Complex-time shredded propagator method for large-scale GW calculations

Minjung Kim, Glenn J. Martyna, and Sohrab Ismail-Beigi

Phys. Rev. B **101**, 035139 — Published 23 January 2020

DOI: 10.1103/PhysRevB.101.035139

Complex time, shredded propagator method for large-scale GW calculations

Minjung Kim,[1] Glenn J. Martyna,[2,3] and Sohrab Ismail-Beigi[1,*]

[1] *Department of Applied Physics, Yale University, New Haven, Connecticut 06520, USA*
[2] *IBM TJ Watson Laboratory, Yorktown Heights, 10598, New York, USA*
[3] *Pimpernel Science, Software and Information Technoglogy, Westchester, NY 10598, USA*

(Dated: October 21, 2019)

The GW method is a many-body electronic structure technique capable of generating accurate quasiparticle properties for realistic systems spanning physics, chemistry, and materials science. Despite its power, GW is not routinely applied to study large, complex assemblies due to the method's high computational overhead and quartic scaling with particle number. Here, the GW equations are recast, exactly, as Fourier-Laplace time integrals over complex time propagators. The propagators are then "shredded" via energy partitioning and the time integrals approximated in a controlled manner using generalized Gaussian quadrature(s) while discrete variable methods are employed to represent the required propagators in real-space. The resulting cubic scaling GW method has a sufficiently small prefactor to outperform standard quartic scaling methods on small systems ($\gtrapprox 10$ atoms) and offers 2-3 order of magnitude improvement in large systems ($\approx 200-300$ atoms). It also represents a substantial improvement over other cubic methods tested for all system sizes studied. The approach can be applied to any theoretical framework containing large sums of terms with energy differences in the denominator.

I. INTRODUCTION

Density Functional Theory (DFT)[1,2] within the local density (LDA) or generalized gradient (GGA)[3,4] approximation provides a solid workhorse capable of realistically modeling an ever increasing number and variety of systems spanning condensed matter physics, materials science, chemistry, and biology. Generally, this approach provides a highly satisfactory description of the total energy, electron density, atomic geometries, vibrational modes, etc. However, DFT is a ground-state theory for electrons and DFT band energies do not have direct physical meaning because DFT is not formally a quasiparticle theory. Therefore, significant failures can arise when DFT band structure is used to predict electronic excitations.[5–7]

The GW approximation to the electron self-energy[8–11] is one of the most accurate fully *ab initio* methods for the prediction of electronic excitations. Despite its power, GW is not routinely applied to complex materials systems due to its unfavorable computational scaling: the cost of a standard GW calculation scales as $\mathcal{O}(N^4)$ where N is the number of atoms in the simulation cell whereas the standard input to a GW study, a Kohn-Sham DFT calculation, scales as $\mathcal{O}(N^3)$.

Reducing the computational overhead of GW calculations has been the subject of much prior research. First, GW methods scaling as $\mathcal{O}(N^4)$ but with smaller prefactors either avoid the use of unoccupied states via iterative matrix inversion[12–18] or use sum rules or energy integration to greatly reduce the number of unoccupied states required for convergence.[19–21] Second, cubic-scaling $\mathcal{O}(N^3)$ methods, including both a spectral representation approach[22] and a space/imaginary time method[23] utilizing analytical continuation from imaginary to real frequencies, have been proposed. Third, a linear scaling GW technique[24] has recently been developed that employs stochastic approaches for the total density of electronic states with the caveat that the non-deterministic stochastic noise must be added to the list of usual convergence parameters.

Here, we present a deterministic, small prefactor, $\mathcal{O}(N^3)$ scaling GW approach that does not require analytic continuation. The GW equations are first recast exactly using Fourier-Laplace identities into the complex time domain where products of propagators expressed in real-space using discrete variable techniques[25] are integrated over time to generate an $\mathcal{O}(N^3)$ GW formalism. However, the time integrals are challenging to perform numerically due to the multiple time scales inherent in the propagators. Second, the time scale challenge is met by shredding the propagators in energy space, again exactly, to allow windows of limited dynamical bandwidth to be treated via generalized Gaussian quadrature numerical integration with low overhead and high accuracy. The unique combination of a (complex) time domain formalism, bandwidth taming propagator partitioning, and discrete variable real-space forms of the propagators permits a fast $\mathcal{O}(N^3)$ to emerge. Last, our approach is easy to implement in standard GW applications[26,27] because the formulae follow naturally from those of the standard approach(es) and much of the existing software can be refactored to utilize our reduced order technique.

The resulting GW formalism is tested to ensure both its accuracy and high performance in comparison to the standard $\mathcal{O}(N^4)$ approach for crystalline silicon, magnesium oxide, and aluminium. The new method's accuracy and performance are compared also to that of reduced overhead quartic scaling methods as well as existing $\mathcal{O}(N^3)$ scaling techniques. Importantly, we provide estimates of the speed-up over conventional GW computations and the memory requirement in the application of the new method to study technologically and scientifically interesting systems consisting of $\lessapprox 200-300$ atoms

– the sweet spot for the approach on today's supercomputers.

II. THEORY

A. Summary of GW

The theoretical object of interest for understanding one-electron properties such as quasiparticle bands and wave functions is the one-electron Green's function $G(x,t,x',t')$, which describes the propagation amplitude of an electron starting at x' at time t' and ending at x at time t:[28]

$$iG(x,t,x',t') = \langle T\left\{\,\hat{\psi}(x,t)\,\hat{\psi}(x',t')^\dagger\,\right\}\rangle\,,$$

where the electron coordinate $x=(r,\sigma)$ specifies electron position (r) and spin (σ). Here, $\hat{\psi}(x,t)$ is the electron annihilation field operator at (x,t), T is the time-ordering operator, and the average is over the statistical ensemble of interest. We focus primarily on the zero-temperature case (i.e., ground-state averaging); however, to treat systems with small gaps, the grand canonical ensemble is invoked. As is standard, henceforth atomic units are employed: $\hbar=1$ and the quantum of charge $e=1$.

The Green's function in the frequency domain obeys Dyson's equation

$$G^{-1}(\omega) = \omega I - [T + V_{ion} + V_H + \Sigma(\omega)]$$

where the x,x' indices have been suppressed; a more compact but complete notation shall be employed henceforth

$$G(\omega)_{x,x'} = G(x,x',\omega)\,.$$

Above, I is the identity operator, T is the electron kinetic operator, V_{ion} is the electron-ion interaction potential operator (or pseudopotential for valence electron only calculations), V_H is the Hartree potential operator, and $\Sigma(\omega)$ is the self-energy operator encoding all the many-body interaction effects on the electron Green's function.

The GW approximation to the self-energy is

$$\Sigma(t)_{x,x'} = iG(t)_{x,x'}W(t^+)_{r,r'}$$

where t^+ is infinitesimally larger than t and $W(t)_{r,r'}$ is the dynamical screened Coulomb interaction between an external test charge at $(r',0)$ and (r,t):

$$W(\omega)_{r,r'} = \int dr''\ \epsilon^{-1}(\omega)_{r,r''}V_{r'',r'}\,.$$

Here, ϵ is the linear response, dynamic and nonlocal microscopic dielectric screening matrix, and $V_{r,r'}=1/|r-r'|$ is the bare Coulomb interaction. The GW self-energy includes the effects due to dynamical and nonlocal screening on the propagation of electrons in a many-body environment. The notation introduced above (to be continued below) is that parametric functional dependencies are placed in parentheses and explicit dependencies are given as subscripts; the alternative notation wherein all variables are in parentheses with explicit dependencies given first followed by parametric dependencies separated by a semicolon is also employed where convenient (e.g., $W(r,r';\omega)\equiv W(\omega)_{r,r'}$).

To provide a closed and complete set of equations, one must approximate ϵ. The most common approach is the random-phase approximation (RPA): one first writes ϵ in terms of the dynamic irreducible polarizability P via

$$\epsilon(\omega)_{r,r'} = \delta(r-r') - \int dr''\ V_{r,r''}P(\omega)_{r'',r} \qquad (1)$$

and P is related to G by the RPA

$$P(t)_{r,r'} = -i\sum_{\sigma,\sigma'} G(t)_{x,x'}G(-t)_{x',x}\,.$$

In the vast majority of GW calculations, including the formalism given here, the Green's function is approximated by an independent electron form (band theory) specified by a complete set of one-particle eigenstates $\psi_n(x)$ (compactified to $\psi_{x,n}$) and eigenvalues E_n

$$G(\omega)_{x,x'} = \sum_n \frac{\psi_{x,n}\psi^*_{x',n}}{\omega - E_n}\,. \qquad (2)$$

The ψ_n and E_n are obtained as eigenstates of a non-interacting one-particle Hamiltonian from a first principles method such as Density Functional Theory,[1,2] although one is not limited to this choice. Although not central to the analysis given here, formally E_n has a small imaginary part that is positive for occupied states (i.e., energies below the chemical potential) and negative for unoccupied states. We have suppressed the non-essential crystal momentum index k in Eq. (2) for simplicity – including it simply amounts to adding the k index to the eigenstates $\psi_{x,n}\to\psi^k_{x,n}$ and energies $E_n\to E^k_n$ and averaging over the k sampled in the first Brillouin zone (BZ).

For our purposes, the frequency domain representations of all quantities are useful. The Green's function G in frequency space is given in Eq. (2) while the frequency dependent polarizability, P, is

$$\begin{aligned} P(\omega)_{r,r'} &= \sum_{c,v,\sigma,\sigma'} \psi_{x,c}\psi^*_{x,v}\psi^*_{x',c}\psi_{x',v}[f(E_v)-f(E_c)] \\ &\quad\times\frac{2(E_c-E_v)}{\omega^2-(E_c-E_v)^2} \\ &= \sum_{c,v,\sigma,\sigma'} \psi_{x,c}\psi^*_{x,v}\psi^*_{x',c}\psi_{x',v}[f(E_v)-f(E_c)] \\ &\quad\times\left[\frac{1}{(\omega-(E_c-E_v))}-\frac{1}{(\omega+(E_c-E_v))}\right] \end{aligned} \qquad (3)$$

Here, v labels occupied (valence) eigenstates while c labels unoccupied (conduction) eigenstates. The occupancy function $f(E)$ required to handle finite temperatures for zero/small gap systems is explicitly included

(see Sec. II D); for gapped systems at zero temperature $f(E_v) = 1$ and $f(E_c) = 0$. (The occupancy $f(E;\beta,\mu)$ formally depends parametrically on two thermodynamic variables: the inverse temperature $\beta = 1/k_B T$ and the chemical potential μ.) We have employed a general, compact notation valid for collinear and non-collinear spin calculations. For collinear spin, non-zero contributions to P only occur when the spin indices σ and σ' of $x = (r,\sigma)$ and $x = (r',\sigma')$ match; for the full spinor (non-collinear) case, we sum over all the spin projections σ,σ' in the usual way.

Of particular practical importance is the zero-frequency or static polarizability $P(\omega = 0)$ (which we also simply denote as P below)

$$P_{r,r'} = -2 \sum_{c,v,\sigma,\sigma'} \frac{\psi_{x,c}\psi^*_{x,v}\psi^*_{x',c}\psi_{x',v}[f(E_v) - f(E_c)]}{E_c - E_v} \tag{4}$$

which is employed both as part of plasmon-pole models of the frequency dependent screening[8–11,29] as well as within the COHSEX approximation[8] (see below). Again, the crystal momentum index has been suppressed for simplicity; including it requires the replacements $P \to P^q$ where q is the momentum transfer, $\psi_{x,v} \to \psi^k_{x,v}$ and $E_v \to E^k_v$, $\psi_{x,c} \to \psi^{k+q}_{x,c}$ and $E_c \to E^{k+q}_c$, and averaging Eqs. (3,4) over k (i.e., Brillouin zone sampling). We note that current numerical methods for computing P based on the sum-over-states formulae, e.g., that of Eq. (4), have an $\mathcal{O}(N^4)$ scaling (e.g., see Ref. [26]).

Formally, the screened interaction W can always be represented as a sum of "plasmon" screening modes indexed by p,

$$\begin{aligned} W(\omega)_{r,r'} &= V_{r,r'} + \sum_p \frac{2\omega_p B^p_{r,r'}}{\omega^2 - \omega_p^2} \\ &= V_{r,r'} + \sum_p B^p_{r,r'} \left[\frac{1}{\omega - \omega_p} - \frac{1}{\omega + \omega_p}\right] \end{aligned} \tag{5}$$

Here B_p is the mode strength for screening mode p and $\omega_p > 0$ is its frequency. This form is directly relevant when making computationally efficient plasmon-pole models for the screened interaction.[29] The self-energy is then given by

$$\begin{aligned} \Sigma(\omega)_{x,x'} &= -\sum_v \psi_{x,v}\psi^*_{x',v} W(\omega - E_v)_{r,r'} \\ &\quad + \sum_{n,p} \frac{\psi_{x,n} B^p_{r,r'} \psi^*_{x',n}}{\omega - E_n - \omega_p} \\ &= -\sum_v \psi_{x,v} V_{r,r'} \psi^*_{x',v} \\ &\quad + \sum_{v,p} \frac{\psi_{x,v} B^p_{r,r'} \psi^*_{x',v}}{\omega - E_v + \omega_p} + \sum_{c,p} \frac{\psi_{x,c} B^p_{r,r'} \psi^*_{x',c}}{\omega - E_c - \omega_p} \end{aligned} \tag{6}$$

where the n-sum is over all bands (i.e., valence and conduction). Inclusion of crystal momentum in Eq. (6) means $\Sigma(\omega)$ carries a k index, $\psi_{x,v} \to \psi^{k-q}_{x,v}$ and $E_v \to E^{k-q}_v$. All screening quantities derived from P^q now also carry a q index, W^q, ω^q_p and B^{pq}, and Eq. (6) is averaged over the q sampling.

Within the COHSEX approximation, when the applicable screening frequencies, ω_p, are much larger than the interband energies of interest, the frequency dependence of Σ can be neglected

$$\begin{aligned} \Sigma^{(\mathrm{COHSEX})}_{x,x'} &= -\sum_v \psi_{x,v}\psi^*_{x',v} W(0)_{r,r'} \\ &\quad + \frac{1}{2}\delta(x - x') \left[W(0)_{r,r'} - V_{r,r'}\right] . \end{aligned} \tag{7}$$

where the label in the supercript is placed in paranthesis to avoid possible confusion - a convention to be followed below. The numerically intensive part of the COHSEX approximation is the computation of the static polarizability, Eq. (4) – once P is on hand, the static $W(0)$ is completely determined by P via matrix multiplication and inversion,

$$W(0) = \epsilon^{-1}(0)V = (I - VP)^{-1}V .$$

Equations (3,4,6,7) are of primary interest, here, as evaluating them scales as $\mathcal{O}(N^4)$ as written. Terms with manifestly cubic scaling terms will not be discussed further. The computation of observables such as ϵ_∞ and the band gap in various approximations, e.g., $E^{(\mathrm{gap,G_0W_0})}$ and $E^{(\mathrm{gap,COHSEX})}$, from the key terms, are described in Refs. 8–11. The superscript on the band gap is employed to distinguish the gap of the input, single particle, spectrum, gap, $E^{(\mathrm{gap})}$, from appropriate corrections to it which we present below to evaluate the performance of the new method. Comparison of the accuracy of different approximations to the gap is not part of this work but is fully described in the above references.

B. Complex time shredded propagator formalism

We now describe the main ideas and merits of our new approach to cubic scaling GW calculations. The resulting formalism is general and can be applied to a broad array of theoretical frameworks whose evaluation involves sums over states with energy differences in denominators.

The analytic structure of the equations central to GW calculations, outlined in the prior section, necessitates the evaluation of terms of the form

$$\chi(\omega)_{r,r'} = \sum_{i=1}^{N_a} \sum_{j=1}^{N_b} \frac{A^i_{r,r'} B^j_{r,r'}}{\omega + a_i - b_j} \tag{8}$$

as can be discerned from Eqs. (3,4,6,7). The input energies a_i and b_j and the matrices A^i and B^j are either direct outputs of the $\mathcal{O}(N^3)$ ground state calculation (i.e., single particle energies and products of wave functions when $\chi = P$), or are obtained from $\mathcal{O}(N^3)$ matrix operations on the frequency dependent polarizability $P(\omega)$, or other such derived quantities.

The analytic form of χ in Eq. (8) arises because we have chosen to work in the frequency or energy representation. However, one can equally well represent such an equation in real, imaginary or complex time by changing the structure of the theory to involve time integrals over propagators. Here, we will effect the change of representation from time to frequency directly through the introduction of Fourier-Laplace identities which allows us to reduce the computational complexity of the GW calculation. This imaginary time formalism has connections to prior work found in Refs. [23, 30, and 31].

In more detail, while the frequency representation has advantages, the evaluation of Eq. (8) scales as $\mathcal{O}(N_a N_b N_r^2)$ because the numerator is separable but the energy denominator is not. This basic structure of the frequency representation leads to the familiar $\mathcal{O}(N^4)$ computational complexity of GW as the number of states or modes (N_a, N_b) and the number real-space points (N_r) required to represent them, here by discrete variable methods, scale as the number of electrons, N. For the widely used plane wave (i.e., Fourier) basis, adopted herein, a uniform grid in r-space that is dual to the finite g-space representation is indicated – fast Fourier transforms (FFTs) switch between the dual spaces, g- and r-space, both efficiently and exactly (without information loss); for other basis sets, appropriate real-space discrete variable representations (DVRs) with similar dual properties can be adopted[25,32,33].

In the following, a time domain formalism that reduces the computational complexity of Eq. (8) by N to achieve $\mathcal{O}((N_a + N_b)N_r^2) \sim \mathcal{O}(N^3)$ scaling, in a controlled and rapidly convergent manner, is developed. This will be accomplished through the introduction of time integrals and associated propagators which we shall then shred (i.e., partition) to tame the multiple time scales inherent to the theory. Again, the resulting formulation is general – it applies to any theory with the structure of Eq. (8).

Reduced scaling is enabled by replacing the energy denominator $1/(\omega + a_i - b_j)$ of Eq. (8) by a separable form through the introduction of the generalized Fourier-Laplace transform

$$F(E;\zeta) = \int_0^\infty d\tau\, h(\tau;\zeta) \exp[-\zeta E \tau] \tag{9}$$

That is, inserting the transform, Eq. (8) becomes

$$\chi(\omega;\zeta)_{r,r'} = \sum_{i=1}^{N_a}\sum_{j=1}^{N_b} F(\omega + a_i - b_j;\zeta) A^i_{r,r'} B^j_{r,r'} \tag{10}$$

Here, ζ is a complex constant with $|\zeta|$ akin to an inverse Planck's constant that sets the energy scale, and $h(\tau;\zeta)$ is a weight function. The desired separability arises from the exponential function in the integrand of $F(E;\zeta)$ and allows us to reduce the computational complexity of GW. In the following the ζ dependence of χ will be suppressed for reasons that will become immediately apparent.

To motivate the utility of Eq. (10), consider the case where $\forall\ i,j$ either $\omega + a_i - b_j > 0$ or $\omega + a_i - b_j < 0$: here, ζ is chosen to be real (positive for the first case and negative for the second), and we set $h(\tau;\zeta) = \zeta$. This corresponds to a textbook Laplace transform[34] and yields an *exact* expression for the energy denominator:

$$\lim_{h(\tau;\zeta)\to\zeta} F(\omega + a_i - b_j;\zeta) = \frac{1}{\omega + a_i - b_j}\,. \tag{11}$$

For this case, the introduction of the transform involves no approximation, and $h(\tau;\zeta) = \zeta$ will be employed to establish and describe our formalism. It is directly applicable to the static limit of $\chi(\omega)$ where $\omega \to 0$ and $a_i - b_j > 0\ \forall\ i,j$ (i.e., gapped systems, c.f. the static polarizability matrix of Eq. (4)). The importance of the actual value of ζ will become clear below. A yet more general treatment, applicable to gapless systems and finite frequencies $\omega \neq 0$, requiring non-trivial $h(\tau;\zeta)$, will then be given, wherein F becomes an approximation to the inverse of the energy denominator within the class of regularization procedures commonly employed in standard GW computations.

Inserting the generalized Fourier-Laplace identity into Eq. (8) yields

$$\begin{aligned}
\chi(0)_{r,r'} &= \int_0^\infty d\tau\, h(\tau;\zeta) \left[\sum_{i=1}^{N_a} A^i_{r,r'} e^{-\zeta(a_i - E^{(\mathrm{off})})\tau}\right] \\
&\qquad \times \left[\sum_{j=1}^{N_b} B^j_{r,r'} e^{-\zeta(E^{(\mathrm{off})} - b_j)\tau}\right] \qquad (12)\\
&= \int_0^\infty d\tau\ h(\tau;\zeta) \rho^{(A)}_{r,r'}(\zeta\tau) \bar{\rho}^{(B)}_{r,r'}(\zeta\tau) \\
&= \int_0^\infty d\tau\, h(\tau;\zeta)\, \tilde{\chi}(\zeta\tau;0)_{r,r'}\ .
\end{aligned}$$

Here, $E^{(\mathrm{off})}$ is a convenient energy offset selected such that all the exponential functions are decaying (e.g., midgap) and

$$\begin{aligned}
\rho^{(A)}(\zeta\tau)_{r,r'} &= \sum_{i=1}^{N_a} A^i_{r,r'} e^{-\zeta(a_i - E^{(\mathrm{off})})\tau} \\
\bar{\rho}^{(B)}(\zeta\tau)_{r,r'} &= \sum_{j=1}^{N_b} B^j_{r,r'} e^{-\zeta(E^{(\mathrm{off})} - b_j)\tau} \\
\tilde{\chi}(\zeta\tau;0)_{r,r'} &= \rho^{(A)}(\zeta\tau)_{r,r'} \bar{\rho}^{(B)}(\zeta\tau)_{r,r'}
\end{aligned} \tag{13}$$

where the $\rho^{(A,B)}(\zeta\tau)$ are imaginary time propagators (manifestly, for $a_i > b_i\ \forall\ i,j$ but the reverse is treated by letting $\zeta \to -\zeta$ and switching the ρ and $\bar{\rho}$ labels). The result is a separable form for $\tilde{\chi}(\tau\zeta;0)_{r,r'}$, a product of A and B propagators, whose zero frequency transform over $h(\tau;\zeta)$ yields the desired $\chi(0)_{r,r'}$. This exact reformulation can be evaluated in $\mathcal{O}(N^3)$ given that an $\mathcal{O}(N^0)$ scaling discretization (i.e., quadrature) of the time integral can be defined.

Consider that the largest energy difference in the argument of the exponential terms defining $\tilde{\chi}(\zeta\tau;0)_{r,r'}$, is the bandwidth $E^{(\mathrm{bw})} = \max(a_i) - \min(b_j)$ while the smallest

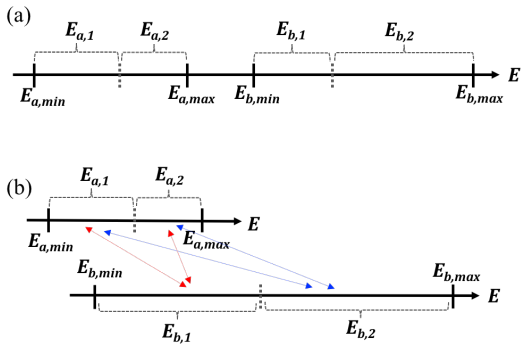

FIG. 1. An example of the proposed energy windowing approach with $N_{a_w} = N_{b_w} = 2$ (a) For gapped systems, the energy ranges of $\{a_i\}$ and $\{b_j\}$ do not overlap. (b) For systems with overlapping energy ranges, energy window pairs arise both with energy crossings, red arrows, and without, blue arrows.

energy difference is the gap $E^{(\mathrm{gap})} = \min(a_i) - \max(b_j)$ which are both known from input. Both energy differences are essentially independent of system size N for large N (exactly so for periodically replicated arrays of atoms in a supercell). Hence the longest and shortest time scales, $\sim \hbar/E^{(\mathrm{bw})}$ and $\sim \hbar/E^{(\mathrm{gap})}$, in $\tilde{\chi}(\tau\zeta;0)_{r,r'}$ are independent of N. Therefore, barring non-analytic behavior in the density of states or modes, a system size independent discretization scheme can be devised to generate $\chi(0)_{r,r'}$ from $\tilde{\chi}(\zeta\tau;0)_{r,r'}$. Of course, the formulation is most useful when the discrete form rapidly approaches the continuous integral with increasing number of discretizations (i.e., quadrature points).

The development of a rapidly convergent discretization scheme is, however, challenged by the large dynamic range present in the electronic structure of most materials systems, $E^{(\mathrm{bw})}/E^{(\mathrm{gap})} \gtrapprox 100$. Simply selecting the free parameter $|\zeta| \approx 1/E^{(\mathrm{bw})}$ to treat such large bandwidths is insufficient to allow a small number of discretizations (i.e., number of quadrature points) to represent the time integrals accurately. Hence, an efficient approach capable of taming the multiple time scale challenge presented by the large dynamic range in the integrand, $\tilde{\chi}(\zeta\tau;0)_{r,r'}$, of $\chi(0)_{r,r'}$, will be given. Once such an approach has been developed for gapped systems, the solution will be generalized to treat gapless systems and response functions at finite frequencies through use of imaginary ζ and non-trivial $h(\tau;\zeta)$.

In order to tame the multiple time scales inherent in the present time domain approach to $\chi(0)_{r,r'}$, the propagators $\rho^{(A,B)}$ must be modified. Borrowing ideas from Feynman's path integral approach, the propagators are "shredded" (sliced into pieces) in energy space. That is, the energy range spanned by a_i is partitioned into N_{a_w} contiguous energy windows indexed by $l = 1, \ldots, N_{a_w}$ and b_j is similarly partitioned into N_{b_w} windows indexed by $m = 1, \ldots, N_{b_w}$; to illustrate this shredding, a 2×2 energy window decomposition for a gapped system is shown in Fig. 1(a) (i.e., $N_{a_w} = N_{b_w} = 2$). Shredding the propagators allows $\tilde{\chi}(\tau\zeta;0)_{r,r'}$ to be recast *exactly* as a sum over window pairs (l,m),

$$\chi(0)_{r,r'} = \sum_{l=1}^{N_{a_w}} \sum_{m=1}^{N_{b_w}} \int_0^\infty d\tau\, h(\tau;\zeta_{lm}) \tilde{\chi}^{lm}(\zeta_{lm}\tau;0)_{r,r'} \,, \tag{14}$$

where for each window pair (l,m),

$$\begin{aligned} \tilde{\chi}^{lm}(\zeta_{lm}\tau;0)_{r,r'} &= \rho^{(A)}_{lm}(\zeta_{lm}\tau)_{r,r'} \bar{\rho}^{(B)}_{lm}(\zeta_{lm}\tau)_{r,r'} \\ \rho^{(A)}_{lm}(\zeta_{lm}\tau)_{r,r'} &= \sum_{\{i\in\mathcal{L}\}} A^i_{r,r'} e^{-\zeta_{lm}(a_i - E^{(\mathrm{off})})\tau} \\ \bar{\rho}^{(B)}_{lm}(\zeta_{lm}\tau)_{r,r'} &= \sum_{\{j\in\mathcal{M}\}} B^j_{r,r'} e^{-\zeta_{lm}(E^{(\mathrm{off})} - b_j)\tau} \,. \end{aligned} \tag{15}$$

Here, $\mathcal{L}$ and $\mathcal{M}$ represent the sets of integer indices of the single particle states that contribute to the l^{th} A-type and m^{th} B-type energy windows, respectively. The energy $E^{(\mathrm{off})}$ is an offset chosen for convenience: e.g., choosing it to be in the gap between the smallest a_i and largest b_j to generate strictly decaying exponential functions. As above, treating $b_j > a_i$ only necessitates reversing the sign of the ζ_{lm} and switching the bar labels on the density matrices. The energy windows need not be equally spaced in energy; in fact, the optimal choice of windows is not equally spaced even for a uniform density of states or modes as shown in Sec. II C.

The shredded form of $\chi(0)_{r,r'}$ given in Eq. (14) has computational complexity of $\mathcal{O}(N^3)$ because the operation count to evaluate it, is

$$N_r^2 \sum_{lm} (L_l^{(A)} + L_m^{(B)}) N_{lm}^{(\tau,h)} \sim \mathcal{O}(N^3) \,, \tag{16}$$

to be compared with the operation count of the standard GW method, $N_a N_b N_r^2 \sim \mathcal{O}(N^4)$. Here the $L_l^{(A)}, L_m^{(B)} \sim \mathcal{O}(N)$ are the number of states or modes in the l^{th} and m^{th} energy windows, respectively, and $N_{lm}^{(\tau,h)} \sim \mathcal{O}(N^0)$ is the number of quadrature points required for accurate integration in a specific window pair (l,m) (see Sec. II C).

The shredded propagator formulation of $\chi(0)_{r,r'}$ has four important advantages. First, every term in the double sum over window pairs (l,m) has its own intrinsic bandwidth which is handled by its own ζ_{lm} while preserving the desired separability. Second, each window pair can be assigned its own quadrature optimized to treat its limited dynamic range. Third, the windows can be selected to minimize the dynamic range in the window pairs which allows small $N_{lm}^{(\tau,h)}$ (i.e., efficient quadrature) to treat all pairs with small fractional error, $\epsilon^{(q)}$. These first three advantages are sufficient to tame the multiple time scale challenge. Fourth, finite frequency expressions for gapped systems as well as gapless systems at finite temperature can be addressed utilizing simple extensions of Eq. (14) as demonstrated below.

The next theoretical issue to tackle is to show that the optimal windows can be found in $\mathcal{O}(N^3)$ or less computational effort given the input energies a_i and b_j. Since

the computationally intensive part of $\chi(0)_{r,r'}$ involves its r, r' spatial dependence, it is best to choose an optimal windowing scheme in the limit $A^i_{r,r'}, B^j_{r,r'} \to 1$ as, within a limited energy range of a window pair, the spatial dependence of the A^i or B^j are to good approximation similar. (Note, the plane-wave basis approach considered here does not exploit spatial locality and full-sized N_r^2 matrices are employed, but other approaches may benefit considering spatial locality in window creation). If the density of states for a_i and b_j is taken to be locally flat, then the optimal number and placement of windows can be determined in $\mathcal{O}(N^0)$; if the actual density of states is taken into account, the scaling remains $\mathcal{O}(N^0)$ as the density of states is an input from the electronic structure computation (typically, KS-DFT). Here, optimal indicates the windows are selected to minimize the operation count, Eq. (16), required to compute Eq. (14) over the number and placement (in energy space) of the windows. In practice, as discussed in Sec. II C, we take $N_{lm}^{(\tau,h)}$ to be the number of quadrature points required to guarantee a prespecified, upper error bound, obeyed by all the time integrals of each window pair; again, each window pair (l, m) has its own tuned quadrature and time scale taming parameter, ζ_{lm}.

The control given by the energy windowed formulation of $\chi(0)_{r,r'}$ in Eq. (14) is the key to extending our efficient $\mathcal{O}(N^3)$ method to gapless systems and to finite frequencies. For gapless systems at zero frequency, there will be some few energy windows pairs (most likely only one) for which $a_i = b_j$ happens at least once. This is not problematic because, e.g., for the case of computing the polarizability matrix of Eqs. (3,4), the occupancy difference $f(E_v) - f(E_c)$ regularizes the singularity of the denominator via L'Hôpital's rule applied to $[f(E_v) - f(E_c)]/(E_c - E_v)$ (the mapping from the general formalism being $a_i \to E_v$, $b_j \to E_c$). Adding the occupancy factors presents no difficulties: all that is required is to take the difference between two terms of the same form as Eq. (8) in the problematic window pair(s) with an overlapping energy range – a small added expense (see Sec. II D). However, a more general approach that can handle finite frequencies, described next, can also be adopted to handle gapless systems.

For the case of finite frequency $\omega \neq 0$, in some window pair(s) the quantity in the denominator, $e_{ij} = \omega + a_i - b_j$, can change sign (see Fig. 1.b). In standard GW implementations, singularities (zeros of e_{ij}) that may arise in these window pairs are tamed by either dropping their contributions to the sum when $|e_{ij}|$ is small[35] or by regularizing $1/e_{ij}$, e.g., replacing $1/e_{ij}$ by $e_{ij}/(e_{ij}^2 + |\zeta|^{-2})$.[36]

Lorentzian regularization can be accommodated easily within our time domain formalism by selecting $h(\tau;\zeta) = |\zeta| \exp(-\tau)$ for the weight function in Eqs. (9-10) and choosing ζ to be a pure imaginary number,

$$\begin{aligned}\frac{e_{ij}}{e_{ij}^2 + |\zeta|^{-2}} &= Im\left[\int_0^\infty d\tau\, |\zeta| e^{-\tau}\, e^{i|\zeta|e_{ij}\tau}\right] \\ &= |\zeta| \int_0^\infty d\tau\, e^{-\tau}\Big[\sin(|\zeta|(\omega - b_j))\cos(|\zeta|a_i) \\ &\quad - \cos(|\zeta|(\omega - b_j))\sin(|\zeta|a_i)\Big] \end{aligned} \tag{17}$$

for the small number of window pairs where e_{ij} changes sign. In order to factorize the complex exponential and expose the separability of i, j in the second line of the above equation, we have chosen to decompose the energy difference as $e_{ij} = (\omega - b_j) + (a_j)$, but the decomposition $e_{ij} = (\omega + a_i) + (-b_j)$ is also possible. Nonetheless, a large number of quadrature points must be taken to accurately discretize the time integral of Eq. (17), in practice.

Alternatively, as will be detailed in Sec. II E, the weight function

$$h(\tau;\zeta) = |\zeta| \exp(-\tau - \tau^2/2)$$

and its transform

$$\begin{aligned} F(e_{ij};\zeta) = |\zeta| Im\left\{\sqrt{\frac{\pi}{2}} \exp\left(-\frac{(e_{ij}|\zeta| + i)^2}{2}\right) \times \right. \\ \left. \left[1 + i\,\mathrm{erfi}\left(\frac{e_{ij}|\zeta| + i}{\sqrt{2}}\right)\right]\right\}, \end{aligned} \tag{18}$$

form a preferable choice of regularization. Importantly, the transform, Eq. (18), approaches $1/e_{ij}$ at large e_{ij}, is well behaved for all e_{ij} but can be generated accurately with fewer time integration quadrature points than required by the Lorentzian. The benefits of the alternative weight function, an asymptotic analysis, and the associated rapidly convergent quadrature are presented in Sec. II E 2 and associated appendices.

Lastly, we note that the new formalism can handle problematic regions / points in the density of states that might need specialized treatment, such as van Hove singularities, by simply assigning them their own window in a Lebesgue-type approach (see Sec. II C 3).[37] As long as the *number* of special regions/ points is independent of systems size, the scaling of the method remains $\mathcal{O}(N^3)$.

In order to convince the reader that the new formalism represents an important improvement, we provide a comparison of our $\mathcal{O}(N^3)$ time domain results to those of the corresponding $\mathcal{O}(N^4)$ direct frequency domain computation in Fig. 2 for two standard test systems, crystalline silicon and magnesium oxide. In the figure, the new method is referred to via the sobriquet complex time shredded propagator (CTSP) method where CTSP-W indicates the use of optimal windowing, and in the discussion to follow, CTSP-1 the use of one window. Even for small unit / supercells, the $\mathcal{O}(N^3)$ computational approach outlined above delivers a significant reduction in computational effort compared to the standard approach (the CTSP error decreases exponentially with the number of time integration quadrature points as given in Sec.

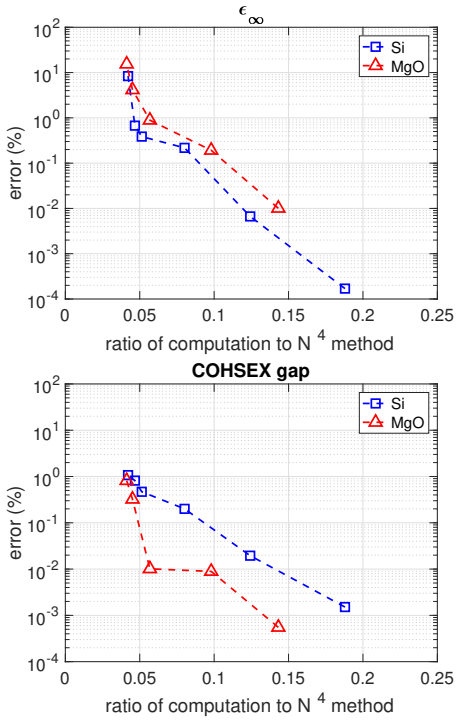

FIG. 2. Numerical error versus computational savings for our cubic scaling formalism, CTSP-W, compared to the standard quartic GW formulation for bulk Si and MgO modeled in a 16 atom supercell. The CTSP-W error decreases and computational work increases as the integration error is decreased (i.e., the number of quadrature points is increased). Computational work is measured by the ratio of operation count, Eq. (16), of the cubic method to the quartic method. Top: Error in the macroscopic optical dielectric constant ($\epsilon_\infty(\text{MgO}) = 6.35$, $\epsilon_\infty(\text{Si}) = 64.85$). Bottom: Error in the COHSEX band gap ($E^{(\text{gap,COHSEX})}(\text{MgO}) = 7.56$ eV, $E^{(\text{gap,COHSEX})}(\text{Si}) = 1.92$ eV).

II.C.2.b and log-linear plots are thereby the natural way to present the data).

The detailed analysis underlying CTSP's reduced scaling with system size *and* high performance is presented in Secs. II C-II F and associated appendices. We also show below that (all) the new method's parameters can be reduced to one, the fractional time integration quadrature error, $\epsilon^{(q)}$, which allows for the easily tunable convergence demonstrated by the results given above (see Fig. 2). The use of the simple operation count as given in Eq. (16) to represent computational work is, also, justified in the following.

C. Static polarization matrix in $\mathcal{O}(N^3)$ for gapped systems

The static polarizability matrix defined in Eq. (4) reduces, for systems with large energy gaps compared to k_BT, to

$$P_{r,r'} = -2\sum_{v}^{N_v}\sum_{c}^{N_c}\frac{\psi^*_{r,v}\psi_{r,c}\psi^*_{r',c}\psi_{r',v}}{E_c - E_v}$$

as the occupation number functions for this special case are zero or one; the occupancies will be reintroduced to treat zero gap systems in Sec. II D. Here, N_v and N_c are the number of valence and conduction states, respectively. Non-essential indices or quantum numbers such as spin σ and Bloch k-vector have been suppressed.

1. *Laplace identity and shredded propagators*

Employing the energy windowing approach of Eqs. (14,15), the energy range of the valence and conduction band is divided into N_{v_w} and N_{c_w} partitions with the valence and conduction partition indexed by l and m ranging from $E_l^{(v,\min)}$ to $E_l^{(v,\max)}$ and $E_m^{(c,\min)}$ to $E_m^{(c,\max)}$, respectively . Thus, the static polarizability can be written as

$$P_{r,r'} = \sum_{l=1}^{N_{v_w}}\sum_{m=1}^{N_{c_w}} P^{lm}_{r,r'} \tag{19}$$

where each window pair (l,m) contributes

$$P^{lm}_{r,r'} = -2\zeta_{lm}\int_0^\infty d\tau\, e^{-\zeta_{lm}E^{(\text{gap})}_{lm}\tau} \times \rho_m(\zeta_{lm}\tau)_{r,r'}\bar{\rho}_l(\zeta_{lm}\tau)_{r',r} \tag{20}$$

via the Laplace identity where the choice $h = \zeta$ generates the desired energy denominator, $1/(E_c - E_v)$ (i.e., $F(x;\zeta) = 1/x$ in Eq. (10)). Each window pair (l,m) has its own energy gap, $E^{(\text{gap})}_{lm} = E_m^{(c,\min)} - E_l^{(v,\max)}$, energy scale, ζ_{lm}, and bandwidth, $E^{(\text{bw})}_{lm} = E_m^{(c,\max)} - E_l^{(v,\min)}$. (To connect directly to the formalism of Eqs. (14-15), the sign of ζ has been reversed and the bar labels on the density matrices have been switched.) The imaginary time density matrices for the windows are given by

$$\rho_m(\tau)_{r,r'} = \sum_{\{c\in\mathcal{M}\}} e^{-\tau\Delta E_{mc}}\psi_{r,c}\psi^*_{r',c}\,, \tag{21}$$

$$\bar{\rho}_l(\tau)_{r,r'} = \sum_{\{v\in\mathcal{L}\}} e^{-\tau\Delta E_{lv}}\psi_{r,v}\psi^*_{r',v}\,. \tag{22}$$

where, again, the integer indices of the single particle states in the m^{th} conduction and l^{th} valence windows are contained in the sets, $\mathcal{M}$ and $\mathcal{L}$, respectively. Here, $\Delta E_{lv} = E_l^{(v,\max)} - E_v$ and $\Delta E_{mc} = E_c - E_m^{(c,\min)}$ are defined with respect to the edges of each energy window. A good choice of windows can significantly reduce the dynamic range, i.e., the bandwidth to band gap ratio $E^{(\text{bw})}_{lm}/E^{(\text{gap})}_{lm}$, for all window pairs. This allows coarse quadrature grids to be employed to approximate the time integrals in all window pairs with controlled accuracy as given next.

2. Discrete approximation to the time integral

The continuous imaginary time integral of Eq. (20) must be discretized in an efficient and error-controlled manner to form an effective numerical method. The natural choice is Gauss-Laguerre (GL) quadrature

$$\int_0^\infty d\tau\ e^{-\tau} s(\tau) \approx \sum_{u=1}^{N^{(\tau,\mathrm{GL})}} w_u\, s(\tau_u)\ . \tag{23}$$

Here, $N^{(\tau,\mathrm{GL})}$ is the number of quadrature points, the u are the integer indices of the points, $s(\tau)$ is the function to be integrated over the exponential function, $\exp(-\tau)$, the $\{w\}$ and $\{\tau\}$ are the $N^{(\tau,\mathrm{GL})}$ member sets of the quadrature weights and nodes[38] whose explicit dependence on $N^{(\tau,\mathrm{GL})}$ has been suppressed for clarity. Inserting the discrete approximation, the contribution from each window pair (l,m) is

$$\begin{aligned} P^{lm}_{r,r'} = -2\zeta_{lm} \sum_{u=1}^{N^{(\tau,\mathrm{GL})}_{lm}} & w_u e^{-\tau_u(\zeta_{lm}E^{(\mathrm{gap})}_{lm}-1)} \\ & \times \ \rho_m(\zeta_{lm}\tau_u)_{r,r'}\bar{\rho}_l(\zeta_{lm}\tau_u)_{r',r}\,. \end{aligned} \tag{24}$$

a. Optimal error-equalizing energy scale factor ζ_{lm}: The energy scale factor ζ_{lm} is selected to equalize the error of all integrals in a window pair. The geometric mean, $\zeta_{lm}^{-1} \approx \sqrt{E^{(\mathrm{bw})}_{lm} E^{(\mathrm{gap})}_{lm}}$, is close to the optimal error matching choice as described in Appendix A 1: the end points of the window range are treated with (nearly) equal accuracy.

b. Estimating the number of quadrature points: For any set of interband transition energies $\{E_m - E_l\}$ in window pair (l,m), the largest quadrature errors occur for the largest interband transition energy $E^{(\mathrm{bw})}_{lm}$ and the smallest interband transition energy $E^{(\mathrm{gap})}_{lm}$. Taking $\zeta_{lm}^{-1} = \sqrt{E^{(\mathrm{bw})}_{lm} E^{(\mathrm{gap})}_{lm}}$ to balance the error across the window pair, the number of quadrature points, $N^{(\tau,\mathrm{GL})}_{lm}$, required to generate the desired fractional error level, scales as $\sim \sqrt{E^{(\mathrm{bw})}_{lm}/E^{(\mathrm{gap})}_{lm}}$ (see Appendix A). Stripping the indices for clarity, we find

$$\begin{aligned} N^{(\tau,\mathrm{GL})}(\alpha;\epsilon^{(q)}) &= \alpha(y - 0.3\ln\epsilon^{(q)}) \\ \alpha &= \sqrt{\frac{E^{(\mathrm{bw})}}{E^{(\mathrm{gap})}}}, \quad y = 0.4 \end{aligned} \tag{25}$$

to be a good approximation, valid for $\epsilon^{(q)} < 0.135$ (see Appendix A). To extend the range to $\epsilon^{(q)} < 1$, we simply set $y = 1$. Importantly, the procedure ensures that $N^{(\tau,\mathrm{GL})}_{lm}$ is chosen such that time integration error for any term in a window pair has upper bound $\epsilon^{(q)}$.

3. Optimal windowing

Given that the number of points required to generate maximal fractional quadrature error $\epsilon^{(q)}$ for a given window pair can be neatly determined, we now consider the construction of the optimal set of windows. This can be accomplished via minimization of the cost to compute the static polarizability over the number of windows, N_{v_w} and N_{c_w}, and the associated N_{v_w} and N_{c_w} member sets, $\{E^{(v,\min)}, E^{(v,\max)}\}$ and $\{E^{(c,\min)}, E^{(c,\max)}\}$ of the window positions in energy space,

$$\begin{aligned} C^{(\mathrm{GL})}(\epsilon^{(q)}) &= \sum_{l=1}^{N_{v_w}}\sum_{m=1}^{N_{c_w}} N^{(\tau,\mathrm{GL})}(\alpha_{lm};\epsilon^{(q)}) \\ &\quad \times \left(\int_{E_l^{(v,\min)}}^{E_l^{(v,\max)}} D(E)dE + \int_{E_m^{(c,\min)}}^{E_m^{(c,\max)}} D(E)dE\right) \\ &= \sum_{l=1}^{N_{v_w}}\sum_{m=1}^{N_{c_w}} C^{(\mathrm{GL})}_{lm}(\epsilon^{(q)}) \end{aligned} \tag{26}$$

which for clarity are omitted from the dependencies of $C^{(\mathrm{GL})}(\epsilon^{(q)})$. Here $N^{(\tau,\mathrm{GL})}(\alpha_{lm};\epsilon^{(q)})$ is given in Eq. (25), and $D(E)$ is the density of states (which will be taken on additional indices when performing k-point sampling as given in Appendix B). The integrals over the density of states, $D(E)$, are simply the number or fraction of states in the appropriate energy window.

For a density of states with problematic points, we assign windows to those regions *a priori* (fixed position in energy space) allowing for fast minimization over the smooth parts of $D(E)$. For example, if there is a special point in the $D(E)$ at energy E_{special}, a window boundary is fixed to bracket this energy, $[E_{\mathrm{special}} - \Delta E/2, E_{\mathrm{special}} + \Delta E/2]$, allowing the minimization to proceed over the smoothly varying regions of the DOS integral in a Lebesgue inspired approach (i.e., the DOS is only required to be Lebesgue integrable)[37].

The cost estimator, Eq. (26), can be minimized straightforwardly, as detailed in Appendix B, once at the start of a GW calculation. The computational complexity of the minimization procedure is negligible $\mathcal{O}(N^0)$ compared to both the $\mathcal{O}(N^3)$ computational complexity of both P and the input band structure. We note that for the form of $N^{(\tau,\mathrm{GL})}(\alpha;\epsilon^{(q)})$ in Eq. (25), the optimal windowing, both the number of windows and their positions in energy, is independent of error level as $N^{(\tau,\mathrm{GL})}(\alpha;\epsilon^{(q)}) = \alpha \cdot U(\epsilon^{(q)})$ is separable. Importantly, all parameters of the method are now completely determined by the usual set (input band structure and a choice of energy cutoff in the conduction band) and *one* new parameter, $\epsilon^{(q)}$, the fractional quadrature error required to accurately transform from the time domain to the frequency domain. The quadrature error will be connected to the error in physical quantities in Sec. III.

D. Static P for gapless systems

The standard approach employed to treat gapless systems is to introduce a smoothed step function $f(E;\mu,\beta)$ for the electron occupation numbers as a function of energy, E, centered on the chemical potential, μ (Fermi level) with "smoothing" parameter or inverse temperature β.[39–41] Examples include the Fermi-Dirac distribution of the grand canonical ensemble

$$f(E) = \frac{1}{1+\exp[\beta(E-\mu)]}$$

where formally, $\beta = 1/k_B T$, or the more rapidly (numerically) convergent and hence convenient

$$f(E) = \frac{1}{2}\mathrm{erfc}(\beta(E-\mu)) \ .$$

Typical literature values of β correspond to temperatures above ambient conditions (e.g., $\beta^{-1} = 0.1$ eV ≈ 1000 K). The static RPA irreducible polarizability matrix including the occupation functions is given in Eq. (4).

To proceed, note that the energy-dependent part of the sum in Eq. (4),

$$J_{cv} = \frac{f(E_v) - f(E_c)}{E_c - E_v} \ , \tag{27}$$

is smooth for all energies and has the finite value $-f'(\mu)$ as $E_v, E_c \to \mu$ (note, $E_c \geq E_v \ \forall \ c, v$). Hence, for a calculation with a small but finite gap, the terms in the sum for P are finite and well behaved such that windowing plus quadrature approach will work well. As before, we split P into a sum over window pairs with the contributions from each window pair now given by

$$P^{lm}_{r,r'} = -2\zeta_{lm} \sum_{u=1}^{N^{(\mathrm{GL})}_{lm}} w_u \, e^{-\tau_u(\zeta_{lm}E^{(\mathrm{gap})}_{lm}-1)} \times \left[S^{lmu}_{r',r} Q^{lmu}_{r,r'} - T^{lmu}_{r',r} Z^{lmu}_{r,r'}\right]$$

where

$$S^{lmu}_{r,r'} = \sum_{\{v\in\mathcal{L}\}} f(E_v) e^{-\tau_u \zeta_{lm}\Delta E_{vl}} \psi_{r,v}\psi^*_{r',v}$$

$$Q^{lmu}_{r,r'} = \sum_{\{c\in\mathcal{M}\}} e^{-\tau_u \zeta_{lm}\Delta E_{m}} \psi_{r,c}\psi^*_{r',c}$$

$$T^{lmu}_{r,r'} = \sum_{\{v\in\mathcal{L}\}} e^{-\tau_u \zeta_{lm}\Delta E_{vl}} \psi_{r,v}\psi^*_{r',v}$$

$$Z^{lmu}_{r,r'} = \sum_{\{c\in\mathcal{M}\}} f(E_c) e^{-\tau_u \zeta_{lm}\Delta E_{m}} \psi_{r,c}\psi^*_{r',c} \, .$$

The 5-index entities S, Q, T, Z can be computed with $O(N_v N_r^2)$ or $O(N_c N_r^2)$ operations (i.e., cubic scaling) where N_r is the number of r grid points (see also Sec. III.C). Since $f(E_c)$ becomes small as a function of increasing E_c, the TZ term need only be computed for the few window pairs where $\beta(E_c - \mu)$ is sufficiently small. Hence, the additional work required to treat gapless systems is, in fact, modest.

Direct application of the cost-optimal energy windowing method for gapped systems in Sec. II C generates infinite quadrature grids in situations where the gap is exactly zero due to degeneracy at the Fermi energy. The solution is straightforward: the key quantity that is to be represented by quadrature is J_{cv} of Eq. (27). For $E_c - E_v \to 0$, $J_{cv} \to -f'(\mu)$ where $-f'(\mu) = \beta/4$ for the Fermi-Dirac distribution and $\beta/\sqrt{2\pi}$ for the erfc form above. Thus, the system has an effective gap of $\sim \beta^{-1}$. For energy window pairs (l,m) that contain degenerate states at the Fermi energy, we manually set their gap to $E^{(\mathrm{gap})}_{lm} = 1/\beta$ via a "scissoring" operation (i.e., shifting the conduction band up by $1/(2\beta)$ and valence bound down by $1/(2\beta)$) in the offending window pair and then applying the method of Sec. II C. Alternatively, the regularization approach of the next subsection can be adopted for zero-gap systems.

E. $\Sigma(\omega)$ in cubic computational complexity

Given poles of the screened interaction $W(\omega)_{r,r'}$, ω_p, with residues, $B^p_{r,r'}$, the dynamic (frequency-dependent) part of the GW self-energy can be expressed as

$$\Sigma(\omega)_{r,r'} = \sum_{p,v} \frac{B^p_{r,r'}[\psi_{rv}\psi^*_{r'v}]}{\omega - E_v + \omega_p} + \sum_{p,c} \frac{B^p_{r,r'}[\psi_{rc}\psi^*_{r'c}]}{\omega - E_c - \omega_p}. \tag{28}$$

(omitting, the static / bare potential term in Eq. (6) as it can be computed in $\mathcal{O}(N^3)$ and is, thus, not of interest here). In the following, we develop a cubic scaling energy window-plus-quadrature technique that delivers Eq. (28) directly[42] for real frequencies ω in such a way that analytical continuation is not required.

1. *Windowing for* $\Sigma(\omega)$

The dynamic self-energy,

$$\Sigma(\omega)_{r,r'} = \Sigma^{(+)}(\omega)_{r,r'} + \Sigma^{(-)}(\omega)_{r,r'} \tag{29}$$

$$\Sigma^{(+)}(\omega)_{r,r'} = \sum_{p,v} \frac{B^p_{r,r'}[\psi_{rv}\psi^*_{r'v}]}{\omega - E_v + \omega_p}$$

$$\Sigma^{(-)}(\omega)_{r,r'} = \sum_{p,c} \frac{B^p_{r,r'}[\psi_{rc}\psi^*_{r'c}]}{\omega - E_c - \omega_p},$$

consists of two terms, labeled $(\pm)$. The $(+)$ term involves the valence single particle states, their shifted energies $(E_v - \omega)$, the plasmon residues and their modes

(ω_p). The ($-$) term involves the conduction single particle states, their shifted energies ($E_c - \omega$), the plasmon residues and their mode complement ($-\omega_p$). An efficient windowed scheme requires independently decomposing the two terms as is now usual,

$$\Sigma^{(+)}(\omega)_{r,r'} = \sum_{m=1}^{N_{v_w}^{(+)}} \sum_{l=1}^{N_{p_w}^{(+)}} \Sigma^{(+)}(\omega; \zeta_{lm}^{(+)})_{r,r'}^{lm}) \tag{30}$$

$$\Sigma^{(-)}(\omega)_{r,r'} = \sum_{m=1}^{N_{c_w}^{(-)}} \sum_{l=1}^{N_{p_w}^{(-)}} \Sigma^{(-)}(\omega; \zeta_{lm}^{(-)})_{r,r'}^{lm}), \tag{31}$$

simply using the shifted single-particle energies and $\pm$ plasmon modes to define the windows. Note, $\zeta_{lm}^{(+)} \neq \zeta_{lm}^{(-)}$, $N_{p_w}^{(+)} \neq N_{p_w}^{(-)}$ and the index sets are also unique to each term, $+$ and $-$. Almost all the window pairs (l, m) in Eqs. (30,31) can be treated using the approach of Sec. II C with GL quadrature because the denominator, $x = \omega - E_n \pm \omega_p$, is finite and does not change sign where $n = v$ for $+$ case and $n = c$ for $-$ case. The difficulty is that, for some few window pairs, the denominator, x, changes sign such that the Eq. (11) does not apply. Thus, a scheme to treat window pairs with energy crossings is required.

2. *Specialized quadrature for energy crossings*

We treat energy window pairs (l, m) with an energy crossing, where $x = \omega - E_n \pm \omega_p$ changes sign as the sum over p and the generalized index, n, in the windows is performed, by replacing $1/x$ by the regularized $F(x; \zeta)$ of Eq. (9),

$$\Sigma^{(\pm)}(\omega)_{r,r'}^{lm} = \sum_{\{p \in \mathcal{L}^{(\pm)}\}} \sum_{\{n \in \mathcal{M}^{(\pm)}\}} B_{r,r'}^{p} [\psi_{rn} \psi_{r'n}^*] \times F(\omega - E_n \pm \omega_p; \zeta) \,. \tag{32}$$

where ζ is same for all windows with a crossing. As discussed in Sec. II B, the two standard regularization strategies in the GW literature are, one, to these zero contributions for small x (i.e., setting $F(x; \zeta) = 0$ for small x), or, two, to use a Lorentzian smoothing function with $\zeta = -i\gamma$, $\gamma > 0$ and $h(t; \zeta) = \gamma e^{-\tau}$, i.e.,

$$F(x; \zeta) = \frac{x}{x^2 + \gamma^{-2}} = Im \int_0^\infty d\tau \, \gamma \, e^{-\tau} e^{i\tau\gamma x} \,.$$

Below we shall eschew ζ and work in terms of γ which is more natural.

As detailed in Appendix C, a better choice for the weight function and resulting transform are

$$h(\tau; \gamma) = \gamma \exp(-\tau - \tau^2/2) \tag{33}$$

$$F(x; \gamma) = \gamma Im \left\{ \sqrt{\frac{\pi}{2}} e^{-\frac{(x\gamma + i)^2}{2}} \left[1 + i \mathrm{erfi}\left(\frac{x\gamma + i}{\sqrt{2}} \right) \right] \right\} .$$

The new weight has a transform that both approaches $1/x$ faster than the Lorentzian in the large x limit (see Appendix C), and is regular for all x. In addition, its transform can be accurately computed via time integration with fewer quadrature points than required by weight that leads to the Lorentzian (i.e., the pure exponential function).

A Gaussian-type quadrature for the new weight function can be generated following the standard procedure[43] to create a set of nodes, $\{\tau\}$, and weights, $\{w\}$, for a given quadrature grid size $N^{(\tau,\mathrm{HGL})}$ (see Appendix H). The superscript HGL denotes Hermite-Gauss-Laguerre quadrature since the weight function has both linear and quadratic terms in the exponent. Inserting the result, the discrete approximation becomes

$$\begin{aligned} F(x; \gamma) &\approx \gamma Im \sum_{u=1}^{N^{(\tau,\mathrm{HGL})}} w_u \, e^{i\tau_u x\gamma} \\ &\approx \gamma \sum_{u=1}^{N^{(\tau,\mathrm{HGL})}} w_u \, \sin(\tau_u x\gamma) \,. \end{aligned} \tag{34}$$

Finally, for the window pairs (l, m) with an energy crossing

$$\begin{aligned} \Sigma^{(\pm)}(\omega)_{r,r'}^{lm} = \gamma \sum_{u=1}^{N_{lm}^{(\tau,\mathrm{HGL},\pm)}} w_u \Big\{ & \left[\sum_{\{p \in \mathcal{L}^{(\pm)}\}} B_{r,r'}^{p} \sin(\pm\tau_u \omega_p \gamma) \right] \times \\ & \left[\sum_{\{n \in \mathcal{M}^{(\pm)}\}} \psi_{rn} \psi_{r'n}^* \cos(\tau_u(\omega - \epsilon_n)\gamma) \right] \\ & + \left[\sum_{\{p \in \mathcal{L}\}} B_{r,r'}^{p} \cos(\pm\tau_u \omega_p \gamma) \right] \times \\ & \left[\sum_{\{n \in \mathcal{M}^{(\pm)}\}} \psi_{rn} \psi_{r'n}^* \sin(\tau_u(\omega - \epsilon_n)\gamma) \right] \Big\} . \end{aligned} \tag{35}$$

which is separable and can be computed in $\mathcal{O}(N^3)$. Again, one value of broadening parameter γ is selected for all windows with energy crossings. The parameter γ is a convergence parameter taken to be as small as possible without effecting results. The number of grid points will vary depending on the bandwidth in the window pair scaled by γ and the desired fractional error.

3. *Quadrature points for specified error level*

For window pairs without an energy crossing, $\omega - E_n \pm \omega_p$ does not change sign, and the GL quadrature previously analyzed is utilized (the general subscript is n is used to denote that either c or v states are possible). For window pairs with energy crossings, the HGL quadrature is required. Appendix D details the construction of

$N^{(\tau,\mathrm{HGL})}(x;\epsilon^{(q)})$,

$$N^{(\tau,\mathrm{HGL})}(x;\epsilon^{(q)}) = c_2(\epsilon^{(q)})x^2 + c_1(\epsilon^{(q)})x + c_0(\epsilon^{(q)})\,, \quad (36)$$

where $x = \gamma(E_{\max} - E_{\min})$ is the bandwidth of the window pair with energy crossings (scaled by γ), and c_2, c_1, and c_0 are low order polynomial functions of $\ln \epsilon^{(q)}$. The values of the coefficients are given in Appendix D.

4. Optimal window choice

We now consider the computational cost to compute $\Sigma(\omega)$ for window pairs with an energy crossing,

$$C_{lm}^{(\mathrm{HGL})}(\epsilon^{(q)}) = 2N_{lm}^{(\tau,\mathrm{HGL})}(x_{lm};\epsilon^{(q)})\times \left(\int_{\omega_m^{(p,\min)}}^{\omega_m^{(p,\max)}} D^{(p)}(\omega)d\omega + \int_{E_l^{(n,\min)}}^{E_l^{(n,\max)}} D(E)dE\right).$$

Here the m^{th} plasmon mode window spans the energy range $[\omega_m^{(p,\min)}, \omega_m^{(p,\max)}]$, the l^{th} band energy window spans the energy range $[E_l^{(n,\min)}, E_l^{(n,\max)}]$, the density of plasmon modes is $D^{(p)}(\omega)$ and the density of band states is $D(E)$. (The explicit dependence of the cost function on the window edges is, again, suppressed.) The parameter x_{lm} is $x_{lm} = \gamma E_{lm}^{(\mathrm{bw})}$ where $\Delta E_{lm}^{(\mathrm{bw})}$ is the absolute value of the maximum energy difference between the single particle and plasmon modes in the window pair. Although potentially discontinuous as the window ranges evolve during minimization, the insertion does not prevent rapid numerical convergence of the cost function to its minimum value. Further discussion can be found in Appendix E.

F. Cubic-scaling $P(\omega)$

The energy window plus time integral quadrature methods developed to compute the static P and the dynamic $\Sigma(\omega)$ can be applied directly and without modification to the computation of the frequency-dependent polarizability $P(\omega)$ of Eq. (3) with $\mathcal{O}(N^3)$ computational effort. The key observation is that $P(\omega)$ can be written as the sum of two simple energy denominator poles:

$$P(\omega)_{r,r'} = \sum_{c,v,\sigma,\sigma'} [\psi_{x,c}\psi^*_{x',c}][\psi_{x',v}\psi^*_{x,v}]\times \left(\frac{1}{\omega - (E_c - E_v)} - \frac{1}{\omega + (E_c - E_v)}\right). \quad (37)$$

Since $P(\omega) = P(-\omega)$, we need only focus on $P(\omega)$ for $\omega > 0$. The second energy denominator $\omega + E_c - E_v$ is always positive definite since $E_c - E_v \geq 0$ and can be evaluated in $\mathcal{O}(N^3)$ with the same GL quadrature methodology developed for evaluating the static P in Sec. II C; the presence of $\omega > 0$ in the second denominator enlarges the effective energy gap and enhances convergence of our method. The first energy denominator $\omega - (E_c - E_v)$ can change sign once ω is larger than the energy gap. However, this term can be evaluated with $\mathcal{O}(N^3)$ effort using the energy crossing quadrature / regularization method developed for $\Sigma(\omega)$ in Sec. II E.

III. RESULTS: STANDARD BENCHMARKS

Here, the application of the new CTSP method to standard benchmark systems is presented. Results for the optical dielectric constant and the energy band gap within the COHSEX approximation are given for crystalline silicon (Si) and magnesium oxide (MgO). Next, studies of the static polarization of crytsalline Al, a gapless systems, are presented. Last, a G_0W_0 computation of the band gap of crystalline Si is given.

A. Optical dielectric constant & COHSEX band gap

In order to evaluate the performance of the new reduced order method, CTSP, we study two standard benchmark materials: Si and MgO. We first perform plane wave pseudopotential DFT calculations for both materials to generate the DFT band structure and then employ the results in the reported GW computations. Appendix G contains the details of the DFT and GW calculations.

Silicon is a prototypical 3-dimensional covalent crystal (diamond structure) with a moderate band gap (0.5 eV in DFT-LDA) while rocksalt MgO is an ionic crystal with a relatively large gap (4.4 eV in DFT-LDA). To judge the performance of CTSP, the convergence of two basic observables are studied: the macroscopic optical dielectric constant ϵ_∞ and the band gap within the COHSEX approximation to the self-energy.[8]

Figure 3 shows the error in ϵ_∞ as a function of the computational savings achieved by both CTSP-W and CTSP-1 $\mathcal{O}(N^3)$ techniques, and the $\mathcal{O}(N^3)$ interpolation method described in Appendix F, relative to the standard $\mathcal{O}(N^4)$ method for 16 atom (periodic) supercells of MgO and Si. Each data point is generated by fixing a fractional quadrature error, $\epsilon^{(q)}$ and then minimizing our cost function for CTSP-W. Figure 4 shows the COHSEX band gap for the CTSP-W and interpolation methods, respectively.

The results demonstrate that the CTSP-W approach shows high performance with respect to both accuracy and efficiency; the improvement over the CTSP-1 and the interpolation methods is clear for Si. The larger MgO gap can be treated with fewer integration points and also leads to functions that are easier to interpolate. For both materials, the CTSP-W method achieves better than 0.1 eV accuracy in the band gap with at least an order of magnitude reduction in computation compared

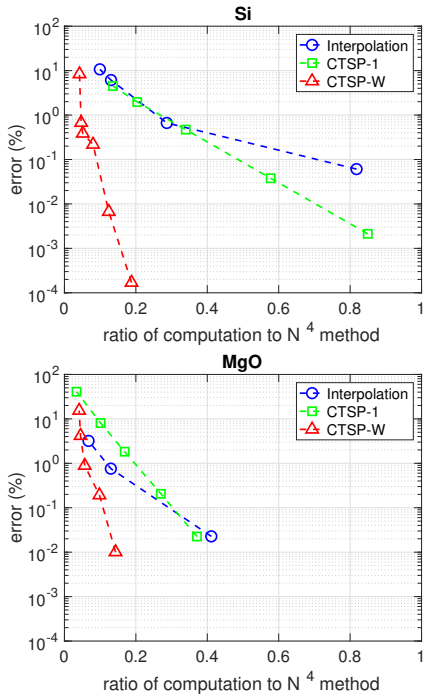

FIG. 3. Error in the macroscopic RPA optical dielectric constant ϵ_∞ for the interpolation, the CTSP-W and the CTSP-1 methods with respect to the prediction of the standard quartic $\mathcal{O}(N^4)$ technique. The horizontal axis is the ratio of computational load of the cubic methods to the standard $\mathcal{O}(N^4)$ method as measured by operation count for a system of 16 Si atoms. Upper: Bulk Si data generated by using input fractional quadrature error $\epsilon^{(q)}$ $\{0.001, 0.01, 0.1, 0.2\}$ for interpolation; $\{0.001, 0.01, 0.1, 0.3, 0.5\}$ for CTSP-1; and $\{0.001, 0.01, 0.1, 0.3, 0.5, 0.8\}$ for CTSP-W. Middle: same for bulk MgO with $\epsilon^{(q)}$ $\{0.001, 0.01, 0.1\}$ for interpolation $\{0.001, 0.01, 0.1, 0.3, 0.7\}$ for CTSP-1 and $\{0.001, 0.01, 0.1, 0.2, 0.4\}$ for CTSP-W. One point on each method's curve is generated per input value $\epsilon^{(q)}$.

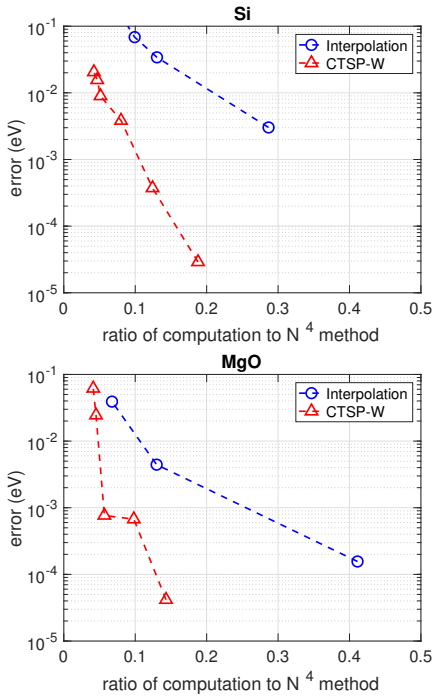

FIG. 4. Error in the COHSEX approximation to the band gap ($\Gamma - X$ gap for Si and Γ for MgO) for the interpolation, and the CTSP-W with respect to the prediction of the standard quartic $\mathcal{O}(N^4)$ technique as function of the ratio of operation count to the standard method. All data sets are computed for a supercell size of 16 atoms. The nomenclature and numerical tolerances are those of Fig. 3.

to the $\mathcal{O}(N^4)$ approach. Note that the computational savings of CTSP-W relative to the standard technique, as measured by operation count, will improve linearly as the number of atoms is increased (beyond $N = 16$).

Figure 5 shows the correlation between the log of the fixed fractional quadrature error, $\epsilon^{(q)}$, and the log of the fractional error in the macroscopic optical dielectric constant given by the application of CTSP-W to Si and MgO. The data indicate the error in ϵ_∞ is at least one order of magnitude smaller than the input fractional quadrature error and the slopes of the log-log curves are approximately unity. Although these results do not represent a rigorous bound, they nonetheless show that the error in physical quantities calculated via the CTSP methods can be controlled by turning one simple "knob" ($\epsilon^{(q)}$) and the integration accuracy required for well converged results is surprisingly modest (see also Sec. III.D).

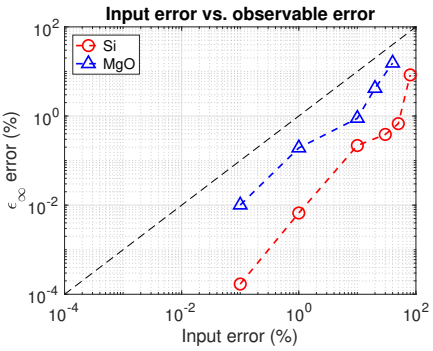

FIG. 5. The relation between the input fractional quadrature error tolerance, $\epsilon^{(q)}$, and the output error in the physical observable, ϵ_∞, for the CTSP-W method applied to Si and MgO. The dotted line represents the situation where quadrature error equals to observable error.

B. Zero gap materials

In order to test the performance of the reduced order approach for static P in zero gap materials, we study crystalline aluminum (Al) (Appendix G contains the details of DFT calculations performed to obtain the band structure). Gaussian broadening ($\beta^{-1} = 0.03$ Ry) is employed to treat the occupation numbers and for simplicity

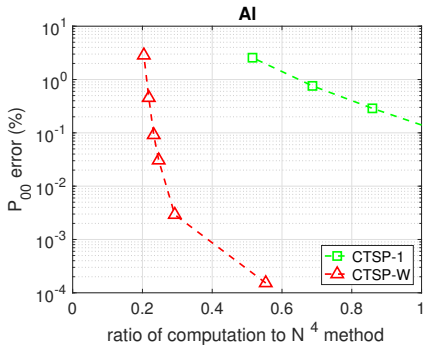

FIG. 6. Error in the $P_{0,0}^{q=(\frac{1}{2},\frac{1}{2},\frac{1}{2})}$ element of bulk Al using CTSP-W. The horizontal axis is the ratio of computational load of the cubic to that of the standard $\mathcal{O}(N^4)$ method for a supercell of 8 Al atoms. A total 400 states were used, and the broadening parameter was set to 0.03 Ry (see Appendix G). The set of fractional quadrature errors $\epsilon^{(q)}$ employed to generate the curve is $\{0.01, 0.1, 0.3, 0.5, 0.7, 0.8\}$.

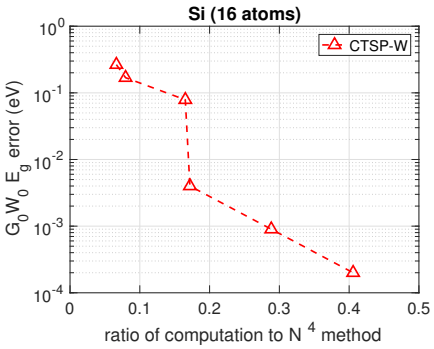

FIG. 7. Error in the bulk G_0W_0 band gap ($\Gamma - X$ gap for Si) for the CTSP-W method as a function of the ratio of computational load to that of the "exact" quartic method (horizontal dashed line). All data were generated using a supercell of 16 Si atoms. The set of fractional integration errors employed to generate the curve is $\{0.001, 0.01, 0.05, 0.1, 0.5, 1\}$. The computed band gap with the $O(N^4)$ method is $E^{(\mathrm{gap},G_0W_0)} = 1.37$ eV.

the occupation numbers are set to 0 or 1 when these differ from their corresponding zero-temperature values by less than 10^{-6}. Although there is no energy gap in a truly extended metallic Al system, calculations performed in a finite periodic supercell will have discrete eigenvalues and a (small) artificial energy gap. However, the new method is robust to zero energy gap as described above and associated appendices.

Figure 6 shows the error in the $P_{0,0}$ element of the RPA irreducible polarizability (where these matrix indices correspond to reciprocal space) for $q = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$. The performance of the CTSP-W method is compared to both the quartic scaling method and the single window limit, CTSP-1 as above. Here, the single window limit is not as efficient as it was for Si or MgO. This is because treating very small $E^{(\mathrm{gap})}$ requires a large quadrature grid to obtain the desired accuracy in the single window limit. However, the CTSP-W method completely removes any trace of difficulties associated with the small $E^{(\mathrm{gap})}$ and delivers accurate results with high efficiency for (nearly) zero-gap systems.

C. G_0W_0 gap

Figure 7 shows the convergence of the G_0W_0 band gap of Si computed via the CTSP-W approach with integration error - the G_0W_0 band gap is determined in the standard way using $\Sigma(\omega)$ computed by the cubic-scaling CTSP-W method as described in Sec. II E. The dynamic behavior of W (i.e., the pole energies ω_p and pole strengths B^p of Eq. (6)) are determined using the generalized plasmon-pole (GPP) model of Hybertsen and Louie[35]. The figure shows that high accuracy is possible with large computational savings when compared to the standard $\mathcal{O}(N^4)$ approach.

D. Single convergence parameter

Finally, compared to standard $\mathcal{O}(N^4)$ GW calculations, our cubic scaling CTSP approach has a single added input parameter which is the *a priori* desired fractional quadrature error, $\epsilon^{(q)}$. Due to the CTSP method's construction, the choice of window parameters and quadrature grids are all determined by this single parameter. As illustrated in Fig. 5, the output error in computed observables is smaller in magnitude than the input quadrature error. Hence, one can estimate quickly the value of the input quadrature error that bounds the desired accuracy in the output observables (although the bound is not rigorous).

A simple quantity that can be computed in $\mathcal{O}(N)$ complexity via CTSP and provides an estimate of the error in physical observables generated by the CTSP method, is the model static polarizability,

$$P^{(\mathrm{model})} = \sum_{cv} \frac{f(E_v) - f(E_c)}{E_c - E_v} \ . \tag{38}$$

That is, Eq. (38), can be computed for a range of quadrature errors at the start of a GW calculation: monitoring the CTSP error in $P^{(\mathrm{model})}$ provides a refined estimate of the $\epsilon^{(q)}$ required to reach the desired error level in the resulting observables specified at input.

If a more quantitative estimate of the error in physical observables is required, we recommend performing a convergence study on a small model system representative of the system of interest (e.g., a small supercell or unit cell of bulk material instead of a large supercell of bulk with defects, an idealized surface with small in-plane lattice parameters instead of a complex surface reconstruction, etc.). This procedure will again necessitate performing a series of cubic scaling CTSP computations at various levels of quadrature error, $\epsilon^{(q)}$, to refine the pa-

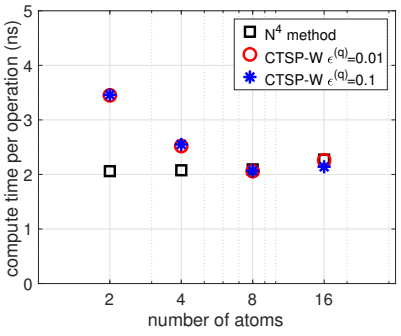

FIG. 8. Compute time per operation for evaluation of the static P of crystalline Si. Black squares indicate the $\mathcal{O}(N^4)$ method, and red circles and blue asterisks indicate the $\mathcal{O}(N^3)$ CTSP-W method with maximum fractional quadrature error ($\epsilon^{(q)}$) of 1% and 10%. A standard workstation with the Linux operating system was employed – the application was not parallelized. The flatness of the curves for both the $\mathcal{O}(N^4)$ and $\mathcal{O}(N^3)$ methods indicate their computation time increase as N^4 and N^3: there is no additional hidden cost to using the $\mathcal{O}(N^3)$ method in an actual calculation. Computational details are given in Appendix G. Issues such as cache utilization, different for different system sizes in our un-tuned application, account for the factors of $< 2\times$ deviation from a horizontal line.

rameter choice through a direct study of the convergence of the physical observables. Due to the small size of the model system, the process will require small or negligible compute time. Finally, a full convergence study of the CTSP prediction of the observables of interest as a function of $\epsilon^{(q)}$ on the (large) system can be performed. This latter procedure retains cubic computational complexity but with increased prefactor. In general, performing convergence studies using CTSP on model systems is likely to provide sufficient error control for validation purposes - the model system approach is commonly used to select the cutoff energy in the conduction band for standard GW computations, for instance.

IV. RESULTS: SCALING ANALYSIS AND COMPARISON TO OTHER METHODS

First, the cubic scaling and small prefactor of the new CTSP-W $\mathcal{O}(N^3)$ method of this paper are verified with actual computations. Next, the ability of the new technique to treat physical systems of scientific and technological interest that heretofore have been too computationally intensive to study routinely, is evaluated. Last, the performance of the new method is compared to small prefactor $\mathcal{O}(N^4)$ methods and other $\mathcal{O}(N^3)$ techniques.

A. Verification of cubic scaling

We verify the scaling of the CTSP-W method in realistic calculations at two input fractional quadrature error levels ($\epsilon^{(q)}$). The actual total computer time required to compute the static P is measured as a function system size (number of atoms in the supercell) and the compute time *per operation* presented in Fig. 8 for crystalline silicon. The number of operations are $N_v N_c N_r^2$ for the $\mathcal{O}(N^4)$ method and $\sum_{l,m} N_{lm}^{(\tau,\mathrm{GL})}(L_l^{(v)} + L_m^{(c)})N_r^2$ for the CTSP-W $\mathcal{O}(N^3)$ method (see Eq. (16)). The result is a flat line – the algorithms scale as they should on a present-day desktop computer.[44]

It is important that the compute times *per operation* are very close to each other indicating that the $\mathcal{O}(N^3)$ method has a prefactor that is comparable to that of the $\mathcal{O}(N^4)$ method even in small systems, $N \lessapprox 10$ atoms. Thus, the reduced order method is highly efficient. These results also validate our use of operation counts as the measure of computational work in the comparisons presented above (and below) – the CTSP method has virtually the same computational overhead as that of the standard $\mathcal{O}(N^4)$ scaling approach per operation.

B. Sizing for large systems

The computational effort required to generate the static $P(0)_{r,r'}$ polarizability matrix with the quartic, CTSP-1 and CTSP-W methods will now be analyzed for two systems: "medium" and "large". The medium-sized system is a 72-atom GaN supercell, while the large system is a 177-atom photovoltaic interfacial system. The number of plane-waves in the basis set is 19,200 and 149,000, and the number of FFT grid points is 39,000 and 319,000 for 72-atom and 177-atom systems, respectively (see Table I). The first material system, GaN, is a III/V semiconductor important in the production of RF (radio frequency) components and LEDs (light emitting diodes). The second material system, a hybrid photovoltaic interfacial system consists of ZnO nanowires covalently bonded to P3HT polymer chains (Poly(3-hexylthiophene)), has been studied previously[45] both experimentally and theoretically at the DFT-level of theory: this type of photovoltaic system combines an organic polymer photoabsorber (here, P3HT) with an inorganic carrier transport channel (here, ZnO) at the nanometer scale to dissociate optically excited excitons into highly mobile carriers.

Table II shows the number of operations[27] required to compute $P(0)_{r,r'}$ for the two systems using the standard quartic scaling method, CTSP-1 and CTSP-W. For the CTSP-W method, the parameters are selected by minimizing the cost function described in Sec. II C. The quadrature grids are chosen to achieve less than 0.1% error in the calculation of $P^{(\mathrm{model})}$. We emphasize that 0.1% error in $\epsilon^{(q)}$ will achieve accurate results as pre-

	Medium system	Large system
N	72	177
N_v	144	529
N_c	2,806	10,600
N_k	8	4
N_{FFT}	39,000	319,000
Gap	2.2 eV	1 eV
Bandwidth	110 eV	103 eV

TABLE I. Representative medium and large physical systems that at present are computationally challenging to approach with existing $\mathcal{O}(N^4)$ GW methods. Here, N_v and N_c are number of valence and conduction bands respectively. N_k is number of k-points to be sampled. N_{FFT} is the number of FFT grids, which is the same as the rank of P matrix. The gap and bandwidth are estimates to be determined by application of our method.

sented in Figs. 3-5. Table II shows that the CTSP-W method yields an efficient computation of $P(0)_{r,r'}$ without sacrificing accuracy – for the medium system, the CTSP-W method delivers about a 40× reduction in operation count while, for the large system, CTSP-W delivers about a 100× reduction (compared to the standard quartic scaling technique). Thus, these technologically interesting problems are now approachable in terms of computer time typically available to users of supercomputer centers throughout the world.

It is important also to consider the memory requirements to store a large matrix such as $P_{r,r'}$ and whether this requirement can be satisfied by today's supercomputers (see Table II). The Blue Waters machine at NCSA[46], a leading HPC platform, has 64 GB of memory per node and the installation has 23K nodes for a total of 1.4 petabytes of memory. The BlueGene/Q installation at Argonne National Laboratory, Mira, has 16 GB of memory per node and 49K nodes for a total of 0.8 petabytes. Thus, using only a fractional allocation of such computers, the P-matrix, even for the large system, can easily be accommodated. More compact representations of the P-matrix are possible and under development. Of course, the effective utilization of distributed memory supercomputers requires a well-parallelized GW software implementation such as that developed by us in Ref. 27 or by others in Ref. 26; both of these software applications can be modified to implement the CTSP methodology straightforwardly.

Last, it is worth noting that 5-index entities such as those presented in Sec. II.D are purely formal devices and never stored in full form. The only large matrix stored is $P_{r,r'}$ and that is computed in N_T^2 tiles of size $(N_r/N_T \times N_r/N_T)$, indexed by $\{T, T'\}$, to reduce the memory footprint as follows:[27] intermediate quantities can be strictly reduced to matrices of size $(N_r/N_T \times N_r/N_T)$ because at each step through the l, m, T, T', u control sum or loop, two such matrices are constructed (i.e., the current tile of $S^{lmu}_{r,r'}$ and $Q^{lmu}_{r,r'}$), their product taken and the result added to the current resolution of the current tile of $P_{r,r'}$. If necessary, Q, Z and their product can be subsequently computed, reusing the memory, and also added to resolution of the current tile of $P_{r,r'}$. In the next iteration of the 5-index sum, the memory is, again, reused and the process repeated until $P_{r,r'}$ is fully resolved (the 5-index sum is completed).

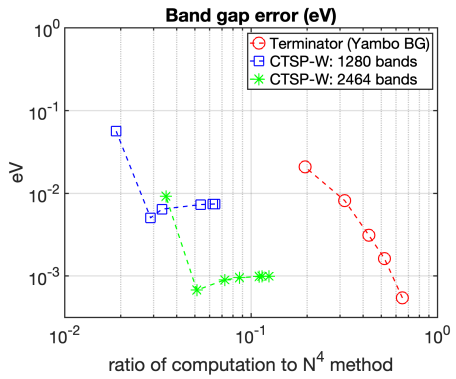

FIG. 9. Comparison of the performance, absolute band gap error versus computational load, for the $\mathcal{O}(N^4)$ GW terminator method of Ref. 47 as implemented in the Yambo software, labeled Yambo BG, and the $\mathcal{O}(N^3)$ CTSP-W method for a 16-atom Si cell with one k-point. The BG band gap is shown as red circles. The CTSP-W 2464 band gap (i.e., computed using 2464 bands) is given by the green stars, and the CTSP-W 1280 band gap (i.e., computed using 1280 bands) is given by the blue squares. See Fig. 3 for other details.

C. Comparison with small prefactor $\mathcal{O}(N^4)$ methods

Next, we compare the CTSP-W method to existing, low prefactor, quartic scaling GW methods. We choose to employ the Yambo GW software (http://www.yambo-code.org/) which implements a quartic scaling sum-over-states technique within the "terminator" acceleration approach of Bruneval and Gonze (BG)[47]. A 16-atom Si supercell with one $k-$point is employed. Figure 9 shows the error in the band gap versus computational savings. The band gap error is referenced to a computation utilizing a large number of bands (3200) that converges the gap to better than 1 meV. Computational savings are referenced to a computation utilizing 2464 bands with the standard quartic scaling method. As expected, the BG method systematically rapidly improves the band gap as the computational load is increased (i.e., more unoccupied bands are explicitly summed over), since it is designed to compute a good approximation to having an infinite number of unoccupied bands. The CTSP-W is run under two realistic conditions: using 1280 bands which leads to a band gap error ~ 10 meV if all 1280 bands are used, and using 2464 bands which leads to a band gap error below 1 meV. We note that a precision of 10 meV for band gaps

		Medium	Large
Operation count $P^q_{r,r'}$	Standard $O(N^4)$	4.92×10^{15}	2.28×10^{18}
	CTSP-1	5.38×10^{14}	2.26×10^{17}
	CTSP-W	1.18×10^{14}	2.67×10^{16}
Memory for $P^q_{r,r'}$		23 GB	1 TB

TABLE II. Number of operations required to calculate the $P^q_{r,r'}$ matrix for the medium and large systems of Table I followed by the memory required to store $P_{r,r'}$ in each case.

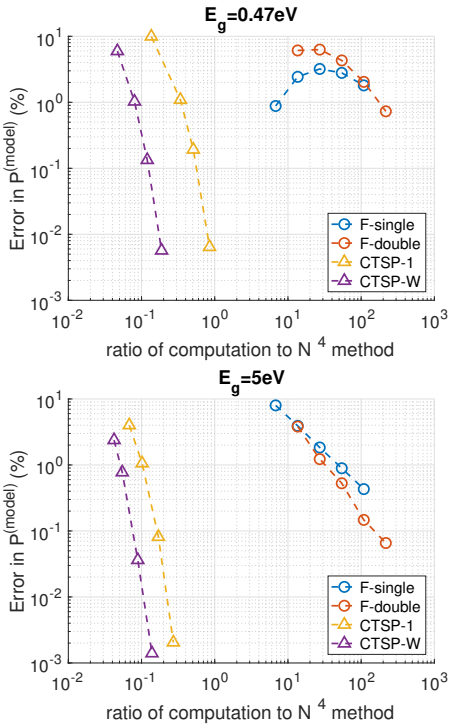

FIG. 10. Comparison between the performance of the cubic scaling Foerster et al.[48] and the CTSP-W method. The y-axis is the error (%) in $P^{(\mathrm{model})}$ (for a 16-atom Si supercell with 399 bands) and the x-axis is the log of the ratio of computational work to that of the quartic method. The standard for the error is the quartic scaling result. For the Foerster method, the operation count is $N_\omega(N_c+N_v)$ where N_ω is the number of frequency grid points, and for the CTSP-W method the workload is defined in Eq. (16) where N_r is set to one. The band gap $E^{(\mathrm{gap})}$ is manually adjusted (i.e., using a "scissors" operation, to investigate its effect on performance. F-single and F-double mean the Foerster method using single and double windows.

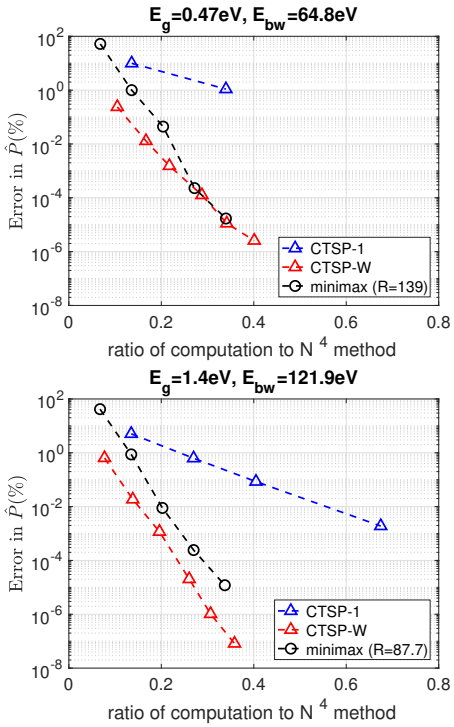

FIG. 11. Comparison of the performance of the cubic scaling minimax method of Liu et al.[49] and the CTSP-W method of this paper, for the computation of $P^{(\mathrm{model})}$. The computational work (x-axis) for the minimax method is $N_\omega(N_c+N_v)$ where N_ω is the number of imaginary time grid points (see also the text). The standard quartic scaling method provides the baseline.

is more than sufficient for GW calculations since the GW approximation itself is not this accurate.

Both the $\mathcal{O}(N^4)$ terminator method and the $\mathcal{O}(N^3)$ CTSP-W method deliver significant savings in the computational workload compared to the standard quartic scaling method for an accuracy of 10 meV in the band gap of the 16-atom Si system. We observe that the CTSP-W method is already more efficient than the terminator method for a 16-atom cell. Since the CTSP-W method scales cubically, its efficiency advantage over the quartic method increases linearly with system size. We thus conclude that the new cubic method has a sufficiently small prefactor to be competitive with existing accelerated quartic scaling GW methods even for simulations cells with as few as ~ 10 atoms.

D. Comparison with other $\mathcal{O}(N^3)$ methods

Last, the performance of the new CTSP-W method is compared to other cubic scaling GW methods – that of Foerster *et al.*[48] and Liu *et al.*[49], respectively – in

Figs. 10 and 11. The comparison is for the model given in Eq. (38). The input energies are taken from a 16-atom crystalline Si supercell with 399 bands generated at the Γ-point of the BZ (The band occupancies are 1 or 0 for this system). The workload is defined as $N_\omega(N_c + N_v)$ where N_ω is a number of frequency grid points used in Eq. (32) of Ref. 48. We calculate $P^{(\mathrm{model})}$ with the single and double window methods of Foerster et al. To investigate the effect of the input single particle band gap on performance, we manually adjust it from $E^{(\mathrm{gap})} = 0.47$ eV to $E^{(\mathrm{gap})} = 5$ eV by uniformly shifting the conduction bands up in energy (a "scissors" operation).

First, for all cases examined, CTSP is more computationally efficient than the method of Foerester *et al.* for the same level of accuracy. We note that the approach of Foerester has not been widely adoption in GW applications due to its large crossover point (large system sizes are required before the technique is more efficient than the standard approach).

Second, Fig. 11 shows a comparison between the minimax grid technique (the approach of Liu *et al.*) and the CTSP-W method for computing the static polarizability, $P^{(\mathrm{model})}$. We used two sets of data, 399 Si eigenvalues from 16-atom cell and 435 MgO eigenvalues from 16-atom cell. For the minimax method, the computational work is defined as $N_\omega(N_c + N_v)$ where N_ω is the number of imaginary time grid points used in the minimax technique. We find that the minimax method is competitive with the CTSP-W method but slightly inferior in performance.

To compare two methods more deeply, we note two points in favor the CTSP-W method. First, the choice of quadrature grids and energy windows is a straightforward and robust exercise, requiring only the minimization of a simple cost function. By contrast, finding the N_ω energy grid points that solve the minmax problem[49] is quite challenging, and we found it required significant (human and computer) effort to do so. Second, CTSP-W computes frequency-dependent spectral quantities such as the polarizability and the self-energy directly on the real ω frequency axis which is the final desired and useful physical representation of any spectral function. Namely, using CTSP-W, there is no need to compute quantities along the imaginary energy or time axis and then analytically continue to real frequencies. This is highly desirable as it avoids (i) the use of analytical continuation methods that are based on assumptions on the analytical form of the functions, and (ii) the numerical instabilities inherent in analytical continuation when high accuracy is desired.[50]

V. CONCLUSION

In summary, the GW equations have been recast, exactly, as Fourier-Laplace time integrals over complex time propagators. The propagators are then partitioned in energy space and the time integrals approximated in a controlled manner using generalized Gaussian quadratures. Coupled with discrete variable methods to represent the propagators in real-space, a cubic scaling GW method emerges. Comparisons show that the new method, CTSP, has sufficiently small prefactor to outperform standard and accelerated quartic scaling methods on small systems ($N \gtrapprox 10$ atoms). For large systems (up to 200-300 atoms), we demonstrate the method fits comfortably in today's supercomputers both in terms of memory requirement and computational load and offers speedups of 2 to 3 orders of magnitude compared to the conventional technique. CTSP's efficiency indicate that it has the potential for wide adoption, and we are currently developing a fine grained parallel version of the method based on our previous work.[27]

Lastly, we discuss possible further development of the CTSP method aimed at reducing its prefactor and/or its computational complexity (scaling). The key CTSP expressions of Eqs. (22) and (35) contain sums over many high energy conduction band states which are computationally expensive both to generate and manipulate. Thus, one may develop a modified terminator method[47] to reduce the number of needed high energy band states significantly, thereby reducing the $\mathcal{O}(N^3)$ prefactor. Reducing the scaling to $\mathcal{O}(N^2)$ or $\mathcal{O}(N^2 \log N)$ is more challenging for a plane-wave basis. However, restricting band state sums to an energy window (as in CTSP-W) is equivalent to summing over all band states with occupancies given by the difference between two different Fermi-Dirac distributions whose chemical potentials are set at the start and end of the window, respectively. Thus, we can, in principle, apply the Fermi Operator Expansion[51,52] approach to describe the Fermi-Dirac functions using polynomials of the Hamiltonian and matrix-vector multiplications without reference to the bands themselves. Such an approach should lead to an $\mathcal{O}(N^2)$ (or $\mathcal{O}(N^2 \log N)$) scaling method (under plane waves and other basis sets), but significant future work is required to realize the concept with both low prefactor and effective error control.

ACKNOWLEDGMENTS

We thank Jack Deslippe, Gian-Marco Rignanese and Dennis Newns for helpful discussions. This work was supported by the NSF via grant ACI-1339804. The present address of GJM is Pimpernel Science, Software and Information Technology from whom GJM acknowledges funding; however, the views expressed are those of the authors and do not reflect Pimpernel policy.

APPENDIX OVERVIEW

In order to improve the readability of the paper, we have chosen to place the detailed analyses in appen-

dices. In Appendix A, the Gauss-Laguerre quadrature for window pairs without energy crossing is discussed. In Appendix B, the determination of the optimal windowing by minimization of the cost function for quantities without energy crossings is described. Appendices C and D discuss the weight function and quadrature employed to treated window pairs with energy crossings, respectively, while Appendix E describes minimization of the cost function for quantities whose evaluation involves treating energy window pairs with energy crossings. In Appendix F an alternative $\mathcal{O}(N^3)$ method based on interpolation is given, and in Appendix G computational details related to the results presented in the main text are described. Last, matlab code to generate the weights and nodes of the Hermite-Gauss-Laguerre quadrature is presented.

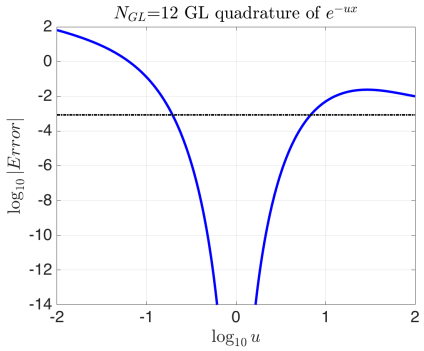

FIG. 12. Gauss-Laguerre (GL) quadrature error in the integration of $e^{-x\tau}$ with 12 quadrature points as a function of $\log_{10} x$, solid blue curve (see Eq. (A.1)). The dashed black horizontal line shows that for $-0.75 \lessapprox \log_{10} x \lessapprox 0.75$ equal error is generated for x and $1/x$

Appendix A: Gauss-Laguerre quadrature optimization

We provide the optimizations required to evaluate energy denominators by discrete approximation to time domain integrals for a set of energies in a window pair.

1. Optimal error matching choice for energy scale ζ

First, we describe the optimal error matching choice of ζ_{lm} for the Gauss-Laguerre (GL) quadrature of Eq. (24). We suppress the energy window index lm and describe why $\zeta^{-1} \approx \sqrt{E^{(\mathrm{bw})}E^{(\mathrm{gap})}}$ is a good choice for the energy scale ζ: it equalizes the error of the GL quadrature across a given window pair.

We seek to optimally approximate the continuous time integral yielding the desired energy denominaor via numerical quadrature,

$$\frac{1}{\Delta} = \zeta \int_0^\infty e^{-\zeta\Delta\tau} d\tau \approx \zeta \sum_{u=1}^{N^{(\tau,\mathrm{GL})}} w_u e^{-\tau_u(\zeta\Delta-1)}$$

for $\Delta = E_c - E_v > 0$. That is, defining the dimensionless quantity, $x = \zeta\Delta$, we wish to minimize the error

$$\frac{\epsilon^{(q)}(x)}{x} = \frac{1}{x} - \sum_{u=1}^{N^{(\tau,\mathrm{GL})}} w_u \exp(-\tau_u(x-1)) \,. \qquad \text{(A1)}$$

for x spanning the scaled range of a given window pair. We first note that the error is exactly zero at $x = 1$ since GL quadrature is exact when integrating $e^{-\tau}$. Figure 12 shows a plot of the error versus x. The the error curve is symmetric around $\ln x = 0$, especially when smaller error values are of interest, which is the case herein. That is, the integration error, to a good approximation, is even in $\ln x$ about $\ln x = 0$.

Second, the interband energies Δ range from $E^{(\mathrm{gap})}$ to $E^{(\mathrm{bw})}$. Examining Fig. 12, the lowest errors are sampled as x ranges from its lowest value of $\zeta E^{(\mathrm{gap})}$ to its highest value of $\zeta E^{(\mathrm{bw})}$. Therefore, it is reasonable to choose ζ such that $x = \zeta E^{(\mathrm{gap})} < 1$ and $x = \zeta E^{(\mathrm{bw})} > 1$ straddle $x = 1$ and have the same error, i.e., optimal error equalization. For a symmetric error function about $\ln x = 0$, this requires $-\ln(\zeta E^{(\mathrm{gap})}) = \ln(\zeta E^{(\mathrm{bw})})$ which yields the geometric mean $\zeta^{-1} = \sqrt{E^{(\mathrm{bw})}E^{(\mathrm{gap})}}$. The geometric mean becomes exactly optimal as $N^{(\tau,\mathrm{GL})}$ is increased as well as when $E^{(\mathrm{bw})}/E^{(\mathrm{gap})}$ is close to unity (the many windows limit).

2. Number of Gauss-Laguerre quadarature points for bounded error

When we fix $\zeta^{-1} = \sqrt{E^{(\mathrm{bw})}E^{(\mathrm{gap})}}$, the maximum fractional error of Eq. (A1), $\epsilon^{(q)}$, occurs at the largest energy transition (i.e., the error in computing the inverse energy $1/E^{(\mathrm{bw})}$ via quadrature). For $N^{(\tau,\mathrm{GL})}$ quadrature points, we have

$$\epsilon^{(q)}(\alpha) = 1 - \alpha \sum_{u=1}^{N^{(\tau,\mathrm{GL})}} w_u \exp\left[(1-\alpha)\tau_u\right]. \qquad \text{(A2)}$$

where $\alpha = \sqrt{E^{(\mathrm{bw})}/E^{(\mathrm{gap})}}$. The analogous equation for the fractional error in the computation of $1/E^{(\mathrm{gap})}$ has α replaced by $1/\alpha$, and is equal to the error of Eq. (A2) due to optimal error-matching choice of ζ. Figure 13 displays a contour plot of the fractional quadrature error, $\epsilon^{(q)}$ of Eq. (A2). The plot demonstrates that $N^{(\tau,\mathrm{GL})}$ is essentially linear in α for any fixed choice of fractional error. Analysis of the contour plot shows that an accurate and compact explicit relation between the variables is

$$N^{(\tau,\mathrm{GL})}(\alpha;\epsilon^{(q)}) = \alpha(y - 0.3\ln\epsilon^{(q)}) \qquad \text{(A3)}$$
$$y = 0.4$$

This equation fits the data well for $\epsilon^{(q)} \leq 0.135$ but the range can be extended to unit $\epsilon^{(q)}$ by taking $y = 1.0$.

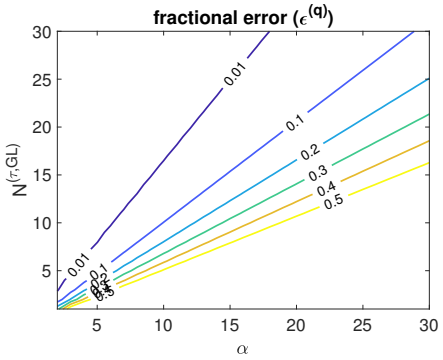

FIG. 13. The fractional error, $\epsilon^{(q)}$, of Gauss-Lauguerre quadrature (for $x = 1/E^{(\mathrm{bw})}$) as a function of α and $N^{(\tau,\mathrm{GL})}$. Here, α is defined to be $\sqrt{E^{(\mathrm{bw})}/E^{(\mathrm{gap})}}$. Each contour is labeled by $\epsilon^{(q)}$. For a fixed fractional error, $N^{(\tau,\mathrm{GL})}$ is linear in α.

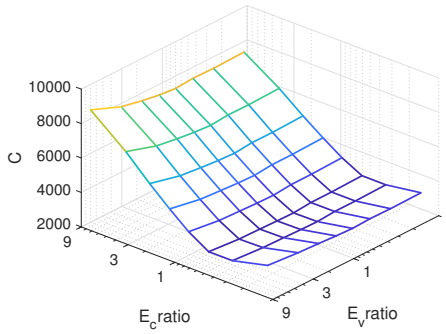

FIG. 14. Computational cost to compute the static P using a $N_{v_w} = 2 \times N_{c_w} = 2$ window scheme as a function of energy window size for a 16-atom Si system with 399 states (32 occupied and 367 unoccupied states) and one k-point (no $k-$point sampling and hence $q = 0$ strictly). The position of the minimum *does not* occur at the equipartition point, $(1, 1)$, but rather at $(E_v^{(\mathrm{ratio})}, E_c^{(\mathrm{ratio})}) = (1.25, 0.29)$.

Note, our choice bounds the integration error: the end points of the energy windows are worst cases and all other transitions are computed more accurately. Hence, the number of quadrature points needed to compute the interband transitions within an energy window pair can be simply estimated so as to ensure a maximal *a priori* fractional error bound, $\epsilon^{(q)}$.

Appendix B: Optimal sets of energy windows

We describe our prescription to determine the optimal number and placement of energy windows in the range of E_c and E_v. This is accomplished by minimizing the computational cost function $C^{(\mathrm{GL})}(\epsilon^{(q)})$ of Eq. (26). In this appendix, we omit the fractional error level $\epsilon^{(q)}$ as it does not affect the optimal set of energy windows. To motivate the discussion, consider a 2×2 window scheme where the two free parameters are the dividing energy values E_v^* and E_c^* in the valence and conduction bands, respectively, that determine the boundaries of the windows. These are converted to dimensionless quantities, $E_c^{(\mathrm{ratio})} = (E_c^* - E_c^{(\mathrm{min})})/(E_c^{(\mathrm{max})} - E_c^*)$ and $E_v^{(\mathrm{ratio})} = (E_v^* - E_v^{(\mathrm{min})})/(E_v^{(\mathrm{max})} - E_v^*)$. Figure 14 shows the dependence of the cost $C^{(\mathrm{GL})}$ on two ratios for the case of flat densities of states. The function, $C^{(\mathrm{GL})}$, is a smooth function of the window boundaries and we find that this smoothness is not confined to 2×2 windowing but carries over to larger number of windows. Note, the position of the minimum in Fig. 14 is nontrivial, occurring at the point, $(E_v^{(\mathrm{ratio})}, E_c^{(\mathrm{ratio})}) = (1.25, 0.29)$.

Since $C^{(\mathrm{GL})}$ is a smooth function of the energy window partitions, for a given number of windows (N_{v_w},N_{c_w}) and some starting set of window partitions (e.g. all equal), we can employ a simple gradient descent algorithm to minimize $C^{(\mathrm{GL})}$ over the positions of the energy window boundaries and to find the minimum value of $C^{(\mathrm{GL})}(N_{v_w},N_{c_w})$. By varying the number of windows N_{v_w}, N_{c_w} over a reasonable range and tabulating the minimized cost function $C^{(\mathrm{GL})}(N_{v_w},N_{c_w})$, the global minimum and the hence the optimal choice of windowing, i.e., the number of windows pairs $\{N_{v_w},N_{c_w}\}$ and their partitioning of the energy ranges, can be found. In practice, varying the number of window from 1 to 9 is sufficient to determine the best windowing choice for all the systems we have considered; so that, 81 small minimization procedures are performed in total. Note, the process is simplified because of the separable nature of $N^{(\tau,\mathrm{GL})}$ of Eq. A3: the partitioning results do not depend on the desired fractional error, $\epsilon^{(q)}$.

Figure 15 illustrates the minimal value the cost function at several $\{N_{v_w},N_{c_w}\}$ for a bulk Si crystal described by 16-atom supercell and 32 valence and 367 conduction band states. Here, the minimal computational load occurs for at the point, $(N_{v_w} = 1, N_{c_w} = 5)$. For $k-$point sampling over the first BZ under the CTSP-W method, the computation of P^q at momentum transfer q is optimized by applying the windowing with cost function minimization procedure to each $k, k+q$ *pair*. That is, the densities of states acquire band indices, $D^k(E), D^{k+q}(E)$, and the number of windows $\{N_{v_w}^k,N_{c_w}^{k+q}\}$ and their partition, the sets $\{E_k^{(v,\mathrm{min})}, E_k^{(v,\mathrm{max})}\}$, $\{E_{k+q}^{(c,\mathrm{min})}, E_{k+q}^{(c,\mathrm{max})}\}$, are optimized for each $(k, k + q)$ pair in the BZ.

Appendix C: Weight function for window pairs with energy crossings

We develop a weight function and associated quadrature for the case when $F(x;\zeta)$ must be evaluated for energy differences x that are both positive and negative within a window pair, i.e., energy crossings occur. A standard choice in the GW literature is to employ a Lorentzian broadening parameter $\gamma > 0$ to regularize the

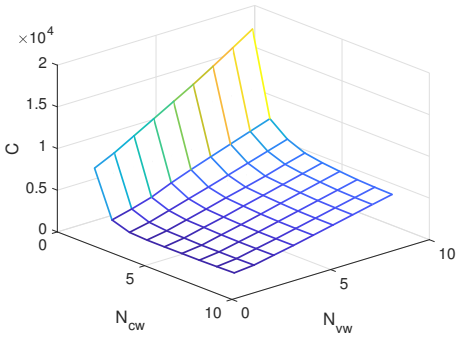

FIG. 15. Minimized computational cost, $C^{(\mathrm{GL})}$, to compute the static P of a 16-atom Si system for window pairs $\{N_{v_w}, N_{c_w}\}$ spanning $(N_{v_w} = 9 \times N_{c_w} = 9)$ (81 total pairs). The total number of bands in the system was taken to be 399 (32 occupied and 367 unoccupied states) and one k-point (no $k-$point sampling and hence $q = 0$ strictly). The position of the minimum is at the point, $(N_{v_w} = 1, N_{c_w} = 5)$.

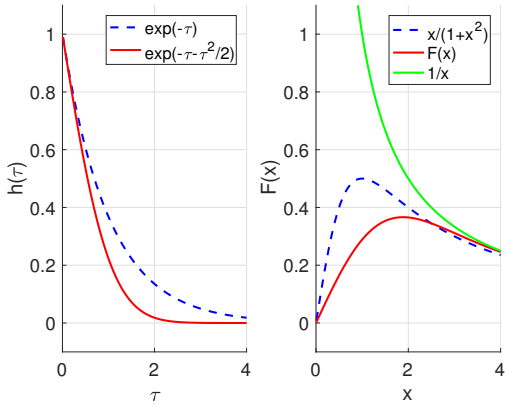

FIG. 16. Left: comparison of the two weight functions described in the text for $\gamma = 1$. The blue dashed curve is the exponential weight $\exp(-\tau)$ associated with Lorentzian broadening; the solid red curve is the new weight function associated with Eq. (C.3). Right: Fourier transforms of the weight functions. The transform of the exponential weight $e^{-\tau}$ (dashed blue) is $x/(1+x^2)$ while the transform of the weight $h(\tau) = \exp(-\tau - \tau^2/2)$ is given by Eq. (C.3) (solid red). For comparison, the target function $1/x$ is shown as well (short dashed green). Equation (C.3) is smoother for small x and approaches $1/x$ more rapidly at large x than $x/(1+x^2)$.

singularity of $1/x$ by replacing it with

$$F(x) = Im\frac{\gamma}{1 - ix\gamma} = \frac{x}{x^2 + \gamma^{-2}}\,. \tag{C1}$$

in the spirit of the additional scattering that typically ameliorates resonances in real materials. This odd function in x is continuous, approximates $1/x$ when $\gamma|x| \gg 1$, and has a separable form as a Fourier integral

$$F(x) = \gamma\, Im \int_0^\infty d\tau\, e^{-\tau} e^{i\tau x\gamma}\,. \tag{C2}$$

The exponential weight function implies that the most appropriate quadrature method for approximating the integral is the simply Gauss-Laguerre quadrature. Hence, this $F(x)$ can be used to separate the sums over n and p when computing $\Sigma(\omega)$.

The difficulties with this choice are practical. First, the quadrature grids needed for reasonable errors can become large. Second, the function approaches $1/x$ only when $|x| \gg \gamma^{-1}$ such that if γ^{-1} is not small compared to the width of the energy windows being employed, there will be sizable errors across window boundaries when we switch from $F(x)$ to $1/x$. On the other hand, if we make γ^{-1} small to avoid this matching error, the steepness of $F(x)$ near the origin, which is directly related to the rapid oscillations versus τ of $e^{-i\gamma x\tau}$ with large γ in the integral form of F in Eq. (C2), requires a large quadrature grid to describe accurately.

We alleviate the above difficulties by taking advantage of the freedom afforded in choosing the functional form of $F(x;\zeta)$ in Eq. (9). Instead of employing the weight function $h(\tau;\zeta) = |\zeta|e^{-\tau}$ with $\zeta = i\gamma$, we propose to use

$$h(\tau;\zeta) = |\zeta| \exp(-\tau - \tau^2/2)$$

which falls off much faster for large τ and will thus generate a much smoother $F(x)$ for small x. However, since its behavior for small τ is the same as the $e^{-\tau}$, the associated $F(x)$ will also approach $1/x$ asymptotically at large x. In addition, choosing the ratio of exactly $1/2$ between the prefactors of the linear and quadratic parts of the exponential defining h is not arbitrary: this choice of ratio guarantees that $F(x;\zeta) = 1/x + O(1/x^5)$ for large x while any other choice $F(x;\zeta) = 1/x + O(1/x^3)$. We also note the tranform $F(x)$ can be written, in closed form, in terms of the generalized error function

$$F(x) = \zeta Im\left\{\sqrt{\frac{\pi}{2}} e^{-\frac{(x\zeta+i)^2}{2}}\left[1 + i\mathrm{erfi}\left(\frac{x\zeta + i}{\sqrt{2}}\right)\right]\right\}. \tag{C3}$$

Figure 16 shows a comparison of the two weight functions and their computed Fourier transforms $F(x)$. The weights and nodes for a Gaussian-type quadrature for the weight function, $\exp(-\tau-\tau^2/2)$, which we term Hermite-Gauss-Laguerre (HGL) quadrature, can be generate using the procedures embodied in the matlab code provided in Appendix H.

It is useful to compare the accuracy of with which the two choices of weight function can be numerically integrated. Table III shows the number of quadrature points required to generate a specified error when using the Lorentzian generating weight $e^{-\tau}$ and improved weight $\exp(-\tau - \tau^2/2)$ for an energy window of unit width. To generate this table, we specify a maximum percentage error and then find γ such that $F(x)$ differs from $1/x$ by less than the specified error when $x = 1$. We then find the size of a quadrature grid $N^{(\tau,\mathrm{HGL})}$ such that the difference between the quadrature approximation of

% error	$N^{(\tau,\mathrm{GL})}$ $(w = e^{-\tau})$	$N^{(\tau,\mathrm{HGL})}$ $(w = e^{-\tau-\tau^2/2})$
5	6	1
1	24	1
0.1	124	5
0.01	547	15
0.001	2216	36

TABLE III. Number of quadrature grid points required to meet the maximum specified percent error for the for the integration of $\exp(i\tau)$ over the two weight functions discussed in this section - the energy window has unit width, i.e., $\gamma x = 1$.

Eq. (35) and the true $F(x)$ is below the same error level for all x in the window (i.e., $0 \le x \le 1$). It is clear that the new weight function and associated quadrature is at least an order of magnitude more efficient in generating its transfrom than the standard choice $e^{-\tau}$.

Appendix D: Hermite-Gauss-Laguerre quadrature grid size at fixed error

A necessary input to the cost function, whose minimization determines optimal window placement, is the number of grid points required to generate a desired error level, $\epsilon^{(q)}$, in the time integrals. Figure 1(b) shows a 2×2 windowing example containing window pairs with an energy crossing. That is, the sign of the denominator changes for the window pairs $\{E_{a,1}, E_{b,1}\}$ and $\{E_{a,2}, E_{b,1}\}$. In order to treat such pairs, we employ the weight function $h(\tau;\zeta) = |\zeta|\exp(-\tau - \tau^2/2)$ and Hermite-Gauss-Laguerre quadrature to discretize the τ integrals. For all window pairs without energy crossing, the time integrals are discretized using Gauss-Laguerre quadrature and the methodology developed for static P computations; these windows are not considered further. We continue below to develop the tools required to treat windows with energy crossings.

We first seek a quantitative relationship between the number of quadrature points $N^{(\tau,\mathrm{HGL})}$, the energy difference $x = E_a - E_b$, and the fractional error of the quadrature for the case of energy windows with an energy crossing. The fractional quadrature error is defined as

$$\epsilon^{(q)} = \frac{|F(x) - \sum_{u=1}^{N^{(\tau,\mathrm{HGL})}} w_u \sin(\tau_u x)|}{|F(x)|} \qquad \text{(D1)}$$

where, again,

$$F(x) = Im \int_0^\infty d\tau e^{(-\tau-\tau^2/2)} e^{i\tau x}$$

and we have standardized the analysis by setting the energy scaling variable to unity ($|\zeta| = \gamma = 1$). Here, $F(x)$ is computed to very high accuracy via numerical integration or evaluation of the generalized error function. Figure 17 displays the function $\epsilon^{(q)}(x, N^{(\tau,\mathrm{HGL})})$: due to the presence of the sine function in $\epsilon^{(q)}$, the quadrature error $\epsilon^{(q)}$ is oscillatory as a function of x and finding a simple relationship between $\epsilon^{(q)}$, x and $N^{(\tau,\mathrm{HGL})}$ is challenging.

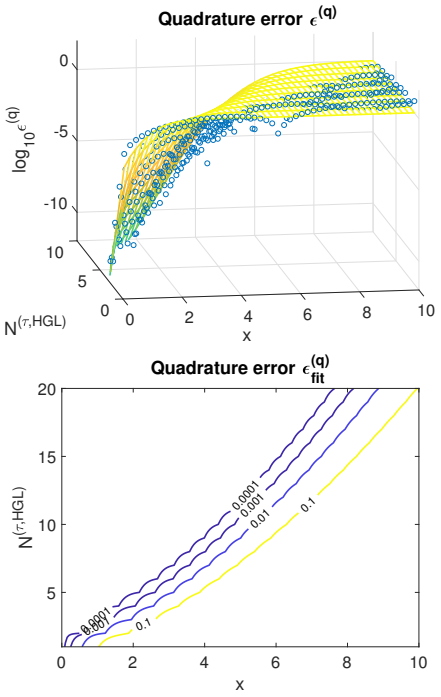

FIG. 17. Hermite-Gauss-Laguerre quadrature error as function of x and $N^{(\tau,\mathrm{HGL})}$ (see Eq. (D.1)). In the upper plot, the fractional quadrature error ($\epsilon^{(q)}$) is indicated as blue dots along with the fit function, $\epsilon^{(q)}_{\mathrm{fit}}$ (see Eq. D2). In the lower plot, the contour lines of $\epsilon^{(q)}_{\mathrm{fit}}$ are shown. Each contour line can be represented with high fidelity using only quadratic function of x (see Eq. D3).

We find that the function, $\epsilon^{(q)}_{\mathrm{fit}}$,

$$\epsilon^{(q)}_{\mathrm{fit}} = \tanh\left(x^{2N^{(\tau,\mathrm{HGL})}}\right) \times \exp\left[-(1 + 3.3N^{(\tau,\mathrm{HGL})})e^{-0.68x^2/N^{(\tau,\mathrm{HGL})}}\right], \qquad \text{(D2)}$$

which is also plotted in Fig. 17, provides a good fit to the data. Direct analytical inversion of Eq. (D2) to obtain $N^{(\tau,\mathrm{HGL})}$ as a function of x and $\epsilon^{(q)}_{\mathrm{fit}}$ is not feasible. However, a good estimate is

$$N^{(\tau,\mathrm{HGL})}(x;\epsilon^{(q)}) = c_2(\epsilon^{(q)})x^2 + c_1(\epsilon^{(q)})x + c_0(\epsilon^{(q)}) \qquad \text{(D3)}$$

where

$$c_2 = -0.0036\ln\epsilon^{(q)} + 0.11\ ,$$
$$c_1 = -0.0043(\ln\epsilon^{(q)})^2 - 0.13\ln\epsilon^{(q)} + 0.54\ ,$$
$$c_0 = -0.204\ln\epsilon^{(q)} - 0.29\ .$$

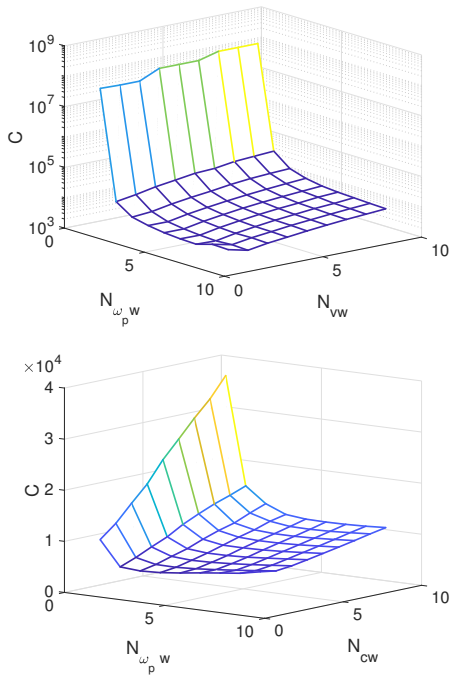

FIG. 18. The optimized cost function to compute the model dynamic $\Sigma^{(\mathrm{model})}(\omega)$ of the text as a function of the number of energy windows for bulk silicon with $\omega = E_{v,min}$ (valence band minimum energy), 16 atoms and 399 bands. The upper and lower plots show the cost to compute the valence and the conduction band contributions to $\Sigma^{(\mathrm{model})}(\omega)$, respectively. The position of the two minima are $(N_{v_w} = 2, N_{p_w} = 7)$, upper, and $(N_{c_w} = 1, N_{p_w} = 3)$, lower.

Appendix E: Treating systems with energy level crossings

Here, the procedure to determine window placement for cases in which there is an energy crossing (e.g. in the computation of the self-energy $\Sigma(\omega)$), is described. In direct analogy with the static P case, we write a cost function with separate energy windows for occupied (valence, v) case $\omega - E_v + \omega_p$ and the unoccupied (conduction, c) case $\omega - E_c - \omega_p$ in the band sums for the self-energy. We then allow the number of energy windows N_{p_w} and N_{v_w} or N_{c_w} to range from 1 to 9, and for each such choice (N_{p_w}, N_{v_w}) or (N_{p_w}, N_{c_w}), the computational cost is minimized via a simple gradient descent method. When a window pair has an energy crossing, we simply employ Eq. (D2) to estimate the quadrature size, while for all other window pairs we employ Eq. (A3) to determine the size of the quadrature grid.

For a concrete example, consider the model self-energy

$$\Sigma^{(\mathrm{model})}(\omega) = \sum_{vp} \frac{1}{\omega - E_v + \omega_p} + \sum_{cp} \frac{1}{\omega - E_c - \omega_p} \,. \tag{E1}$$

using energies and plasmon frequencies from an 8-atom crystalline Si supercell cell. A total of 32 valence bands, 382 conduction bands and and 425 plasmon modes are employed. The valence band ranges from -0.21 to 0.23 Ha, the conduction band from 0.25 to 2.29 Ha, and the plasmon modes from 0.31 to 45.5 Ha. Selecting $\omega = -0.21$, only the valence band sum for $\Sigma(\omega)$ has the sign-change requiring the use of HGL quadrature. The conduction contribution to $\Sigma(\omega)$ does not change sign, and we simply utilize GL quadrature for all $\{\omega_p, E_c\}$ pairs. In Figure 18, we present the cost function minimized for 81 $\{N_{v_w}, N_{p_w}\}$ pairs (upper) and 81 $\{N_{c_w}, N_{p_w}\}$ pairs (lower) at error, $\epsilon^{(q)} = 0.01$. For the valence band sum, the optimal number of windows is $(N_{v_w} = 2, N_{p_w} = 7)$, while for the conduction band, the optimal number of windows is $(N_{c_w} = 1, N_{p_w} = 3)$ (i.e., the position of the minimum in the upper and lower curves of Fig. 18, respectively).

Appendix F: Interpolation method

1. Theory

In real space, the static random phase approximation (RPA) irreducible polarizability matrix is

$$P_{r,r'} = -2 \sum_{v}^{N_v} \sum_{c}^{N_c} \frac{\psi^*_{r,v} \psi_{r,c} \psi^*_{r',c} \psi_{r',v}}{E_c - E_v} \tag{F1}$$

One advantage of working in a real-space basis is that the sum over products of wave functions is separable so one can come up with cubic scaling algorithms if one can make separable approximations to the energy denominator. We begin by rewriting P as

$$P_{r,r'} = -2 \sum_{v} \psi^*_{r,v} \, A(E_v)_{r,r'} \, \psi_{r',v}$$

where the matrix A is defined as

$$A(z)_{r,r'} = \sum_{c} \psi_{r,c} \psi^*_{r',c} / (E_c - z) \,.$$

For a system with an energy gap $E^{(\mathrm{gap})}$, the denominator $E_c - E_v$ is always positive with a minimum value of the gap $E^{(\mathrm{gap})}$. Furthermore, the matrix A must be evaluated only for energies z within the range of valence band energies E_v. Hence, the calculation of P uses $A(z)$ for values of z where it is smooth in z. This means we can use interpolation: we first tabulate $A(z)$ for a range of z values ranging over the valence band energies. This tabulation costs $N_z N_c N_r^2$ which is cubic since the valence bandwidth is an intensive quantity and the number of points N_z needed for a fixed accuracy is a fixed, intensive number. Next, to compute P, we sum over v, and for each E_v we interpolate A to that energy by using the tabulated A. This calculation is also cubic and costs $N_i N_v N_r^2$ where $N_i \leq N_z$ is the number of tabulated z values needed for interpolation (e.g. $N_i = 2$ for linear interpolation).

An efficient interpolation scheme should require a small number of $z-$points N_z as well as a modest interpolation cost N_i. In our case, the energy dependence requiring interpolation is given by $1/(E_c - z)$ which is most rapidly changing for the largest values of z near the top of the valence $E_v^{(\max)}$ band and when E_c takes on its smallest value at the conduction band minimum $E_c^{(\min)}$. Hence, an efficient interpolation scheme will use a non-uniform z grid that appropriately concentrates sampling points near $E_v^{(\max)}$.

The next section below describes the approach we use to find optimal interpolation grids z_j for the case of linear interpolation (i.e., two-point nearest neighbor interpolation with $N_i = 2$) when sampling over the entire range of valence band energies. We note higher order interpolation schemes with $N_i > 2$ can be used as well that will reduce the number of grid points needed for a fixed error but require more work to perform the interpolation. In our experience, the higher order interpolations do not in the end improve performance at the same level of error when compared to the simpler linear interpolation method.

Regardless of the precise interpolation scheme used, all such interpolation methods will have errors that decrease as a power of the number of grid points, n. As the data presented in the main text shows, the Fourier-Laplace transform based methods turn out to have superior error properties (their errors fall off exponentially in n).

2. Energy grids for interpolation

The function of z that we wish to interpolate over z is

$$A(z)_{r,r'} = \sum_c^{N_c} \frac{\psi_{r,c}\psi^*_{r',c}}{E_c - z}\,.$$

The function is steepest versus z close to the top of the valence band $E_v^{(\max)}$ when the energy difference in the denominator is small. In fact, we will consider the worse case scenario and focus on the stiffest and steepest term in the entire sum which is for the case $E_c = E_c^{(\min)}$, the conduction band minimum energy. Hence the most difficult to interpolate term is given by the dimensionless function

$$f(z) = \frac{E_{\rm gap}}{E_c^{(\min)} - z} \equiv \frac{1}{1+x},$$

where $z = E_v^{(\max)} - xE^{(\rm gap)}$, and the scaled energy variable x satisfies $0 \le x \le (E_v^{(\max)} - E_v^{(\min)})/E^{(\rm gap)}$.

The question is how to pick a grid of $\{x_j\}$ values with n points where $x_1 = 0$ and $x_n = (E_v^{(\max)} - E_v^{(\min)})/E^{(\rm gap)}$. For simplicity, we will be using linear interpolation, so that given some x between two grid points $x_j \le x \le x_{j+1}$, the linear interpolation is $f^l(x) = [f(x_j)(x_{j+1} - x) + f(x_{j+1})(x - x_j)]/\Delta x_j$ where $\Delta x_j = x_{j+1} - x_j$. Calculus then provides an analytical expression for the maximum error $f^l(x) - f(x)$ in the interval $x_j \le x \le x_{j+1}$. For large n and thus small spacings Δx_j, the lowest order term for the error is

$$(f^I - f)_{\max} \approx \frac{(\Delta x_j)^2}{4(1+x_j)^3}\,.$$

We wish to bound this error by a fixed fractional error tolerance, $\epsilon^{(q)}$, for all j,

$$\frac{(\Delta x_j)^2}{4(1+x_j)^3} \le \epsilon^{(q)}\,. \tag{F2}$$

which then in principle determines the grid points x_j. In practice, exact solution of this equation is very difficult, so we again appeal to the large n limit where x_j can be viewed as a function $x(j)$ of a continuous argument j so we approximate $\Delta x_j \approx dx/dj$. Then Eq. (F2) turns into an ordinary differential equation with specified boundary conditions. The solution is

$$x(j) = \frac{1}{(1 - (j-1)\sqrt{\epsilon^{(q)}})^2} - 1\,.$$

Since $x(n) = (E_v^{(\max)} - E_v^{(\min)})/E^{(\rm gap)}$ is known, this determines n for each $\epsilon^{(q)}$. And finally we have $z_j = E_v^{(\max)} - x_j E^{(\rm gap)}$.

The above choice of grid bounds the error when evaluating the function once. However, when using the interpolation to compute P from A, we will be evaluating the interpolation over many values across the valence band which approximate an integral. Hence, a more appropriate error control scheme will not only consider the error in interpolating $f(x)$ but also the fact that narrower intervals of x will be sampled less often (assuming a smooth and roughly flat density of states). Hence we should instead bound the error in the function times the size of the interval:

$$\Delta x_j \times \frac{(\Delta x_j)^2}{4(1+x_j)^3} \le \epsilon^{(q)}$$

Repeating the above exercise, the grid appropriate to this error bound is given by

$$x(j) = \exp\left([4\epsilon^{(q)}]^{1/3}(j-1)\right) - 1\,. \tag{F3}$$

As before, the fixed value of $x(n)$ then determines n at fixed $\epsilon^{(q)}$, and we use the x_j to get the energy grid points z_j. The results in the main text are based on use of this second (exponential) grid of Eq. (F3).

Appendix G: Details of KS-DFT and GW computations

We have performed DFT simulations to obtain the single particle wave functions and energies employed as input to the GW calculations reported in the main text.

The plane-wave, non-local, norm-conserving pseudopotential, supercell approach was employed as implemented in the Quantum Espresso software application[53].

To study cystalline Si, we employed the local density approximation (LDA) for exchange and correlation as parameterized by Perdew and Zunger[54]. The norm-conserving pseudopotential for Si was generated with the valence configuration of $3s^2 3p^2 3d^0$ with cutoff radii of 1.75, 1.93, and 2.07 a.u. for s, p, and d channels, respectively. The plane wave cutoff was taken to be 25 Ry, and the lattice parameter was set to the experimental value of 5.43 Å.

To study cystalline MgO, the GGA-PBE exchange-correlation functional[55] was employed. Both Mg and O were represented by norm-conserving pseudopotentials generated with valence configuration $3s^2$ and $2s^2 2p^4 3d^0 4f^0$ for Mg and O, respectively. The plane wave cutoff was taken to be 50 Ry, and the lattice parameter was set to 8.42 Å.

To generate ϵ_∞ and the COHSEX band gap for crystalline Si and MgO, we sampled the Γ-point of the BZ in a 16 atom supercell for both cases. The reference G_0W_0 prediction of the band gap of Si was obtained using a 4×4×4 sampling of the primitive cell – equivalent to a 2×2×2 sampling of the 16 atom supercell. The total number of bands in the 16 atom supercell was taken to be 399 and 433 for Si and MgO, respectively. For the CTSP-W results, $\{N_{v_w} = 1, N_{c_w} = 4\}$ and $\{N_{v_w} = 1, N_{c_w} = 4\}$ was employed to treat both MgO and Si.

To create the data on computational load versus the number of atoms (Fig. 8 of the main text), we studied Si with the following k-point sampling and bands: 52 bands and 8 k-points for the 2-atom cell, 104 bands with 4 k-points for the 4-atom cell, 208 bands with 2 k-points for the 8-atom cell, and 416 bands with 1 k-point for 16-atom cell. For the CTSP-W method, $\{N_{v_w} = 1, N_{c_w} = 5\}$ were employed for all simulations.

To study crystalline Al, we employed the LDA for exchange and correlation as parameterized by Perdew and Zunger[54]. The plane wave cutoff was taken to be 50 Ry, and the lattice parameter was set to 3.99 Å. To obtain $P_{0,0}$, we employed a 16 atom supercell, sampled 2 k-points and included a total 400 bands. Gaussian smearing was used to represent the occupation numbers with $\beta^{-1} = 0.03$ Ry. For the CTSP-W results, $\{N_{v_w} = 1, N_{c_w} = 7\}$ was employed in all cases.

Appendix H: Hermite-Gauss-Laguerre Quadrature

The nodes and weights for the Hermite-Gauss-Laguerre (HGL) quadrature described in the main text can be obtained by employing the matlab functions provided below:

```
function [x,w]=GLquad(n)
% function [x,w]=GLagIntP(n)
% Gauss-Laguerre integration: return nodes x
% and weights w for a
% quadrature grid with n points

% This is basically the Golub-Welsch method
J=diag(1:2:2*n-1)+diag(1:n-1,1)+diag(1:n-1,-1);
[v,l]=eig(J);
[x,ix]=sort(diag(l));
w=v(1,ix)'.^2;
return

function [xmat,wmat] = myweightquad(n)
%function [xmat,wmat] = myweightquad(n)
% Return all nodes (xmat) and weights (wmat)
% for quadratures up to % n points for weight
% w(x)=exp(-x-x^2/2).  These are organized in
% matrices. xmat are the nodes and wmat
% are the weights.  Each column is for a
% quadrature size going from
% 1 to n (left to right).  Thus the lower
% triangle is padded with zeros.

% Figure out number of grid points
% so that the biggest moment (2n)
% is well converged.  We do
% Gauss-Laguerre quadrature to
% do these integrals over the weights!
Iold = 0;
for nx=round(10.^[1:.2:7])
    [xq,wq] = \rrGLquad(nx);
    weight = exp(-xq.^2/2);
    I = sum(wq.*weight.*xq.^(2*n));
    if Iold>0
        err = (I-Iold)/I;
        if abs(err)<1e-14
            break
        end
    else
    end
    Iold = I;
end

% Build polynomials as we go
% and figure out the recursion
% relation coefficients as we go
p = zeros(length(xq),n+1);
p(:,1) = 1;
a = zeros(n,1);
b = zeros(n,1);
for j=1:n
    xpp = sum(wq.*xq.*weight.*p(:,j).^2);
    pp = sum(wq.*weight.*p(:,j).^2);
    a(j) = xpp/pp;
    if j>1
        ppm1 = sum(wq.*weight.*p(:,j-1).^2);
        b(j) = pp/ppm1;
    end
    if j>1
        p(:,j+1) = ...
```

```
            (xq-a(j)).*p(:,j)-b(j)*p(:,j-1);
    else
        p(:,j+1) = (xq-a(j)).*p(:,j);
    end
end

% Prepare for Golub-Welsch
b = b(2:end);
b = sqrt(b);
mu0 = sum(wq.*weight);

% Build Golub-Welsch J matrix,
% eigen decompose it, and get weights and
% nodes for each value of j=1,...,n
% (i.e., all weights and nodes for
% quadratures up to size n)
J = diag(a) + diag(b,1) + diag(b,-1);
xmat = zeros(n,n);
wmat = zeros(n,n);
for j=1:n
    Jcut = J(1:j,1:j);
    [v,d] = eig(Jcut);
    d = diag(d);
    [~,is] = sort(d);
    d = d(is);
    v = v(:,is);
    x = d;
    w = v(1,:).^2*mu0;
    w = w';
    xmat(:,j) = [x' zeros(1,n-j)]';
    wmat(:,j) = [w' zeros(1,n-j)]';
end

return
```

* sohrab.ismail-beigi@yale.edu

[1] P. Hohenberg and W. Kohn, Physical Review **136**, B864 (1964).
[2] W. Kohn and L. J. Sham, Physical Review **140**, A1133 (1965).
[3] J. P. Perdew and A. Zunger, Physical Review B **23**, 5048 (1981).
[4] J. P. Perdew, J. A. Chevary, S. H. Vosko, K. A. Jackson, M. R. Pederson, D. J. Singh, and C. Fiolhais, Physical Review B **46**, 6671 (1992).
[5] J. P. Perdew, R. G. Parr, M. Levy, and J. L. Balduz, Physical Review Letters **49**, 1691 (1982).
[6] S. Lundqvist and N. H. March, *Theory of the Inhomogeneous Electron Gas* (Springer, 2013).
[7] V. I. Anisimov, F. Aryasetiawan, and A. I. Lichtenstein, Journal of Physics: Condensed Matter **9**, 767 (1997).
[8] L. Hedin, Physical Review **139**, A796 (1965).
[9] M. S. Hybertsen and S. G. Louie, Physical Review B **34**, 5390 (1986).
[10] F. Aryasetiawan and O. Gunnarsson, Reports on Progress in Physics **61**, 237 (1998).
[11] G. Onida, L. Reining, and A. Rubio, Reviews of Modern Physics **74**, 601 (2002).
[12] H. F. Wilson, F. Gygi, and G. Galli, Physical Review B (Condensed Matter and Materials Physics) **78**, 113303 (2008).
[13] H. F. Wilson, D. Lu, F. Gygi, and G. Galli, Physical Review B **79**, 245106 (2009).
[14] D. Rocca, D. Lu, and G. Galli, The Journal of Chemical Physics **133**, 164109 (2010).
[15] D. Lu, F. Gygi, and G. Galli, Physical Review Letters **100**, 147601 (2008).
[16] F. Giustino, M. L. Cohen, and S. G. Louie, Physical Review B **81**, 115105 (2010).
[17] P. Umari, G. Stenuit, and S. Baroni, Physical Review B **81**, 115104 (2010).
[18] M. Govoni and G. Galli, Journal of Chemical Theory and Computation **11**, 2680 (2015).
[19] F. Bruneval and X. Gonze, Physical Review B **78**, 085125 (2008).
[20] J. A. Berger, L. Reining, and F. Sottile, Physical Review B **82**, 041103 (2010).
[21] W. Gao, W. Xia, X. Gao, and P. Zhang, Scientific Reports **6**, 36849 (2016).
[22] D. Foerster, P. Koval, and D. Sanchez-Portal, The Journal of Chemical Physics **135**, 074105 (2011).
[23] P. Liu, M. Kaltak, J. Klimeš, and G. Kresse, Physical Review B **94**, 165109 (2016).
[24] D. Neuhauser, Y. Gao, C. Arntsen, C. Karshenas, E. Rabani, and R. Baer, Physical Review Letters **113**, 076402 (2014).
[25] J. C. Light and T. Carrington, "Discrete-variable representations and their utilization," in *Advances in Chemical Physics* (John Wiley and Sons, Inc., 2007) pp. 263–310.
[26] J. Deslippe, G. Samsonidze, D. A. Strubbe, M. Jain, M. L. Cohen, and S. G. Louie, Computer Physics Communications **183**, 1269 (2012).
[27] M. Kim, S. Mandal, E. Mikida, K. Chandrasekar, E. Bohm, N. Jain, Q. Li, R. Kanakagiri, G. J. Martyna, L. Kale, and S. Ismail-Beigi, Computer Physics Communications (2019), https://doi.org/10.1016/j.cpc.2019.05.020.
[28] J. W. Negele and H. Orland, *Quantum Many-particle Systems* (Westview Press, 1998).
[29] L. Hedin and S. Lundqvist, in *Advances in Research and Applications*, Vol. Volume 23 (Academic Press, 1970) pp. 1–181.
[30] M. M. Rieger, L. Steinbeck, I. D. White, H. N. Rojas, and R. W. Godby, Computer Physics Communications **117**, 211 (1999).
[31] M. Kaltak, J. Klimeš, and G. Kresse, Journal of Chemical Theory and Computation **10**, 2498 (2014).
[32] D. Baye and P.-H. Heenen, Journal of Physics A: Mathematical and General **19**, 2041 (1986).
[33] R. A. Friesner, The Journal of Chemical Physics **85**, 1462 (1986).
[34] "$s^{-1} = \int_0^\infty dt \exp(-st) = \zeta \int_0^\infty dt \exp(-\zeta st)$,".
[35] M. S. Hybertsen and S. G. Louie, Physical Review B **34**, 5390 (1986).
[36] J. Deslippe, G. Samsonidze, D. A. Strubbe, M. Jain, M. L.

Cohen, and S. G. Louie, Computer Physics Communications **183**, 1269 (2012).
[37] J. Williamson, *Lebesgue Integration: Dover Books on Mathematics* (Dover, 2014).
[38] M. Abramowitz and Stegun, eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 10th ed. (U.S. Government Printing Office, 1972).
[39] C. L. Fu and K. M. Ho, Physical Review B **28**, 5480 (1983).
[40] R. J. Needs, R. M. Martin, and O. H. Nielsen, Physical Review B **33**, 3778 (1986).
[41] M. J. Gillan, Journal of Physics: Condensed Matter **1**, 689 (1989).
[42] To avoid excessive memory use, one can compute the large matrix $\Sigma(\omega)_{r,r'}$ for a fixed ω and then compute and only store the much smaller number of desired matrix elements $< n|\Sigma(\omega)|n' >$ before moving to the next ω value.
[43] A. Gil, J. Segura, and N. M. Temme, *Numerical Methods for Special Functions* (SIAM, 2007).
[44] For systems with a small number of atoms, the CTSP-W runs slightly slower per operation, $< 2\times$, due to the inefficient caching and pipelining of our untuned software.
[45] S. Zhang, C. I. Pelligra, G. Keskar, J. Jiang, P. W. Majewski, A. D. Taylor, S. Ismail-Beigi, L. D. Pfefferle, and C. O. Osuji, Advanced Materials **24**, 82 (2012).
[46] https://bluewaters.ncsa.illinois.edu/.
[47] F. Bruneval and X. Gonze, Physical Review Letters **78**, 085125 (2008).
[48] D. Foerster, P. Koval, and D. Sanchez-Portal, Journal of Chemical Physics **135**, 074105 (2011).
[49] P. Liu, M. Kaltak, J. Klimes, and G. Kresse, Physical Review B **94**, 165109 (2016).
[50] K. S. D. Beach, R. J. Gooding, and F. Marsiglio, Physical Review B **61**, 5147 (2000).
[51] S. Goedecker and L. Colombo, Physical Review Letters **73**, 122 (1994).
[52] S. Goedecker, Reviews of Modern Physics **71**, 1085 (1999).
[53] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, and I. Dabo, Journal of Physics: Condensed Matter **21**, 395502 (2009).
[54] J. P. Perdew and A. Zunger, Physical Review B **23**, 5048 (1981).
[55] J. P. Perdew, K. Burke, and M. Ernzerhof, Physical Review Letters **77**, 3865 (1996).