Stochastic density functional theory

Marcel David Fabian*
Fritz Haber Center for Molecular Dynamics and Institute of Chemistry,
The Hebrew University of Jerusalem, Jerusalem 9190401, Israel

Ben Shpiro*
Fritz Haber Center for Molecular Dynamics and Institute of Chemistry,
The Hebrew University of Jerusalem, Jerusalem 9190401, Israel

Eran Rabani†
Department of Chemistry, University of California, Berkeley,
California 94720, USA, and Materials Science Division,
Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA,
and The Sackler Center for Computational Molecular and Materials Science, Tel Aviv University, Tel Aviv 69978, Israel

Daniel Neuhauser‡
Department of Chemistry and Biochemistry, University of California, Los Angeles California 90095, USA

Roi Baer§
Fritz Haber Center for Molecular Dynamics and Institute of Chemistry,
The Hebrew University of Jerusalem, Jerusalem 9190401, Israel

Linear-scaling implementations of density functional theory (DFT) reach their intended efficiency regime only when applied to systems having a physical size larger than the range of their Kohn-Sham density matrix (DM). This causes a problem since many types of large systems of interest have a rather broad DM range and are therefore not amenable to analysis using DFT methods. For this reason, the recently proposed stochastic DFT (sDFT), avoiding exhaustive DM evaluations, is emerging as an attractive alternative linear-scaling approach. This review develops a general formulation of sDFT in terms of a (non)orthogonal basis representation and offers an analysis of the statistical errors (SEs) involved in the calculation. Using a new Gaussian-type basis-set implementation of sDFT, applied to water clusters and silicon nanocrystals, it demonstrates and explains how the standard deviation and the bias depend on the sampling rate and the system size in various types of calculations. We also develop basis-set embedded-fragments theory, demonstrating its utility for reducing the SEs for energy, density of states and nuclear force calculations. Finally, we discuss the algorithmic complexity of sDFT, showing it has CPU wall-time linear-scaling. The method parallelizes well over distributed processors with good scalability and therefore may find use in the upcoming exascale computing architectures.

I. INTRODUCTION

Density functional theory (DFT) is emerging as a usefully-accurate general-purpose computational platform for predicting from first principles the ground-state structure and properties of systems spanning a wide range of length scales, from single atoms and gas-phase molecules, through macromolecules, proteins, nanocrystals, nanosheets, nanoribbons, surfaces, interfaces up to periodic or amorphous homogeneous or heterogeneous materials [1–5]. Significant efforts have been diverted towards the development of numerical and computational methods enabling the use of DFT for studying extensive molecular systems. Several routes have been suggested: linear-scaling approaches [6–32], relying on the sparsity of the density matrix [33], DFT-based tight-binding (DFTB) methods [34–36] which reduce the numerical scaling using model Hamiltonians. Moreover, significant efforts have gone towards developing orbital-free DFT [36, 37] approaches using density-dependent kinetic energy functionals. The first two types of approaches mentioned above are designed to answer questions typically asked about molecules, while for materials and other large scale systems, we are more interested in coarse-grained properties. For example, with molecules, one is interested in bond orders, bond lengths spectral lines; while for large systems we are more interested in atomic densities, pair-correlation distributions (measured using neutron scattering) as well as charge/spin densities, polarizabilities and optical and electrical conductivity. In molecules, we strive to understand each occupied/unoccupied Kohn-Sham eigenstate while in large systems we are concerned with the density of hole and electron states.

Of course, detailed "molecular type" questions can also arise in large systems, primarily when the processes of interest occur in small pockets or localized regions — for example, biochemical processes in proteins, localized cat-

alytic events on a surface, impurities in solids, etc. Here, a combination of methods, where the small subsystem can be embedded in the larger environment is required.

In this advanced review, we will focus on the stochastic DFT (sDFT) approach, developed using grids and plane-waves in recent years [38–42] but also based on ideas taken from works starting in the early 1990's, mainly within the tight-binding electronic structure framework [43–50]. We make the point that the efficiency of sDFT results from its adherence to answering the coarse-grained "large system questions" mentioned above, rather than those asked for molecules.

The new viewpoint taken here is that of stochastic DFT using non-orthogonal localized basis-sets. The primary motivation behind choosing local basis-sets is that they are considerably more compact than plane-waves and therefore may enable studying significantly larger systems. Deterministic calculations using local basis-sets are more readily applicable to large systems, and thus can generate useful benchmarks with which the statistical errors and other properties characterizing sDFT can be studied in detail.

The review includes three additional sections, further divided into subsections, to be described later. Section II reviews the theory and techniques used for non-orthogonal sDFT and studies in detail the statistical errors and their dependence on sampling and system size. In section III we explain the use of embedded fragments and show their efficacy in reducing the stochastic errors of sDFT. Section IV summarizes and discusses the findings.

II. THEORY AND METHODS

In this section, we discuss three formulations of KS-DFT represented in non-orthogonal basis-sets. Since the issue of algorithmic scaling is at the heart of developing DFT methods for large systems, we emphasize for each formulation the associated algorithmic complexity (so-called system-size scaling). We start with the traditional basis-set formulation of the Kohn-Sham equations leading to standard cubic-scaling (subsection II A). Then, showing how, by focusing on observables and exploiting the sparsity of the matrices, a quadratic-scaling approach can be developed with no essential loss of rigor or accuracy (subsection II B). Most of the discussion will revolve around the third and final approach, stochastic DFT, which estimates expectation values using stochastic sampling methods, as described in subsection II C. This latter approach leads, to linear-scaling complexity.

A. Traditional basis-set formulation of Kohn-Sham equations with cubic scaling

The Kohn-Sham (KS) density functional theory (KS-DFT) is a molecular orbitals (MOs) approach which can be applied to a molecular system of N_e electrons using a basis-set of atom-centered orbitals $\phi_\alpha(\boldsymbol{r})$, $\alpha = 1, \ldots, K$. The basis functions were developed to describe the electronic structure of the parent atom, and for molecules they are the building blocks from which the orthonormal MOs are built as superpositions:

$$\psi_n(\boldsymbol{r}) = \sum_{\alpha=1}^{K} \phi_\alpha(\boldsymbol{r}) C_{\alpha n}, \, n = 1, \ldots, K. \qquad (1)$$

In the simplest "population" model, each MO can either "occupy" two electrons (of opposing spin) or be empty. The occupied MOs (indexed as the first $N_{occ} = N_e/2$ MOs) are used to form the total electron density:

$$n(\boldsymbol{r}) = 2 \times \sum_{n}^{N_{occ}} |\psi_n(\boldsymbol{r})|^2. \qquad (2)$$

The coefficient matrix C in Eq. (1) can be obtained from the variational principle applied to the Schrödinger equation, leading to the Roothaan-Hall generalized eigenvalue equations [51, 52] (we follow the notations in refs. [53–55]):

$$FC = SCE. \qquad (3)$$

Here, $F = T + V^{en} + J[n] + V^{xc}[n]$ is the $K \times K$ KS Fock matrix, $S_{\alpha\alpha'} = \langle \phi_\alpha | \phi_{\alpha'} \rangle$ is the overlap matrix of the AO's and E is a diagonal matrix containing the MO energies, $\varepsilon_1, \ldots, \varepsilon_K$. The Fock matrix $F_{\alpha\alpha'}$ includes the kinetic energy integrals, $T_{\alpha\alpha'} = \langle \phi_\alpha | -\frac{1}{2}\nabla^2 | \phi_{\alpha'} \rangle$, the nuclear attraction integrals $V_{\alpha\alpha'}^{en} = \langle \phi_\alpha | \hat{v}_{en} | \phi_{\alpha'} \rangle$, where $\hat{v}_{en}$ is the electron-nuclear interaction operator, the Coulomb integrals $J_{\alpha\alpha'} = \langle \phi_\alpha | v_H[n](\hat{\boldsymbol{r}}) | \phi_{\alpha'} \rangle$, where $v_H[n](\boldsymbol{r}) = \int \frac{n(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|} d^3 r'$ is the Hartree potential, and finally, the exchange-correlation integrals, $V_{\alpha\alpha'}^{xc} = \langle \phi_\alpha | v_{xc}[n](\hat{\boldsymbol{r}}) | \phi_{\alpha'} \rangle$ where $v_{xc}[n](\hat{\boldsymbol{r}})$ is the exchange correlation potential.

In KS theory, the Fock matrix F and the electron density $n(\boldsymbol{r})$ are mutually dependent on each other and must be obtained self-consistently. This is usually achieved by converging an iterative procedure,

$$\cdots \longrightarrow n(\boldsymbol{r}) \longrightarrow \{v_H[n](\boldsymbol{r}), v_{xc}[n](\boldsymbol{r})\} \longrightarrow \qquad (4)$$

$$\longrightarrow F \xrightarrow{O(K^3)} \{C, E\} \longrightarrow n(\boldsymbol{r}) \longrightarrow \ldots,$$

where in each iteration, a previous density iterate $n(\boldsymbol{r})$ is used to generate the Hartree $v_H[n](\boldsymbol{r})$ and exchange-correlation $v_{xc}[n](\boldsymbol{r})$ potentials from which we construct

the Fock matrix F. Then, by solving Eq. (3) the coefficient matrix C is obtained from which a new density iterate $n(\boldsymbol{r})$ is generated via Eqs. (1)-(2). The iterations continue until convergence (density stops changing with a predetermined threshold), and a *self-consistent field solution* is thus obtained.

This implementation of the basis-set based approach becomes computationally expensive for very large systems due to the cubic scaling of solving the algebraic Roothan-Hall equations (Eq. (3)). This cubic-scaling step is marked by placing $O\left(K^3\right)$ on the corresponding arrow in Eq. (4). The Coulomb integral calculation has a much lower scaling and can be completed in a $O\left(K \log K\right)$ scaling effort, either using continuous fast-multipole methods [56, 57] or fast-Fourier transforms on grids, as done here.

B. Equivalent trace-based formulation with quadratic scaling

In order to lower the scaling, we can take advantage of the fact that both F and S are very sparse matrices in the AO representation. The complication, however, is that the C matrix of Eq. (3) is non-sparse and therefore should be circumvented. This is challenging since the C matrix of Eq. (3) is used to extract both the eigenvalues ε_n and at the same time to enforce the MO orthogonalization, both described by the matrix equations:

$$C^T F C = E, \text{ and } C^T S C = I. \tag{5}$$

The first step in circumventing the calculation of the C matrix introduces the *density matrix* (DM) formally defined as

$$P = C f\left(E; T, \mu\right) C^T, \tag{6}$$

where $f\left(E; T, \mu\right)$ is the diagonal matrix obtained by plugging E instead of ε in the Fermi-Dirac distribution function:

$$f\left(\varepsilon; T, \mu\right) \equiv \frac{1}{1 + e^{(\varepsilon-\mu)/k_B T}}. \tag{7}$$

The diagonal matrix elements, $2f\left(\varepsilon_n\right)$ (we omit designating the temperature T and chemical μ in f when no confusion is expected) represent the level occupation of the MO $\psi_n\left(\boldsymbol{r}\right)$ (which typically holds a spin-up and a spin-down electron, hence the factor of 2). T can be a real finite temperature or a very low fictitious one. In the latter case, the $T \to 0$ limit of Eq. (7) yields $f\left(\varepsilon_n\right) = 1$ for $n \leq N_{\text{occ}}$ and 0 otherwise, assuming that the chemical μ has been chosen such that $N_e = 2\sum_n f\left(\varepsilon_n\right)$.

In contrast to the formal definition in Eq. (5) of P as a matrix, in sDFT regards P as an *operator* expressed in terms of F and S through the relation

$$P = f\left(S^{-1}F; T, \mu\right) S^{-1}. \tag{8}$$

Here, $S^{-1}F$ is "plugged" in place of ε into the function f of Eq. (7) [58]. Just like P is an operator, our method also views S^{-1} as an *operator* which is applied to any vector u with linear-scaling cost using a preconditioned conjugate gradient method [59,60]). The *operator* P, applied to an arbitrary vector u, uses a Chebyshev expansion [9,17,44,61] of length N_C: $Pu = \sum_{l=0}^{\check{N}_C} a_l\left(T, \mu\right) u^l$ where a_l are the expansion coefficients and $u^0 = S^{-1}u$, $u^1 = Hu^0$ and then $u^{l+1} = 2Hu^l - u^{l-1}$, $l = 2, 3, \ldots$. In this expansion the operator H is a shifted-scaled version of the operator $S^{-1}F$ bringing its eigenvalue spectrum into the $[-1, 1]$ interval. Every operation Pu, which involves repeated applications of H to various vectors is automatically linear-scaling due to the fact that F and S are sparse. Clearly, the numerical effort in the application of P to u depends on the length N_C of the expansion. When the calculation involves a finite physical temperature T, $N_C = 2\left(\frac{E_{max}-E_{min}}{k_B T}\right)$, where E_{max} (E_{min}) is the largest (smallest) eigenvalue of H. Since N_C is inversely proportional to T, the numerical effort of sDFT reduces as T^{-1} in contrast to deterministic KS-DFT approaches where it rises as T^3 [41]. For zero temperature calculations one still uses a finite temperature but chooses it according to the criterion $k_B T \ll \varepsilon_g$ where ε_g is the KS energy gap. For metals it is common to take a fictitious low temperatures.

The above analysis shows then, that the application of P to a vector can be performed in a linear-scaling cost without constructing P. We use this insight in combination with the fact that the expectation value of one-body observables $\hat{O} = \sum_{n=1}^{N_e} \hat{o}_n$ (where $\hat{o}$ is the underlying single electron operator and the sum is over all electrons) can be achieved as a matrix trace with P:

$$\left\langle \hat{O} \right\rangle = 2\text{Tr}\left[PO\right], \tag{9}$$

where $O_{\alpha\alpha'} = \left\langle \phi_\alpha |\hat{o}| \phi_{\alpha'} \right\rangle$ is the matrix representation of the operator within the atomic basis. Eq. (9) can be used to express various expectation values, such as the electron number

$$N_e = 2\text{Tr}\left[PS\right] \tag{10}$$
$$= 2Tr\left[f\left(S^{-1}F; T, \mu\right)\right],$$

the orbital energy

$$E_{orb} = 2Tr\left[PF\right] \tag{11}$$
$$= 2\text{Tr}\left[e\left(S^{-1}F; T, \mu\right)\right],$$

where, $e\left(\varepsilon\right) = f\left(\varepsilon\right)\varepsilon$ and the fermionic entropy

$$\Sigma_F = -2k_B\text{Tr}\left[PS \ln PS + (I - PS)\ln(I - PS)\right] \tag{12}$$
$$= 2\text{Tr}\left[\sigma_F\left(S^{-1}F; \beta, \mu\right)\right]$$

where $\sigma_F = -k_B\left(f \ln f + (1 - f)\ln(1 - f)\right)$. The expectation value of another observable, the density of

states $\rho_s(E) = \sum_n \delta(E - \varepsilon_n)$ can also be written as a trace [59]:

$$\rho_s(\varepsilon) = \pi^{-1} \lim_{\eta \to 0} \text{Im Tr}\left[(\varepsilon S - F - i\eta S)^{-1} S\right], \quad (13)$$

$$= \pi^{-1} \text{Im Tr}\left[g\left(S^{-1}F; \varepsilon\right)\right]$$

where $g(\varepsilon'; \varepsilon) = \lim_{\eta \to 0} \frac{1}{\varepsilon - \varepsilon' - i\eta}$.

Since the density matrix is an operator in the present approach, the trace in Eq. (9) can be evaluated by introducing the unit column vectors $u^{(\alpha')}$ $(\alpha' = 1, \ldots, K)$ and operating with P on them, and the trace becomes:

$$\left\langle \hat{O} \right\rangle = 2 \sum_{\alpha, \alpha'=1}^{K} \left(Pu^{(\alpha')}\right)_\alpha O_{\alpha\alpha'}. \quad (14)$$

Evaluating this equation requires *quadratic-scaling* computational complexity since it involves K applications of P to unit vectors $u^{(\alpha')}$. One important use of Eq. (9) is to compute the electron density at spatial point $\boldsymbol{r}$:

$$n(\boldsymbol{r}) = 2\text{Tr}[PN(\boldsymbol{r})], \quad (15)$$

where $N_{\alpha\alpha'}(\boldsymbol{r}) = \phi_\alpha(\boldsymbol{r})\phi_{\alpha'}(\boldsymbol{r})$ is the overlap distribution matrix, leading to the expression

$$n(\boldsymbol{r}) = 2 \sum_{\alpha, \alpha'=1}^{K} \left(Pu^{(\alpha')}\right)_\alpha \phi_\alpha(\boldsymbol{r})\phi_{\alpha'}(\boldsymbol{r}). \quad (16)$$

Here, given $\boldsymbol{r}$, only a finite (system-size independent) number of α and α' pairs must be summed over. Hence, the calculation of the density at just this point involves a linear-scaling effort because of the need to apply P to a finite number of $u^{(\alpha')}$'s. It follows, that the density function $n(\boldsymbol{r})$ on the entire grid can be obtained in quadratic scaling effort [60]. This allows us to change the SCF schema of Eq. (4) to:

$$\cdots \longrightarrow n(\boldsymbol{r}) \longrightarrow \{v_H[n](\boldsymbol{r}), v_{xc}[n](\boldsymbol{r})\} \longrightarrow \quad (17)$$

$$\longrightarrow F \xrightarrow{O(K^2)} n(\boldsymbol{r}) \longrightarrow \ldots,$$

where the quadratic step is marked $O(K^2)$.

Summarizing, we have shown an alternative trace-based formulation of Kohn Sham theory which focuses on the ability to apply the DM to vectors in a linear-scaling way, without actually calculating the matrix P itself. This leads to a deterministic implementation of KS-DFT theory of quadratic scaling complexity.

C. Basis-set stochastic density functional theory with linear-scaling

The first report of linear-scaling stochastic DFT (sDFT) [38] used a grid-based implementation and focused on the standard deviation error. Other developments of sDFT included implementation of a stochastic

approach to exact exchange in range-separated hybrid functionals [42] and periodic plane-waves applications to warm dense matter [41] and materials science [61]. These developments were all done using orthogonal or grid representations and included limited discussions of the statistical errors.

Here, sDFT is presented in a general way (subsection II C 1), applicable to any basis, orthogonal or not. We then present a theoretical investigation of the variance (subsection II C 3) and bias (subsection II C 4) errors, and using our Gaussian-type basis code, bsInbar, we actually calculate these SEs in water clusters [62] (by direct comparison to the deterministic results) and study their behavior with sampling and system size. Finally, in subsection II C 5 we discuss the scaling and the scalability of the method.

1. sDFT formulation

Having described the quadratic scaling in the previous section, we are but a step away from understanding the way sDFT works. The basic idea is to evaluate the trace expressions (Eqs. (9)-(16)) using the *stochastic trace formula* [63]:

$$\text{Tr}[M] = \mathbf{E}\left\{\sum_{\alpha\alpha'}^{K} \chi_\alpha M_{\alpha\alpha'} \chi_{\alpha'}\right\} \equiv \mathbf{E}\left\{\chi^T M \chi\right\}, \quad (18)$$

where M is an arbitrary matrix, χ_α are K random variables taking the values ± 1 and $\mathbf{E}\left\{\chi^T M \chi\right\}$ symbolizes the statistical expected value of the functional $\chi^T M \chi$. One should notice that Eq. (18) is an identity, since we actually take the expected value. However, in practice we must take a finite sample of only I independent random vectors χ's. This gives an *approximate* practical way of calculating the trace of M:

$$\text{Tr}[M] \approx \text{Tr}_I[M] \equiv \frac{1}{I} \sum_{i=1}^{I} \left(\chi^i\right)^T M \chi^i. \quad (19)$$

From the central limit theorem, this trace evaluation introduces a fluctuation error equal to

$$\text{Var}\left(\text{Tr}_I[M]\right) = \frac{\Sigma_M^2}{I}, \quad (20)$$

where $\Sigma_M^2 = \mathbf{Var}\left(\text{Tr}_1[M]\right)$ is the variance of $\sum_{\alpha\alpha'}^{K} \chi_\alpha M_{\alpha\alpha'} \chi_{\alpha'}$ (discussed in detail in below). This allows to balance between statistical fluctuations and numerical effort, a trade-off which we exploit in sDFT.

With this stochastic technique, the expectation value of an operator $\hat{O}$ becomes (c.f. Eq. (14)):

$$\left\langle \hat{O} \right\rangle = 2\mathbf{E}\left\{(P\chi)^T (O\chi)\right\}, \quad (21)$$

where the application of P to the random vector χ is performed in the same manner as described above for u (see the text immediately after Eq. (8)). This gives the electronic density (see Eq. (16)):

$$n\left(\boldsymbol{r}\right) = 2\mathbf{E}\left\{\psi_{P\chi}\left(\boldsymbol{r}\right)\psi_{\chi}\left(\boldsymbol{r}\right)\right\}, \qquad (22)$$

yielding a vector (called a grid-vector) of density values $n\left(\boldsymbol{r}\right)$ at each grid-point. This involves producing two grid-vectors, $\psi_{\chi}\left(\boldsymbol{r}\right) = \chi_{\alpha}\phi_{\alpha}\left(\boldsymbol{r}\right)$ and $\psi_{P\chi}\left(\boldsymbol{r}\right) = \left(P\chi\right)_{\alpha'}\phi_{\alpha'}\left(\boldsymbol{r}\right)$ and then multiplying them point by point and averaging on the I random vectors.

2. sDFT calculation detail in the basis-set formalism

It is perhaps worthwhile discussing one trick-of-the-trade allowing the efficient calculation of expectation values of some observables, such as N_e, E_{orb}, Σ_F and ρ_s, see Eqs. (10) - (13). These are all expressed as traces over a function $z\left(\varepsilon\right)$, respectively $f\left(\varepsilon\right)$, $\varepsilon f\left(\varepsilon\right)$, $\sigma_F\left(\varepsilon\right)$ and $\rho_e\left(\varepsilon\right)$. As a result, all calculations of such expectation values can be expressed as

$$\mathrm{Tr}\left[z\left(FS^{-1}\right)\right] = \sum_{l=0}^{N_C} a_l m_l, \qquad (23)$$

where a_l are the Chebyshev expansion coefficients (defined above, in subsection II B), easily calculable, depending on the function z and:

$$m_l = \mathrm{Tr}\left[\chi^T T_l \chi\right] = \mathbf{E}\left\{\chi^T T_l \chi\right\}, \qquad (24)$$

are the Chebyshev moments [46], where T_l is the l'th Chebyshev polynomial. The computationally expensive part of the calculation, evaluating the moments m_l, is done once and then used repeatedly for all relevant expectation values. One frequent use of this moments method involves repeated evaluation of the number of electrons N_e until the proper value of the chemical potential is determined.

We should note that many types of expectation values cannot be calculated directly from the moments m_l. For example, the density, the kinetic and potential energies. For these a full stochastic evaluation is needed.

Let us digress a little to explain how we make the calculations, presented in this review, that enable us to study the properties of sDFT and compare them to deterministic calculations. The code we have written for that purpose is called *bs-Inbar* [65], and implements both the deterministic Kohn-Sham DFT approach described in the present and previous sections as well as the stochastic DFT to be discussed below. Following previous works [23, 66], we use an auxiliary equally-spaced (grid spacing $\Delta x = 0.5a_0$) Cartesian grid for calculating the electron-nuclear interaction integrals $V_{\alpha\alpha'}^{en}$, the Coulomb repulsion
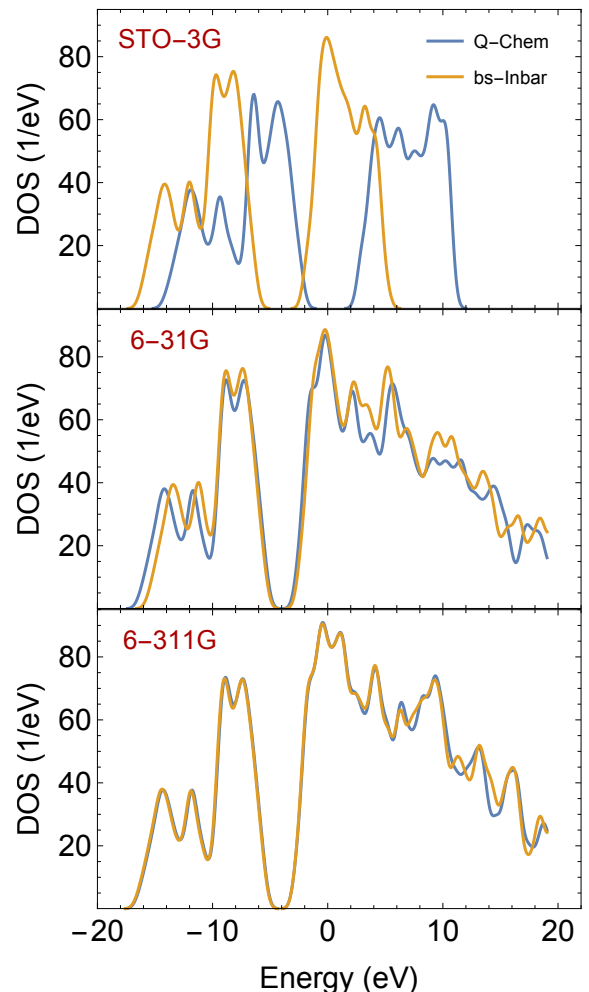

Figure 1. The DOS as a function of energy for a hydrogen-saturated silicon cluster ($Si_{87}H_{76}$) calculated using the all-electron Q-CHEM [64] and the the bs-Inbar codes. Comparison is made for three standard Gaussian basis-sets as indicated in the panels. We used the local density approximation (LDA) for the exchange-correlation energy. Both calculations plot the DOS of Eq. (13) using $k_B T = 0.01 E_h$.

integrals $J_{\alpha\alpha'}$, built from the grid vector representing the density $n\left(\boldsymbol{r}\right)$ using fast Fourier transform techniques, and the exchange correlation integrals $V_{\alpha\alpha'}^{xc}$. This is the $n\left(\boldsymbol{r}\right) \rightarrow \{v_H\left[n\right]\left(\boldsymbol{r}\right), v_{xc}\left[n\right]\left(\boldsymbol{r}\right)\} \rightarrow F$ step of Eq. (4). We developed efficient methods to represent the basis functions on the grid to quickly generate molecular orbitals of the type of Eq. (1) on the grid. These techniques are necessary for the step $F \rightarrow n\left(\boldsymbol{r}\right)$ of Eq. (17) for generating the density $n(\boldsymbol{r})$ from the DM Eq. (16). There are some technical details, such as the effects of core electrons, which cannot be treated efficiently on the grid, and thus are taken into account using norm-conserving pseudopotentials techniques [67, 68], and the deleterious Coulomb/Ewald images which are screened out using the method of Ref. 69. Additional technical elements con-

cerning the bs-Inbar implementation will be presented elsewhere. In Fig. 1 we demonstrate the validity of the deterministic bs-Inbar implementation by comparing its $Si_{87}H_{76}$ DOS function to that obtained from the eigenvalues of an all-electron calculation within the same basis-set (using the Q-CHEM program [64]). For the largest basis-set (triple zeta 6-311G) the two codes produce almost identical DOS (with small difference at high energies), while for the smallest basis (STO-3G) the all electron result shifts strongly to higher energies. Clearly, the bs-Inbar results are less sensitive to the basis-set, likely due to the use of pseudopotentials instead of treating core electrons explicitly.

Having demonstrated the validity of our deterministic numerical implementation by comparing to deterministic DFT results of Q-CHEM, let us now turn our attention to demonstrating the validity of the sDFT calculation when comparing it to deterministic calculation under the same conditions. In Fig. 2 (top panel) where we plot, for water clusters of three indicated sizes, the energy per electron as a function of $1/I$, where I is the number of random vector $\chi's$ used for the stochastic trace formulas (Eq. (22)-(24)). As the the number of random vectors I grows (and $1/I$ drops) the results converge to the deterministic values (shown in the figure as stars at $1/I = 0$). We repeated the calculations 10 times with different random number generator seeds and used the scatter of results for estimating the standard deviation σ and the expected value μ (these are represented, respectively, as error bars and their midpoints in the figure). It is seen that the standard deviation in the energy per particle drops as I increases and in Fig. 3 it is demonstrated that the standard deviation drops as $I^{-1/2}$, in accordance with the central limit theorem. The average values of the energy per particle in Fig. 2 drop steadily towards the converged deterministic values (stars). The fact that the average is always larger than the exact energy, as opposed to fluctuating around it, is a manifestation of a bias δE in the method. When δE is larger than σ it drops in proportion to I^{-1}. In subsections (II C 3)-(II C 4) we will discuss and explain this behavior.
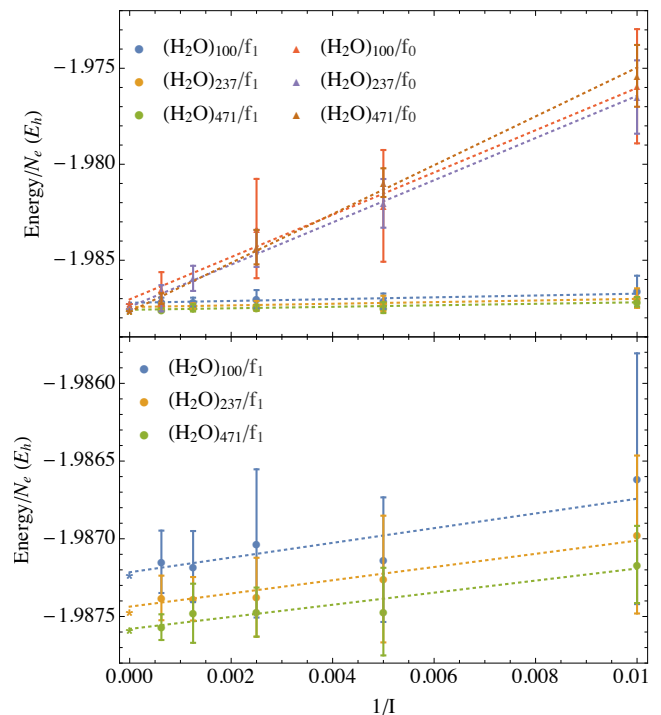

Figure 2. Top panel: The estimate of energy per electron as a function of the inverse number of random vectors $(1/I)$ for water molecule clusters of indicated sizes, without fragments $(/f_0)$ and with fragments (discussed in section (III)) of single H_2O molecules $(/f_1)$. The dotted lines are linear fit to the data (weighted by the inverse error bar length). The deterministic results are represented at $1/I = 0$ by star symbols. Bottom panel: a zoomed view of the $/f_1$ results. These results were calculated using the STO-3G basis-set within the LDA.
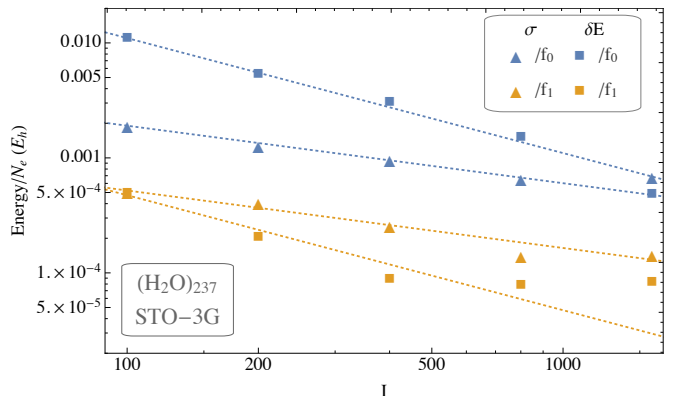

Figure 3. The standard deviation (σ,circles) and errors (δE, squares) of the stochastic estimate of the energy per electron as a function of the number of random vectors (I) in $(H_2O)_{237}$ without fragments ($/f_0$, blue) and with H_2O fragments ($/f_1$, yellow, discussed in section (III)) . The dashed lines are best fit functions αI^{-n} to the data, where $n = 1/2$ for fitting the standard deviations and $n = 1$ for the bias. These results were calculated using the STO-3G basis-set within the LDA.

3. Statistical fluctuations

It is straightforward to show that the variance of the trace formula Eq. (18) is:

$$\Sigma_M^2 \equiv \mathbf{Var} \left\{ \sum_{\alpha\alpha'}^K \chi_\alpha M_{\alpha\alpha'} \chi_{\alpha'} \right\} \quad (25)$$

$$= \frac{1}{2} \sum_{\alpha\neq\alpha'}^K \left(M_{\alpha\alpha'} + M_{\alpha'\alpha} \right)^2$$

$$= (\text{sym}) \, 2 \sum_{\alpha\neq\alpha'}^K M_{\alpha\alpha'}^2, \quad (26)$$

where (sym) marks an equality when M is a symmetric matrix. Therefore, from Eq. (15) the variance in the density $\hat{n}(\boldsymbol{r})$ is

$$\mathbf{Var}_I \left\{ \hat{n}(\boldsymbol{r}) \right\} = \frac{8}{I} \sum_{\alpha\neq\alpha'}^K \left[\sum_\beta^K P_{\alpha\beta} \phi_\beta(\boldsymbol{r}) \phi_{\alpha'}(\boldsymbol{r}) \right]^2. \quad (27)$$

The quantity inside the square brackets involves a limited number, independent of system size, of α'-β index pairs $[(PS)_{\alpha\alpha'}]^2$ and since $tr[PS] = N_e$ we can assume that the magnitude of the brackets squared is $O\left(\frac{N_e}{K}\right)^2$, i.e. independent of system size. Summing over α introduces a system size dependence, hence we conclude that $\mathbf{Var}_I\{\hat{n}(\boldsymbol{r})\}$ has magnitude of $O\left(\frac{N_e}{I}\right)$. When the system is large enough P becomes sparse and then $\mathbf{Var}_I(n(\boldsymbol{r}))$ will tend to become of the magnitude $O\left(\frac{1}{I}\right)$, i.e. system-size independent. The same kind of analysis applies to any single electron observable $\hat{O}$ with sparse matrix representation:

$$\text{Var}_I \left\{ \hat{O} \right\} \propto \frac{N_e}{I}. \quad (28)$$

and independent of system size once P localizes. Since intensive properties are obtained by dividing the related extensive properties by N_e, the standard deviation per electron of intensive properties will evaluate as:

$$\sigma_{\text{intensive}} \propto \frac{\sqrt{\text{Var}_I\{\hat{n}(\boldsymbol{r})\}}}{N_e} \propto \frac{1}{\sqrt{I}N_e}. \quad (29)$$

The decay of the sDFT fluctuations with system size, first pointed out in Ref. 38, is compatible with the fact that fluctuations in intensive variables decay to zero in the thermodynamic limit [70]. For non-metallic systems P becomes sparse as system size grows. Once this sparsity kicks in, $\sigma_{\text{intensive}}$ is expected to decay as $1/\left(\sqrt{I}N_e\right)$. A numerical demonstration of Eq. (29) is given in Fig. 4, for systems of varying numbers N_{waters} of water molecules (all using $I = 100$ random vectors χ), where the standard deviation σ in the energy per particle (blue triangles) indeed drops with system size roughly as $N_{water}^{-1/2}$.
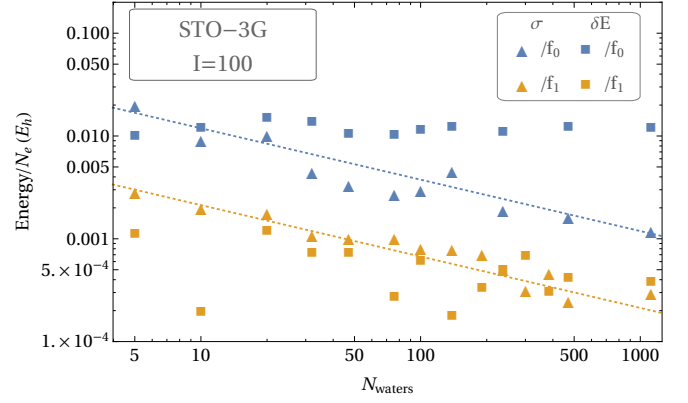

Figure 4. The standard deviation (σ, triangles) and errors (δE squares) of the stochastic estimate of the energy per electron as a function of the number of water molecules N_{water} using no fragments ($/f_0$, blue markers) and single water molecule fragments ($/f_1$, yellow markers). The dotted lines are $1/\sqrt{N_{water}}$ fits to the σ values. These results were calculated using the STO-3G basis-set within the LDA and employed $I = 100$ random vectors.

4. Bias due to nonlinearities

In sDFT, the Hamiltonian $H = S^{-1}F$ is estimated using a random density, and therefore it too, is a random variable with an expected value $\bar{H} = \mathbf{E}\{H\}$ and a fluctuation due to the a covariance matrix $\sigma_{ij;kl}^2 = \mathbf{E}\{H_{ij}H_{kl}\} - \bar{H}_{kl}$. Consider an observable $\hat{O}$ with an exact expectation value $\left\langle \hat{O} \right\rangle_{\bar{H}} = \text{Tr}\left(f\left(\bar{H};T,\mu\right) S^{-1}O \right)$ (Eqs. (8) and (21)). We note, that even when $\bar{H}$ is the exact (deterministic) Hamiltonian, the expectation values $\left\langle \hat{O} \right\rangle_H$ will *not* average to the exact value $\left\langle \hat{O} \right\rangle_{\bar{H}}$, simply because the function of the average of a random variable is distinct or "biased" from the average of the function: $\mathbf{E}\left\{ \left\langle \hat{O} \right\rangle_H \right\} \neq \left\langle \hat{O} \right\rangle_{\bar{H}}$. Clearly, the extent of this bias stems from the how $\mathbf{E}\{f(H;T,\mu)\}$ deviates from $f(\mathbf{E}\{H\};T,\mu)$ and using Taylor's theorem this can be estimated as

$$\mathbf{E}\{f(H;T,\mu)\} - f\left(\bar{H};T,\mu\right) = \quad (30)$$

$$\frac{1}{2} \sum_{i,j,k,l} \sigma_{ij;kl}^2 \frac{\partial}{\partial H_{ij}} \frac{\partial}{\partial H_{kl}} f\left(\bar{H};T,\mu\right).$$

There are three lessons from this analysis: 1) all expectation values $\left\langle \hat{O} \right\rangle$ based on I random vectors in the sDFT method suffer a *bias* $\delta\left\langle \hat{O} \right\rangle \propto \text{coVar}_I\{H\} \propto \text{Var}_I\left\{\hat{O}\right\}$; 2) from Eq. (28) this bias in the *intensive* value $\left\langle \hat{O} \right\rangle/N_e$ is proportional to I^{-1} but independent from system size; and lastly: the double derivative of f on the right hand side of Eq. (30) (called the "Hessian") is related in a complicated way to the curvature $f''(\varepsilon;T,\mu)$ of the Fermi-

Dirac function. This curvature is practically zero for almost all ε except near $\varepsilon \approx \mu \pm k_B T$, and for sufficiently small temperatures, the large Fermi-Dirac curvature regions are safely tucked into the HOMO-LUMO gap, so that indeed the bias can be small.

Summarizing, we find the following trends in the SEs of intensive quantities:

$$\sigma_{\text{intensive}} \propto \left(\frac{1}{N_e I} \right)^{1/2}, \qquad (31)$$

$$\delta_{\text{intensive}} \propto \frac{1}{I}. \qquad (32)$$

Numerical demonstrations of Eq. (32) are given in Fig. 3 (blue squares) where the bias $\delta E / N_e$ is seen to drop as I^{-1} and in Fig. 4 (blue squares) where the bias is seen to be independent of the system size.

5. *Scaling and scalability*

In the left panel of Fig. 5 we show, using a series of water clusters how wall times scale as a function of system size for the sDFT calculation. Because the evaluation of the Hartree potential is made with fast Fourier transforms, the effort is expected to scale as $x = N \log N$ where N indicates the number of water molecules. When considering a single SCF iteration we find this near-linear-scaling as expected. When considering the entire calculation until SCF convergence (which is achieved when the change in the total energy per electron is smaller than $10^{-5} E_h$), we find the number of SCF iterations growing gently with system size and the scaling seems to be near $O\left(x^{1.16}\right)$.

As demonstrated in Fig. 6, we see excellent scalability with number of processors with a mere 8% decrease from the ideal speedup when the number of cores was increased by a factor of 8. This is a result of assigning to each thread a smaller number of random vectors. Ideal wall times are achieved when the there is but one stochastic orbital per thread [71]. For the systems studied Fig. 5 it is an hour for a full SCF calculation of the water 1100 system (ca. 9000 electrons, 13000 orbitals).

Under these conditions, the sDFT wall-times can be significantly lower than those of "conventional" basis-set DFT calculations, as shown in the right panel of Fig. 5. This happens despite the fact that the program used, Q-CHEM, was remarkably still showing quadratic scaling since the cubic scaling component was not yet dominant.

III. EMBEDDED FRAGMENTS METHOD

A. Theory

The notion of fragments, developed first in Ref. 39 was to break up the system into disjoint pieces called

fragments labeled by the index f, and for each fragment compute a DM P^f, such that to a good approximation we can write:

$$P \approx \sum_f P^f. \qquad (33)$$

Clearly, the coherences between different fragments are also missing from $\sum_f P^f$ and these too are assumed small but not totally negligible. From Eq. (9), the expectation value of an arbitrary one-electron operator $\hat{O}$ can be expressed as a contribution of two terms, $\left\langle \hat{O} \right\rangle = 2\text{Tr}\left[\sum_f P^f O\right] + 2\text{Tr}\left[\left(P - \sum_f P^f\right) O\right]$, where the first is the "fragment expected value" and the second is a *correction*, expressed as a small trace to be evaluated using the stochastic trace formula. Applying the stochastic trace formula to just a small trace obviously lowers the SEs when compared to using it for a full trace.

Ref. 39 considered two types of fragmentation procedures. The first was to used natural fragments which could just be considered separately, for example, a single water molecule in a water cluster or a single C_{60} molecule in a cluster of C_{60}'s. Since the molecules are not covalently bonded they are weakly interacting and Eq. (33) is expected to be satisfied to a good degree (however, adjacent water molecules can interact via hydrogen bonds and this may reduce the efficacy of the single-molecule fragments, as discussed below).

The efficiency of the fragments depends entirely on the closeness of the approximation in Eq. (33) and therefore significant effort has to go to developing techniques for constructing fragments. One can probably make good use of the experience gained by the biological and materials embedding methods [72–78].

The notion of *saturated* fragments was developed further in Ref. 40 and used in silicon clusters where covalent bonds were cut when forming the *bare fragment*. The *dangling bonds* on the surface of the bare fragments were then saturated with foreign H or Si atoms. This produced a saturated fragment (see Fig. 7) and a special algebraic technique was developed for carving out the bare fragment DM P^f. The results facilitated what seems to be nearly unbiased force evaluations for the atoms in large nanocrystals, with the structure studied using Langevin molecular dynamics.

B. Efficiency of the embedded fragments

To asses the utility of fragments that do not strictly require saturation, such as water fragments in water clusters, consider first Fig. 2, where we compare the energy per particle of $(H_2O)_n$, with $n = 100$, 237 and 471, estimated using sDFT with no fragments (denoted $/f_0$) and using fragments of just one water molecule ($/f_1$). It is seen that there is a dramatic decrease in the the standard deviation and in the bias. In Fig. 3 we study in
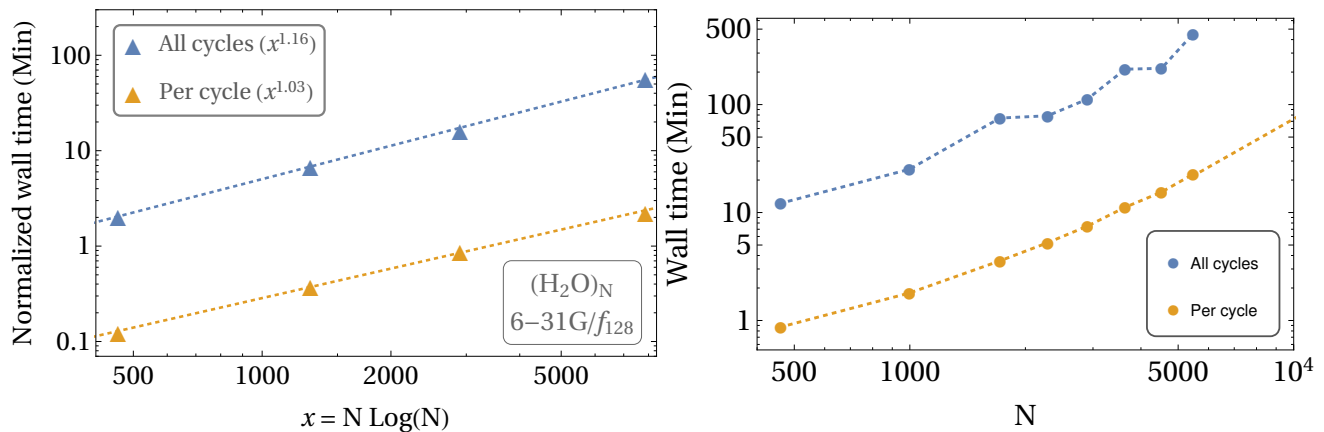
Figure 5. The timing of DFT calculations of $(H_2O)_N$ water clusters using the 6-31G basis-set within the LDA. Left panel: The sDFT wall time as function of $x = N \log N$ normalized to one random orbital per thread for a full SCF calculation (blue symbols) and for a single SCF cycle (orange symbols). Dashed lines are functions $t = Ax^n$, where n is best-fitted to the data and shown in the legend. Right panel: Wall time of a conventional SCF calculation (using Q-CHEM [64]), performed on a single node, as a function of N for a full SCF calculation (blue symbols) and for a single SCF cycle (orange symbols). The Calculations were run on an Intel Xeon CPU E3-1230 v5 @ 3.40GHz 64 GB RAM (*without* Infiniband networking). Each processor supports 8 threads. The sDFT results were calculated with 800 random vectors and fragments of a representative size of 128 water molecules (denoted $/f_{128}$).
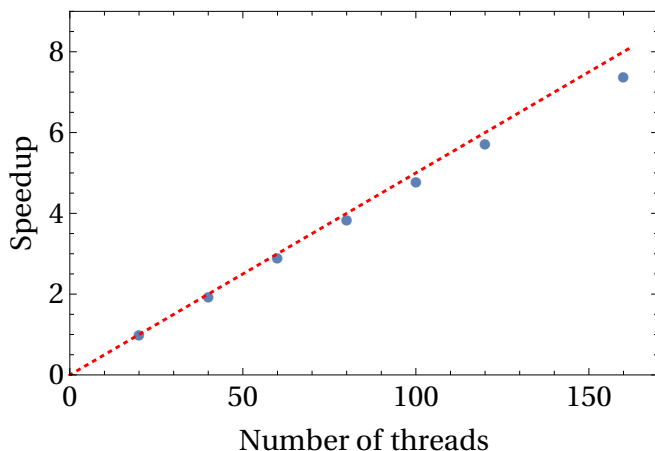

Figure 6. Scalability plot of the calculation, showing the speedup as a function of the number of threads used when calculating a SCF iteration of $(H_2O)_{1120}$ (at the 6-31G basis-set level within LDA) using a total of 2400 random vectors. Calculations were performed on several 2.30GHz Intel Xeon E5-2650 v3 with 252 GB and Infiniband networking.
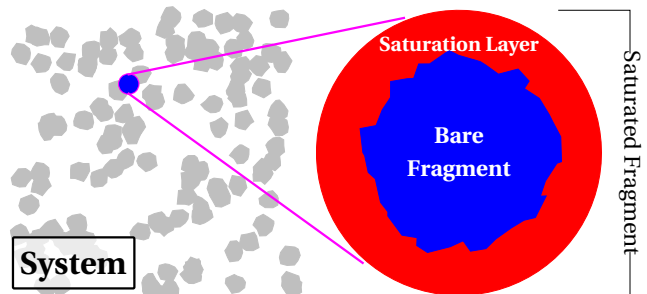

Figure 7. A schematic depiction of a bare fragment (blue region) as a localized set of atoms or molecules within the large system. The fragment is first saturated by coating it with capping atoms (red region), its *saturated*-DM is calculated using a deterministic DFT calculation, from which a bare DM P^f is is "carved" out by an algebraic procedure.

more detail $(H_2O)_n$, finding that with no fragments we are in a bias dominated regime while the use of fragments allows us to move to a regime controlled by fluctuations. Evidently, in the latter case, the large fluctuations mask the linear decrease of the bias with $1/I$, which was so clearly visible in the former one. In Fig. 4, we study the SEs as a function of system size N, comparing the calculations with and without fragments. We see that while fragments help reducing SEs, they do not change the fact that the bias is largely independent of N.

The use of fragments greatly benefits other types of sDFT observables. Consider, for example, the density of states function $\rho_e(E)$ of water [39]. In the left panels of Fig. 8 we plot the DOS for a $(H_2O)_{1120}$ cluster described using the 6-31G basis-set comparing to the deterministic result under an identical setup. We see in the top left panel, that by using $I = 400$ random vectors and small single-molecule fragments $(/f_1)$, the sDFT DOS generally follows that of the exact calculation quite closely. However, a zoom into the frontier orbital gap shows, that even though the stochastic-based calculation exhibits, as it should, a very low DOS in the frontier gap region, there is clearly room for further improvement, since the gap is not sufficiently-well described. Increasing the number of
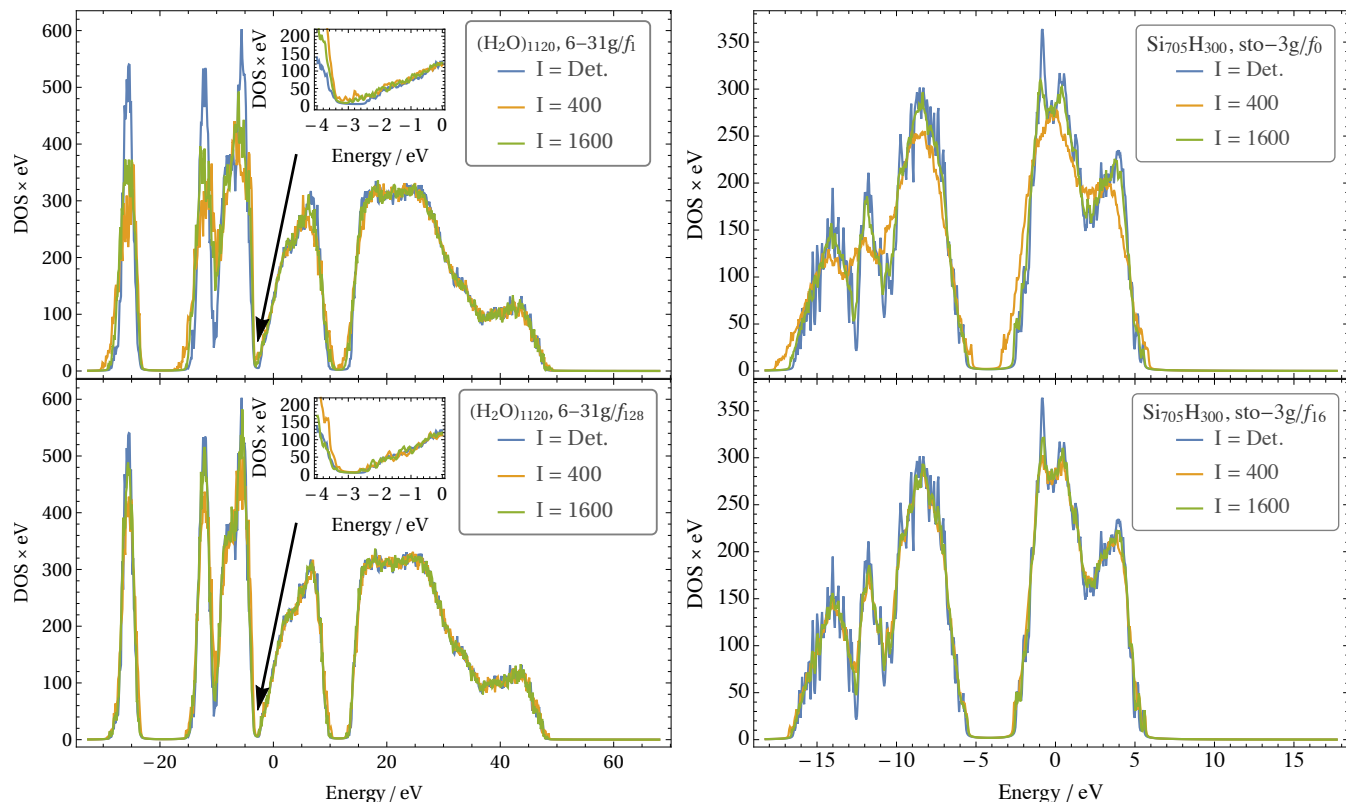
Figure 8. Left panels: The LDA DOS of a cluster of 1120 water molecules using the 6-31G basis-set computed with $I = 400$ and $I = 1600$ random vectors and using single-molecule fragments ($/f_1$ top panel) and 128 molecule fragments ($/f_{128}$ bottom left panel). The insets zoom on the region of the band gap. Right panels: The LDA DOS of $Si_{705}H_{300}$, a hydrogen-terminated silicon nanocrystal, using the STO-3G basis-set computed with $I = 400$ and $I = 1600$ random vectors and using no fragments ($/f_0$ top panel) and 16 atom fragments ($/f_{16}$ bottom panel). In all panels the results are compared to deterministic calculations under the same conditions.

random vectors used from $I = 400$ to $I = 1600$ improves the overall accuracy but increasing the fragment size to 128 water molecules ($/f_{128}$) is even more advantageous, as can be seen in the lower left panel of the figure. It is evident from this description that it is crucial to develop methods that enable better fragments (in the sense that the approximation in Eq. (33) is as tight as possible). Despite the obvious utility of the fragments for the water cluster systems, there is a need to reach quite large fragments for high accuracy. Perhaps this is due to the fact that we do not saturate the bare fragments with their neighboring molecules, as first suggested in recent unpublished work [61]. Future work will test this hypothesis.

Finally we also show in Fig. 8 (right panels) the effect of fragments on the DOS of a large silicon cluster. Here, we must use saturated fragments, as was done in Ref. 40. The density of states, compared to a deterministic calculation is again greatly improved when fragments of size 16 silicon atoms are used (bottom left panel).

C.   Localized energy changes

So far, we have dealt with two types of observables: intensive properties (such as energy per electron) which is a highly averaged quantity, and density of states which, due to the tall number of levels in large systems can be smeared, i.e. locally average, with little loss of essential accuracy. We now demonstrate the possibility of calculating forces on a small atom or molecule within the large system, using stochastic DFT. Previous works concerning this issue [38, 40] demonstrated that the Hellman-Feynman force $F^a = -\int n\left(\boldsymbol{r}\right)\frac{\partial}{\partial R^a}v_{eN}\left(\boldsymbol{r}\right)d\boldsymbol{r} + \sum_{a\neq a}\frac{Z_a Z_{a'}\left(\boldsymbol{R}^a-\boldsymbol{R}^b\right)}{\left|\boldsymbol{R}^a-\boldsymbol{R}^b\right|^3}$ involves a controlled variance and small bias.

Here we consider a related but different question, the possibility of systematically reducing the bias in forces on nuclei within a localized region of interest in a large system. This is useful when modeling reactions in biomolecular systems, as often done using the QM/MM approach, where quantum chemistry forces are used for simulating chemical reactions and other electronic processes (charge

transfer or excitation) while force fields are used for the rest of the system [72–74, 79–82].

For this, we take a fragment which encapsulates the region of interest and "embed" it into the system using sDFT. We study such a process in Fig. 9 where the force F exerted on a certain, marked, water molecule in a larger $(H_2O)_{237}$ cluster is calculated, first by deterministic DFT (shown as a dashed red line in the figure) and then by sDFT as a function of the number of random vectors I $(F(I))$, using two types of fragment sizes: 12-molecule fragments $(/f_{12})$, on the left, and larger 32-molecule fragments $(/f_{32})$ on the right. We note, that $F(I = 0)$ is the deterministic force felt by the molecule in its parent fragment. In the right panel we show the case of a parent fragment which fully encloses the marked molecule. At $I = 0$ the force is already very close to the deterministic value, indicating $\sum_f P_f$ is an excellent approximation for P. When embedded by $I > 0$ stochastic iterations, we find that fluctuations are introduced, but the error bars (marking 95% chance that $\mathbf{E}\{F(I)\}$ always include the exact value) indicate a small bias (such that the error is not dominated by the bias). If we repeat this calculation, but use small fragments which do not encapsulate the marked molecule, the $F(I = 0)$ is very different from the deterministic exact force ($\sum_f P_f$ is a deficient approximation for P). When embedded by $I > 0$ stochastic iterations, the bias is gradually removed as I grows, in accordance with the steady diminishing of the bias discussed in subsection II C 4.

We may conclude from this computational experiment that sDFT may be especially useful for studying chemical processes in small subsystems which can be encapsulated in fragments. Without using fragments, this is also possible an increase in the number of samplings I needs to be employed in order to remove the bias.

IV. SUMMARY AND DISCUSSION

The sDFT approach has been used in various means and for a selection of applications [38–42, 61]. The common thread for all the previous sDFT works was its formulation using an orthogonal basis (grids or plane-waves representation). In this review, we have focused on studying sDFT in the perspective of a local non-orthogonal basis-set. One advantage of the localized basis-set method is that even for large systems the deterministic calculation can still be performed allowing to study in detail errors and their dependence on system size.

The sDFT theory was described using three stages, starting from the standard basis-set formulation of DFT, leading to cubic scaling. Next, we developed a deterministic trace-based calculation, exploiting the sparsity of the Fock and overlap matrices, which lead to a quadratic approach but remained numerically accurate. Finally,

came the sDFT which uses stochastic sampling to evaluate the trace-based calculations, thereby lowering the scaling to linear. The price to pay is the introduction of statistical errors, which one can mitigate by increasing the sampling rate. In order to study and demonstrate the sDFT properties, we developed a basis-set DFT approach using an auxiliary grid for constructing the Hartree and exchange-correlation matrices. Based on this code we also developed the stochastic sDFT implementation. We also developed a basis-set-based fragment method and tested its utility

Using the code, we analyzed the statistical errors associated with the stochastic calculations and their dependence on the number of stochastic samples I, the system size, N (one can take the number of electrons N_e or the basis-set size K as N), and the fragment size. As in previous sDFT papers, the results demonstrated a $I^{-1/2}$ and $N^{-1/2}$ dependence of the statistical fluctuations. Furthermore, we were able to explore the nature of the systematic errors in the sDFT calculation. The bias errors in stochastic methods, have been discussed before in [41, 83]. In sDFT we show that they do not grow with system size and that they decay as I^{-1}. We also developed an analytical model to explain these observations.

It has also been shown that using fragments the noise in the results can be significantly reduced reaching a regime where the statistical fluctuations are the dominating contributions to the error (rather than the bias). These conclusions are in line with previous studies [39, 40, 84], By implementing the fragments we were able to calculate other observables (such as the density of states, or forces) in a much more accurate fashion for a very similar cost.

We demonstrated that our sDFT implementation displays system-size linear-scaling CPU time (Figure 5) and that it is efficacious in parallel architectures (Figure 6). Indeed, it seems to reach its full utility in CPU-abundant architectures, suggesting it may be suitable for Exascale computing.

Future work in the sDFT implementation is required for speeding up the calculations on each node, this can be achieved by shared-memory or GPU parallelization. Further development is also needed for improved fragments which will reduce the variance and bias errors as well as reduce the number of SCF iterations. Finally, as mentioned above, using the sDFT code to drive a Langevin sampling of the nuclear configurations [40] will allow us to compute observables related to the thermal-nuclear structure of the molecular systems.
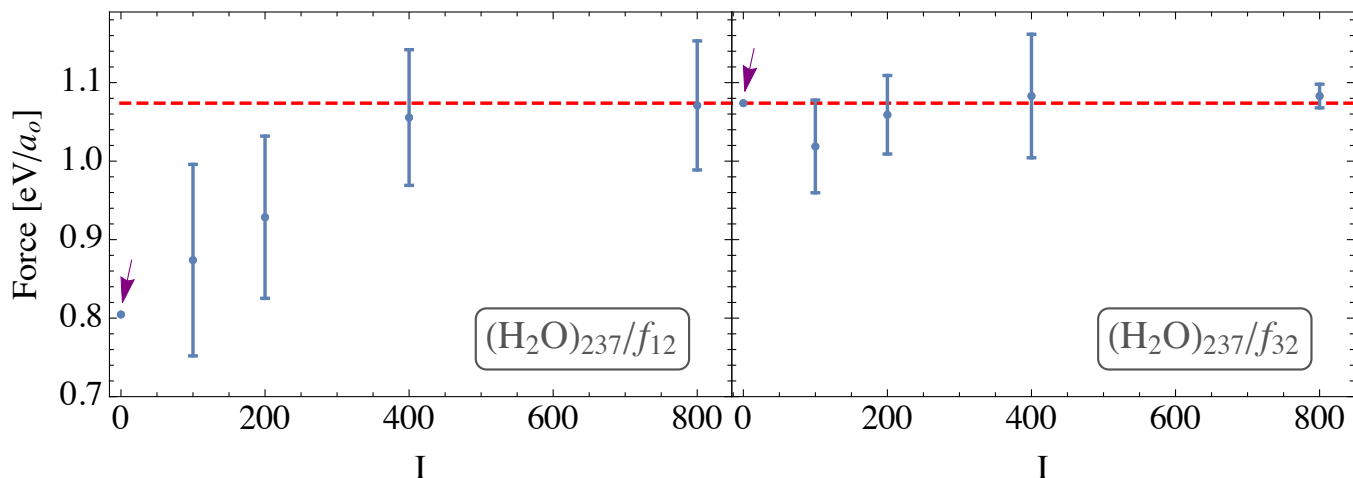
Figure 9. The force on a marked water molecule in $(H_2O)_{237}$ (red dashed line is the deterministic DFT value) calculated as $-\delta E/\delta x$ where δE is the energy difference between two positions of the molecule displaced by a distance $\delta x = 0.05a_0$. On the left (right) panel we present $/f_{12}$ ($/f_{32}$) results. The arrow points to the force $F(I=0)$ (the deterministic force on the molecule when in the parent fragment). The "error bars" are 95% confidence intervals for $\mathbf{E}\{F(I)\}$. These results were calculated using the STO-3G basis-set within the LDA.

---

* These authors contributed equally
† eran.rabani@berkeley.edu
‡ dxn@ucla.edu
§ roi.baer@huji.ac.il

[1] L. Ratcliff, *Optical absorption spectra calculated using linear-scaling density-functional theory* (Springer Heidelberg, 2013).

[2] T. Tsuneda, *Density functional theory in quantum chemistry* (Springer, 2014).

[3] J. Morin and J. M. Pelletier, *Density Functional Theory: Principles, Applications and Analysis* (Nova Science Publishers, Incorporated, 2013).

[4] E. Engel and R. M. Dreizler, *Density functional theory: an advanced course* (Springer Science & Business Media, 2011).

[5] F. Graziani, M. P. Desjarlais, R. Redmer, and S. B. Trickey, *Frontiers and Challenges in Warm Dense Matter*, Vol. 96 (Springer Science & Business, 2014).

[6] W. T. Yang, Phys. Rev. Lett. **66**, 1438 (1991).

[7] X. Li, W. Nunes, and D. Vanderbilt, Phys. Rev. B **47**, 10891 (1993).

[8] P. Ordejon, D. A. Drabold, M. P. Grumbach, and R. M. Martin, Physical Review B-Condensed Matter **48**, 14646 (1993).

[9] S. Goedecker and L. Colombo, Phys. Rev. Lett. **73**, 122 (1994).

[10] R. W. Nunes and D. Vanderbilt, Physical Review B-Condensed Matter **50**, 17611 (1994).

[11] Y. Wang, G. M. Stocks, W. A. Shelton, D. M. C. Nicholson, Z. Szotek, and W. M. Temmerman, Phys. Rev. Lett. **75**, 2867 (1995).

[12] E. Hernandez and M. J. Gillan, Physical Review B-Condensed Matter **51**, 10157 (1995).

[13] S. Goedecker, Journal of Computational Physics **118**, 261 (1995).

[14] P. Ordejon, E. Artacho, and J. M. Soler, Physical Review B-Condensed Matter **53**, 10441 (1996).

[15] D. R. Bowler, M. Aoki, C. M. Goringe, A. P. Horsfield, and D. G. Pettifor, Modell. Simul. Mater. Sci. Eng. **5**, 199 (1997).

[16] R. Baer and M. Head-Gordon, Phys. Rev. Lett. **79**, 3962 (1997).

[17] R. Baer and M. Head-Gordon, J. Chem. Phys. **107**, 10003 (1997).

[18] A. H. R. Palser and D. E. Manolopoulos, Physical Review B-Condensed Matter **58**, 12704 (1998).

[19] S. Goedecker, Rev. Mod. Phys. **71**, 1085 (1999).

[20] G. E. Scuseria, J. Phys. Chem. A **103**, 4782 (1999).

[21] G. Galli, Physica Status Solidi B-Basic Research **217**, 231 (2000).

[22] S. Adhikari and R. Baer, J. Chem. Phys. **115**, 11 (2001).

[23] J. M. Soler, E. Artacho, J. D. Gale, A. Garcia, J. Junquera, P. Ordejon, and D. Sanchez-Portal, J. Phys. C **14**, 2745 (2002).

[24] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi, and M. C. Payne, J. Chem. Phys. **122**, 084119 (2005).

[25] M. J. Gillan, D. R. Bowler, A. S. Torralba, and T. Miyazaki, Comput. Phys. Commun. **177**, 14 (2007).

[26] C. Ochsenfeld, J. Kussmann, and D. S. Lambrecht, "Linear-scaling methods in quantum chemistry," in *Reviews in Computational Chemistry* (Wiley-Blackwell, 2007) Chap. 1, pp. 1–82.

[27] V. Havu, V. Blum, P. Havu, and M. Scheffler, Journal of Computational Physics **228**, 8367 (2009).

[28] L. Lin, J. Lu, L. Ying, and E. Weinan, Chinese Annals of Mathematics, Series B **30**, 729 (2009).

[29] T. Ozaki, Phys. Rev. B **82**, 075131 (2010).

[30] D. Bowler and T. Miyazaki, Reports on Progress in Physics **75**, 036503 (2012).

[31] J. E. Moussa, J. Chem. Phys. **145**, 164108 (2016).

[32] L. E. Ratcliff, S. Mohr, G. Huhs, T. Deutsch, M. Masella, and L. Genovese, Wiley Interdisciplinary Reviews: Computational Molecular Science **7**, e1290 (2017).

[33] W. Kohn, Phys. Rev. Lett. **76**, 3168 (1996).

[34] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, Physical Review B **58**, 7260 (1998).

[35] B. Aradi, B. Hourahine, and T. Frauenheim, The Journal of Physical Chemistry A **111**, 5678 (2007).

[36] V. V. Karasiev and S. B. Trickey, in *Advances in Quantum Chemistry*, Vol. 71 (Elsevier, 2015) pp. 221–245.

[37] W. C. Witt, G. Beatriz, J. M. Dieterich, and E. A. Carter, Journal of Materials Research **33**, 777 (2018).

[38] R. Baer, D. Neuhauser, and E. Rabani, Phys. Rev. Lett. **111**, 106402 (2013).

[39] D. Neuhauser, R. Baer, and E. Rabani, J. Chem. Phys. **141**, 041102 (2014).

[40] E. Arnon, E. Rabani, D. Neuhauser, and R. Baer, J. Chem. Phys. **146**, 224111 (2017).

[41] Y. Cytter, E. Rabani, D. Neuhauser, and R. Baer, Phys. Rev. B **97**, 115207 (2018).

[42] D. Neuhauser, E. Rabani, Y. Cytter, and R. Baer, J. Phys. Chem. A **120**, 3071 (2015).

[43] D. A. Drabold and O. F. Sankey, Phys. Rev. Lett. **70**, 3631 (1993).

[44] O. F. Sankey, D. A. Drabold, and A. Gibson, Phys. Rev. B **50**, 1376 (1994).

[45] L.-W. Wang, Physical Review B **49**, 10154 (1994).

[46] H. Röder, R. Silver, D. Drabold, and J. J. Dong, Phys. Rev. B **55**, 15382 (1997).

[47] A. Weiße, G. Wellein, A. Alvermann, and H. Fehske, Reviews of modern physics **78**, 275 (2006).

[48] C. Bekas, E. Kokiopoulou, and Y. Saad, Applied numerical mathematics **57**, 1214 (2007).

[49] L. Lin, Y. Saad, and C. Yang, SIAM review **58**, 34 (2016).

[50] Z. Wang, G.-W. Chern, C. D. Batista, and K. Barros, The Journal of Chemical Physics **148**, 094107 (2018).

[51] C. C. J. Roothaan, Rev. Mod. Phys. **23**, 69 (1951).

[52] G. G. Hall, Proc. R. Soc. Lond. A **205**, 541 (1951).

[53] A. Szabo and N. S. Ostlund, *Modern quantum chemistry: introduction to advanced electronic structure theory* (Courier Corporation, 1996).

[54] W. Koch and M. Holthausen, *A Chemist's Guide to Density Functional Theory* (Wiley, Heidelberg, 2001).

[55] R. Abrol and A. Kuppermann, J. Chem. Phys. **116**, 1035 (2002).

[56] Y. Y. Shao and M. Head-Gordon, Chem. Phys. Lett. **323**, 425 (2000).

[57] C. A. White, B. G. Johnson, P. M. Gill, and M. Head-Gordon, Chemical physics letters **230**, 8 (1994).

[58] This relation can be proved by plugging $E = C^T F C$ from Eq. (5) into Eq. (6), giving $P = Cf\left(C^T FC; T, \mu\right)C^T$, then using the rule $Af\left(XA\right) = f\left(AX\right)A$ (valid for functions that can be represented as power series and square matrices) obtain $P = f\left(CC^T F; T, \mu\right)CC^T$ and finally using $CC^T = S^{-1}$ from from Eq. (5).

[59] K. Gross Eberhard, E. Runge, and O. Heinonen, A. Hilger (1991).

[60] Note that when the DM P is sparse, the evaluation of the density of Eq. 16 can be performed in linear-scaling complexity. The stochastic method (explained in Subsec-tion II C 1) does not exploit this sparsity explicitly.

[61] C. Ming, , D. Neuhauser, R. Baer, and E. Rabani, to be published (2018).

[62] The clusters we used were produced by Daniel Spångberg at Uppsala University, Department of Materials Chemistry, and retrieved from the ergoscf webpage http://www.ergoscf.org/xyz/h2o.php.

[63] M. F. Hutchinson, Commun Stat Simul Comput. **19**, 433 (1990).

[64] Y. Shao, Z. Gan, E. Epifanovsky, A. T. Gilbert, M. Wormit, J. Kussmann, A. W. Lange, A. Behn, J. Deng, X. Feng, D. Ghosh, M. Goldey, P. R. Horn, L. D. Jacobson, I. Kaliman, R. Z. Khaliullin, T. Kus, A. Landau, J. Liu, E. I. Proynov, Y. M. Rhee, R. M. Richard, M. A. Rohrdanz, R. P. Steele, E. J. Sundstrom, H. L. W. III, P. M. Zimmerman, D. Zuev, B. Albrecht, E. Alguire, B. Austin, G. J. O. Beran, Y. A. Bernard, E. Berquist, K. Brandhorst, K. B. Bravaya, S. T. Brown, D. Casanova, C.-M. Chang, Y. Chen, S. H. Chien, K. D. Closser, D. L. Crittenden, M. Diedenhofen, R. A. D. Jr., H. Do, A. D. Dutoi, R. G. Edgar, S. Fatehi, L. Fusti-Molnar, A. Ghysels, A. Golubeva-Zadorozhnaya, J. Gomes, M. W. Hanson-Heine, P. H. Harbach, A. W. Hauser, E. G. Hohenstein, Z. C. Holden, T.-C. Jagau, H. Ji, B. Kaduk, K. Khistyaev, J. Kim, J. Kim, R. A. King, P. Klunzinger, D. Kosenkov, T. Kowalczyk, C. M. Krauter, K. U. Lao, A. D. Laurent, K. V. Lawler, S. V. Levchenko, C. Y. Lin, F. Liu, E. Livshits, R. C. Lochan, A. Luenser, P. Manohar, S. F. Manzer, S.-P. Mao, N. Mardirossian, A. V. Marenich, S. A. Maurer, N. J. Mayhall, E. Neuscamman, C. M. Oana, R. Olivares-Amaya, D. P. O'Neill, J. A. Parkhill, T. M. Perrine, R. Peverati, A. Prociuk, D. R. Rehn, E. Rosta, N. J. Russ, S. M. Sharada, S. Sharma, D. W. Small, A. Sodt, T. Stein, D. Stock, Y.-C. Su, A. J. Thom, T. Tsuchimochi, V. Vanovschi, L. Vogt, O. Vydrov, T. Wang, M. A. Watson, J. Wenzel, A. White, C. F. Williams, J. Yang, S. Yeganeh, S. R. Yost, Z.-Q. You, I. Y. Zhang, X. Zhang, Y. Zhao, B. R. Brooks, G. K. Chan, D. M. Chipman, C. J. Cramer, W. A. G. III, M. S. Gordon, W. J. Hehre, A. Klamt, H. F. S. III, M. W. Schmidt, C. D. Sherrill, D. G. Truhlar, A. Warshel, X. Xu, A. Aspuru-Guzik, R. Baer, A. T. Bell, N. A. Besley, J.-D. Chai, A. Dreuw, B. D. Dunietz, T. R. Furlani, S. R. Gwaltney, C.-P. Hsu, Y. Jung, J. Kong, D. S. Lambrecht, W. Liang, C. Ochsenfeld, V. A. Rassolov, L. V. Slipchenko, J. E. Subotnik, T. V. Voorhis, J. M. Herbert, A. I. Krylov, P. M. Gill, and M. Head-Gordon, Mol. Phys. **113**, 184 (2015).

[65] Bs-Inbar is the basis-set version of our electronic structure program named "Inbar". *Inbar* is the Hebrew equivalent of the Greek ἤλεκτρον, i.e. electron, which like Ambar and Amber, of Perisan/Arabic origins, refers to the yellowish glowing fossilized tree resin.

[66] J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing, and J. Hutter, Comput. Phys. Commun. **167**, 103 (2005).

[67] N. Troullier and J. L. Martins, Phys. Rev. B **43**, 1993 (1991).

[68] L. Kleinman and D. Bylander, Phys. Rev. Lett. **48**, 1425 (1982).

[69] G. J. Martyna and M. E. Tuckerman, J. Chem. Phys. **110**, 2810 (1999).

[70] J. W. Gibbs, *Elementary Principles in Statistical Me-*

chanics (Yale University Press, New Haven, 1902).

[71] Once we apply one stochastic orbital per thread, further gain from parallelization needs to be obtained from other sources, such as open MP techniques. This has not been implemented yet.

[72] V. Barone, M. Biczysko, and G. Brancato, Adv. Quantum Chem. **59**, 17 (2010).

[73] M. W. van der Kamp and A. J. Mulholland, Biochemistry (Mosc.) **52**, 2708 (2013).

[74] B. Sabin, John R, Erkki, and S. Canuto, *Combining quantum mechanics and molecular mechanics. Some recent progresses in QM/MM methods*, Advances in QUANTUM CHEMISTRY, Vol. 59 (Academic Press, Amsterdam, 2010).

[75] C. Huang and E. A. Carter, J. Chem. Phys. **135**, 194104 (2011).

[76] T. N. Lan, A. A. Kananenka, and D. Zgid, "Communication: Towards ab initio self-energy embedding theory in quantum chemistry," (2015).

[77] D. Ghosh, D. Kosenkov, V. Vanovschi, C. F. Williams, J. M. Herbert, M. S. Gordon, M. W. Schmidt, L. V. Slipchenko, and A. I. Krylov, J. Phys. Chem. A **114**, 12739 (2010).

[78] P. Elliott, K. Burke, M. H. Cohen, and A. Wasserman, Phys. Rev. A **82**, 024501 (2010).

[79] A. E. Cho, V. Guallar, B. J. Berne, and R. Friesner, J. Comput. Chem. **26**, 915 (2005).

[80] D. Svozil and P. Jungwirth, J. Phys. Chem. A **110**, 9194 (2006).

[81] H. Lin and D. G. Truhlar, Theor. Chem. Acc. **117**, 185 (2007).

[82] A. Sharir-Ivry, H. A. Crown, W. Wu, and A. Shurki, J. Phys. Chem. A **112**, 2489 (2008).

[83] D. Neuhauser, R. Baer, and D. Zgid, J. Chem. Theory Comput. **13**, 5396 (2017).

[84] M. Chen, R. Baer, D. Neuhauser, and E. Rabani, arXiv:1811.02107 [physics] (2018), arXiv: 1811.02107.