Swift GW beyond 10,000 electrons using sparse stochastic compression

Vojtěch Vlček, Wenfei Li, Roi Baer, Eran Rabani, and Daniel Neuhauser

Phys. Rev. B **98**, 075107 — Published 6 August 2018

DOI: 10.1103/PhysRevB.98.075107

Swift GW beyond $10,000$ electrons using sparse stochastic compression

Vojtěch Vlček,[1,2,*] Wenfei Li,[1,†] Roi Baer,[3,‡] Eran Rabani,[4,5,§] and Daniel Neuhauser[1,¶]

[1] *Department of Chemistry and Biochemistry, University of California, Los Angeles California 90095, USA*
[2] *After July 1 2018: Department of Chemistry and Biochemistry, University of California, Santa Barbara California 93106, USA*
[3] *Fritz Haber Center for Molecular Dynamics, Institute of Chemistry, The Hebrew University of Jerusalem, Jerusalem 91904, Israel*
[4] *Department of Chemistry, University of California and Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*
[5] *The Raymond and Beverly Sackler Center for Computational Molecular and Materials Science, Tel Aviv University, Tel Aviv, Israel 69978*

(Dated: July 18, 2018)

We introduce the concept of sparse stochastic compression (SSC), an efficient stochastic sampling of any general function. The technique uses sparse stochastic orbitals (SSOs), short vectors that sample a small number of space points. As a first demonstration, SSOs are applied in conjunction with simple direct-projection to accelerate our recent stochastic GW technique; the new developments enable accurate prediction of G_0W_0 quasiparticle energies and gaps for systems with up to $N_e > 10,000$ electrons, with small statistical errors of ± 0.05 eV and using less than 2000 core CPU hours. Overall, stochastic GW scales now linearly (and often sub-linearly) with N_e.

I. INTRODUCTION

Fundamental band gaps and quasiparticle (QP) energies determine the electronic properties of molecules and solids, but their first principles calculations are nontrivial. Density functional theory (DFT) [1] is usually used for ground state charge densities and atomic geometries, but gives wrong QP energies.[2–4] Going beyond DFT is computationally demanding. For small molecules, configuration interaction [5–7] and the equation of motion coupled cluster technique [8–10] yield accurate QP energies, but scale very steeply with the number of electrons.

In recent years, the GW approximation [4, 7, 11] became the predominant framework for QP calculations. The method describes all many-body effects through the single-particle self-energy, approximated as $\Sigma = GW$, where G is the single particle Green's function and W is the screened Coulomb interaction. GW provides accurate ionization energies and electron affinities for both molecules and solids, at a steep scaling.[12–18] Most computational improvements focus on reducing the prefactor rather than lowering the overall scaling.[14, 17]

We recently introduced a stochastic formulation of GW [19] that expresses the self-energy as a statistical quantity, averaged over random samplings. The spirit behind the stochastic paradigm is that quasiparticles (electrons and holes) are viewed, whenever possible, as being in time-dependent states that are random combination of occupied (or unoccupied) states. All quantities, such as the Green's function and the effective interaction, are expressed, in one form or another, in terms of such evolving random combinations. There is an elegance in this viewpoint, in that physically one should not need to know the detailed eigenstates for large systems. The resulting stochastic GW method reproduces the results of deterministic GW [20] but is very fast so it is applicable to large systems with thousands of valance electrons.[19, 21, 22]

Here, two major improvements of stochastic GW are introduced, and together they enable routine calculations of QP energies for systems with $N_e > 10,000$. The first is simple and technical, a direct projection of the occupied states (which formally increases the scaling of the method with system size but in practice reduces the effort below hundreds of thousands of electrons).

The second improvement is more fundamental, and relates to a general feature of the stochastic method, a representation of functions by a random basis-set. Originally, we used a stochastic resolution of the identity, SRI, but it turns out that for large systems the required number of stochastic basis functions grows with the number of electrons, destroying the overall linear scaling for large systems. To circumvent this, we develop a new approach based on sparse vectors, which we label as sparse stochastic resolution of the identity (SSRI); the method does not lower the accuracy but is much cheaper, thereby enabling the treatment of very large systems with $N_e > 10,000$. SSRI has potentially a large number of applications, and we use stochastic GW here to demonstrate its efficiency.

With direct projection and SSRI, stochastic GW is efficient and scales very gently. The overall method, as reviewed here, is fundamentally quite simple – once one accepts the concept of evolving random states as the basic tool, the formulation is quite transparent. The method is demonstrated here first for finite molecules (acenes and C_{60} molecules); in addition we apply for the first time stochastic GW for periodic systems with very large su-

* vlcek@ucsb.edu
† liwenfei@chem.ucla.edu
‡ roi.baer@huji.ac.il
§ eran.rabani@berkeley.edu
¶ dxn@ucla.edu

percells.

The paper is organized as follows: Deterministic GW is reviewed in Sec. II. In Sec. III we qualitatively overview the ingredients in Stochastic GW. Sec. IV details (following Refs. 19 and 20) how the stochastic expansion of G converts GW into the action of W on a source term. Sec. V reviews the use of linear response with deterministic or stochastic TDH (time-dependent Hartree) for acting with W^R, the causal (retarded) effective interaction. In Sec. VI sparse orbitals are introduced and used to convert the application of W^R to W. Results for molecules and solids are shown in Sec. VII, followed by conclusions in Sec. VIII.

II. THE GW METHOD

We first outline deterministic GW. The starting point is a specific real-valued KS (Kohn-Sham) orbital ϕ (typically the HOMO or LUMO) and associated eigenvalue ε^{KS} that fulfill $H_0\phi = \varepsilon^{KS}\phi$. Here the KS-DFT Hamiltonian is (using atomic units, and treating closed-shell systems):

$$H_0 = -\frac{1}{2}\nabla^2 + v_{\rm nuc}[n_0] + v_{\rm H}[n_0] + v_{\rm xc}[n_0],$$

and we introduced the ground state density ($n_0(\mathbf{r})$) and the nuclear and exchange-correlation potentials, while the Hartree potential is $v_{\rm H}[n](\mathbf{r}) = \int \nu(\mathbf{r},\mathbf{r}')\, n(\mathbf{r}')d\mathbf{r}'$ with $\nu(\mathbf{r},\mathbf{r}') = |\mathbf{r}-\mathbf{r}'|^{-1}$. In the diagonal approximation, the associated QP energy fulfills:[4]

$$\varepsilon^{QP} = \varepsilon^{KS} + \left\langle \phi \left| X + \Sigma\left(\omega = \varepsilon^{QP}\right) - v_{\rm xc} \right| \phi \right\rangle, \qquad (1)$$

where X is the Fock exchange-operator and Σ refers throughout to the polarization self-energy.

The frequency-resolved matrix element of the polarization self energy is obtained from the time-dependent form, $\langle \phi | \Sigma(\omega) | \phi \rangle \equiv \int \langle \phi | \Sigma(t) | \phi \rangle\, e^{-\frac{1}{2}\gamma^2 t^2} e^{i\omega t} dt$ where γ is an energy broadening term for converging the time integration.[23] The required polarization self-energy $\Sigma(t)$ has a very simple direct product form in the GW approximation:[4]

$$\Sigma(\mathbf{r},\mathbf{r}',t) = iG(\mathbf{r},\mathbf{r}',t)W(\mathbf{r},\mathbf{r}',t), \qquad (2)$$

where G is the Green's function (detailed below), and W the effective polarization interaction. We use here the one-shot G_0W_0 approximation, but omit the 0 subscript throughout, as well as the P (polarization) subscript on Σ and W. Despite its elegance, it is expensive to directly calculate $\langle \phi | \Sigma(t) | \phi \rangle$ using Eq. (2) and the first goal of stochastic GW is to convert the direct product to an initial value expression as detailed below.

III. STOCHASTIC PARADIGM FOR GW: GENERAL DISCUSSION

A. Overview

The general deterministic expression, as reviewed above, involves three main steps, i.e., the evaluation of two matrices (or formally operators) – the Green's function and the polarization interaction, and the conversion of the causal (retarded) polarization interaction W^R (discussed later) to the time-ordered one W.

In deterministic GW the last step is trivial but the first two are quite expensive since they involve the construction of frequency dependent matrices so the overall method scales practically as $\mathrm{O}(N_e^3) - \mathrm{O}(N_e^4)$ with a large prefactor related to the number of frequencies needed, the grid size needed to represent the basis-functions, etc.

The stochastic paradigm views GW quite differently. In essence, it converts operators into correlation functions. Thus, the Green's function is replaced, as proved later, by a correlation function between a random function at time 0 and an evolved random function at a later time. This paradigm is quite intuitive, in that physically, for large systems, the dynamics is fast so one should not be required to know the detailed energy resolved eigenstates over the whole occupied and unoccupied manifold for getting the correction to the HOMO/LUMO states.

For W, the stochastic paradigm plays out in an even more interesting fashion. In the usual deterministic paradigm it is well-known that the action of the retarded effective action W^R is evaluated most efficiently by perturbing the sea of occupied states and letting all the perturbed occupied states evolve in time. For large systems, therefore, a large number of propagations is needed. But in the stochastic paradigm one takes just a few random combination of all these states and propagates them. For example, a system with 10,000 electrons would typically be represented by only about 8 [*sic!*] random combination of occupied states that are propagated; conceptually, think of each of these states as carrying about one eighth of the total electronic charge, i.e., as if they represent an artificial "particle" with a charge of $\approx 1250e^-$! It is almost counterintuitive that such a simplification would work, but as we have tested it works better and better as the system gets larger, due to self-averaging. Heuristically, the response of W is collective, plasmon-like, and the stochastic method excels at describing such collective properties.

In addition to the stochastic resolution of the G and W^R, the stochastic paradigm plays out in the conversion from the action of W^R to that of W. The required procedure, detailed later, could be done directly but would then require many terabytes of storage. To circumvent it, we have used what could be labeled as stochastic compression – i.e., spanning the detailed space-dependent function by stochastic vectors. Here we improve the method by developing sparse stochastic compression, where each stochastic

function is defined over only a small number of space points. Here also a counterintuitive point emerges – the quality of the stochastic compression is independent of the size of each sparse vector. Therefore, a large number of very sparse vectors can be used – with little numerical effort (as each vector involves only a small number of grid points) but with high fidelity (due to the large number of such vectors).

In the remainder of the section we outline the basis premise of the stochastic technique, the stochastic resolution of identity (SRI); in the following sections SRI is used to develop the different ingredients in stochastic GW.

B. Stochastic resolution of identity

Our starting point is a set of random functions on a grid, each labeled $\bar{\zeta}(\mathbf{r})$. The simplest choice is real discrete stochastic functions that have a random sign at each point:

$$\bar{\zeta}(\mathbf{r}) = \pm(dV)^{-\frac{1}{2}}$$

(dV is the grid volume element). The stochastic functions fulfill $\{\bar{\zeta}(\mathbf{r})\bar{\zeta}(\mathbf{r}')\} = (dV)^{-1}\delta_{\mathbf{r},\mathbf{r}'}$, where $\delta_{\mathbf{r},\mathbf{r}'}$ is a Kronecker delta and the $\{\cdots\}$ refers to a statistical average over all stochastic functions. This implies a resolution of the identity relation, $\mathcal{I} = \{|\bar{\zeta}\rangle\langle\bar{\zeta}|\}$. In practice we need to use a finite number (labeled $N_{\bar{\zeta}}$) of stochastic functions and the resolution becomes approximate

$$\mathcal{I} \simeq \frac{1}{N_{\bar{\zeta}}}\sum_{\bar{\zeta}} |\bar{\zeta}\rangle\langle\bar{\zeta}|. \tag{3}$$

IV. STOCHASTIC PARADIGM FOR RESOLVING G

In this section we review the use of the general stochastic paradigm for the first part of the GW formalism, i.e., a stochastic representation of the zero-order Green's function. As we show below the Green's function becomes a stochastic average correlation function, whereby a random vector is correlated to the same vector when projected to the occupied (or unoccupied) space and propagated in time.

A. Separable expression for the Green's function

It is easy to show that the Kohn-Sham Green's function is given by the operator form $iG(t) = e^{-iH_0 t}\left((\mathcal{I}-\mathcal{P})\theta(t) - \mathcal{P}\theta(-t)\right)$ where the projection operator to the $N_{\rm occ}$ occupied states is $\mathcal{P} = \sum_{n\le N_{\rm occ}} |\phi_n\rangle\langle\phi_n|$. To make a separable form, multiply $iG(t)$ by Eq. (3), leading to:

$$iG(\mathbf{r},\mathbf{r}',t) \simeq \frac{1}{N_{\bar{\zeta}}}\sum_{\bar{\zeta}} \zeta(\mathbf{r},t)\bar{\zeta}(\mathbf{r}'), \tag{4}$$

where $|\zeta(t)\rangle \equiv iG(t)|\bar{\zeta}\rangle$. Eq. (4) is the main ingredient of stochastic GW, reformulating the Green's function as a sum over separable terms.

To evaluate $|\zeta(t)\rangle$, start with the negative-time Green's function which is a propagator of the occupied states, $iG(t<0) = -e^{-iH_0 t}\mathcal{P}$, so:

$$|\zeta(t<0)\rangle = -e^{-iH_0 t}|\zeta^v\rangle, \tag{5}$$

where we define a stochastic occupied (valence) state $|\zeta^v\rangle = \mathcal{P}|\bar{\zeta}\rangle$. Similarly for positive times the Green's function is the propagator of unoccupied (conduction) states, $iG(t>0) = e^{-iH_0 t}(\mathcal{I}-\mathcal{P})$, so:

$$|\zeta(t>0)\rangle = e^{-iH_0 t}|\zeta^c\rangle, \tag{6}$$

where $|\zeta^c\rangle = (\mathcal{I}-\mathcal{P})|\bar{\zeta}\rangle = |\bar{\zeta}-\zeta^v\rangle$.

B. Projective filtering

The next stage is therefore to calculate $\mathcal{P}|\bar{\zeta}\rangle$. Previously we used Chebyshev filtering which scales linearly with system size, but with a large prefactor. Therefore as long as the occupied states are available (i.e., for systems with up to tens of thousands of electrons) it is faster to use projective filtering, i.e.,

$$\zeta^v(\mathbf{r}) = \langle\mathbf{r}|\mathcal{P}\bar{\zeta}\rangle = \sum_{n\le N_{\rm occ}} \phi_n(\mathbf{r})\langle\phi_n|\bar{\zeta}\rangle. \tag{7}$$

The cost of Eq. (7) is the number of occupied states ($N_{\rm occ}$), times the number of grid point (N_g), times the number of stochastic functions ($N_{\bar{\zeta}}$). The first two factors are proportional to the system size; however, as we show later, for a given accuracy, the number of required stochastic states decreases (or at least does not increase) with system size. Therefore the scaling of Eq. (7) is at most $\mathrm{O}(N^2)$, and even that part has a small prefactor so the dominant scaling is quasi-linear, as detailed below.

Once $\zeta^v(\mathbf{r})$ is prepared, the time-dependent orbitals of Eqs. (5) and (6) are evaluated by a Trotter (split-operator) propagation, $|\zeta(t\pm dt)\rangle = e^{\mp iH_0 dt}|\zeta(t)\rangle$, for positive or negative times respectively.

C. Separable expression for $\langle\Sigma\rangle$

Given Eq. (2) and the separable form of Eq. (4) it immediately follows that

$$\langle\phi|\Sigma\left(t\right)|\phi\rangle\simeq\frac{1}{N_{\bar{\zeta}}}\sum_{\bar{\zeta}}\int\phi\left(\mathbf{r}\right)\zeta\left(\mathbf{r},t\right)u\left(\mathbf{r},t\right)d\mathbf{r}, \quad (8)$$

where

$$u\left(\mathbf{r},t\right)=\int W\left(\mathbf{r},\mathbf{r}',t\right)\bar{\zeta}(\mathbf{r}')\phi(\mathbf{r}')d\mathbf{r}'. \quad (9)$$

V. ACTING WITH THE RETARDED POLARIZATION POTENTIAL

To calculate $u\left(\mathbf{r},t\right)$ in Eq. (9), one needs to act with $W\left(\mathbf{r},\mathbf{r}',t\right)$ on the product $\bar{\zeta}(\mathbf{r}')\phi(\mathbf{r}')$. This will be done in two stages: First, we will calculate the action of the retarded (causal) effective-interaction:

$$u^{R}\left(\mathbf{r},t\right)=\int W^{R}\left(\mathbf{r},\mathbf{r}',t\right)\bar{\zeta}(\mathbf{r}')\phi(\mathbf{r}')d\mathbf{r}'. \quad (10)$$

Physically, $u^{R}\left(\mathbf{r},t\right)$ is the effective potential at time t due to a stochastic external charge $(\bar{\zeta}\phi)$ applied at time 0.

The next section then explains how to convert the causal $u^{R}\left(\mathbf{r},t\right)$ function to the time-ordered one $u\left(\mathbf{r},t\right)$.

A. Deterministic W^R

It is well-known (Ref. 7 and 24) that linear-response TDH can be used to calculate the action of W^R on any initial function. In our context, this amounts to perturbing all occupied states,

$$\phi_{n}^{\lambda}(\mathbf{r},t=0)=e^{-i\lambda v_{\mathrm{pert}}(\mathbf{r})}\phi_{n}(\mathbf{r}),\quad n\leq N_{\mathrm{occ}}$$

where λ is small (typically $10^{-4}E_h^{-1}$) and $v_{\mathrm{pert}}(\mathbf{r})\equiv\int\nu\left(\mathbf{r},\mathbf{r}'\right)\bar{\zeta}(\mathbf{r}')\phi(\mathbf{r}')d\mathbf{r}'$. Then one propagates simultaneously all occupied states, $|\phi_{n}^{\lambda}\left(t+dt\right)\rangle=e^{-iH^{\lambda}(t)dt}|\phi_{n}^{\lambda}\left(t\right)\rangle$ using a time-dependent Hamiltonian:

$$H^{\lambda}(t)=H_{0}+v_{\mathrm{H}}^{\lambda}(\mathbf{r},t)-v_{\mathrm{H}}(\mathbf{r}), \quad (11)$$

where $v_{\mathrm{H}}^{\lambda}(\mathbf{r},t)\equiv v_{\mathrm{H}}[n^{\lambda}(t)](\mathbf{r})$, $v_{\mathrm{H}}(\mathbf{r})\equiv v_{\mathrm{H}}[n_{0}](\mathbf{r})$ and $n^{\lambda}(\mathbf{r},t)$ is the density associated with these occupied orbitals (including the spin factor), $n^{\lambda}(\mathbf{r},t)=2\sum_{n\leq N_{\mathrm{occ}}}|\phi_{n}^{\lambda}\left(\mathbf{r},t\right)|^{2}$. The causal response of Eq. (10) is then

$$u^{R}\left(\mathbf{r},t\right)=\frac{v_{\mathrm{H}}^{\lambda}(\mathbf{r},t)-v_{\mathrm{H}}(\mathbf{r})}{\lambda}. \quad (12)$$

B. Stochastic W^R

Deterministic linear-response TDH is expensive for large systems since all occupied states are propagated. We have therefore developed and applied a very cheap alternative, stochastic TDH.[19, 25, 26] Each time we choose a random vector $|\bar{\zeta}\rangle$ we also choose and propagate a small set ($N_\eta\sim5-30$) of occupied stochastic functions formally defined as:

$$\eta_{l}(\mathbf{r})=\sum_{n\leq N_{\mathrm{occ}}}\eta_{nl}\phi_{n}(\mathbf{r}),\quad l=1,...,N_{\eta}, \quad (13)$$

where the coefficients can be real or complex, and are either specified directly (e.g., $\eta_{nl}=\pm1$) or based on a projection of a random vector $\bar{\eta}_{l}(\mathbf{r})$, i.e., $|\eta_{l}\rangle=P|\bar{\eta}_{l}\rangle$ (see Ref. 27).

Once the occupied stochastic functions are prepared then, completely analogously to the deterministic case, they are perturbed

$$\eta_{l}^{\lambda}(\mathbf{r},t=0)=e^{-i\lambda v_{\mathrm{pert}}(\mathbf{r})}\eta_{l}(\mathbf{r}), \quad (14)$$

and propagated, $\eta_{l}^{\lambda}\left(t+dt\right)\rangle=e^{-iH^{\lambda}(t)dt}|\eta_{l}^{\lambda}\left(t\right)\rangle$, and the time-dependent Hamiltonian is constructed again using Eq. (11) but now the Hartree potential $v_{\mathrm{H}}^{\lambda}(\mathbf{r},t)$ is based on the density of the propagated stochastic-occupied orbitals,

$$n^{\lambda}(\mathbf{r},t)=C_{\mathrm{norm}}\frac{2}{N_{\eta}}\sum_{l\leq N_{\eta}}\left|\eta_{l}^{\lambda}(\mathbf{r},t)\right|^{2}, \quad (15)$$

where C_{norm} is a normalization constant ensuring the correct total number of electrons ($\int n^{\lambda}(\mathbf{r},t)d\mathbf{r}=N_e$). Also note that unlike the deterministic case it is necessary now to repeat the calculation with $\lambda=0$ and the response is then the difference of the time-dependent potentials: [28]

$$u^{R}\left(\mathbf{r},t\right)=\frac{v_{\mathrm{H}}^{\lambda}(\mathbf{r},t)-v_{\mathrm{H}}^{\lambda=0}(\mathbf{r},t)}{\lambda}. \quad (16)$$

To conclude this section note that while the occupied $\eta_l(\mathbf{r})$ functions are used to make the effective interaction W^R that acts on $|\bar{\zeta}\rangle$, they are not by themselves a function of $|\bar{\zeta}\rangle$, except in the sense that, to improve the statistical accuracy, a different set of $\eta_l(\mathbf{r})$ is chosen for each choice of $|\bar{\zeta}\rangle$. It is actually important to choose stochastic orbitals that are independent of each other (i.e., η_l and ζ that are chosen independently) to avoid numerical bias – in the same way that in deterministic nested summations the indices are independent of each other.

VI. SPARSE STOCHASTIC ORBITALS AND THE CAUSAL TO TIME-ORDERED TRANSFORMATION

W and W^R are related in Fourier space – they are the same for positive frequencies and are complex-conjugates

at negative frequencies.[29] The same properties are true for u and u^R, as long as the source term $(\bar{\zeta}\phi)$ in Eq. (10) is real. Practically, this gives a recipe for obtaining u from u^R, which we label as $u = \mathcal{T}(u^R)$, meaning: FFT u^R from time to frequency, conjugate at negative frequencies and inverse FFT the result back to time

$$\begin{aligned} u^R(\mathbf{r},t) \rightarrow u^R(\mathbf{r},\omega) &= \int_0^\infty e^{-\frac{1}{2}\gamma^2 t^2} e^{i\omega t} u^R(\mathbf{r},t) dt \\ \rightarrow u(\mathbf{r},\omega) &= \begin{cases} u^R(\mathbf{r},\omega) & \omega \geq 0 \\ \left(u^R(\mathbf{r},\omega)\right)^* & \omega < 0 \end{cases} \\ \rightarrow u(\mathbf{r},t) &= \frac{1}{2\pi}\int_{-\infty}^\infty e^{-i\omega t} u(\mathbf{r},\omega) d\omega. \end{aligned} \tag{17}$$

This procedure is, however, storage intensive since the whole $u^R(\mathbf{r},t)$ from each core needs to be stored on disk.

A. Stochastic basis

Our previous approach (Ref. 19) to solving the storage issue was based on a stochastic resolution of identity (Eq. (3)) inserted between ζ and $u(\mathbf{r},t)$ in Eq. 8. Specifically, a third set of random functions $\xi(\mathbf{r}) = \pm(dV)^{-0.5}$ was absorbed into definition of $u(\mathbf{r},t)$:

$$u(\mathbf{r},t) \simeq u_{\text{aprx}}(\mathbf{r},t) \equiv \frac{1}{N_\xi}\sum \xi(\mathbf{r}) u_\xi(t), \tag{18}$$

Here $u_\xi(t)$ are obtained by time-ordering $\left(u_\xi = \mathcal{T}(u_\xi^R)\right)$ the causal coefficients $u_\xi^R(t) \equiv \lambda^{-1}\left(v_\xi^\lambda - v_\xi^{\lambda=0}\right)$ where $v_\xi^\lambda = \langle\xi|v_{\text{H}}^\lambda(t)\rangle$ (see Eq. (16)).

In Appendix A, we prove that the accuracy of the stochastic expansion decreases with system size, unless N_ξ is increased. We previously (Refs. 19 and 20) used $N_\xi = 100 - 300$, but for the very large systems studied here N_ξ needs to be increased to avoid large statistical errors. For large N_ξ, however, the overlaps $\langle\xi|u^R(t)\rangle$ dominate the computational cost, altering the linear scaling with system size (see [30]).

B. Sparse stochastic basis:

In order to overcome the drawback of the stochastic basis representation we use random functions in Eq. (18) that are non-zero only over small number of points (labeled N_s) rather than extending over the full grid; we label them as a "sparse" stochastic basis. Note that the label "sparse" is used rather "local" since the points formally do need not be close to one another, even if in practice it is easiest numerically to have them contiguous.

The sparse-stochastic basis expansion is then:

$$u(\mathbf{r},t) \simeq \frac{L}{N_\xi}\sum_{\xi\in\text{sparse}} \xi(\mathbf{r}) u_\xi(t), \tag{19}$$

where $L \equiv \frac{N_g}{N_s}$ and typically we use $L \sim 100 - 1000$. Equivalently we can write Eq. (19) as a formal sparse-stochastic resolution of the identity, SSRI:

$$\mathcal{I} \simeq \frac{L}{N_\xi}\sum_{\xi\in\text{sparse}} |\xi\rangle\langle\xi|. \tag{20}$$

In Appendix B we develop this concept in more detail and show that the numerical error in SSRI is independent of the vectors length N_s as long as sufficiently many vectors are used. The total cost in the expansion is (for each time step) only $N_s N_\xi$, vs. $N_g N_\xi$ in the original stochastic expansion (Eq. (18)). Therefore, a much larger N_ξ $(10,000 - 50,000)$ can now be used keeping the error in Eq. (19) in check.

The use of a sparse stochastic basis completes the stochastic *GW* algorithm, which is detailed in Appendix C.

VII. SIMULATIONS AND RESULTS

The stochastic *GW* simulations were run on a large cluster.[31] The implementation is trivially parallelized with speedup efficiency that is always greater than 80%.

All simulations used uniform grids with isotropic spacing $dx = dy = dz$. For both molecules and periodic solids, the KS-LDA ground state was computed using Troullier-Martins pseudopotentials,[32] and a kinetic energy cutoff of 28 E_h. For molecules, the Martyna-Tuckerman approach [33] was used to avoid the effect of periodic images.

A. Finite systems

The new stochastic *GW* implementation was first tested on acenes up to hexacene as well as a C_{60} molecule. Table I lists the parameters used for each system, and see also [34]. To isolate the influence of the number of stochastic TDH functions, N_η, we studied the QP energies of the set of molecules with deterministic and stochastic TDH propagation (the latter with $N_\eta = 16$). In both cases $N_{\bar{\zeta}}$ was increased till the resulting statistical error for the HOMO and LUMO QP energies is ≤ 0.05 eV. Fig. 1 shows that the stochastic and deterministic calculations require similar $N_{\bar{\zeta}}$, so the residual statistical error due to the use of stochastic TDH is small.The deterministic version scales quadratically with the size of the system so as shown in Fig. 1 it quickly becomes much more expensive than a constant-N_η fully stochastic treatment. Beyond tetracene the CPU time of the fully stochastic approach (with a constant N_η) scales linearly with a slope of less than 2 core-hours per electron.

Further, for large systems the number of propagated stochastic orbitals N_η can be reduced without increasing the stochastic error. This is illustrated for C_{60} where

System	N_e	N_g	N_η	HOMO	LUMO
Benzene	30	$(48)^3$	16	-9.18 ±0.09	0.73 ±0.09
Naphtalene	48	$48 \cdot 52 \cdot 60$	16	-8.12 ±0.09	-0.60 ±0.09
Tetracene	84	$48 \cdot 52 \cdot 82$	16	-6.82 ±0.08	-1.80 ±0.06
Pentacene	102	$48 \cdot 52 \cdot 94$	16	-6.49 ±0.07	-2.28 ±0.06
Hexacene	120	$48 \cdot 52 \cdot 104$	16	-6.18 ±0.06	-2.42 ±0.06
C_{60}	240	$(88)^3$	8	-7.80 ±0.04	-3.27 ±0.04
			16	-7.78 ±0.04	-3.30 ±0.04

TABLE I. Estimated QP energies (eV) for a set of finite systems with a fully stochastic approach. The calculations used $dx = 0.35a_0$, $N_{\bar{\zeta}} = 750$, $N_\xi = 20000$, and each sparse stochastic function extended over only $L^{-1} = 1\%$ of the total grid.

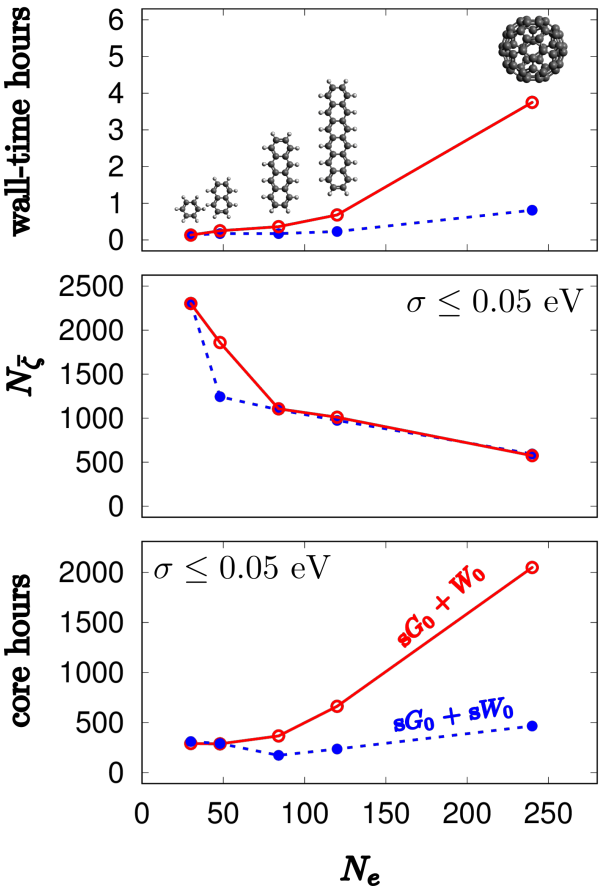

FIG. 1. Resources needed to reduce the QP energy stochastic error to 0.05 eV for acenes and C_{60}. Top panel: Wall time (hours); the dashed and solid lines refer to deterministic and stochastic (with $N_\eta = 16$) TDH propagations. Middle panel: Number of stochastic vectors $N_{\bar{\zeta}}$; Bottom panel: required CPU core hours. All calculations used $N_\xi = 20,000$ and $L^{-1} = 1\%$.

$N_\eta = 8$ and $N_\eta = 16$ (Table I) give an almost identical stochastic error.

.

N_{cells}	N_e	N_g	$E_g\,(eV)$			
			Diamond		Silicon	
8	256	$(42)^3$	5.36	±0.09	1.17	±0.06
27	864	$(60)^3$	5.28	±0.07	1.35	±0.05
64	2048	$(80)^3$	5.40	±0.06	1.29	±0.04
216	6912	$(120)^3$	5.55	±0.04	1.24	±0.04
343	10978	$(140)^3$	5.51	±0.04	1.24	±0.03

TABLE II. Estimated QP gaps for bulk carbon and silicon using $N_{\bar{\zeta}} = 400$, $N_\eta = 8$, $N_\xi = 20,000$ and $L = 100$. N_{cells} is the number of conventional cells in a supercell, N_e the total number of valence electrons, and N_g is the total number of grid points.

B. Periodic solids

We next studied the performance of stochastic *GW* for periodic systems employing large real space grids (equivalent to Γ-point sampling of large supercells in planewave codes). Specifically, we studied the scaling of stochastic GW for silicon and diamond supercells including several unit cells with lattice constants taken from experiment.[35, 36]; for details see [37]. Although the systems were large, most time was still spent on the TDH stage, as mentioned earlier [38].

We generally used $N_\eta = 8$ propagated stochastic orbitals for periodic systems. Higher values barely change the predicted QP energies (by 0.01 eV or less). Table II, obtained with a fixed $N_{\bar{\zeta}} = 400$, shows that the stochastic error of E_g (the gap between the bottom of the conduction band and the top of the valance band) decreases rapidly with system size. Further, the number of stochastic vectors $N_{\bar{\zeta}}$ required to decrease the error below 0.05 eV is plotted in Fig. 2. The lower panel shows that the total CPU time then scales at worst linearly with N_e. Specifically, calculations for diamond and silicon supercells with 10978 valence electrons consumed only about 1900 and 1000 core hours!

Per-electron the periodic calculations were much faster (up to almost 20 times!) than for finite systems. One obvious reason is that it is much easier to pack electrons in a periodic system, so, for example, the largest supercell of silicon or diamond has 50 times more electrons than C_{60} but its grid is only ~ 4 times bigger. In addition, the large periodic systems have much more electrons so they required fewer samples ($N_{\bar{\zeta}}$).

VIII. DISCUSSION AND CONCLUSIONS

Our calculations show very favorable scaling of the statistical error in all three types of stochastic samplings: Sampling of the Green's function, the stochastic TDH, and SSRI for efficient sampling and converting the causal to time-ordered quantity. Specifically

- The stochastic error depends on the number of vectors used to sample the Green's function, $N_{\bar{\zeta}}$. In-

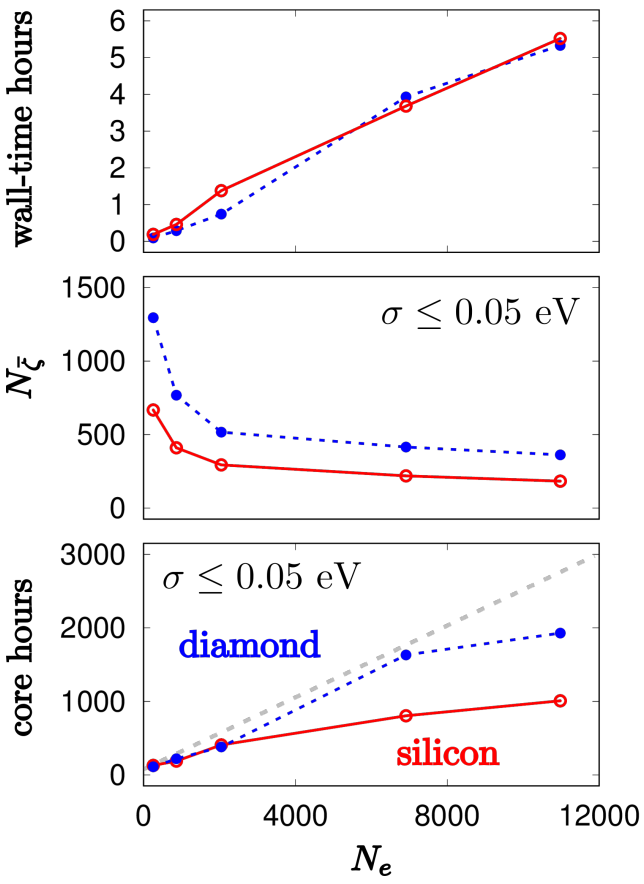

FIG. 2. Resources required to lower the stochastic error in E_g to 0.05 eV for silicon and diamond supercells (red and blue, respectively). All calculations used $N_\eta = 8$, $N_\xi = 20,000$ and $L^{-1} = 1\%$. Top: CPU hours; Middle: Required $N_{\bar{\zeta}}$; and bottom: Total CPU core hours.

terestingly, to obtain a low error of 0.05 eV in the quasiparticle energies, it is enough to sample the Green's function with circa 2000 vectors for small systems and this number even decreases with system size so for C_{60} it is only 600 and for large periodic supercells it decreases to a few hundreds. The reason for this is most likely self-averaging, i.e., for big systems the contributions of different regions average. It is very likely for for systems with defects and gap states there would be less self averaging so $N_{\bar{\zeta}}$ should not decrease with system size.

- Similarly, the method benefits from the remarkable fact that very few propagated stochastic orbitals η are needed for the TDH propagation. We used $N_\eta = 8$ for large molecules and and periodic solids, i.e., in each simulation (i.e., for each choice of a random vector representing the Green's function), it is enough to propagate with only 8 vectors that represent the motion of the whole perturbed occupied electron sea!

- Finally, the main new ingredient in the paper is SSRI, that makes it possible to easily increase the number of such sampling vectors (N_ξ) by 100-fold or more, from hundreds to tens of thousands. The key feature is that the accuracy is independent of the size of the sparse stochastic vectors as long as each grid point is sufficiently sampled (i.e., as long as $L \ll N_\xi$). The SSRI expansion (Eq. (19)) adds only a tiny stochastic error (less than 0.01 eV) and its cost is negligible. Interestingly, SSRI, is potentially useful for a large number of applications that are unrelated to stochastic GW, including long-range exchange, stochastic MP2 (direct and exchange), and stochastic resolution of the identity.[27, 39, 40]

Taken together, we find a very favorable scaling. Ordered supercells with 10978 valence electrons require less than 2000 core hours (and only 6 wall hours) to yield QP energies with statistical errors below 0.05 eV. Our method thus makes it possible to calculate QP energies of extremely large systems with thousands of atoms on small computer clusters. The overall algorithm is straightforward, and an open-source software (StochasticGW) is freely available.[41]

The present method has two ingredients which formally scale non-linearly. As mentioned, we use occupied-projection, which scales as $O(N_e^2)$; this by itself however is not a major issue since it will not be the dominant part of the calculation until we would reach $N_e \gg 100,000$. But more importantly, in most DFT codes the extraction of the occupied states scales as $O(N_e^3)$ and is prohibitive for very large systems. We therefore anticipate that when simulating systems with $N_e > 50,000$ it may be necessary to switch back to the Chebyshev projection approach that avoids the eigenstates altogether, as long as the underlying DFT potential could be obtained by either linear scaling DFT [42–45] or stochastic DFT.[46, 47]

Finally, we note that the stochastic GW algorithm is not restricted to the common RPA-based G_0W_0 approximation which we focused on. Two simple variants can be trivially implemented. The first is replacing TDH by time-dependent DFT (TDDFT), [48, 49] which is equivalent to modifying the kernel W and tends to improve the LUMO energies;[19, 50]); the second is a post-processing stretched-scissors modification of the Green's function (see Ref. 51 for details).

ACKNOWLEDGMENTS

We are grateful for support by the Center for Computational Study of Excited State Phenomena in Energy Materials (C2SEPEM) 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. V. V. greatly appreciates helpful discussion with Gabriel Kotliar and

Mark Hybertsen. The calculations were performed as part of the XSEDE computational Project No. TG-CHE170058.[52]

APPENDIX A: STATISTICAL ERROR OF A STOCHASTIC BASIS EXPANSION

Given a stochastic expansion of a general function, analogous to Eq. (18),

$$|f\rangle \simeq |f_{\rm aprx}\rangle \equiv \frac{1}{N_\xi}\sum |\xi\rangle\langle\xi|f\rangle, \tag{A.1}$$

we show here that the average relative error in the representation of f is proportional to the number of grid points. Specifically, define

$$\begin{aligned}\sigma^2(f) &= \{\langle f_{\rm aprx}|f_{\rm aprx}\rangle\} - \langle f|f\rangle \\ &= \frac{1}{N_\xi^2}\left\{\sum_{\xi,\xi'}\langle\xi|\xi'\rangle\langle f|\xi\rangle\langle\xi'|f\rangle\right\} - \langle f|f\rangle,\end{aligned} \tag{A.2}$$

where all functions are assumed real. Separating yields

$$\sigma^2(f) = J_1 + J_2 - \langle f|f\rangle. \tag{A.3}$$

Here J_1 is the $\xi=\xi'$ contribution

$$\begin{aligned}J_1 &= \frac{1}{N_\xi^2}\left\{\sum_{\xi'=\xi}\langle\xi|\xi\rangle\langle f|\xi\rangle\langle\xi|f\rangle\right\} \\ &= \frac{N_g}{N_\xi^2}\left\{\sum_{\xi}\langle f|\xi\rangle\langle\xi|f\rangle\right\} = \frac{N_g}{N_\xi^2}\left\{\sum_{\xi}\langle f|\xi\rangle\langle\xi|f\rangle\right\},\end{aligned}$$

where the definition $\xi(\mathbf{r}) = \pm(dV)^{-0.5}$ implies that $\langle\xi|\xi\rangle = N_g$ (always, not just as an average). The resulting expression for J_1 simply involves a resolution of the identity $\{|\xi\rangle\langle\xi|\} = I$, so

$$J_1 = \frac{N_g}{N_\xi^2}\sum_\xi\langle f|f\rangle = \frac{N_g}{N_\xi^2}N_\xi\langle f|f\rangle. \tag{A.4}$$

Similarly, J_2 is the $\xi'\neq\xi$ contribution

$$J_2 = \frac{1}{N_\xi^2}\left\{\sum_\xi\sum_{\xi'\neq\xi}\langle\xi|\xi'\rangle\langle\xi'|f\rangle\langle f|\xi\rangle\right\},$$

and since the condition $\xi'\neq\xi$ does not restrict ξ', the resolution of the identity $I=\{|\xi'\rangle\langle\xi'|\}$ is still valid, so

$$\begin{aligned}J_2 &= \frac{1}{N_\xi^2}\left\{\sum_\xi\sum_{\xi'\neq\xi}\langle\xi|f\rangle\langle f|\xi\rangle\right\} \\ &= \frac{1}{N_\xi^2}\sum_\xi\sum_{\xi'\neq\xi}\langle f|f\rangle = \frac{\langle f|f\rangle}{N_\xi^2}N_\xi\left(N_\xi-1\right).\end{aligned}$$

Adding the terms gives

$$\frac{\sigma^2(f)}{\langle f|f\rangle} = \frac{(N_g-1)}{N_\xi} \simeq \frac{N_g}{N_\xi}, \tag{A.5}$$

as stipulated.

APPENDIX B: SPARSE STOCHASTIC EXPANSION: EXAMPLE AND DETAILS

A simple example clarifies the concept of sparse stochastic orbitals.

Break the N_g grid points to two sets A, B, each with $N_g^A = N_g^B = \frac{1}{2}N_g$ points. Apply the stochastic resolution again with N_ξ functions, but now the first half of the functions (ξ_A) are non-zero only over the A-set points, and the other half are non vanishing over the B set. Then, in an obvious notation:

$$u(\mathbf{r},t) \simeq \begin{cases}\frac{1}{N_g^A}\sum_{\xi_A}\xi_A(\mathbf{r})u_{\xi,A}(t) & \mathbf{r}\in A \\ \frac{1}{N_g^B}\sum_{\xi_B}\xi_B(\mathbf{r})u_{\xi,B}(t) & \mathbf{r}\in B,\end{cases} \tag{B.1}$$

where $u_{\xi,A}(t) \equiv \langle\xi_A|u\rangle_A \equiv dV\cdot\sum_{\mathbf{r}\in A}\xi_A(\mathbf{r})u(\mathbf{r})$ and analogously for B.

The cost of calculating each $u_{\xi,A}(t)$ is half that of calculating the original $u_\xi(t)$, since the summation includes half the grid points. But the squared standard deviation of u is unchanged!

$$\begin{aligned}\sigma^2(u) &= \sigma_A^2(u) + \sigma_B^2(u) = \\ &= \frac{N_g^A\langle u|u\rangle_A}{\frac{1}{2}N_\xi} + \frac{N_g^B\langle u|u\rangle_B}{\frac{1}{2}N_\xi} = \frac{N_g\langle u|u\rangle}{N_\xi},\end{aligned}$$

where we used $\langle u|u\rangle_A + \langle u|u\rangle_B = \langle u|u\rangle$. This implies that the use of Eq. (18) instead of Eq. (B.1) reduces the numerical effort by a factor of two without affecting the statistical error.

Obviously, we could continue with this process of using smaller and smaller segments further, resulting in Eqs. (19),(20). In practice, we pick here a small segment size $N_s \sim 0.001N_g - 0.01N_g$, so that the ratio of total grid length and the segment length is $L \approx 100-1000$. For simplicity, we do not even require the segments to be non-overlapping. The only requirement is to ensure that each point has the same L^{-1} probability to be sampled, i.e., to have a sparse basis function that includes it.

Note that:

1. The segments need to sufficiently sample each point; each grid point has a probability L^{-1} of being sampled by each of the N_ξ functions so it is important to have $1 \ll L^{-1}N_\xi$, i.e., $L \ll N_\xi$.. Put differently, the segment size cannot be too small.

2. If a segment starting point is chosen near the first or last point in the grid, then either the function should be wrapped (so a portion of the segment is near the end of the grid and another portion is near the beginning of the grid) or it should be padded (at the beginning or end) with zeros, to guarantee that all points are equally sampled.

3. As mentioned, one could envision (although we have not done it here) that each segment would be non-contiguous, i.e., made from N_s random points from the full grid. We do not even have to ensure that the points in each segment are all different from each other, as long as they are randomly selected!

APPENDIX C: FINAL *GW* ALGORITHM SUMMARY AND COSTS

The final stochastic *GW* algorithm is simple:

Choose $N_{\bar{\zeta}} \sim 200-2000$ stochastic functions (the wall time is minimized if $N_{\bar{\zeta}}$ CPU cores are used, i.e., one per $\bar{\zeta}$). A larger $N_{\bar{\zeta}}$ reduces the stochastic errors. Then, for each choice of $\bar{\zeta}$:

1. Choose a set of $N_\xi \sim 5,000-50,000$ fractured random functions $\xi(\mathbf{r})$, each with $\frac{N_g}{L}$ grid points. Typically $L \sim 100-1000$.

2. Choose a set of $N_\eta \sim 5-30$ stochastic-occupied functions $\eta_l(\mathbf{r})$ (Eq. (13)).

3. Calculate $v_{\rm pert}(\mathbf{r})$ and perturb the $\eta_l(\mathbf{r})$ per Eq. (14).

4. Propagate the perturbed $\eta_l^\lambda(\mathbf{r},t)$ calculating at each time step $v_{\rm H}^\lambda(\mathbf{r},t)$ and constructing and storing in memory the set of $v_\xi^\lambda(t)$.

5. Repeat Step 4 for unperturbed functions (using $\lambda = 0$), storing $v_\xi^{\lambda=0}(t)$ along the propagation. Then at the end of the propagation calculate $u_\xi^R(t)$ and apply a time-ordering operation $u_\xi = \mathcal{T}(u_\xi^R)$ (analogous to Eq. (17)).

6. Then calculate $\zeta(\mathbf{r},t)$ for negative and positive times (Eqs. (5) and (6)) and use with Eq. (19) to accumulate the matrix element of the self-energy (Eqs. (8) and (9)).

Once steps 1-6 are finished average the resulting $\langle\phi|\Sigma(t)|\phi\rangle$ from each core, Fourier transform the result and solve Eq. (1).

The algorithm above, using stochastic TDH, is the most efficient version for large systems. If deterministic TDH is used, the steps are similar except that instead of the stochastic occupied states η_l one perturbs and propagates the deterministic occupied states $\phi_n(\mathbf{r})$ (and then there is no need to calculate $v_\xi^{\lambda=0}(t)$, which is obtained directly from the ground state density $n(\mathbf{r})$).

Most of the cost in the fully stochastic*GW* algorithm is associated with propagating the N_η vectors in time $\eta_l^\lambda(\mathbf{r},t)$ with and without perturbation. The effort is essentially proportional to the number of stochastic samples $N_{\bar{\zeta}}$, the number of stochastic-occupied functions (N_η), number of grid points (N_g), and the number of time steps N_τ, with a prefactor associated with the calculation of the kinetic energy operator at each time steps. The only one of this factors which increases with system size is N_g, since, as shown in the results section the necessary number of stochastic functions (N_η, $N_{\bar{\zeta}}$) does not increase with system size. Therefore this CPU-dominant part scales at worst linearly with N_e.

[1] P. Hohenberg and W. Kohn, Phys. Rev. **136**, 864 (1964).
[2] R. M. Dreizler and E. K. U. Gross, *Density Functional Theory: An Approach to the Quantum Many-Body Problem* (Springer Science & Business Media, 1990).
[3] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods* (Cambridge University Press, 2004) p. 624.
[4] F. Aryasetiawan and O. Gunnarsson, Reports Prog. Phys. **61**, 237 (1998).
[5] I. Shavitt, Mol. Phys. **94**, 3 (1998).
[6] C. D. Sherrill and H. F. Schaefer III, in *Advances in quantum chemistry*, Vol. 34 (Elsevier, 1999) pp. 143–269.
[7] R. M. Martin, L. Reining, and D. M. Ceperley, *Interacting Electrons* (Cambridge University Press, 2016).
[8] D. Rowe, Rev. Mod. Phys. **40**, 153 (1968).
[9] J. F. Stanton and R. J. Bartlett, J. Chem. Phys. **98**, 7029 (1993).
[10] A. I. Krylov, Annu. Rev. Phys. Chem. **59** (2008).
[11] L. Hedin, Phys. Rev. **139**, A796 (1965).
[12] M. S. Hybertsen and S. G. Louie, Phys. Rev. B **34**, 5390 (1986).
[13] P. Umari, G. Stenuit, and S. Baroni, Phys. Rev. B **79**, 201104 (2009).
[14] P. Umari, G. Stenuit, and S. Baroni, Phys. Rev. B **81**, 115104 (2010).
[15] J. Deslippe, G. Samsonidze, D. A. Strubbe, M. Jain, M. L. Cohen, and S. G. Louie, Comput. Phys. Commun. **183**, 1269 (2012).
[16] M. P. Ljungberg, P. Koval, F. Ferrari, D. Foerster, and D. Sanchez-Portal, Phys. Rev. B **92**, 075422 (2015).
[17] M. Govoni and G. Galli, J. Chem. Theory Comput. **11**, 2680 (2015).
[18] J. Wilhelm, D. Golze, L. Talirz, J. Hutter, and C. A. Pignedoli, The journal of physical chemistry letters **9**,

306 (2018).
[19] D. Neuhauser, Y. Gao, C. Arntsen, C. Karshenas, E. Rabani, and R. Baer, Phys. Rev. Lett. **113**, 076402 (2014).
[20] V. Vlček, E. Rabani, D. Neuhauser, and R. Baer, J. Chem. Theory Comput. **13**, 4997 (2017).
[21] V. Vlček, H. R. Eisenberg, G. Steinle-Neumann, D. Neuhauser, E. Rabani, and R. Baer, Phys. Rev. Lett. **116**, 186401 (2016).
[22] V. Vlček, E. Rabani, and D. Neuhauser, Phys Rev Mater **2**, 030801 (2018).
[23] Note that generally quantities in time and frequency use the same symbol, and the specifics are clear from the argument.
[24] L. Hedin, Journal of Physics: Condensed Matter **11**, R489 (1999).
[25] Y. Gao, D. Neuhauser, R. Baer, and E. Rabani, J. Chem. Phys. **142**, 034106 (2015).
[26] E. Rabani, R. Baer, and D. Neuhauser, Phys. Rev. B **91**, 235302 (2015).
[27] D. Neuhauser, E. Rabani, Y. Cytter, and R. Baer, J. Phys. Chem. A **120**, 3071 (2015).
[28] Note that this is not needed in the deterministic case where $v_{\rm H}^{\lambda=0}(\mathbf{r},t) = v_{\rm H}(\mathbf{r})$ does not change in time; but even without perturbation the stochastic TDH (or TDDFT) orbitals are not eigenstates and change in time leading to fluctuations in the density, so Eq. (16) is required to ensure that the response is indeed in the linear regime.
[29] A. L. Fetter and J. D. Walecka, *Quantum Thoery of Many Particle Systems* (McGraw-Hill, New York, 1971) p. 299.
[30] Note that this unfavorable scaling of the stochastic representation of u with system size (which is solved later by SSRI) is in contract to the favorable scaling of the stochastic representation of G and of W^R . We can only speculate that this stems, apparently, from the fact that G and W^R relate to global quantities (Green's function or plasmon-like response) with a time-dependence that is smooth with energy, while the general stochastic basis representation of u is unrelated to its physical properties.
[31] We used the Comet cluster with Intel Xeon E5-2680v3 processors (2.5 GHz clock speed), with up to 1728 cores on 144 CPUs. In all calculations reported here all 12 cores on each CPU were used.
[32] N. Troullier and J. L. Martins, Phys. Rev. B **43**, 1993 (1991).
[33] G. J. Martyna and M. E. Tuckerman, J. Chem. Phys. **110**, 2810 (1999).
[34] The uniform real-space grid spacing dx is sufficiently small to converge the LDA eigenvalues to < 10 meV. Further, the QP shifts are generally less sensitive to dx than the LDA eigenvalues. The damping parameter γ cannot be too high to avoid over-broadening. For finite systems, $\gamma = 0.1\,E_{\rm h}$ (cf., Eq. (17)) was sufficient to converge ε^{QP} (for a given $N_{\bar{\zeta}}$) to better than 0.01 eV, although we used an even more conservative value of $\gamma = 0.06\,E_{\rm h}$.
[35] T. Yamanaka and S. Morimoto, Acta Crystallogr., Sect. B: Struct. Sci **52**, 232 (1996).
[36] A. D. Elliot, Acta Crystallogr., Sect. B: Struct. Sci **66**, 271 (2010).
[37] The DFT eigenvalues were converged with respect to grid size to < 5 meV, with grid spacings of around 0.45 and 0.35 a_0 in all directions for silicon and diamond, respectively. As in the molecular case, an energy-broadening parameter of $\gamma = 0.06\,E_{\rm h}$ was sufficient for convergence.
[38] The initial projection and preparation of the stochastic occupied orbitals, $\bar{\zeta}(\mathbf{r})$ and $\eta_l(\mathbf{r})$, took at most 2% of the CPU time. In addition, the SSRI stage (converting u^R to u, Sec. VI B) took less than 5% of the total time when using $L = 100$ (so each sparse orbital covers only 1% of the grid) and $N_\xi = 20,000$. With these parameters the component of the stochastic error in the QP shifts due to the SSRI is tiny, less than 0.01 eV.
[39] D. Neuhauser, R. Baer, and D. Zgid, J. Chem. Theory Comput. **13**, 5396 (2017).
[40] T. Y. Takeshita, W. A. de Jong, D. Neuhauser, R. Baer, and E. Rabani, J. Chem. Theory Comput. **13**, 4605 (2017).
[41] The StochasticGW code is available under GPL at www.stochasticgw.com.
[42] W. Yang, Physical Review Letters **66**, 1438 (1991).
[43] E. Hernández and M. Gillan, Physical Review B **51**, 10157 (1995).
[44] S. Mohr, L. E. Ratcliff, L. Genovese, D. Caliste, P. Boulanger, S. Goedecker, and T. Deutsch, Physical Chemistry Chemical Physics **17**, 31360 (2015).
[45] J. VandeVondele, U. Borstnik, and J. Hutter, Journal of chemical theory and computation **8**, 3565 (2012).
[46] R. Baer, D. Neuhauser, and E. Rabani, Phys. Rev. Lett. **111**, 106402 (2013).
[47] D. Neuhauser, R. Baer, and E. Rabani, J. Chem. Phys. **141**, 041102 (2014).
[48] R. Del Sole, L. Reining, and R. Godby, Phys. Rev. B **49**, 8024 (1994).
[49] A. Grüneis, G. Kresse, Y. Hinuma, and F. Oba, Phys. Rev. Lett. **112**, 096401 (2014).
[50] L. Hung, F. H. da Jornada, J. Souto-Casares, J. R. Chelikowsky, S. G. Louie, and S. Öğüt, Phys. Rev. B **94**, 085125 (2016).
[51] V. Vlček, R. Baer, E. Rabani, and D. Neuhauser, arXiv preprint arXiv:1701.02023 (2017).
[52] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, *et al.*, Computing in Science & Engineering **16**, 62 (2014).