Efficient Full-frequency GW Calculations using a Lanczos Method

Weiwei Gao,[1] Zhao Tang,[2] Jijun Zhao,[3,*] and James R. Chelikowsky[2,4,5,†]

[1] *Key Laboratory of Materials Modification by Laser, Ion and Electron Beams, Ministry of Education, Dalian University of Technology, Dalian 116024, China*

[2] *Center for Computational Materials, Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX 78712*

[3] *Key Laboratory of Atomic and Subatomic Structure and Quantum Control (Ministry of Education), Guangdong Basic Research Center of Excellence for Structure and Fundamental Interactions of Matter, School of Physics, South China Normal University, Guangzhou 510006, China*

[4] *Department of Physics, The University of Texas at Austin, Austin, TX 78712*

[5] *McKetta Department of Chemical Engineering, The University of Texas at Austin, Austin, TX 78712*

The GW approximation is widely used for reliable and accurate modeling of single-particle excitations. It also serves as a starting point for many theoretical methods, such as its use in the Bethe-Salpeter equation (BSE) and dynamical mean-field theory. However, full-frequency GW calculations for large systems with hundreds of atoms remain computationally challenging, even after years of efforts to reduce the prefactor and improve scaling. We propose a method that reformulates the correlation part of the GW self-energy as a resolvent of a Hermitian matrix, which can be efficiently and accurately computed using the standard Lanczos method. This method enables full-frequency GW calculations of material systems with a few hundred atoms on a single computing workstation. We further demonstrate the efficiency of the method by calculating the defect-state energies of silicon quantum dots with diameters up to 4 nm and nearly 2,000 silicon atoms using only 20 computational nodes.

As a first-principles approach based on many-body perturbation theory, the GW approximation has been successfully applied to accurately compute quasiparticle excitation in weakly and moderately correlated materials [1–3]. The approximation also plays an essential role in the first-principles calculations of excitonic effects using the GW+BSE approach [4, 5] and is used in conjunction with other methods [6–11]. The computational scaling of different implementations of GW approximation ranges from $O(N)$ to $O(N^6)$ and typically has a much larger pre-factor compared to density functional theory (DFT) calculations with semi-local exchange-correlation functionals [12, 13].

During the last decade, different formulations and algorithms have been proposed and implemented for accelerating GW calculations to meet the challenge of modeling large and complex materials [12, 13]. A few seminal papers have demonstrated GW calculations of quasiparticle energies of large systems with the number of atoms ranging from 1,000 to around 2,700 [14–18]. These large-scale GW calculations rely on well-crafted numerical optimization and large computation resources, which are of limited accessibility. Even with notable advancements, GW calculations for systems with a few hundred atoms, which are typically required for computationally studying a point defect in solids or small quantum dots, cannot be performed routinely. Owing to the significant expense of data curation, GW calculations are rarely used in wide-reaching data-driven research, such as constructing large databases of material properties and training supervised machine-learning models.

In GW calculations, one of the most computationally expensive steps is calculating the frequency-dependent screened Coulomb potential W. In many implementations, the irreducible polarizability function and then the inverse dielectric function are calculated to compute W [19, 20]. Such a procedure deals with the frequency dependence of W using approximations like plasmon-pole models or numerical tools such as contour deformation or analytical continuations [12, 21]. Alternatively, another approach computes the reducible polarizability and W by solving the Casida equation derived in linear-response time-dependent density functional theory [22–26]. Once all the eigenvalues and eigenvectors of the Caisda equation are solved, the frequency-dependent W and GW quasiparticle self-energies can be written and computed in a closed form [23–25]. While this approach is formally simple and works efficiently for small systems with less than 40 atoms, it becomes numerically intractable for large systems due to the high cost of solving the Casida equation.

Here, we propose a method that avoids solving the Casida equation while still allowing us to perform full-frequency GW calculations analytically and efficiently. To better illustrate the concept, we discuss our method applied to finite systems, for which real-valued wave functions and simplified notations can be used. Initially, we perform DFT calculations to obtain Kohn-Sham orbitals $|\phi_m\rangle$ and their corresponding energies ϵ_m, which are used as an initial approximation for quasiparticle wave functions and energies, respectively. Next, the Casida equation can be constructed [22–24, 26, 27]

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ -\mathbf{B} & -\mathbf{A} \end{pmatrix} \begin{pmatrix} X^s \\ Y^s \end{pmatrix} = \begin{pmatrix} X^s \\ Y^s \end{pmatrix} \Omega_s, \tag{1}$$

The dimension of matrices $\mathbf{A}$ and $\mathbf{B}$ is N_{vc}^2, where $N_{vc} = N_v \cdot N_c$ scales as $O(N^2)$ with respect to system size N. Here N_v and N_c represent the number of occupied and empty orbitals, respectively. With the random-phase approximation (RPA) used in the GW approximation, the matrix elements of $\mathbf{A}$ and $\mathbf{B}$ are given as

$$\mathbf{A}_{vc,v'c'} = (\epsilon_c - \epsilon_v)\delta_{vv'}\delta_{cc'} + (vc|v'c') \quad (2)$$

$$\mathbf{B}_{vc,v'c'} = (vc|c'v') = \int\int \frac{\phi_v(\mathbf{r})\phi_c(\mathbf{r})\phi_{c'}(\mathbf{r}')\phi_{v'}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d^3r d^3r' \quad (3)$$

We use indices v and v' for occupied states, c and c' for empty states, and k, l, n, and m for general orbitals, respectively. For finite systems, the Casida equation can be reformulated as a smaller eigenvalue problem [23, 24]

$$\mathbf{C}Z^s = Z^s\Omega_s^2 \quad (4)$$

where $\mathbf{C} = (\mathbf{A}-\mathbf{B})^{1/2}(\mathbf{A}+\mathbf{B})(\mathbf{A}-\mathbf{B})^{1/2}$ is a symmetric matrix of dimension N_{vc}^2. After solving Eq. 4 for the eigenpairs (Z^s, Ω_s^2) of $\mathbf{C}$, one can compute the full-frequency GW self-energy

$$\langle\phi_m|\Sigma^{\mathbf{GW}}(\omega)|\phi_m\rangle = \Sigma_{mm}^{\mathrm{ex}} + \Sigma_{mm}^{\mathrm{corr}}(\omega), \quad (5)$$

$$\Sigma_{mm}^{\mathrm{ex}} = -\sum_v (vm|vm), \quad (6)$$

$$\Sigma_{mm}^{\mathrm{corr}}(\omega) = \sum_n^{N_v+N_c}\sum_s^{N_{vc}} \frac{W_{nm}^s W_{nm}^s}{\omega - \epsilon_n + \eta_n(\Omega_s - i\delta)} \quad (7)$$

where η_n is 1 for occupied orbitals and -1 for empty orbitals, and δ is a positive infinitesimal number to avoid singularity. The matrix elements W_{nm}^s are

$$W_{nm}^s = \sum_{vc}(nm|vc)\sqrt{\frac{\epsilon_c - \epsilon_v}{\Omega_s}} Z_{vc}^s. \quad (8)$$

The exchange part of the self-energy $\Sigma_{mm}^{\mathrm{ex}}$ is independent of frequency and relatively easy to compute, while the correlation part $\Sigma_{mm}^{\mathbf{corr}}(\omega)$ includes the frequency-dependent screening effects of dielectric responses. The poles of frequency-dependent screened effects can be determined by the eigenvalues of $\mathbf{C}$. As a result, the most expensive step of the aforementioned method is diagonalizing the Casida equation, as the computational cost scales as $O(N^6)$. To make further progress, we intend to avoid this costly step by defining a vector $|P_{nm}\rangle$ of dimension N_{vc}, which has elements given by $(P_{nm})_{vc} = (nm|vc)(\epsilon_c - \epsilon_v)^{1/2}$. Then W_{nm}^s becomes

$$W_{nm}^s = \langle P_{nm}|Z^s\rangle\Omega_s^{-\frac{1}{2}}. \quad (9)$$

$\Sigma_{mm}^{\mathrm{corr}}$ can be rewritten as

$$\Sigma_{mm}^{\mathrm{corr}}(\omega) = \sum_{n=1}^{N_v+N_c} \Sigma_{mm}^{\mathrm{corr}}(\omega, n), \quad (10)$$

where

$$\Sigma_{mm}^{\mathrm{corr}}(\omega, n) = \frac{1}{z_n}\sum_{s=1}^{N_{vc}}\langle P_{nm}|Z^s\rangle\langle Z^s|P_{nm}\rangle \times \left[\frac{1}{\Omega_s} - \frac{1}{\Omega_s + \eta_n z_n}\right], \quad (11)$$

where $z_n = \omega - \epsilon_n - i\eta_n\delta$.

Examining Eq. 11, we note the formula for $\Sigma_{mm}^{\mathrm{corr}}(\omega, n)$ is similar to a general resolvent matrix element of the form $\sum_k \langle\star|k\rangle\langle k|\star\rangle/(z-\lambda_k) = \langle\star|1/(z-\mathbf{H})|\star\rangle$, where $\mathbf{H}$ is a general Hermitian matrix with eigenvalues λ_k and eigenvectors $|k\rangle$, z is a complex number, and $|\star\rangle$ is a ket. Motivated by this observation, we reformulate Eq. 11 as the resolvent of a symmetric matrix $\mathbf{D}$

$$\Sigma_{mm}^{\mathrm{corr}}(\omega, n) = \frac{1}{z_n}\langle P_{nm}|\frac{1}{\mathbf{D}} - \frac{1}{\mathbf{D} + \eta_n z_n}|P_{nm}\rangle. \quad (12)$$

Matrix $\mathbf{D}$ satisfies $\mathbf{D}^2 = \mathbf{C}$ and its eigenvalues are the square root of those of matrix C, i.e., $\mathbf{D}Z^s = Z^s\Omega_s$. We use a g-th degree polynomial function p_g to fit the square root function $p_g(x) = \sum_{k=0}^g a_k x^k \approx \sqrt{x}$ within $x \in [\min\Omega_s^2, \max\Omega_s^2]$, which is the range between minimum and maximum eigenvalues of matrix $\mathbf{C}$. Accordingly, $\mathbf{D}$ can be approximated by $\mathbf{D} = p_g(\mathbf{C}) + \Delta_g \approx p_g(\mathbf{C}) = a_0\mathbf{I} + \sum_{k=1}^g a_k\mathbf{C}^k$, where $\mathbf{I}$ is an identity matrix and the fitting error Δ_g can be controlled via the degree g of the polynomial function and fitting procedures. More discussions on Eq. 10 to Eq. 12 are presented in Section 1 of the Supplemental Material [28].

Given Eq. 12 and matrix $\mathbf{D}$, the Lanczos method can then be applied to efficiently compute the resolvent of matrix $\mathbf{D}$, which is an important step in calculating $\Sigma_{mm}^{\mathrm{corr}}(\omega, n)$. In the calculation of $\Sigma_{mm}^{\mathrm{corr}}(\omega)$, we prepare $|P_{nm}\rangle$ for each state n in the summation of Eq. 10, where $|P_{nm}\rangle$ is used as the starting vector for the Lanczos tridiagonalization procedure of the symmetric matrix $\mathbf{D}$. With L steps of Lanczos iterations, one can construct a tridiagonal matrix $\mathbf{D}_L$ with dimension L in the following form:

$$\mathbf{D}_L = \begin{pmatrix} a_0 & b_1 & 0 & \dots & 0 & 0 \\ b_1 & a_1 & b_2 & \dots & 0 & 0 \\ 0 & b_2 & a_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{L-1} & b_L \\ 0 & 0 & 0 & \dots & b_L & a_L \end{pmatrix}. \quad (13)$$

Once the tridiagonal matrix $\mathbf{D}_L$ is obtained, a resolvent matrix element (such as Eq. 12) can be computed using

the continuous fraction

$$\langle P_{nm}|\frac{1}{z-\mathbf{D}}|P_{nm}\rangle = \cfrac{1}{z-a_0-\cfrac{b_1^2}{z-a_1-\cfrac{b_2^2}{a_2-\ldots}}}, \qquad (14)$$

which is also known as the Haydock method [29]. The computation of Eq. 14 is efficient, and one can easily calculate the quasiparticle energies for a series of frequencies by varying z in Eq. 14. When applied to eigenvalue problems, the Lanczos algorithm can lead to ghost eigenvalues. However, applying the Lanczos method to calculate resolvent is free of such a numerical problem [30].

There are several advantages to using the Lanczos method for computing $\Sigma_{mm}^{\rm corr}$. Solving the eigenvalue problem of the Casida matrix $\mathbf{C}$ is avoided, and the resulting full-frequency GW calculations become more efficient than the conventional method represented by Eq. 7, which explicitly requires the eigenpairs of $\mathbf{C}$. Frequency grids, analytical continuation, and approximations like plasmon-pole models are not required, as the frequency dependence of W and $\Sigma_{mm}^{\rm corr}$ are implicitly treated via Lanczos iterations. Moreover, the method is in principle applicable to any basis sets of wave functions, as our derivation does not rely on any features of specific basis functions.

As a general-purpose algorithm, Lanczos-based methods have been used in computational material science, such as computing Green's function [29], optical absorption spectra with linear-response time-dependent density functional theory [31–35] and Bethe-Salpeter equation [36]. Earlier work [37–39] applied Lanczos methods for solving the Sternheimer equation to obtain frequency-dependent screened Coulomb potential. Recently, several new methods [40–44] have been explored to achieve efficient full-frequency GW calculations. For example, Scott et al. [41] adopted a block Lanczos algorithm to solve an effective Hamiltonian whose eigenvalues systematically approximate the excitation energies of GW theory. Bintrim and Berkelbach [40, 45] proposed a method that does not require integration over the frequency grids. Instead, the GW quasiparticle energies are obtained by solving the eigenvalues of an effective Hamiltonian, which follows the algebraic diagrammatic construction [46]. Compared to these methods, our method does not solve the Sternheimer equation or obtain GW quasiparticle energies from the eigenvalues of an effective Hamiltonian. Instead, the frequency-dependent screened Coulomb potential is found using linear-response TDDFT within the Casida formalism, and the GW quasiparticle energies are computed from a summation of resolvent elements given by Eq. 12.

The accuracy of the Lanczos-based method for the *GW* approximation is checked by calculating the highest occupied and lowest unoccupied molecule orbital (HOMO/LUMO) energies of the GW100 set [47–50], which include 100 small close-shell molecules for benchmarking different implementations of the GW approximation. G_0W_0-level calculations are carried out throughout this work. As studied in previous work, G_0W_0-level calculations depend on the starting point, while quasiparticle self-consistent GW (QSGW) and fully self-consistent GW can alleviate the dependence of calculation results on the starting points [51–54]. Our new Lanczos method is compatible with QSGW [55] because the accelerated steps (i.e., bypassing the diagonalization of the Casida equation and using the Lanczos method to compute the correlation part of the self-energy) do not interfere with the self-consistent iterations. The Lanczos method only requires the updated quasiparticle energies and wave functions of the current iteration to start the next iteration of GW calculation. A real-space-based pseudo-potential DFT code `PARSEC` is used in our implementation to efficiently obtain Kohn-Sham orbitals for large finite systems [56, 57]. More details of our computations are presented in Section 2 of the Supplemental Material [28, 58, 59]. Fig. 1 (a) shows the results computed with the Lanczos method and the reference agrees well for all GW100 molecules. The mean

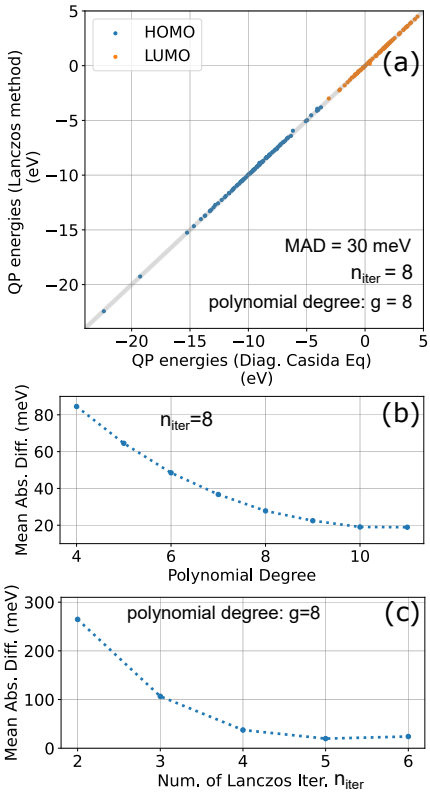

Figure 1. (a). Comparison between the HOMO and LUMO energies calculated with the reference method and the Lanczos method. (b). The mean absolute differences (MAD) between the quasiparticle energies of one hundred small molecules calculated with the reference approach and the Lanczos method with different degrees of polynomial functions; (c). MAD of quasiparticle energies calculated with the reference method and the Lanczos method using different numbers of iterations $N_{\rm iter}$;

average difference (MAD) between the results calculated using the reference method, which finds the eigenpairs of the Casida equation explicitly, and the Lanczos method is within 20 meV. Our tests also show the Lanczos-based formalism converges fast to the degree g of polynomial p_g and the number of Lanczos iterations $N_{\rm iter}$. As shown in Fig. 1 (b) and (c), the MAD is below 30 meV when the polynomial degree $g \geq 8$ and $N_{\rm iter} \geq 5$.

The computationally expensive steps in our method are: (1) calculating electron-repulsion integrals $(kl|nm)$ and (2) calculating the matrix-vector product $\mathbf{D}|\star\rangle$, where $|\star\rangle$ is a general vector. One can use suitable low-rank approximation methods, such as resolution-of-identity or density-fitting methods [60–62], to speed up these computations. Density-fitting methods exploit the rank deficiency of orbital pair products $\phi_n(\mathbf{r})\phi_m(\mathbf{r})$ and use a set of auxiliary basis functions $\zeta_\mu(\mathbf{r})$ to fit these orbital pairs

$$\phi_n(\mathbf{r})\phi_m(\mathbf{r}) \approx \sum_{\mu=1}^{N_\mu} \zeta_\mu(\mathbf{r}) \mathcal{C}^\mu_{nm} \tag{15}$$

where the required number of auxiliary basis functions N_μ for accurately representing the orbital pairs is expected to be small and scale as $O(N)$ and $\mathcal{C}^\mu_{nm}$ are fitting coefficients. With the approximation given in Eq. 15, one can calculate integrals $(kl|nm)$, which contribute to the elements of $\mathbf{C}$ and $\mathbf{D}$, with the following equations

$$(kl|nm) \approx \sum_{\alpha=1}^{N_\mu}\sum_{\beta=1}^{N_\mu} (\alpha|\beta)\mathcal{C}^\alpha_{kl}\mathcal{C}^\beta_{nm} \tag{16}$$

$$(\alpha|\beta) = \int \frac{\zeta_\alpha(\mathbf{r})\zeta_\beta(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d^3r d^3r'. \tag{17}$$

These methods reduce four-center integrals to two-center integrals and also factorize $\mathbf{C}$ and $\mathbf{D}$ as products of small matrices to accelerate the matrix-vector products. Here we used the interpolative separable density fitting (ISDF) method to efficiently construct low-rank approximations of orbital pairs [13, 62, 63]. Additionally, one can further exploit the point-group symmetries to make the Casida matrix $\mathbf{C}$ block diagonal and simplify the calculation of matrix-vector products [64].

Combined with the ISDF method [13], our method is efficient and enables large-scale G_0W_0 computations with modest computing resources. To demonstrate the efficiency of our method, we performed calculations for hydrogen-passivated silicon clusters. Silicon clusters have attracted research interest as prototypical semiconducting clusters for studying the fundamental physical properties of zero-dimensional systems [65] and their applications in many fields [66–69]. Defects in passivated silicon nanocrystals can introduce mid-gap defect levels as potential sources of photoluminescence [70, 71], while their electronic structures are rarely studied by GW

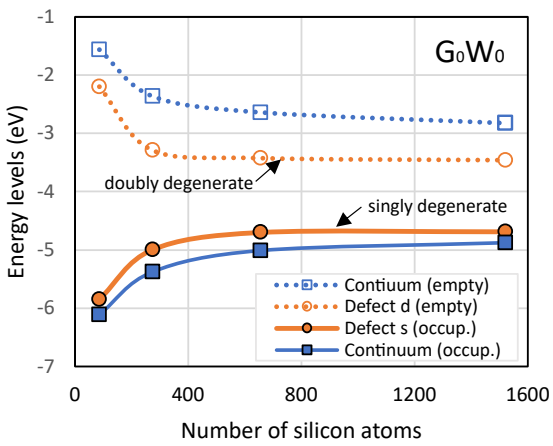

Figure 2. Evolution of "continuum" states and silicon-vacancy defect states in different-sized silicon clusters.

Nanocluster	N_v	$N_c + N_v$	N_μ	$N_{\rm node}$	$t_{\rm wall}$ (hr)
$Si_{86}H_{76}$	210	1730	7000	1	0.02
$Si_{274}H_{172}$	634	5200	20000	2	0.3
$Si_{452}H_{228}$	1018	8200	32000	2	0.7
$Si_{656}H_{300}$	1462	12000	48000	2	1.5
$Si_{1522}H_{524}$	3306	27000	108000	10	8.5
$Si_{1947}H_{604}$	4196	33400	133600	20	9.2

Table I. The running time $t_{\rm wall}$ and the number of compute nodes $N_{\rm node}$ for calculating the GW quasiparticle energy of one quasiparticle state. Each compute node has 64 cores and 4 graphic processing units (GPU). $t_{\rm wall}$ includes the running time for performing the ISDF method and computing $\Sigma^{\mathbf{GW}}(\omega)$ using the Lanczos method. $N_{\mathbf{iter}} = 8$ and $g = 7$ are used for Lanczos iterations and polynomial functions, respectively.

calculations. We computed the defect energy levels of charge-neutral silicon vacancies in silicon clusters of different sizes. The ground state of a silicon vacancy has zero net spin. Different from nano-diamondoids, where surface states are located in the gap, the surface states of silicon nanoclusters are mixed with continuum states and the mid-gap states originate from defects. In the single-particle level, an occupied singlet and a pair of unoccupied doubly degenerate defect states are located inside the gap [72]. As shown in Fig. 2, when the size of silicon clusters increases, the energies of defect states evolve at a similar rate as the continuum states. For the $Si_{1522}H_{524}$ cluster, the band gap of continuum states is around 2.3 eV, still far from the bulk silicon band gap of 1.1 eV. We also computed the HOMO-LUMO gap of non-defective silicon nanocrystals, and our results agree well with previous calculations [37, 73] (see Section 3 in the Supplemental Material [28] for more details).

The main calculation parameters and required computation resources for these calculations are shown in Table I. Notably, only two computing nodes are required

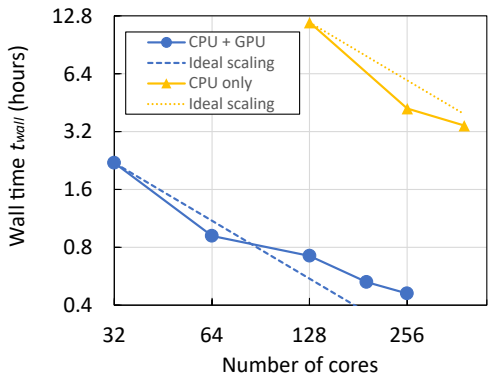

Figure 3. The wall time for computing the GW quasiparticle energy of one state of $Si_{452}H_{228}$ cluster with different numbers of CPU cores and GPUs (Nvidia A100). Every computer node has an EPYC 7763 CPU with 64 cores. For calculations accelerated with GPUs, every 16 CPU cores share 1 GPU processor.

for $Si_{656}H_{300}$. For the largest system, $Si_{1953}H_{604}$ with a diameter of around 4 nm, we used 20 nodes to accomplish the full-frequency GW calculation. The computational cost of our algorithm has a theoretical scaling of $O(N^4)$. In the benchmarks of silicon clusters, we observe a practical scaling of roughly $O(N^{2.3})$ for systems with less than 600 silicon atoms (see Section 4 of the Supplemental Material [28] for a detailed analysis of the computational costs). As shown in Fig. 3, we compared the running time for full-frequency GW calculations of $Si_{453}H_{228}$ using different numbers of nodes. For our Lanczos-based method, the most time-consuming steps are matrix-matrix and matrix-vector multiplications, which are suitable for massive parallelization and acceleration with GPUs in heterogeneous supercomputers. Fig. 3 demonstrates the reasonably good strong scaling with computation resources. When the calculations are accelerated with GPUs, the speed-up factors compared to CPU-only calculations are around 20.

In summary, a full-frequency GW formalism based on a Lanczos method is proposed to realize efficient modeling of hundreds of atoms with modest resources. This method can be used for highly efficient full-frequency GW calculations of large finite systems, such as semiconductor quantum dots and ligand-protected superatomic clusters with a few hundred atoms. Our method can also facilitate the construction of computational databases with quasiparticle-energy data, which were challenging to accomplish with limited computational costs before. This method is ready to generalize to extended systems, for which complex-valued wave functions are required. If the Tamm-Dancoff approximation is used (i.e., setting matrix $\mathbf{B} = 0$ in the RPA Casida matrix) for extended systems, then the calculation is greatly simplified as only a Hermitian matrix $\mathbf{A}$ remains, and a standard Lanczos algorithm for Hermitian matrix can be used. On the other hand, if RPA is used, then a Lanczos-based method designed for pseudo-Hermitian matrices [34, 36, 74] can be adopted.

We are grateful for the discussion with Peihong Zhang. This work is supported by the National Natural Science Foundation of China (12104080, 91961204), GHfund A (2022201, ghfund202202012538), the Fundamental Research Funds for the Central Universities (DUT22LK04 and DUT22ZD103), and a sub-award from the Center for Computational Study of Excited-State Phenomena in Energy Materials at the Lawrence Berkeley National Laboratory, which is funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division under Contract No. DEAC02-05CH11231, as part of the Computational Materials Sciences Program. Computational resources are provided by the Perlmutter supercomputer cluster of the National Energy Research Scientific Computing Center (NERSC), the Sugon Supercomputer Center at Wuzhen, and the Sugon Supercomputer Center at Kunshan.

* zhaojj@dlut.edu.cn
† jrc@utexas.edu

[1] L. Hedin, New method for calculating the one-particle green's function with application to the electron-gas problem, Phys. Rev. **139**, A796 (1965).
[2] M. S. Hybertsen and S. G. Louie, Electron correlation in semiconductors and insulators: Band gaps and quasiparticle energies, Phys. Rev. B **34**, 5390 (1986).
[3] R. W. Godby, M. Schlüter, and L. J. Sham, Self-energy operators and exchange-correlation potentials in semiconductors, Phys. Rev. B **37**, 10159 (1988).
[4] M. Rohlfing and S. G. Louie, Electron-hole excitations in semiconductors and insulators, Phys. Rev. Lett. **81**, 2312 (1998).
[5] S. Albrecht, L. Reining, R. Del Sole, and G. Onida, Ab initio calculation of excitonic effects in the optical spectra of semiconductors, Phys. Rev. Lett. **80**, 4510 (1998).
[6] Y.-H. Chan, D. Y. Qiu, F. H. da Jornada, and S. G. Louie, Giant exciton-enhanced shift currents and direct current conduction with subbandgap photo excitations produced by many-electron interactions, Proceedings of the National Academy of Sciences **118**, e1906938118 (2021).
[7] S. Biermann, F. Aryasetiawan, and A. Georges, First-principles approach to the electronic structure of strongly correlated systems: Combining the *GW* approximation and dynamical mean-field theory, Phys. Rev. Lett. **90**, 086402 (2003).
[8] P. Sun and G. Kotliar, Extended dynamical mean-field theory and GW method, Phys. Rev. B **66**, 085120 (2002).
[9] T. Zhu and G. K.-L. Chan, Ab initio full cell $GW+$DMFT for correlated materials, Phys. Rev. X **11**, 021006 (2021).
[10] Z. Li, G. Antonius, M. Wu, F. H. da Jornada, and S. G. Louie, Electron-phonon coupling from ab initio linear-response theory within the *gw* method: Correlation-enhanced interactions and superconductivity in $Ba_{1-x}K_xBiO_3$, Phys. Rev. Lett. **122**, 186402 (2019).
[11] Z. Li, M. Wu, Y.-H. Chan, and S. G. Louie, Unmasking the origin of kinks in the photoemission spectra of cuprate superconductors, Phys. Rev. Lett. **126**, 146401 (2021).

[12] D. Golze, M. Dvorak, and P. Rinke, The GW compendium: A practical guide to theoretical photoemission spectroscopy, Frontiers in Chemistry **7**, 377 (2019).

[13] W. Gao, W. Xia, P. Zhang, J. R. Chelikowsky, and J. Zhao, Numerical methods for efficient GW calculations and the applications in low-dimensional systems, Electronic Structure **4**, 023003 (2022).

[14] J. Wilhelm, D. Golze, L. Talirz, J. Hutter, and C. A. Pignedoli, Toward GW calculations on thousands of atoms, The Journal of Physical Chemistry Letters **9**, 306 (2018).

[15] V. Vlček, W. Li, R. Baer, E. Rabani, and D. Neuhauser, Swift GW beyond 10,000 electrons using sparse stochastic compression, Phys. Rev. B **98**, 075107 (2018).

[16] M. D. Ben, C. Yang, Z. Li, F. H. d. Jornada, S. G. Louie, and J. Deslippe, Accelerating large-scale excited-state gw calculations on leadership HPC systems, in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis* (2020) pp. 1–11.

[17] I. Duchemin and X. Blase, Cubic-scaling all-electron GW calculations with a separable density-fitting space–time approach, Journal of Chemical Theory and Computation **17**, 2383 (2021).

[18] V. W.-z. Yu and M. Govoni, GPU acceleration of large-scale full-frequency GW calculations, Journal of Chemical Theory and Computation **18**, 4690 (2022).

[19] J. Deslippe, G. Samsonidze, D. A. Strubbe, M. Jain, M. L. Cohen, and S. G. Louie, `BerkeleyGW`: A massively parallel computer package for the calculation of the quasiparticle and optical properties of materials and nanostructures, Computer Physics Communications **183**, 1269 (2012).

[20] D. Sangalli, A. Ferretti, H. Miranda, C. Attaccalite, I. Marri, E. Cannuccia, P. Melo, M. Marsili, F. Paleari, A. Marrazzo, G. Prandini, P. Bonfà, M. O. Atambo, F. Affinito, M. Palummo, A. Molina-Sánchez, C. Hogan, M. Grüning, D. Varsano, and A. Marini, Many-body perturbation theory calculations using the yambo code, Journal of Physics: Condensed Matter **31**, 325902 (2019).

[21] L. Hedin, On correlation effects in electron spectroscopies and the GW approximation, Journal of Physics: Condensed Matter **11**, R489 (1999).

[22] M. E. Casida, Time-dependent density functional response theory for molecules, in *Recent Advances in Density Functional Methods*, edited by D. P. Chong (World Scientific, 1995) pp. 155–192.

[23] F. Bruneval, T. Rangel, S. M. Hamed, M. Shao, C. Yang, and J. B. Neaton, `MOLGW` 1: Many-body perturbation theory software for atoms, molecules, and clusters, Computer Physics Communications **208**, 149 (2016).

[24] M. L. Tiago and J. R. Chelikowsky, Optical excitations in organic molecules, clusters, and defects studied by first-principles green's function methods, Phys. Rev. B **73**, 205334 (2006).

[25] M. J. van Setten, F. Weigend, and F. Evers, The gw-method for quantum chemistry applications: Theory and implementation, Journal of Chemical Theory and Computation **9**, 232 (2013).

[26] D. Mejia-Rodriguez, A. Kunitsa, E. Aprà, and N. Govind, Scalable molecular GW calculations: Valence and core spectra, Journal of Chemical Theory and Computation **17**, 7504 (2021).

[27] G. Onida, L. Reining, and A. Rubio, Electronic excitations: Density-functional versus many-body green's-function approaches, Rev. Mod. Phys. **74**, 601 (2002).

[28] See Supplemental Material at [URL-will-be-inserted-by-publisher] for the details of Eq. (9) - (12), computational details, HOMO-LUMO gaps of silicon nanoclusters, and analysis of the scaling of computation costs.

[29] R. Haydock, The recursive solution of the schrödinger equation, Computer Physics Communications **20**, 11 (1980).

[30] H. Meyer and S. Pal, A band-Lanczos method for computing matrix elements of a resolvent, The Journal of Chemical Physics **91**, 6195 (1989).

[31] O. B. Malcıoğlu, R. Gebauer, D. Rocca, and S. Baroni, turbotddft – a code for the simulation of molecular spectra using the liouville–lanczos approach to time-dependent density-functional perturbation theory, Computer Physics Communications **182**, 1744 (2011).

[32] B. Walker, A. M. Saitta, R. Gebauer, and S. Baroni, Efficient approach to time-dependent density-functional perturbation theory for optical spectroscopy, Phys. Rev. Lett. **96**, 113001 (2006).

[33] L. Zamok, S. Coriani, and S. P. A. Sauer, A tale of two vectors: A Lanczos algorithm for calculating RPA mean excitation energies, The Journal of Chemical Physics **156**, 014102 (2022).

[34] M. Grüning, A. Marini, and X. Gonze, Implementation and testing of lanczos-based algorithms for random-phase approximation eigenproblems, Computational Materials Science **50**, 2148 (2011).

[35] L. X. Benedict, E. L. Shirley, and R. B. Bohn, Theory of optical absorption in diamond, Si, Ge, and GaAs, Phys. Rev. B **57**, R9385 (1998).

[36] M. Shao, F. H. da Jornada, L. Lin, C. Yang, J. Deslippe, and S. G. Louie, A structure preserving lanczos algorithm for computing the optical absorption spectrum, SIAM Journal on Matrix Analysis and Applications **39**, 683 (2018).

[37] M. Govoni and G. Galli, Large scale GW calculations, Journal of Chemical Theory and Computation **11**, 2680 (2015).

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

[39] J. Laflamme Janssen, B. Rousseau, and M. Côté, Efficient dielectric matrix calculations using the lanczos algorithm for fast many-body G_0W_0 implementations, Phys. Rev. B **91**, 125120 (2015).

[40] S. J. Bintrim and T. C. Berkelbach, Full-frequency GW without frequency, The Journal of Chemical Physics **154**, 041101 (2021).

[41] C. J. C. Scott, O. J. Backhouse, and G. H. Booth, A "moment-conserving" reformulation of GW theory, The Journal of Chemical Physics **158**, 124102 (2023).

[42] O. J. Backhouse, A. Santana-Bonilla, and G. H. Booth, Scalable and predictive spectra of correlated molecules with moment truncated iterated perturbation theory, The Journal of Physical Chemistry Letters **12**, 7650 (2021).

[43] T. Chiarotti, N. Marzari, and A. Ferretti, Unified green's function approach for spectral and thermodynamic properties from algorithmic inversion of dynamical potentials, Phys. Rev. Res. **4**, 013242 (2022).

[44] D. A. Leon, A. Ferretti, D. Varsano, E. Molinari, and C. Cardoso, Efficient full frequency GW for metals using a multipole approach for the dielectric screening, Phys. Rev. B **107**, 155130 (2023).

[45] S. J. Bintrim and T. C. Berkelbach, Full-frequency dy-

namical Bethe–Salpeter equation without frequency and a study of double excitations, The Journal of Chemical Physics **156**, 044114 (2022).
[46] J. Schirmer, L. S. Cederbaum, and O. Walter, New approach to the one-particle green's function for finite fermi systems, Phys. Rev. A **28**, 1237 (1983).
[47] M. J. van Setten, F. Caruso, S. Sharifzadeh, X. Ren, M. Scheffler, F. Liu, J. Lischner, L. Lin, J. R. Deslippe, S. G. Louie, C. Yang, F. Weigend, J. B. Neaton, F. Evers, and P. Rinke, GW100: Benchmarking G_0W_0 for molecular systems, Journal of Chemical Theory and Computation **11**, 5665 (2015).
[48] M. Govoni and G. Galli, GW100: Comparison of methods and accuracy of results obtained with the west code, Journal of Chemical Theory and Computation **14**, 1895 (2018).
[49] A. Förster and L. Visscher, GW100: A slater-type orbital perspective, Journal of Chemical Theory and Computation **17**, 5080 (2021).
[50] W. Gao and J. R. Chelikowsky, Real-space based benchmark of G_0W_0 calculations on GW100: Effects of semicore orbitals and orbital reordering, Journal of Chemical Theory and Computation **15**, 5299 (2019).
[51] F. Caruso, P. Rinke, X. Ren, A. Rubio, and M. Scheffler, Self-consistent GW: All-electron implementation with localized basis functions, Phys. Rev. B **88**, 075105 (2013).
[52] N. Marom, F. Caruso, X. Ren, O. T. Hofmann, T. Körzdörfer, J. R. Chelikowsky, A. Rubio, M. Scheffler, and P. Rinke, Benchmark of GW methods for azabenzenes, Phys. Rev. B **86**, 245127 (2012).
[53] P. Rinke, A. Qteish, J. Neugebauer, C. Freysoldt, and M. Scheffler, Combining GW calculations with exact-exchange density-functional theory: an analysis of valence-band photoemission for compound semiconductors, New Journal of Physics **7**, 126 (2005).
[54] M. Dauth, F. Caruso, S. Kümmel, and P. Rinke, Piecewise linearity in the *GW* approximation for accurate quasiparticle energy predictions, Phys. Rev. B **93**, 121115 (2016).
[55] L. Hung, F. H. da Jornada, J. Souto-Casares, J. R. Chelikowsky, S. G. Louie, and S. Öğüt, Excitation spectra of aromatic molecules within a real-space *GW*-BSE formalism: Role of self-consistency and vertex corrections, Phys. Rev. B **94**, 085125 (2016).
[56] L. Kronik, A. Makmal, M. L. Tiago, M. M. G. Alemany, M. Jain, X. Huang, Y. Saad, and J. R. Chelikowsky, PARSEC – the pseudopotential algorithm for real-space electronic structure calculations: recent advances and novel applications to nano-structures, Phys. Status Solidi B **243**, 1063 (2006).
[57] M. Dogan, K.-H. Liou, and J. R. Chelikowsky, Solving the electronic structure problem for over 100 000 atoms in real space, Phys. Rev. Mater. **7**, L063001 (2023).
[58] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. **77**, 3865 (1996).
[59] N. Troullier and J. L. Martins, Efficient pseudopotentials for plane-wave calculations, Phys. Rev. B **43**, 1993 (1991).
[60] F. Weigend, A fully direct RI-HF algorithm: Implementation, optimised auxiliary basis sets, demonstration of accuracy and efficiency, Phys. Chem. Chem. Phys. **4**, 4285 (2002).
[61] X. Ren, P. Rinke, V. Blum, J. Wieferink, A. Tkatchenko, A. Sanfilippo, K. Reuter, and M. Scheffler, Resolution-of-identity approach to Hartree-Fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions, New Journal of Physics **14**, 053020 (2012).
[62] W. Hu, J. Liu, Y. Li, Z. Ding, C. Yang, and J. Yang, Accelerating excitation energy computation in molecules and solids within linear-response time-dependent density functional theory via interpolative separable density fitting decomposition, Journal of Chemical Theory and Computation **16**, 964 (2020).
[63] J. Lu and L. Ying, Compression of the electron repulsion integral tensor in tensor hypercontraction format with cubic scaling cost, Journal of Computational Physics **302**, 329 (2015).
[64] W. Gao and J. R. Chelikowsky, Accelerating time-dependent density functional theory and GW calculations for molecules and nanoclusters with symmetry adapted interpolative separable density fitting, Journal of Chemical Theory and Computation **16**, 2216 (2020).
[65] E. G. Barbagiovanni, D. J. Lockwood, P. J. Simpson, and L. V. Goncharova, Quantum confinement in Si and Ge nanostructures: Theory and experiment, Applied Physics Reviews **1**, 011302 (2014).
[66] J. Liang, C. Huang, and X. Gong, Silicon nanocrystals and their composites: Syntheses, fluorescence mechanisms, and biological applications, ACS Sustainable Chemistry & Engineering **7**, 18213 (2019).
[67] K. Dohnalová, A. N. Poddubny, A. A. Prokofiev, W. D. de Boer, C. P. Umesh, J. M. Paulusse, H. Zuilhof, and T. Gregorkiewicz, Surface brightens up Si quantum dots: direct bandgap-like size-tunable emission, Light: Science & Applications **2**, e47 (2013).
[68] D. Neiner, H. W. Chiu, and S. M. Kauzlarich, Low-temperature solution route to macroscopic amounts of hydrogen terminated silicon nanoparticles, Journal of the American Chemical Society **128**, 11016 (2006).
[69] C. Huan and S. Shu-Qing, Silicon nanoparticles: Preparation, properties, and applications, Chinese Physics B **23**, 088102 (2014).
[70] S. Godefroo, M. Hayne, M. Jivanescu, A. Stesmans, M. Zacharias, O. I. Lebedev, G. Van Tendeloo, and V. V. Moshchalkov, Classification and control of the origin of photoluminescence from Si nanocrystals, Nature Nanotechnology **3**, 174 (2008).
[71] F. Priolo, T. Gregorkiewicz, M. Galli, and T. F. Krauss, Silicon nanostructures for photonics and photovoltaics, Nature Nanotechnology **9**, 19 (2014).
[72] T. Liao, K.-H. Liou, and J. R. Chelikowsky, Dielectric screening and vacancy formation for large neutral and charged Si_nH_m ($n > 1500$) nanocrystals using real-space pseudopotentials, Phys. Rev. Mater. **6**, 054603 (2022).
[73] D. Neuhauser, Y. Gao, C. Arntsen, C. Karshenas, E. Rabani, and R. Baer, Breaking the theoretical scaling limit for predicting quasiparticle energies: The stochastic GW approach, Phys. Rev. Lett. **113**, 076402 (2014).
[74] J. Brabec, L. Lin, M. Shao, N. Govind, C. Yang, Y. Saad, and E. G. Ng, Efficient algorithms for estimating the absorption spectrum within linear response TDDFT, Journal of Chemical Theory and Computation **11**, 5197 (2015).

Supplemental Material for "Efficient Full-frequency GW Calculations using a Lanczos Method"

Weiwei Gao[1], Zhao Tang[2], Jijun Zhao[3] and James R. Chelikowsky[2,4,5]

1. Key Laboratory of Materials Modification by Laser, Ion and Electron Beams, Ministry of Education, Dalian University of Technology, Dalian 116024, China

2. Center for Computational Materials, Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX 78712

3. Key Laboratory of Atomic and Subatomic Structure and Quantum Control (Ministry of Education), Guangdong Basic Research Center of Excellence for Structure and Fundamental Interactions of Matter, School of Physics, South China Normal University, Guangzhou 510006, China

4. Department of Physics, The University of Texas at Austin, Austin, TX 78712

5. McKetta Department of Chemical Engineering, The University of Texas at Austin, Austin, TX 78712

1. Details on Eq. (9) – (12) of the main text

We start from W^s_{nm} defined in Eq. (8) of the main text:

$$W^s_{nm} = \sum_{vc} (nm|vc)\sqrt{\frac{\epsilon_c - \epsilon_v}{\Omega_s}} Z^s_{vc}$$

$$= \sum_{vc} \left[(nm|vc)(\epsilon_c - \epsilon_v)^{\frac{1}{2}} \right] Z^s_{vc} \Omega_s^{-\frac{1}{2}} = \langle P_{nm}|Z^s\rangle \Omega_s^{-\frac{1}{2}}.$$

Here we introduce an auxiliary vector P_{nm} of dimension $N_v N_c$ to get Eq. (9) in the main text. The elements of P_{nm} are defined by $(P_{nm})_{vc} = (nm|vc)(\epsilon_c - \epsilon_v)^{\frac{1}{2}}$. By comparing Eq. (10) and Eq. (7), one can find the expression of $\Sigma^{corr}_{mm}(\omega, n)$

$$\Sigma^{corr}_{mm}(\omega, n) = \sum_{s}^{N_{vc}} \frac{W^s_{nm} W^s_{nm}}{\omega - \epsilon_n + \eta_n(\Omega_s - i\delta)}.$$

Using Eq. (9), $\Sigma^{corr}_{mm}(\omega, n)$ can be further simplified as below

$$\Sigma^{corr}_{mm}(\omega, n) = \sum_{s}^{N_{vc}} \frac{\langle P_{nm}|Z^s\rangle \Omega_s^{-\frac{1}{2}} \langle Z^s|P_{nm}\rangle \Omega_s^{-\frac{1}{2}}}{\omega - \epsilon_n + \eta_n(\Omega_s - i\delta)}$$

$$= \sum_{s}^{N_{vc}} \frac{\langle P_{nm}|Z^s\rangle\langle Z^s|P_{nm}\rangle}{[\ \omega - \epsilon_n + \eta_n(\Omega_s - i\delta)\ ]\Omega_s}$$

$$= \sum_{s}^{N_{vc}} \langle P_{nm}|Z^s\rangle\langle Z^s|P_{nm}\rangle \times \left[\frac{1}{\Omega_s} - \frac{1}{\eta_n(\omega - \epsilon_n + \eta_n\Omega_s - i\eta_n\delta)} \right] \frac{1}{(\omega - \epsilon_n - i\eta_n\delta)}$$

$$= \frac{1}{z_n}\sum_{s}^{N_{vc}} \langle P_{nm}|Z^s\rangle\langle Z^s|P_{nm}\rangle \left[\frac{1}{\Omega_s} - \frac{1}{\Omega_s + \eta_n z_n}\right],$$

which is Eq. (11) of the main text. In the last step, we use $\eta_n^2 = 1$ and introduce a complex variable $z_n = \omega - \epsilon_n - i\eta_n\delta$. Considering matrix $\mathbf{D}$, which satisfies $\mathbf{D}Z^s = \Omega^s Z^s$ and $\sum_s^N |Z^s\rangle\langle Z^s| = \mathbf{I}$, we rewrite Eq. (11) as follow (Eq. (12)):

$$\Sigma_{mm}^{corr}(\omega, n) = \frac{1}{z_n}\left[\langle P_{nm}|\mathbf{D^{-1}}|P_{nm}\rangle - \langle P_{nm}|(\mathbf{D} + \eta_n z_n)^{-\mathbf{1}}|P_{nm}\rangle\right]$$

Since the Casida matrix satisfies $\mathbf{C}Z^s = Z^s\Omega_s^2$, one can use a polynomial function of $\mathbf{C}$ to approximate matrix $\mathbf{D}$ which has eigenvalues equal to the square root of Ω_s^2. We use a least square fitting procedure to find a polynomial approximation of the square root function $\sqrt{\boldsymbol{x}}$ in the range of $[\min(\Omega_s^2), \max(\Omega_s^2)]$. The fitting points are chosen as the Chebyshev nodes of the fitting range. One can use standard Lanczos algorithm to determine the extremal eigenvalues $\min(\Omega_s^2)$ and $\max(\Omega_s^2)$ of $\mathbf{C}$**.** In our current implementation, we simply use $[0.8\min(\epsilon_c - \epsilon_v)^2, 1.25\min(\epsilon_c - \epsilon_v)^2]$ as the fitting range, which works well for the systems studied in this work.

Different from finite systems, where real-numbered wave functions can be adopted to simplify calculations, the study of periodic systems usually requires complex wave functions. In this case, the correlation part of the GW self-energy reads

$$\Sigma_{m\boldsymbol{k},m\boldsymbol{k}}^{corr}(\omega) = \frac{2}{N_{\boldsymbol{q}}N_{\boldsymbol{k}}}\sum_{n}^{N_v+N_c}\sum_{\boldsymbol{q}}^{N_{\boldsymbol{q}}}\sum_{s}^{N_vN_cN_{\boldsymbol{k}}} \frac{|W_{n\boldsymbol{k}-\boldsymbol{q},m\boldsymbol{k}}^s|^2}{\omega - \epsilon_{n\boldsymbol{k}-\boldsymbol{q}} + \eta_{n\boldsymbol{k}-\boldsymbol{q}}(\Omega_s(\boldsymbol{q}) - i\delta)},$$

where

$$W_{n\boldsymbol{k}-\boldsymbol{q},m\boldsymbol{k}}^s = \sum_{v'c'\boldsymbol{k}'}\big[(n\boldsymbol{k}-\boldsymbol{q}, m\boldsymbol{k}|v'\boldsymbol{k}', c'\boldsymbol{k}'+\boldsymbol{q})\mathbf{X}(\boldsymbol{q})_{v'c'\boldsymbol{k}'}^s + (n\boldsymbol{k}-\boldsymbol{q}, m\boldsymbol{k}|c'\boldsymbol{k}'+\boldsymbol{q}, v'\boldsymbol{k}')\mathbf{Y}(\boldsymbol{q})_{v'c'\boldsymbol{k}'}^s\big].$$

Different from finite systems discussed in the main text, we need to construct a Casida equation for each $\boldsymbol{q}$-point with the random phase approximation

$$\mathrm{C}^{\mathrm{RPA}}(\boldsymbol{q}) = \begin{bmatrix}\mathbf{A}(\boldsymbol{q}) & \mathbf{B}(\boldsymbol{q}) \\ -\mathbf{B}(\boldsymbol{q})^* & -\mathbf{A}(\boldsymbol{q})^*\end{bmatrix},$$

where the matrix elements are

$$\mathbf{A}(\boldsymbol{q})_{vc\boldsymbol{k},v'c'\boldsymbol{k}'} = (\epsilon_{c\boldsymbol{k}+\boldsymbol{q}} - \epsilon_{v\boldsymbol{k}})\delta_{vv'}\delta_{cc'}\delta_{\boldsymbol{k}\boldsymbol{k}'} + (v\boldsymbol{k}, c\boldsymbol{k}+\boldsymbol{q}|v'\boldsymbol{k}', c'\boldsymbol{k}'+\boldsymbol{q})$$
$$= (\epsilon_{c\boldsymbol{k}+\boldsymbol{q}} - \epsilon_{v\boldsymbol{k}})\delta_{vv'}\delta_{cc'}\delta_{\boldsymbol{k}\boldsymbol{k}'} + \int d^3r d^3r' \frac{\psi_{c\boldsymbol{k}+\boldsymbol{q}}^*(\boldsymbol{r})\psi_{v\boldsymbol{k}}(\boldsymbol{r})\psi_{c'\boldsymbol{k}'+\boldsymbol{q}}(\boldsymbol{r}')\psi_{v'\boldsymbol{k}'}^*(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|},$$
$$\mathbf{B}(\boldsymbol{q})_{vc\boldsymbol{k},v'c'\boldsymbol{k}'} = (v\boldsymbol{k}, c\boldsymbol{k}+\boldsymbol{q}|c'\boldsymbol{k}'+\boldsymbol{q}, v'\boldsymbol{k}')$$
$$= \int d^3r d^3r' \frac{\psi_{c\boldsymbol{k}+\boldsymbol{q}}^*(\boldsymbol{r})\psi_{v\boldsymbol{k}}(\boldsymbol{r})\psi_{c'\boldsymbol{k}'+\boldsymbol{q}}^*(\boldsymbol{r}')\psi_{v'\boldsymbol{k}'}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}.$$

One can see that $\mathbf{A}(\boldsymbol{q})_{vc\boldsymbol{k},v'c'\boldsymbol{k}'} = \mathbf{A}(\boldsymbol{q})^*{}_{v'c'\boldsymbol{k}',vc\boldsymbol{k}}$ and $\mathbf{B}(\boldsymbol{q})_{vc\boldsymbol{k},v'c'\boldsymbol{k}'} = \mathbf{B}(\boldsymbol{q})_{v'c'\boldsymbol{k}',vc\boldsymbol{k}}$. This means that $\mathbf{A}(\boldsymbol{q})$ is a Hermitian matrix and $\mathbf{B}(\boldsymbol{q})$ is a symmetric matrix.

The vectors and eigenvalues of $\mathbf{C}^{\mathrm{RPA}}(\boldsymbol{q})$ take the following form

$$\mathbf{C}^{\mathrm{RPA}}(\boldsymbol{q})\begin{bmatrix}\mathbf{X}(\boldsymbol{q}) & \mathbf{Y}(\boldsymbol{q})^* \\ \mathbf{Y}(\boldsymbol{q}) & \mathbf{X}(\boldsymbol{q})^*\end{bmatrix} = \begin{bmatrix}\mathbf{X}(\boldsymbol{q}) & \mathbf{Y}(\boldsymbol{q})^* \\ \mathbf{Y}(\boldsymbol{q}) & \mathbf{X}(\boldsymbol{q})^*\end{bmatrix}\begin{bmatrix}\Omega(\mathbf{q}) & 0 \\ 0 & -\Omega(\mathbf{q})\end{bmatrix}.$$

If one uses the Tamm-Dancoff approximation, which is commonly used for studying the optical properties of periodic systems, the RPA Casida matrix $\mathbf{C}^{RPA}(\boldsymbol{q})$ is reduced to $\mathbf{C}^{TD}(\boldsymbol{q}) = \begin{bmatrix} \mathbf{A}(\boldsymbol{q}) & 0 \\ 0 & -\mathbf{A}(\boldsymbol{q})^* \end{bmatrix}$ and $\Sigma^{corr}_{\boldsymbol{mk},\boldsymbol{mk}}(\omega)$ can be rewritten as

$$\Sigma^{corr}_{\boldsymbol{mk},\boldsymbol{mk}}(\omega) = \frac{2}{N_q N_k} \sum_{n}^{N_v+N_c} \sum_{\boldsymbol{q}}^{N_q} \left\langle P_{nm\boldsymbol{k}}(\boldsymbol{q}) \left| \frac{1}{\omega - \epsilon_{n\boldsymbol{k}-\boldsymbol{q}} + \eta_{n\boldsymbol{k}-\boldsymbol{q}}(\mathbf{A}(\boldsymbol{q}) - i\delta)} \right| P_{nm\boldsymbol{k}}(\boldsymbol{q}) \right\rangle.$$

The resolvent $\langle P_{nm\boldsymbol{k}}(\boldsymbol{q}) \left| \frac{1}{\omega - \epsilon_{n\boldsymbol{k}-\boldsymbol{q}} + \eta_{n\boldsymbol{k}-\boldsymbol{q}}(A(\boldsymbol{q}) - i\delta)} \right| P_{nm\boldsymbol{k}}(\boldsymbol{q}) \rangle$ can be computed using the standard Lanczos algorithm for Hermitian matrix $\mathbf{A}(\boldsymbol{q})$. The $(v'c'\boldsymbol{k}')$ vector elements of $P_{nm\boldsymbol{k}}(\boldsymbol{q})$ is given by $P_{nm\boldsymbol{k}}(\boldsymbol{q})_{v'c'\boldsymbol{k}'} = (n\boldsymbol{k} - \boldsymbol{q}, m\boldsymbol{k} | v'\boldsymbol{k}', c'\boldsymbol{k}' + \boldsymbol{q})$.

2. Computational Details

The structures of hydrogen-passivated silicon clusters are relaxed using a real-space based pseudo-potential density functional theory code PARSEC. Perdew-Burke-Ernzerhof exchange-correlation functional [1] and Troullier-Martin norm-conserving pseudopotentials [2] are used in this code. For the GW calculations, the wave functions are sampled with uniform real-space grids with grid size h = 0.6 Bohr. The Lanczos method is implemented in the NanoGW code. We used polynomials with a degree of 7 and the number of Lanczos iterations of 8 for the GW calculations of silicon nanoclusters. For polynomial interpolation of square root functions, we used a linear square fitting method. The Chebyshev nodes were chosen as the fitting points. For G_0W_0 quasiparticle energies, we computed the quasiparticle energy $E_n^{GW}(\omega) = E_n^{DFT} + \langle n|\hat{\Sigma}^{\mathrm{GW}}(\omega) - \hat{V}_{xc}|n\rangle$ on a frequency grid. Then the Newton-Raphson algorithm is used to get a graphical solution of the equation $E_n^{GW}(\omega) = \omega$ for the quasiparticle solution. For the positive infinitesimal number δ appear in Eq. (7) of the main text, we used $\delta = 0.05$ Rydberg.

We use graphic processing units (GPU) to accelerate the Lanczos iteration steps. The most expensive step in Lanczos algorithm is computing the matrix-vector multiplication $|V'\rangle = \mathbf{D}|V\rangle = (a_0\mathbf{I} + \sum_{k=1}^{g} a_k \mathbf{C}^k)\,|V\rangle$, which can be reduced to a series of computations of $\mathbf{C}\,|V\rangle$**.** In our current code implementation, only the computation of matrix-vector products $\mathbf{C}\,|V\rangle$ is accelerated using `cuBLAS` or `rocBLAS` library, while the interpolative separable density fitting method and the calculation of two-center integrals $(\alpha|\beta) = \int \frac{\zeta_\alpha(r)\zeta_\beta(r)}{|r-r'|} d^3r d^3r'$ have not yet been accelerated with GPUs. The code is available at https://gitlab.com/real-space/nanogw/-/tree/Lanczos/. Calculation files are uploaded to https://zenodo.org/records/10627764.

(*Technical notes 1*) In the main text, we compare the HOMO and LUMO energies of the GW100 dataset for the Lanczos-based method and the reference, the difference between the two methods appears to plateau at 20 meV. The source of discrepancy between our new method and the reference method mainly comes from the polynomial fitting of square-root function $\sqrt{x} \approx a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots$, which is not exact and results in small numerical errors. To further demonstrate our statement (i.e., the error mostly comes from the polynomial fitting), we have examined all the GW100 molecules that show large mean absolute differences between the results calculated with Lanczos-based and conventional GW methods. The eigenvalues Ω_s^2

of Casida matrix of these molecule span a relatively wide range $[\min \Omega_s^2, \max \Omega_s^2]$. If one fits the square-root function in a wide range using a polynomial, the fitted polynomial deviates from square-root function near $\min \Omega_s^2$, which is similar to the Runge phenomenon. To show this, we plot the following figures (Fig. S1 ~ Fig. S3), which compare the fitting polynomial and square-root function within three different fitting ranges.

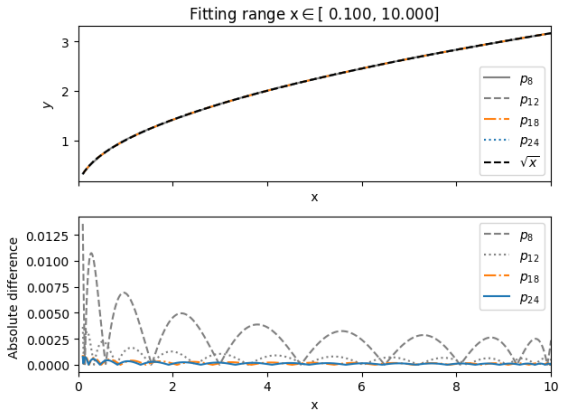

Fig S1. Fitting $\sqrt{x}$ in the range [0.1, 10]. Top panel: comparison between the square-root function and fitting polynomials with different degrees. Bottom panel: the absolute difference between polynomials and the square-root function.

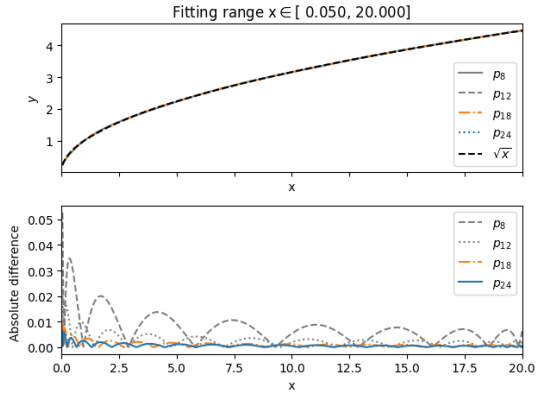

Fig S2. Fitting $\sqrt{x}$ in the range [0.05, 20]

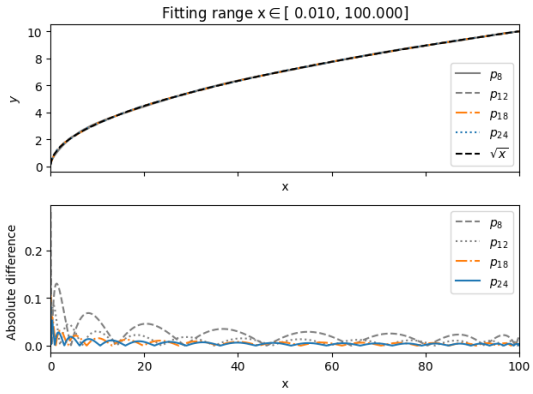

Fig S3. Fitting $\sqrt{x}$ in the range [0.01, 100]

As shown in Fig. S1, when we fit the square-root function $\sqrt{x}$ for $x \in [0.1, 10]$, the absolute difference between the polynomial and $\sqrt{x}$ is small. If we gradually increase the fitting range to $[0.05, 20]$ and $[0.01, 100]$, the absolute difference progressively grows larger, as shown in Fig. S2 and S3. The largest absolute difference appears close to $x = 0$. Additionally, we can see the absolute differences do not effectively decrease if we further increase the polynomial degree g from 18 to 24. These results suggest the errors mostly come from the molecules which have a wide range of Ω_s^2 and the polynomial fitting can not be systematically improved by increasing the polynomial degree.

(*Technical note 2*) The Lanczos-based method proposed in the main text also work well for semi-core states. Fig S4 ~ S7 show some calculation results of low-energy states (including Ar atom 3s orbital, Ne atom 2s orbital, Ge 3d orbital in the GeH_4 molecule, Al 2p orbital in the AlF_3 molecule). The frequency dependent quasiparticle energy $E^{GW}(\omega)$ calculated with the Lanczos+ISDF method and with the reference method (denoted in "Ref." in the figures) are compared. We can see the results with two methods agree well.

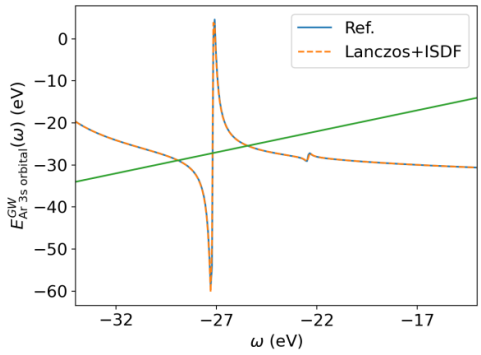

Fig S4. The frequency-dependent quasiparticle energy of Ar atom 3s orbital

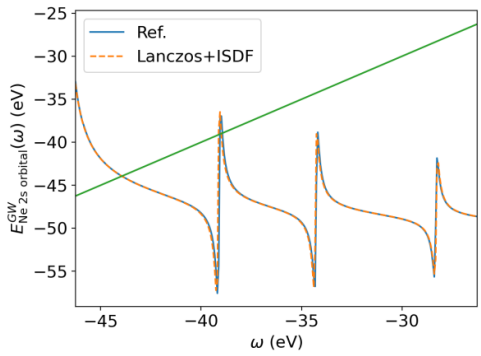

Fig S5. The frequency-dependent quasiparticle energy of Ne atom 2s orbital

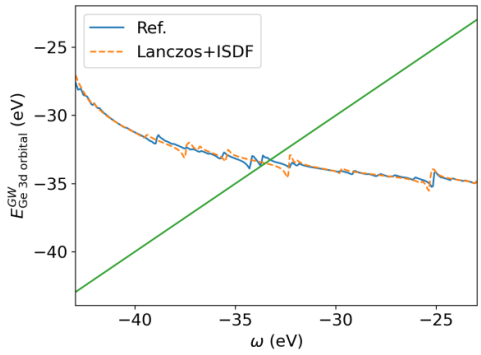

Fig S6. The frequency-dependent quasiparticle energy of Ge 3d orbital in the GeH_4 molecule

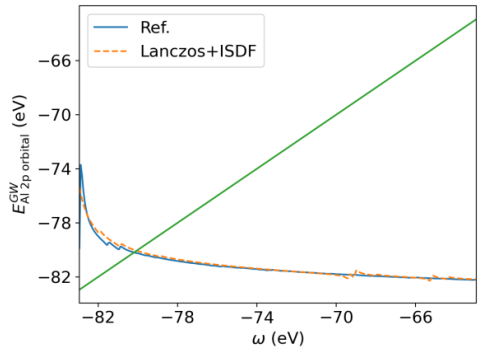

Fig S7. The frequency-dependent quasiparticle energy of Al 2p orbital in the AlF_3 molecule

3. HOMO-LUMO Gaps of Silicon Nanoclusters

We compared the HOMO-LUMO gaps of hydrogen-passivated silicon clusters (without defects) calculated with our implementation and those with previous work, as shown in Table S1. The results agree well with previous calculations. The main calculation parameters for our calculations are listed in Table S2.

Table S1. HOMO-LUMO gaps of non-defective silicon nanoclusters

Cluster	$Si_{87}H_{76}$	$Si_{147}H_{100}$	$S_{293}H_{172}$	$S_{353}H_{196}$	$Si_{705}H_{300}$
This work	4.7	4.1	3.1	3.0	2.4
Ref [3]	4.8	4.1	-	3.0	2.2
Ref [4]	4.77	4.21	3.46	-	-

Table S2. Calculation parameters. N_v is the number of occupied states. N_c is the number of unoccupied states. N_μ is the number of auxiliary basis functions used in the interpolative separable density fitting method.

Cluster	$Si_{87}H_{76}$	$Si_{147}H_{100}$	$S_{293}H_{172}$	$S_{353}H_{196}$	$Si_{705}H_{300}$
N_v	212	344	672	804	1560
$N_v + N_c$	1750	2850	5600	6700	13000
N_μ	7000	12000	22400	27000	52000

4. The scaling of computation costs

To analyze the theoretical scaling of our method, we examine the formula of $\Sigma_m^{corr}(\omega)$

$$\Sigma_m^{corr}(\omega) = \langle \phi_m | \hat{\Sigma}^{corr}(\omega) | \phi_m \rangle = \sum_n^{N_v+N_c} [\Sigma_m^{corr}(\omega, n)].$$

$\Sigma_m^{corr}(\omega)$ involves a summation over $N_v + N_c$ of $\Sigma_m^{corr}(\omega, n)$ terms, where N_v and N_c are the number of valence states and unoccupied states, respectively. The calculation of each $\Sigma_m^{corr}(\omega, n)$ term requires L Lanczos iterations. Each Lanczos iteration involves the calculation of $2g$ matrix-vector multiplications that scales as $O(gN_\mu N_c N_v)$, as illustrated in Fig. S8. Therefore, the total scaling of the calculation is $O((N_v + N_c)LgN_\mu N_c N_v)$. Here N_v, N_c, g, and N_μ are the number of valence states, the number of empty states, the degree of the fitting polynomial, and the number of auxiliary basis functions used in interpolative separable density fitting, respectively. Since g and L scale as constants, while N_v, N_c, and N_μ scales as $O(N)$, the total scaling is $O(N^4)$. The cost for interpolative separable density fitting procedure is $O(N^3)$.

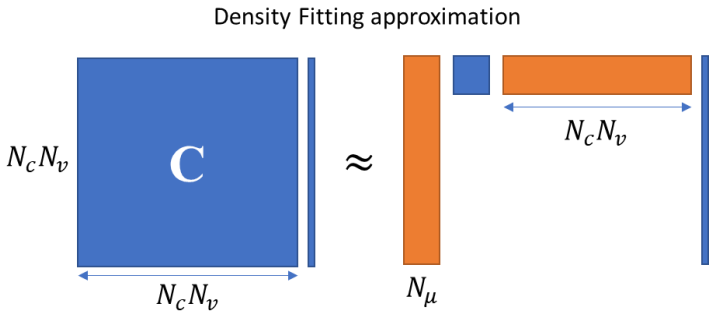

Fig S8. With interpolative separable density fitting method, the cost of a matrix-vector multiplication is significantly reduced and scales as $O(N_\mu N_c N_v)$.

To show the computational scaling of the Lanczos-based method applied to clusters of different sizes, we fitted the computation costs with curves of order $N^{2.3}$, N^3, and N^4. As shown in Fig. S9, the blue diamonds are the node hours (1 node hour = 1 node running for an hour) spent for calculating a GW energy for the corresponding silicon cluster. The computational costs include the computation time for interpolative separable density fitting procedure and all the steps in GW calculations. For smaller clusters with less than 600 silicon atoms (with around 3000 valence electrons), the scaling of our method is between $O(N^2)$ and $O(N^3)$. We start to observe a $O(N^4)$ scaling for silicon clusters with more than 600 silicon atoms.

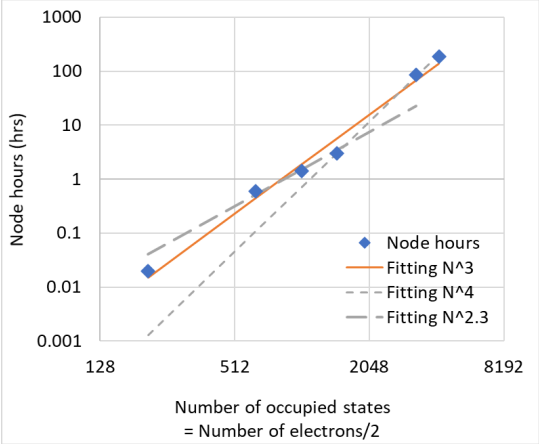

Fig S9. Computational costs (in node hours) of GW calculations for silicon clusters. The costs are fitted with different curves to show the computational scaling for different system sizes.

[1] J.P. Perdew, K. Burke, and M. Ernzerhof, Physical Review Letters, 77, 3865-3868 (1996)
[2] N. Troullier and J. L. Martins, Phys. Rev. B 43, 1993 (1991)
[3] D. Neuhauser, Y. Gao, C. Arntsen, C. Karshenas, E. Rabani, and R. Baer, Phys. Rev. Lett. 113, 076402 (2014).
[4] M. Govoni and G. Galli, Journal of Chemical Theory and Computation 11, 2680 (2015).