^k)^T$. Equation (3.14) indicates that if the magnitude of $\chi_0 v_{\mathrm{hxc}}$ is small, then $\widetilde{U}^0 = B$ is a good initial guess to start the iteration. Then we can reformulate (3.15) and obtain the following iteration scheme:

$$\widetilde{U}^{k+1} = \left(I - W^k (\Pi^k)^T v_{\mathrm{hxc}}\right)^{-1} B = B + W^k \left(I - (\Pi^k)^T v_{\mathrm{hxc}} W^k\right)^{-1} (\Pi^k)^T v_{\mathrm{hxc}} B. \tag{3.16}$$

The second equality in (3.16) uses the Sherman–Morrison–Woodbury identity for computing the inverse. The cost of the inversion is $\mathcal{O}(N_e^3)$ due to the low rank structure of $\widetilde{\chi}_0[\widetilde{U}^k]$. Numerical results indicate that the iteration scheme (3.16) can converge much more rapidly compared to the fixed point iteration for (2.15). In fact often two to four iterations are enough to obtain results that are sufficiently accurate. Algorithm 4 describes the algorithm for using ACP to compute χG.

Algorithm 4 Computing χG with adaptively compressed polarizability operator.

```
Input:
Vectors $\{g_j\}$. Stopping criterion $\delta$.
Eigenpairs corresponding to occupied orbitals $\{\psi_i, \varepsilon_i\}$
Output: $U \approx \chi G$
1. Compute $\widetilde{U}^0 = B = v_{\mathrm{hxc}}^{-1} G$. $k \leftarrow 0$.
2. Do
    (a) Use Algorithm 3 by replacing $G$ with $v_{\mathrm{hxc}} \widetilde{U}^k$ to obtain $W^k$ and $\Pi^k$, and obtain $\widetilde{\chi}_0^k = W^k (\Pi^k)^T$.
    (b) Update $\widetilde{U}^{k+1}$ according to (3.16).
    (c) $k \leftarrow k + 1$
  until $\|\widetilde{U}^k - \widetilde{U}^{k-1}\| < \delta$ or maximum number of iterations is reached.
3. Compute $U \leftarrow \widetilde{U}^k - B$.
```

4. Numerical examples. In this section, we demonstrate the performance of ACP, proposed in the previous section, and compare it with the density functional perturbation theory (DFPT) and with the finite difference approach (FD) through three examples. The first example consists of a 1D reduced Hartree–Fock model problem that can be tuned to resemble an insulating or a semiconducting system. The second example is a 2D model problem with a periodic triangular lattice structure. The third example is a 2D triangular lattice with defects and random perturbations of the atomic positions. Since our computational domain involve a large number of atoms, our treatment of using a system of finite size with periodic boundary conditions is equivalent to the Gamma point sampling strategy of the Brillouin zone for an infinite-sized system [33]. All results are performed on a single computational core of a 1.4 GHz processor with 256 GB memory using MATLAB.

4.1. 1D reduced Hartree–Fock model. The 1D reduced Hartree–Fock model was introduced by Solovej [42] and has been used for analyzing defects in solids; see, e.g., [8, 9]. The simplified 1D model neglects the contribution of the exchange-correlation term. As discussed in previous sections, the presence of exchange-correlation functionals at the LDA/GGA level does not lead to essential difficulties in phonon calculations.

The Hamiltonian in our 1D reduced Hartree–Fock model is given by

$$H[\rho] = -\frac{1}{2}\frac{d^2}{dx^2} + \int K(x,y)\left(\rho(y) + m(y)\right) \mathrm{d}y. \tag{4.1}$$

Here $m(x) = \sum_I m_I(x - R_I)$ is the summation of pseudocharges. Each function $m_I(x)$ takes the form of a 1D Gaussian,

$$m_I(x) = -\frac{Z_I}{\sqrt{2\pi\sigma_I^2}} \exp\left(-\frac{x^2}{2\sigma_I^2}\right), \tag{4.2}$$

where Z_I is an integer representing the charge of the Ith nucleus. In our numerical simulation, we choose all σ_I to be the same.

Instead of using a bare Coulomb interaction which diverges in one dimension when x is large, we use a Yukawa kernel as the regularized Coulomb kernel,

$$K(x,y) = \frac{2\pi e^{-\kappa|x-y|}}{\kappa\epsilon_0}, \tag{4.3}$$

which satisfies the equation

$$-\frac{d^2}{dx^2}K(x,y)+\kappa^2 K(x,y)=\frac{4\pi}{\epsilon_0}\delta(x-y). \tag{4.4}$$

As $\kappa \to 0$, the Yukawa kernel approaches the bare Coulomb interaction given by the Poisson equation. The parameter ϵ_0 is used so that the magnitude of the electron static contribution is comparable to that of the kinetic energy. The ion-ion repulsion energy E_{II} is also computed using the Yukawa interaction K in the model systems.

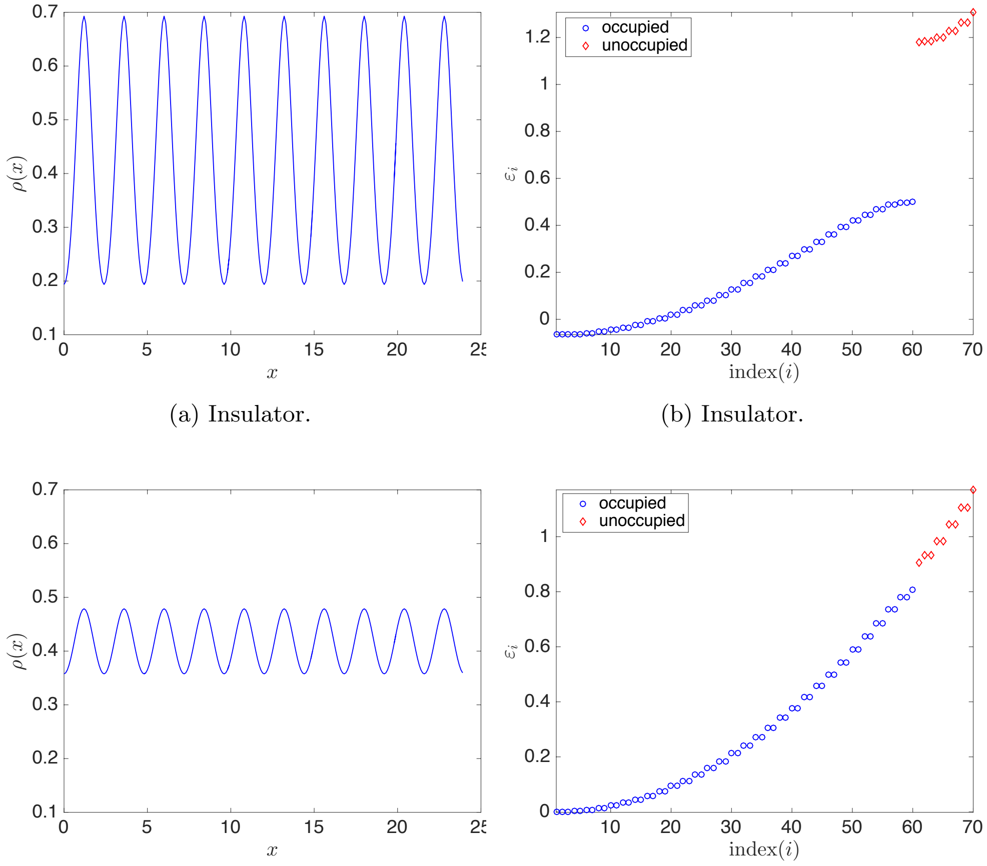
The parameters used in the model are chosen as follows. Atomic units are used throughout the discussion unless otherwise mentioned. For all systems tested in this subsection, the distance between each atom and its nearest neighbor is set to 2.4 a.u. The Yukawa parameter $\kappa = 0.1$. The nuclear charge Z_I is set to 1 for all atoms, and σ_I is set to be 0.3. The Hamiltonian operator is represented in a plane wave basis set.

By adjusting the parameter $\epsilon_0 = 1.0$ or 10, the reduced Hartree–Fock model can be tuned to resemble an insulator or a semiconductor, respectively. We apply ACP to both cases. We use Anderson mixing for self-consistent field (SCF) iterations, and the linearized eigenvalue problems are solved by using the locally optimal block preconditioned conjugate gradient (LOBPCG) solver [25].

For systems of size $N_A = 60$, the converged electron density ρ associated with the two 1D test cases as well as the 70 smallest eigenvalues associated with the Hamiltonian defined by the converged ρ are shown in Figure 3. For the insulator case, the electron density fluctuates between 0.1935 and 0.6927. There is a finite HOMO-LUMO gap, $\varepsilon_g = \varepsilon_{61} - \varepsilon_{60} = 0.6763$. The electron density associated with the semiconductor case is relatively uniform in the entire domain, with the fluctuation between 0.3576 and 0.4788. The corresponding band gap is 0.1012. Figure 3 is obtained by a system with 60 atoms, and we find that systems with different sizes show similar patterns in the band structure for both insulating and semiconducting systems, respectively.

All numerical results of the ACP method below are benchmarked with results obtained from DFPT. In order to demonstrate the effectiveness of the ACP formulation for compressing $U = \chi G$, we directly measure the relative L^2 error, defined as $\|U - U^{\mathrm{ACP}}\|_2/\|U\|_2$, where U^{ACP} is obtained from Algorithm 4. We also report the error of the phonon spectrum by computing the L^∞ error of the phonon frequencies $\{\omega_k\}$. Due to the presence of acoustic phonon modes for which ω_k is close to 0, we report the absolute error instead of the relative error for the phonon frequencies. In DFPT, we use MINRES [36] to solve the Sternheimer equations iteratively. The initial guess vectors for the solutions are obtained from previous iterations in the Dyson equation to reduce the number of matrix-vector multiplications. The same strategy for choosing the initial guess is implemented for the ACP formulation as well. Anderson mixing is used to accelerate the convergence of Dyson equations in DFPT. In ACP we find that the fixed point iteration (3.16) is sufficient for fast convergence.

In Tables 1 and 2, we calibrate the accuracy of our algorithm with different numbers of Chebyshev points N_c and different numbers of columns N_μ, for both insulating and semiconducting systems, respectively. We choose $N_\mu = lN_e$, where $l = 3, 4, \dots, 8$. Tables 1 and 2 show that for both insulating and semiconducting systems, with fixed Chebyshev interpolation points N_c, the numerical accuracy increases monotonically with respect to N_μ, until limited by the accuracy of the Chebyshev interpolation procedure. Note that when the accuracy is limited by the Chebyshev interpolation, the error can saturate, as shown in each column in both Tables 1 and 2. Similarly the increase of Chebyshev interpolation reduces the error until being limited by the choice of N_μ. When both N_μ and N_c are large enough, the error of $\chi_0 G$ can be less than 10^{-6}.



(a) Insulator.

(b) Insulator.

(c) Semiconductor.

(d) Semiconductor.

Fig. 3. *The electron density $\rho(x)$ of the 60-atom* (a) *insulating and* (c) *semiconducting systems in the left panel. The corresponding occupied (blue circles) and unoccupied eigenvalues (red diamonds) are shown in the right panel in* (b), (d)*, respectively. (Color available online.)*

Table 1
The relative L^2 error $|\widetilde{\chi}_0 G - \chi_0 G|/|\chi_0 G|$ for the insulating system with $\varepsilon_g/|\mathcal{I}| = 1.1911$.

N_c \ N_μ	$3N_e$	$4N_e$	$5N_e$	$6N_e$	$7N_e$	$8N_e$
5	6.90E-03	8.91E-04	7.56E-05	9.17E-06	8.49E-06	8.45E-06
10	6.83E-03	7.83E-04	7.31E-05	2.32E-06	4.65E-07	3.40E-07
15	7.84E-03	8.66E-04	4.92E-05	2.59E-06	3.22E-07	1.11E-07
20	7.53E-03	8.20E-04	5.61E-05	2.68E-06	2.93E-07	2.89E-07
25	8.77E-03	8.80E-04	5.48E-05	2.41E-06	3.95E-07	1.23E-07
30	1.08E-02	8.04E-04	5.71E-05	2.95E-06	3.36E-07	2.76E-07

In order to show how N_c scales with respect to the size of $\varepsilon_g/|\mathcal{I}|$, we adjust the parameter $\epsilon_0 = 1.8^{i-1}$, $i = 1, \ldots, 6$, to get systems with different band gaps. The result is reported in Figure 4. By selecting $N_\mu = 6N_e$, the number of Chebyshev nodes which is required to get relative L^2 error $|\widetilde{\chi}_0 G - \chi_0 G|/|\chi_0 G| \sim 10^{-5}$ scales as $\sqrt{\varepsilon_g/|\mathcal{I}|}$, which matches the analysis in section 3.1.

TABLE 2
The relative L^2 error $|\widetilde{\chi}_0 G - \chi_0 G|/|\chi_0 G|$ for the semiconducting system with $\varepsilon_g/|\mathcal{I}| = 0.1253$.

N_μ / N_c	$3N_e$	$4N_e$	$5N_e$	$6N_e$	$7N_e$	$8N_e$
5	2.87E-02	1.02E-02	1.01E-02	1.01E-02	1.01E-02	1.01E-02
10	3.96E-02	4.99E-04	1.82E-04	1.71E-04	1.71E-04	1.71E-04
15	1.75E-02	6.30E-04	6.07E-05	9.33E-06	5.59E-06	4.78E-06
20	4.84E-02	4.47E-04	7.04E-05	8.45E-06	3.24E-06	4.05E-07
25	3.04E-02	5.08E-04	6.95E-05	6.71E-06	2.55E-06	4.57E-07
30	2.29E-02	5.50E-04	5.95E-05	9.66E-06	2.56E-06	3.69E-07

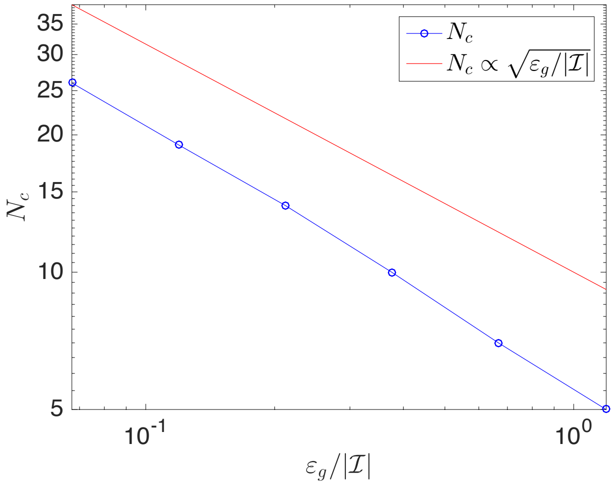


FIG. 4. *Scale of N_c with respect to the size of $\varepsilon_g/|\mathcal{I}|$, compared to the theoretical square root scaling. System size $N_A = 60$. $N_\mu = 6N_e$. The relative L^2 error $|\widetilde{\chi}_0 G - \chi_0 G|/|\chi_0 G| \sim 10^{-5}$.*

In Table 3, we choose N_μ based on the entries of $\widetilde{R}$, as is shown in Algorithm 2, and compare the results to those obtained from the FD approach. In the FD approach, we set the convergence tolerance for LOBPCG to be 10^{-6}, and the SCF tolerance to be 10^{-8}. $\delta = 0.01$ denotes the deviation of atom positions to their equilibrium ones. We remark that the varying of δ from 0.01 to 0.0001 does not change too much in the phonon spectrum. The same parameters for SCF and LOBPCG are used to converge the ground state calculation in the ACP formulation. The absolute error of the phonon spectrum is smaller than 10^{-3}. The ACP formulation can lead to a very accurate phonon spectrum by solving a relatively small number of equations.

In order to demonstrate the effectiveness of the adaptive compression strategy, the relative L^2 error for the approximation of $U = \chi G$ with respect to the iteration in Algorithm 4 is given in Figure 5. For $\epsilon = 10^{-5}$, the error is around 10^{-5} after four iterations. For $\epsilon = 10^{-3}$, the error is reduced to 0.0008. In this case, if we stop after the first iteration in Algorithm 4, the relative error of χG is 0.0339. Numerical results show significant improvement after two to four iterations. This indicates that the self-consistent solution of the Dyson equation is crucial for the accurate computation of phonon spectrum.

We perform phonon calculations for systems of size from 30 to 150 for both insulating and semiconducting systems. In terms of accuracy, Figure 6 shows that as

Table 3

L^∞ error of the phonon spectrum. System is insulating with size $N_A = 60$. Chebyshev nodes $N_c = 20$. N_μ is determined such that $|\widetilde{R}_{N_\mu+1,N_\mu+1}| < \epsilon|\widetilde{R}_{1,1}| \leq |\widetilde{R}_{N_\mu,N_\mu}|$ in Algorithm 2.

Method and parameter	L^∞-norm error
FD $\delta = 0.01$	5.6779e-04
ACP $N_\mu = 241$ for $\epsilon = 10^{-3}$	3.6436e-04
ACP $N_\mu = 359$ for $\epsilon = 10^{-5}$	2.7380e-06

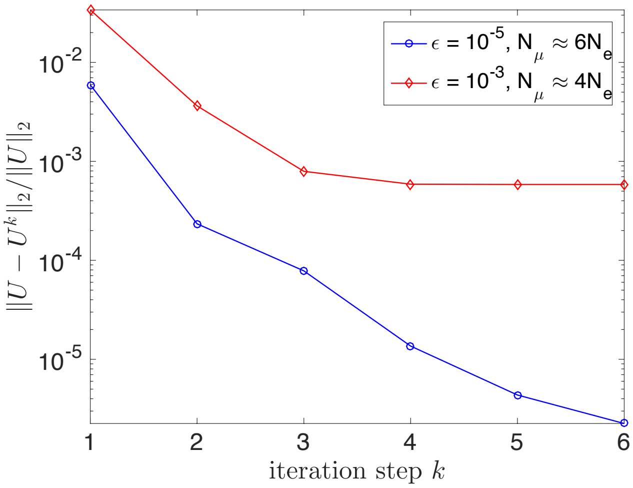


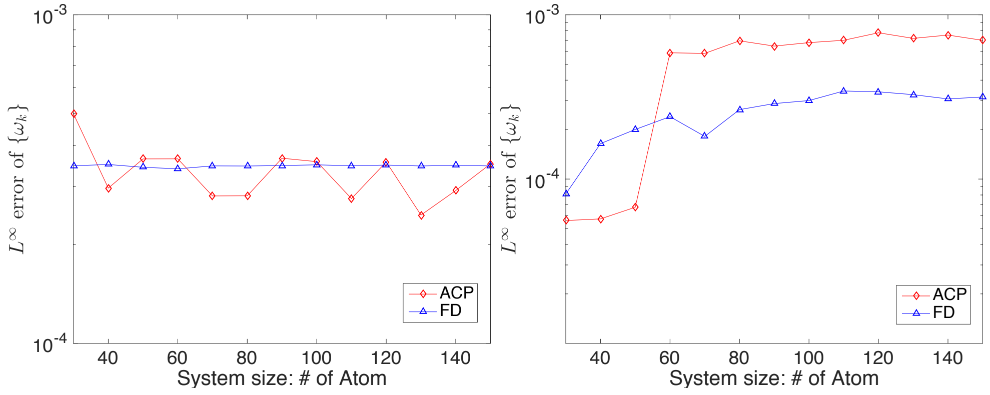
Fig. 5. *Convergence of adaptive compression.*

N_μ increases linearly with respect to the system size, the accuracy of phonon spectrum (L^∞ error) remains roughly the same, which is empirically around the $\epsilon = 10^{-3}$ used to determine N_μ in Algorithm 2. For the computation of the phonon frequency, we find that $N_c = 20$ and $N_\mu = 4N_e$ is sufficient to achieve error around 10^{-3}, and the phonon spectrum is already indistinguishable from that obtained from DFPT. Figure 7 reports the phonon spectrum ϱ_D for systems of size $N_A = 150$. We remark that Figure 7 plots the density of states ϱ by replacing the Dirac-δ distribution in (2.10) with regularized delta function

$$\delta_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}.$$

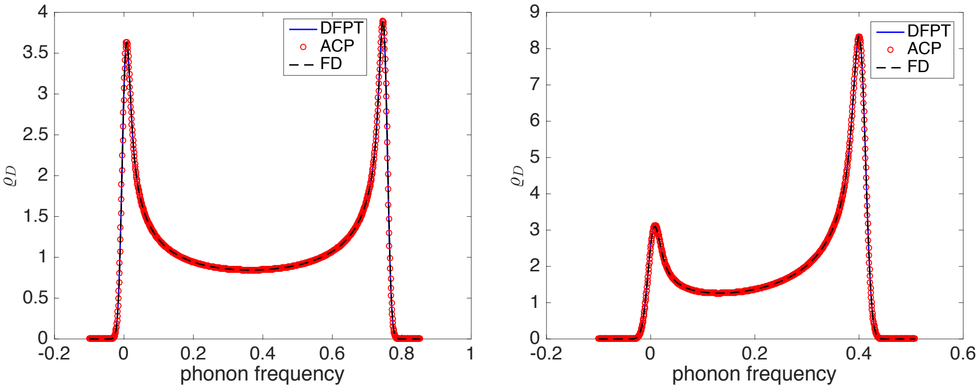
Here the smear parameter σ is chosen as 0.01.

To demonstrate the efficiency of the ACP algorithm, Figure 8 compares the computational time of ACP, DFPT, and FD, respectively. We observe that the computational cost of DFPT matches that of FD due to the choice of initial guess of Sternheimer equations and the Anderson mixing strategy for solving the Dyson equation. Compared to DFPT, the ACP formulation benefits both from the fact that it solves fewer Sternheimer equations and from the fact that the Sherman–Morrison–Woodbury procedure (3.16) is more efficient than Anderson mixing for solving the Dyson equation. In fact for all systems, Algorithm 4 converges within four iterations, while the Anderson mixing in DFPT may require 20 iterations or more for systems of all sizes. Hence for both insulating and semiconducting systems, the ACP formu-



(a) Insulator. (b) Semiconductor.

Fig. 6. *L^∞ error of the phonon frequencies $\{\omega_k\}$ obtained from ACP and FD. For ACP formulation $N_c = 20$. $N_\mu \approx 4N_e$.*



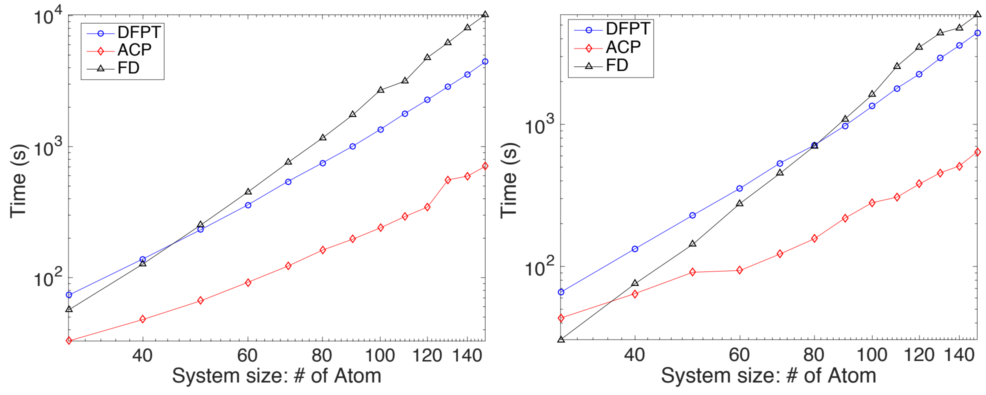
(a) Insulator. (b) Semiconductor.

Fig. 7. *Phonon spectrum for the* 1D *systems computed using ACP, DFPT, and FD for both* (a) *insulating and* (b) *semiconducting systems.*

lation becomes more advantageous than DFPT and FD for systems merely beyond 40 atoms. For the largest system with 150 atoms, ACP is 6.28 and 6.87 times faster than DFPT for insulating and semiconducting systems, respectively.

Table 4 measures the slope of the computational cost with respect to system sizes from $N_A = 90$ to $N_A = 150$. In theory, the asymptotic computational cost of FD and DFPT should be $\mathcal{O}(N_e^4)$, and the cost of ACP should be $\mathcal{O}(N_e^3)$. Numerically we observe that for the 1D examples up to $N_A = 150$, the computational scaling is still in the preasymptotic regime.

4.2. 2D lattice model. In the previous section, we have validated the accuracy of ACP compared to both FD and DFPT. We also find that the efficiency of FD and DFPT can be comparable. Hence for the 2D model, we only compare the efficiency and accuracy of ACP with respect to DFPT. Our first example is a periodic triangular



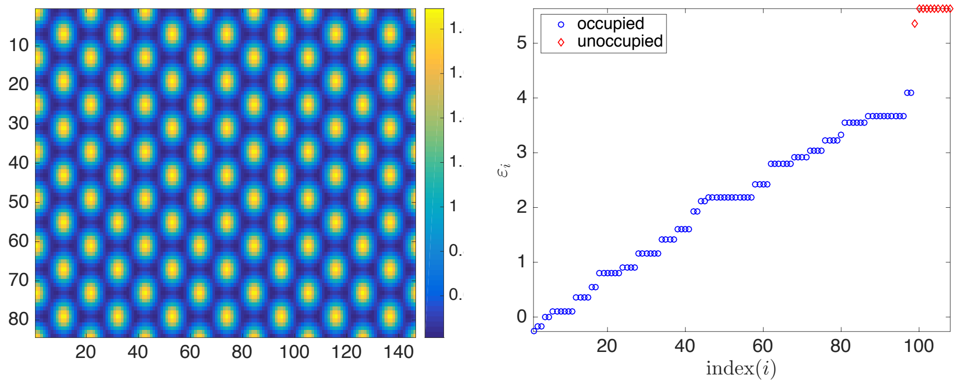
(a) Insulator.

(b) Semiconductor.

Fig. 8. *Computational time of* 1*D examples. Comparison among DFPT, ACP, and FD for* (a) *insulating and* (b) *semiconducting systems, respectively.*

Table 4
Computational scaling measured from $N_A = 90$ to $N_A = 150$.

Method	Insulator	Semiconductor
FD	3.4403	3.3047
DFPT	2.8997	2.9459
ACP	2.5040	2.1065



(a)

(b)

Fig. 9. *The electron density ρ of the* 98*-atom insulating system* (a)*, and the occupied and unoccupied eigenvalues* (b)*.*

lattice relaxed to the equilibrium position. The distance between each atom and its nearest neighbor is set to be 1.2 a.u., and $\epsilon_0 = 0.05$. The nuclear charge Z_I is set to 1 for all atoms, and σ_I is set to be 0.24.

For a system of size $N_A = 98$, the converged electron density ρ as well as the 108 smallest eigenvalues associated with the Hamiltonian at the converged ρ are shown in Figure 9. There is a finite HOMO-LUMO gap, $\varepsilon_g = \varepsilon_{99} - \varepsilon_{98} = 1.2637$, which

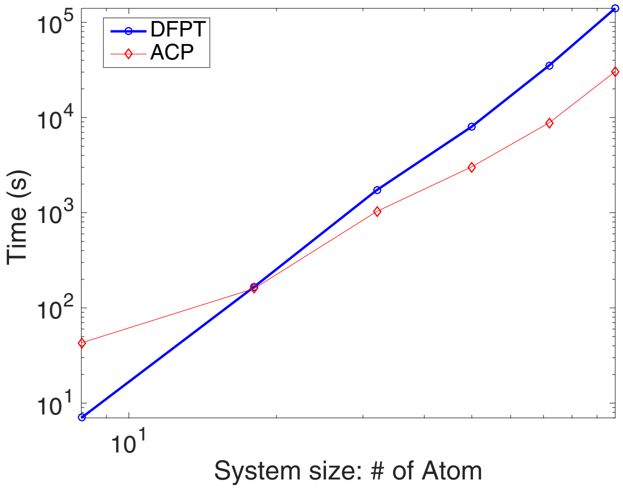


Fig. 10. *Computational time. Comparison of ACP to ACP for the 2D periodic lattice model.*

Table 5

Computational scaling measured from $N_A = 32$ to $N_A = 98$.

Method	Slope
DFPT	3.9295
ACP	3.0249

suggests that the system is an insulator. Figure 10 shows the computational time of ACP and DFPT as the system size grows from 2×2^2 to 2×7^2. We choose $\epsilon = 10^{-3}$ in Algorithm 2, and we find that this amounts to around $N_\mu = 14N_e$ columns selected in the ACP procedure. We choose $N_c = 30$, and hence the ACP formulation solves $420N_e$ equations, compared to the $2N_AN_e = 2N_e^2$ equations needed for DFPT. We iterate Algorithm 4 for two iterations. We find that when the system size increases beyond $N_A = 18$, the ACP formulation becomes more advantageous compared to DFPT. For the largest system $N_A = 98$, ACP is 4.61 times faster than DFPT. Table 5 measures the computational scaling from $N_A = 32$ to $N_A = 98$, which matches closely with the $\mathcal{O}(N_e^3)$ and $\mathcal{O}(N_e^4)$ theoretical scaling of ACP and DFPT, respectively. Figure 11 reports the phonon spectrum ϱ_D for the system of size $N_A = 98$. Here the smear parameter $\sigma = 0.08$.

4.3. 2D model with random vacancies. Our final example is a 2D triangular lattice with defects. We start from a periodic system with $N_A = 72$ atoms, randomly remove three atoms, and then perform structural relaxation for 15 steps. We terminate the structural relaxation before the system reaches its equilibrium position to obtain a disordered structure.

The converged electron density ρ and the smallest 79 eigenvalues are shown in Figure 12. For this system, there is a finite gap $\varepsilon_g = \varepsilon_{70} - \varepsilon_{69} = 1.3500$. Figure 13 shows the phonon spectrum computed from ACP and DFPT, plotted with the smear parameter $\sigma = 0.08$. The computational time for DFPT is 53883 sec, that for ACP is 8741 sec, and the speedup factor for ACP is 6.16. We observe that for the disordered structure, DFPT requires more iterations to converge, while the number of iterations for ACP to converge can remain 2. More specifically, compared to the periodic struc-

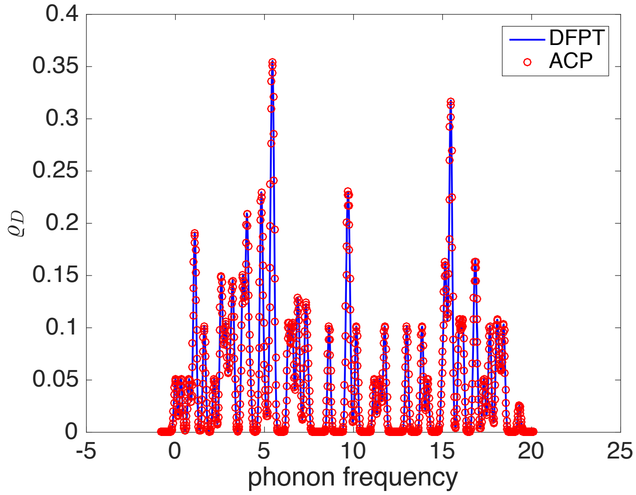


FIG. 11. *Phonon spectrum for the 2D periodic lattice. System size* $N_A = 98$, $\epsilon = 10^{-3}$, $N_\mu \approx 14N_e$.

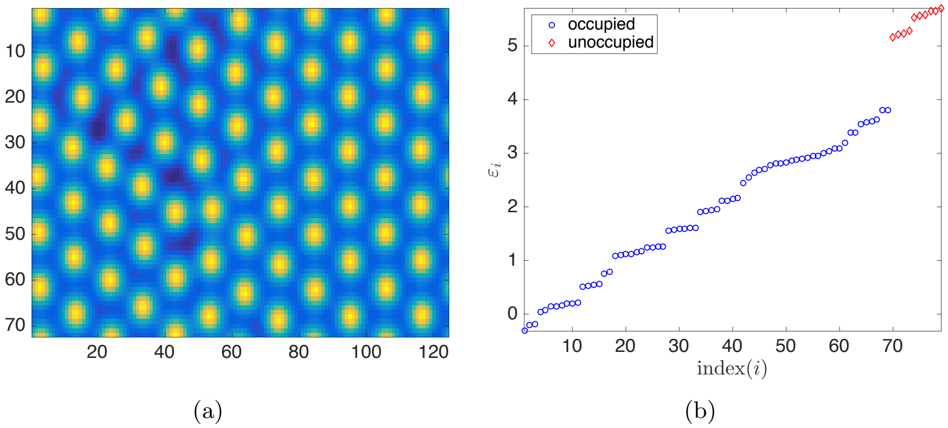


FIG. 12. *The electron density* ρ *of the 2D system with defects* (a), *and the occupied and unoccupied eigenvalues* (b).

ture with $N_A = 72$, the computational time for DFPT is 35501 sec, while that for ACP is 8854 sec.

5. Conclusion.

We have introduced the adaptively compressed polarizability operator (ACP) formulation. To the best of our knowledge, the ACP formulation is the first to reduce the computational complexity of phonon calculations from $\mathcal{O}(N_e^4)$ to $\mathcal{O}(N_e^3)$. This is achieved by reducing the $\mathcal{O}(N_e^2)$ equations in density functional perturbation theory (DFPT) to $\mathcal{O}(N_e)$ equations with systematic control of accuracy. Moreover, the accuracy of the ACP formulation depends weakly on the size of the gap, and hence can be applied to both insulator and semiconductor systems. Our numerical results for model problems indicate that the computational advantage of

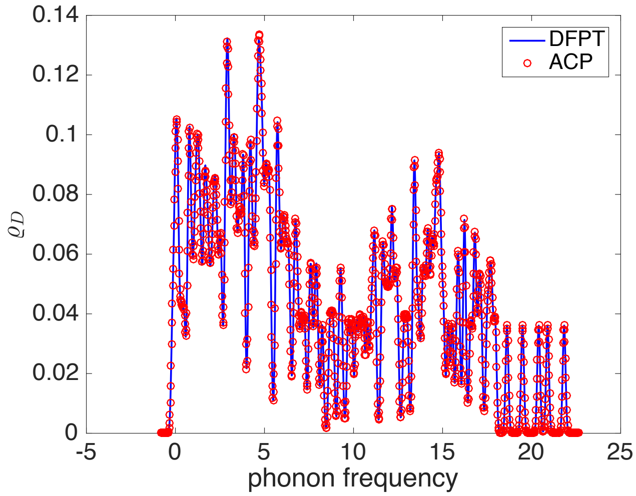


FIG. 13. *Phonon spectrum for the 2D system with defects. System size $N_A = 69$, $\epsilon = 10^{-3}$, $N_\mu \approx 14N_e$.*

the ACP formulation can be clearly observed compared to DFPT and finite difference, even for systems of relatively small sizes. While for simplicity our model problems do not include several components of KSDFT calculations for real materials, such as the nonlocal pseudopotentials and the exchange-correlation kernels, they do capture the essential difficulty in phonon calculations, and we expect that the asymptotic behavior of the ACP method for model problems is transferable to real materials simulation. We have also tuned the parameters of our model problem (such as the lattice constants and the band gaps) to mimic those of real materials. In a future work, we will present the ACP formulation in the presence of the nonlocal pseudopotential, and its application for computing the phonon spectrum for real materials.

The availability of fast phonon calculations provides a possible way to accelerate structural relaxation optimization of large scale molecules and solids. In this work we have restricted ourselves to zero temperature calculations. We plan to extend the ACP formulation to treat systems at finite temperature and hence metallic systems. We have used phonon calculation as an example to demonstrate the effectiveness of the compressed polarizability operator. The same strategy can be applied to applications of DFPT other than phonon calculations. In the future, we also plan to extend the ACP formulation to treat frequency-dependent polarizability operators that arise from many-body perturbation theories.

Acknowledgment. L. L. thanks Martin Head-Gordon for useful discussions.

REFERENCES

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

[2] D. G. Anderson, *Iterative procedures for nonlinear integral equations*, J. Assoc. Comput. Mach., 12 (1965), pp. 547–560.

[3] S. Baroni, S. de Gironcoli, A. Dal Corso, and P. Giannozzi, *Phonons and related crystal properties from density-functional perturbation theory*, Rev. Modern Phys., 73 (2001), pp. 515–562.

[4] S. Baroni, P. Giannozzi, and A. Testa, *Green's-function approach to linear response in solids*, Phys. Rev. Lett., 58 (1987), pp. 1861–1864.

[5] S. Baroni and R. Resta, *Ab initio calculation of the low-frequency Raman cross section in silicon*, Phys. Rev. B, 33 (1986), pp. 5969–5971.

[6] A. D. Becke, *Density-functional exchange-energy approximation with correct asymptotic behavior*, Phys. Rev. A, 38 (1988), pp. 3098–3100.

[7] D. R. Bowler and T. Miyazaki, *$O(N)$ methods in electronic structure calculations*, Rep. Prog. Phys., 75 (2012), 036503.

[8] E. Cancès, A. Deleurence, and M. Lewin, *A new approach to the modeling of local defects in crystals: The reduced Hartree-Fock case*, Commun. Math. Phys., 281 (2008), pp. 129–177.

[9] E. Cancès, A. Deleurence, and M. Lewin, *Non-perturbative embedding of local defects in crystalline materials*, J. Phys. Condens. Matter, 20 (2008), 294213.

[10] E. Cances and N. Mourad, *A mathematical perspective on density functional perturbation theory*, Nonlinearity, 27 (2014), pp. 1999–2033.

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

[12] T. F. Chan and P. C. Hansen, *Computing truncated singular value decomposition least squares solutions by rank revealing QR-factorizations*, SIAM J. Sci. Stat. Comput., 11 (1990), pp. 519–530, https://doi.org/10.1137/0911029.

[13] H. Cheng, Z. Gimbutas, P. G. Martinsson, and V. Rokhlin, *On the compression of low rank matrices*, SIAM J. Sci. Comput., 26 (2005), pp. 1389–1404, https://doi.org/10.1137/030602678.

[14] A. Damle, L. Lin, and L. Ying, *Compressed representation of Kohn–Sham orbitals via selected columns of the density matrix*, J. Chem. Theory Comput., 11 (2015), pp. 1463–1469.

[15] D. Foerster, *Elimination, in electronic structure calculations, of redundant orbital products*, J. Chem. Phys., 128 (2008), 034108.

[16] F. Giustino, M. L. Cohen, and S. G. Louie, *GW method with the self-consistent Sternheimer equation*, Phys. Rev. B, 81 (2010), 115105.

[17] S. Goedecker, *Linear scaling electronic structure methods*, Rev. Modern Phys., 71 (1999), pp. 1085–1123.

[18] X. Gonze, *Adiabatic density-functional perturbation theory*, Phys. Rev. A, 52 (1995), pp. 1096–1114.

[19] X. Gonze and C. Lee, *Dynamical matrices, Born effective charges, dielectric permittivity tensors, and interatomic force constants from density-functional perturbation theory*, Phys. Rev. B, 55 (1997), 10355.

[20] X. Gonze and J.-P. Vigneron, *Density-functional approach to nonlinear-response coefficients of solids*, Phys. Rev. B, 39 (1989), 13120.

[21] M. Gu and S. C. Eisenstat, *Efficient algorithms for computing a strong rank-revealing QR factorization*, SIAM J. Sci. Comput., 17 (1996), pp. 848–869, https://doi.org/10.1137/0917055.

[22] N. Halko, P. G. Martinsson, and J. A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), pp. 217–288, https://doi.org/10.1137/090771806.

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

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

[25] A. V. Knyazev, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput., 23 (2001), pp. 517–541, https://doi.org/10.1137/S1064827500366124.

[26] W. Kohn, *Density functional and density matrix method scaling linearly with the number of atoms*, Phys. Rev. Lett., 76 (1996), pp. 3168–3171.

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

[28] C. Lee, W. Yang, and R. G. Parr, *Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density*, Phys. Rev. B, 37 (1988), pp. 785–789.

[29] L. Lin, *Adaptively compressed exchange operator*, J. Chem. Theory Comput., 12 (2016), pp. 2242–2249.

[30] L. Lin, Y. Saad, and C. Yang, *Approximating spectral densities of large matrices*, SIAM Rev., 58 (2016), pp. 34–65, https://doi.org/10.1137/130934283.

[31] L. Lin and C. Yang, *Elliptic preconditioner for accelerating self-consistent field iteration in Kohn–Sham density functional theory*, SIAM J. Sci. Comput., 35 (2013), pp. S277–S298, https://doi.org/10.1137/120880604.

[32] J. Lu and L. Ying, *Compression of the electron repulsion integral tensor in tensor hypercontraction format with cubic scaling cost*, J. Comput. Phys., 302 (2015), p. 329.

[33] R. Martin, *Electronic Structure: Basic Theory and Practical Methods*, Cambridge University Press, West Nyack, NY, 2004.

[34] H.-V. Nguyen, T. A. Pham, D. Rocca, and G. Galli, *Improving accuracy and efficiency of calculations of photoemission spectra within the many-body perturbation theory*, Phys. Rev. B, 85 (2012), 081101.

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

[36] C. C. Paige and M. A Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629, https://doi.org/10.1137/0712047.

[37] J. E. Pask and P. A. Sterne, *Real-space formulation of the electrostatic potential and total energy of solids*, Phys. Rev. B, 71 (2005), 113101.

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

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

[40] X. Ren, P. Rinke, V. Blum, J. Wieferink, A. Tkatchenko, A. Sanfilippo, K. Reuter, and M. Scheffler, *Resolution-of-identity approach to Hartree–Fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions*, New J. Phys., 14 (2012), 053020.

[41] A. Sodt, J. E. Subotnik, and M. Head-Gordon, *Linear scaling density fitting*, J. Chem. Phys., 125 (2006), 194109.

[42] J. P. Solovej, *Proof of the ionization conjecture in a reduced Hartree-Fock model*, Invent. Math., 104 (1991), pp. 291–311.

[43] L. N Trefethen, *Is Gauss quadrature better than Clenshaw–Curtis?*, SIAM Rev., 50 (2008), pp. 67–87, https://doi.org/10.1137/060659831.

[44] P. Umari, G. Stenuit, and S. Baroni, *Optimal representation of the polarization propagator for large-scale GW calculations*, Phys. Rev. B, 79 (2009), 201104.

[45] P. Umari, G. Stenuit, and S. Baroni, *GW quasiparticle spectra from occupied states only*, Phys. Rev. B, 81 (2010), 115104.

[46] F. Weigend, *A fully direct RI-HF algorithm: Implementation, optimised auxiliary basis sets, demonstration of accuracy and efficiency*, Phys. Chem. Chem. Phys., 4 (2002), pp. 4285–4291.

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

[48] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert, *A fast randomized algorithm for the approximation of matrices*, Appl. Comput. Harmon. Anal., 25 (2008), pp. 335–366.

[49] M. T. Yin and M. L. Cohen, *Theory of static structural properties, crystal stability, and phase transformations: Application to Si and Ge*, Phys. Rev. B, 26 (1982), 5668.