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 Technology, Westchester, New York 10598, USA*

(Received 23 April 2019; published 23 January 2020)

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.

DOI: 10.1103/PhysRevB.101.035139

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 nondeterministic 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 timescales inherent in the propagators. Second, the timescale 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.

*sohrab.ismail-beigi@yale.edu

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 $\lesssim$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]:

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

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

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

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, $V_{\rm ion}$ 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

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

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

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

The ψ_n and E_n are obtained as eigenstates of a noninteracting 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 nonessential 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

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

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 $\beta=1/k_BT$ and the chemical potential μ.] We have employed a general, compact notation valid for collinear and noncollinear spin calculations. For collinear spin, nonzero 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 (noncollinear) 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)

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

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) and (4) over k (i.e., Brillouin zone sampling). We note that current numerical methods for computing P based on the sum-over-states formulas, 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,

$$W(\omega)_{r,r'} = V_{r,r'} + \sum_p \frac{2\omega_p B^p_{r,r'}}{\omega^2 - \omega^2_p} = V_{r,r'} + \sum_p B^p_{r,r'}\left(\frac{1}{\omega - \omega_p} - \frac{1}{\omega + \omega_p}\right). \quad (5)$$

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

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

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

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

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), and (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

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

as can be discerned from Eqs. (3), (4), (6), and (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. [23,30,31].

In more detail, while the frequency representation has advantages, the evaluation of Eq. (8) scales as $\mathcal{O}(N_a N_b N^2_r)$ 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 timescales 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, cf. 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 nontrivial $h(\tau;\zeta)$, 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

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

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 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 timescales, $\sim\hbar/E^{(\mathrm{bw})}$ and $\sim\hbar/E^{(\mathrm{gap})}$, in $\tilde{\chi}(\tau\zeta;0)_{r,r'}$ are independent of N. Therefore, barring nonanalytic 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 timescale 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

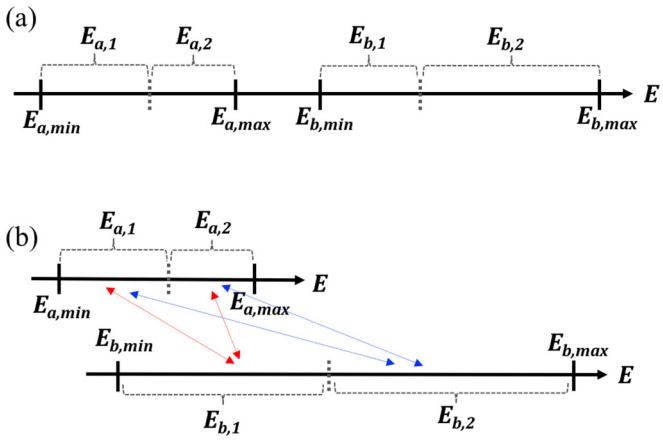

FIG. 1. An example of the proposed energy windowing approach with $N_{a_w} = N_{b_w} = 2$ (a) For gapped systems, the energy ranges of $\{a_i\}$ and $\{b_j\}$ do not overlap. (b) For systems with overlapping energy ranges, energy window pairs arise both with energy crossings, red arrows, and without,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

will be generalized to treat gapless systems and response functions at finite frequencies through use of imaginary ζ and nontrivial $h(\tau;\zeta)$.

In order to tame the multiple timescales 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),

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

where for each window pair (l, m),

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

Here, $\mathcal{L}$ and $\mathcal{M}$ represent the sets of integer indices of the single particle states that contribute to the lth A-type and mth 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

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

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 lth and mth 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 timescale 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 timescale 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) and (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 [9] or by regularizing $1/e_{ij}$, e.g., replacing $1/e_{ij}$ by $e_{ij}/(e_{ij}^2 + |\zeta|^{-2})$ [26].

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) and (10) and choosing ζ to be a pure imaginary number,

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

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

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

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) [35]. As long as the *number* of special regions/points is independent of systems size, the scaling of the method remains $\mathcal{O}(N^3)$.

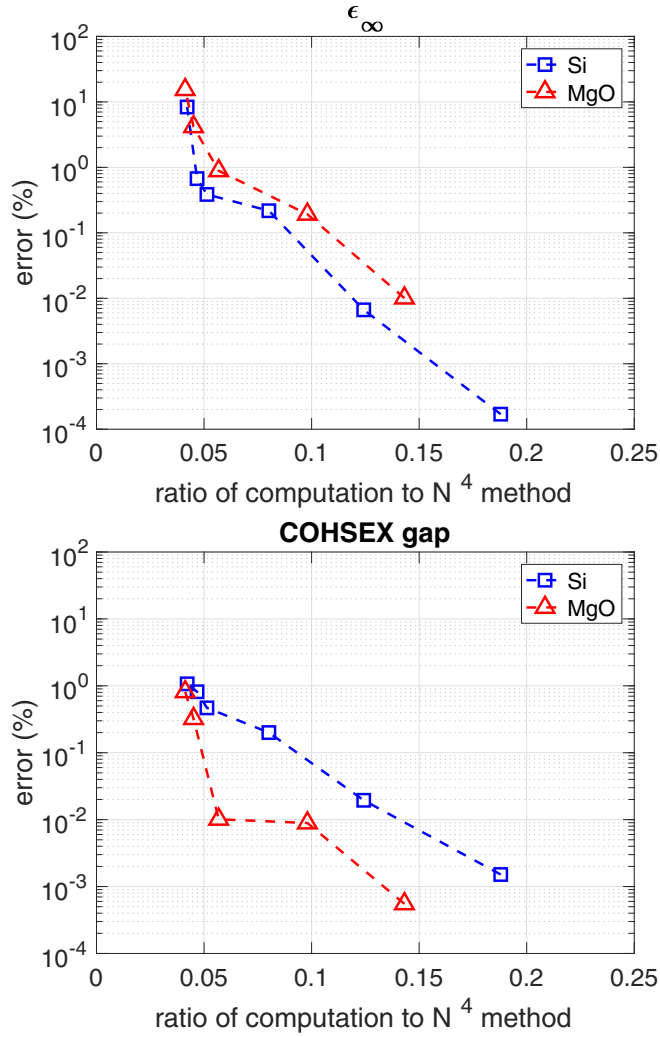

FIG. 2. Numerical error vs computational savings for our cubic scaling formalism, CTSP-W, compared to the standard quartic *GW* formulation for bulk Si and MgO modeled in a 16 atom supercell. The CTSP-W error decreases and computational work increases as the integration error is decreased (i.e., the 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 [ϵ_∞(MgO) = 6.35, ϵ_∞(Si) = 64.85]. (Bottom) Error in the COHSEX band gap [$E^{(\mathrm{gap,COHSEX})}$(MgO) = 7.56 eV, $E^{(\mathrm{gap,COHSEX})}$(Si) = 1.92 eV].

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. II C 2 b and logarithm-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. Nonessential 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) and (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

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

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, $E^{(\mathrm{gap})}_{lm} = E^{(c,\min)}_m - E^{(v,\max)}_l$, energy scale, ζ_{lm}, and bandwidth, $E^{(\mathrm{bw})}_{lm} = E^{(c,\max)}_m - E^{(v,\min)}_l$. [To connect directly to the formalism of Eqs. (14) and (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 mth conduction and lth 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 $E^{(\mathrm{bw})}_{lm}/E^{(\mathrm{gap})}_{lm}$, 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 [36] 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

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

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

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

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,

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

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) [35].

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 [37–39]. 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

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

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, $E_c \geqslant E_v \ \forall\ c, v$). 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

$$\begin{aligned} P_{r,r'}^{lm} &= -2\zeta_{lm} \sum_{u=1}^{N_{lm}^{(\mathrm{GL})}} w_u\, e^{-\tau_u(\zeta_{lm}E_{lm}^{(\mathrm{gap})}-1)} \\ &\quad \times \left[S_{r',r}^{lmu} Q_{r,r'}^{lmu} - T_{r',r}^{lmu} Z_{r,r'}^{lmu} \right], \end{aligned}$$

where

$$\begin{aligned} S_{r,r'}^{lmu} &= \sum_{\{v\in\mathcal{L}\}} f(E_v) e^{-\tau_u \zeta_{lm} \Delta E_{vl}} \psi_{r,v} \psi_{r',v}^*, \\ Q_{r,r'}^{lmu} &= \sum_{\{c\in\mathcal{M}\}} e^{-\tau_u \zeta_{lm} \Delta E_m} \psi_{r,c} \psi_{r',c}^*, \\ T_{r,r'}^{lmu} &= \sum_{\{v\in\mathcal{L}\}} e^{-\tau_u \zeta_{lm} \Delta E_{vl}} \psi_{r,v} \psi_{r',v}^*, \\ Z_{r,r'}^{lmu} &= \sum_{\{c\in\mathcal{M}\}} f(E_c) e^{-\tau_u \zeta_{lm} \Delta E_m} \psi_{r,c} \psi_{r',c}^* . \end{aligned}$$

The five-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 $E_{lm}^{(\mathrm{gap})} = 1/\beta$ 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 section 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

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

[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[1] for real frequencies ω in such a way that analytical continuation is not required.

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

The dynamic self-energy,

$$\begin{aligned}
\Sigma(\omega)_{r,r'} &= \Sigma^{(+)}(\omega)_{r,r'} + \Sigma^{(-)}(\omega)_{r,r'},\\
\Sigma^{(+)}(\omega)_{r,r'} &= \sum_{p,v} \frac{B^p_{r,r'}[\psi_{rv}\psi^*_{r'v}]}{\omega - E_v + \omega_p},\\
\Sigma^{(-)}(\omega)_{r,r'} &= \sum_{p,c} \frac{B^p_{r,r'}[\psi_{rc}\psi^*_{r'c}]}{\omega - E_c - \omega_p},
\end{aligned} \quad (29)$$

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,

$$\Sigma^{(+)}(\omega)_{r,r'} = \sum_{m=1}^{N^{(+)}_{v_w}} \sum_{l=1}^{N^{(+)}_{p_w}} \Sigma^{(+)}\left(\omega; \zeta^{(+)}_{lm}\right)^{lm}_{r,r'}, \quad (30)$$

$$\Sigma^{(-)}(\omega)_{r,r'} = \sum_{m=1}^{N^{(-)}_{c_w}} \sum_{l=1}^{N^{(-)}_{p_w}} \Sigma^{(-)}\left(\omega; \zeta^{(-)}_{lm}\right)^{lm}_{r,r'}, \quad (31)$$

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) and (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),

$$\begin{aligned}
\Sigma^{(\pm)}(\omega)^{lm}_{r,r'} = &\sum_{\{p\in\mathcal{L}^{(\pm)}\}} \sum_{\{n\in\mathcal{M}^{(\pm)}\}} B^p_{r,r'}[\psi_{rn}\psi^*_{r'n}]\\
&\times F(\omega - E_n \pm \omega_p; \zeta)\,.
\end{aligned} \quad (32)$$

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 (1) to these zero contributions for small x [i.e., setting $F(x;\zeta) = 0$ for small x] or (2) 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

$$\begin{aligned}
h(\tau;\gamma) &= \gamma \exp(-\tau - \tau^2/2),\\
F(x;\gamma) &= \gamma Im\left\{\sqrt{\frac{\pi}{2}} e^{-\frac{(x\gamma+i)^2}{2}}\left[1 + i\mathrm{erfi}\left(\frac{x\gamma + i}{\sqrt{2}}\right)\right]\right\}.
\end{aligned} \quad (33)$$

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 [40] 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

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

[1]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 $\langle n|\Sigma(\omega)|n'\rangle$ before moving to the next ω value.

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

$$\Sigma^{(\pm)}(\omega)^{lm}_{r,r'} = \gamma \sum_{u=1}^{N^{(\tau,\mathrm{HGL},\pm)}_{lm}} w_u \left\{ \left[\sum_{\{p\in\mathcal{L}^{(\pm)}\}} B^p_{r,r'} \sin(\pm\tau_u\omega_p\gamma) \right] \times \left[\sum_{\{n\in\mathcal{M}^{(\pm)}\}} \psi_{rn}\psi^*_{r'n} \cos(\tau_u(\omega-\epsilon_n)\gamma) \right] + \left[\sum_{\{p\in\mathcal{L}\}} B^p_{r,r'} \cos(\pm\tau_u\omega_p\gamma) \right] \times \left[\sum_{\{n\in\mathcal{M}^{(\pm)}\}} \psi_{rn}\psi^*_{r'n} \sin(\tau_u(\omega-\epsilon_n)\gamma) \right] \right\}, \quad (35)$$

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

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

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,

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

Here, the mth plasmon mode window spans the energy range $[\omega_m^{(p,\min)}, \omega_m^{(p,\max)}]$, the lth 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 $x_{lm} = \gamma E^{(\mathrm{bw})}_{lm}$ where $\Delta E^{(\mathrm{bw})}_{lm}$ 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 $E_c - E_v \geqslant 0$ 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 crystalline 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 and 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 three-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.