Linear-scaling calculation of Hartree-Fock exchange energy with non-orthogonal generalised Wannier functions

J. Dziedzic, Q. Hill, and C.-K. Skylaris

Citation: J. Chem. Phys. **139**, 214103 (2013); doi: 10.1063/1.4832338
View online: http://dx.doi.org/10.1063/1.4832338
View Table of Contents: http://jcp.aip.org/resource/1/JCPSA6/v139/i21
Published by the AIP Publishing LLC.

Additional information on J. Chem. Phys.
Journal Homepage: http://jcp.aip.org/
Journal Information: http://jcp.aip.org/about/about_the_journal
Top downloads: http://jcp.aip.org/features/most_downloaded
Information for Authors: http://jcp.aip.org/authors

Re-register for Table of Content Alerts

Create a profile.

Sign up today!

Linear-scaling calculation of Hartree-Fock exchange energy with non-orthogonal generalised Wannier functions

J. Dziedzic,[a] Q. Hill,[b] and C.-K. Skylaris[c]
School of Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom

(Received 24 April 2013; accepted 6 November 2013; published online 2 December 2013)

We present a method for the calculation of four-centre two-electron repulsion integrals in terms of localised non-orthogonal generalised Wannier functions (NGWFs). Our method has been implemented in the ONETEP program and is used to compute the Hartree-Fock exchange energy component of Hartree-Fock and Density Functional Theory (DFT) calculations with hybrid exchange-correlation functionals. As the NGWFs are optimised *in situ* in terms of a systematically improvable basis set which is equivalent to plane waves, it is possible to achieve large basis set accuracy in routine calculations. The spatial localisation of the NGWFs allows us to exploit the exponential decay of the density matrix in systems with a band gap in order to compute the exchange energy with a computational effort that increases linearly with the number of atoms. We describe the implementation of this approach in the ONETEP program for linear-scaling first principles quantum mechanical calculations. We present extensive numerical validation of all the steps in our method. Furthermore, we find excellent agreement in energies and structures for a wide variety of molecules when comparing with other codes. We use our method to perform calculations with the B3LYP exchange-correlation functional for models of myoglobin systems bound with O_2 and CO ligands and confirm that the same qualitative behaviour is obtained as when the same myoglobin models are studied with the DFT+U approach which is also available in ONETEP. Finally, we confirm the linear-scaling capability of our method by performing calculations on polyethylene and polyacetylene chains of increasing length.
© *2013 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4832338]

I. INTRODUCTION

Density Functional Theory (DFT)[1] as formulated by Kohn and Sham[2] is widely and routinely used for computational electronic structure simulations due to its favourable balance between computational speed and accuracy. The accuracy of DFT, however, depends on the choice of the approximation for the exchange-correlation functional. Within the hierarchy of approximations, often described as the "Jacob's ladder"[3] of exchange-correlation functionals, the so-called hybrid exchange-correlation functionals, which include a fraction of Hartree-Fock exchange, are amongst the most accurate, as they reduce the error due to self-interaction. Hartree-Fock theory is often considered as the starting point for the development of *ab initio* approximations based on the wavefunction. The exchange energy component in Hartree-Fock theory is

$$E_{\mathrm{HFx}} = -\sum_{i=1}^{N_{\mathrm{MO}}}\sum_{j=1}^{N_{\mathrm{MO}}} z_i z_j \iint \psi_i^*(\mathbf{r})\psi_j^*(\mathbf{r}')\frac{1}{|\mathbf{r}-\mathbf{r}'|}\psi_j(\mathbf{r})\psi_i(\mathbf{r}')d\mathbf{r}d\mathbf{r}', \tag{1}$$

where $\{\psi_i\}$ are the canonical molecular orbitals (MOs), z_i are their occupancies, and N_{MO} is the total number of molecular orbitals included in the calculation.

A MO can be expanded in terms of a set of non-orthogonal localised functions as follows:

$$\psi_i(\mathbf{r}) = \varphi_\alpha(\mathbf{r})M^\alpha{}_i, \tag{2}$$

where we have assumed a summation over repeated Greek indices and we are using tensor notation, distinguishing between contravariant (superscript indices) and covariant quantities (subscript indices).[4]

By inserting (2) into (1), we obtain

$$E_{\mathrm{HFx}} = -\sum_{i=1}^{N_{\mathrm{MO}}} M_i^\beta z_i M_i^{\dagger\alpha} \iint \varphi_\alpha^*(\mathbf{r})\varphi_\gamma^*(\mathbf{r}')\frac{1}{|\mathbf{r}-\mathbf{r}'|}\varphi_\delta(\mathbf{r})\varphi_\beta(\mathbf{r}')d\mathbf{r}d\mathbf{r}' \times \sum_{j=1}^{N_{\mathrm{MO}}} M_j^\delta z_j M_j^{\dagger\gamma} \tag{3}$$

$$= -K^{\beta\alpha}\iint \varphi_\alpha^*(\mathbf{r})\varphi_\delta(\mathbf{r})\frac{1}{|\mathbf{r}-\mathbf{r}'|}\varphi_\beta(\mathbf{r}')\varphi_\gamma^*(\mathbf{r}')d\mathbf{r}d\mathbf{r}'K^{\delta\gamma}. \tag{4}$$

Therefore, the exchange energy is expressed as

$$E_{\mathrm{HFx}} = -K^{\beta\alpha}(\varphi_\alpha\varphi_\delta|\varphi_\beta\varphi_\gamma)K^{\delta\gamma} = -K^{\beta\alpha}X_{\alpha\beta}, \tag{5}$$

where $(\varphi_\alpha\varphi_\delta|\varphi_\beta\varphi_\gamma)$ is a two-electron (four-centre) electron repulsion integral (ERI) and

$$X_{\alpha\beta} = (\varphi_\alpha\varphi_\delta|\varphi_\beta\varphi_\gamma)K^{\delta\gamma} \tag{6}$$

is the exchange matrix. The matrix $\mathbf{K}$ is the density kernel and is the representation of the one-particle density matrix in the duals of the $\{\varphi_\alpha\}$. In the quantum chemistry literature, it

[a] Also at Faculty of Applied Physics and Mathematics, Gdansk University of Technology, Gdansk, Poland.
[b] Current address: Arqiva, Crawley Court, Winchester SO21 2QA, United Kingdom.
[c] Author to whom correspondence should be addressed. Electronic mail: c.skylaris@soton.ac.uk

 © 2013 AIP Publishing LLC

is often simply referred to as the density matrix. The calculation of the exchange energy is computationally demanding and this is clearly demonstrated by the form of Eq. (5). Formally, the number of ERIs that need to be evaluated scales asymptotically as $\sim N_{\mathrm{at}}^4$ though in practice it can range between $\sim N_{\mathrm{at}}^2$ and $\sim N_{\mathrm{at}}^4$ depending on the level of localisation of the $\{\varphi_\alpha\}$ functions. The presence of the Coulomb operator in the ERIs means that these integrals are long-ranged, even when the functions φ_α are highly localised, so the exchange energy is fully non-local as a consequence of the non-locality of the exchange operator in Hartree-Fock theory.

Sophisticated computational techniques have been developed which aim to reduce the computational scaling, as well as the prefactor, for the calculation of the exchange energy. Often, these techniques aim to achieve linear-scaling computational cost for systems with non-zero band gap, by avoiding calculation of ERIs that are either zero or below a certain threshold and by taking advantage of the localisation of the density kernel. Such linear-scaling techniques have been developed mainly for the case where the $\{\varphi_\alpha\}$ are Gaussian basis functions, and include the ONX[5] and LinK[6] methods which are based on prescreening of the ERIs and the density kernel. In more recent, state-of-the-art approaches, it has been recognised that the requirement for rigorous upper bounds for neglecting integrals can be relaxed in favour of more tight estimates of integral values which allow finer control in the precision with which the exchange energy is calculated.[7] Novel ways of reducing memory usage by predicting or determining *a priori* the sparsity pattern of the exchange matrix were recently proposed,[8] with some of the techniques optimised for general-purpose GPU processing.[9] Another very interesting approach is that of using a truncated Coulomb (TC) operator to evaluate the ERIs which can therefore be made very short-ranged, and then adding the long range contribution as a correction with systematically improvable approximations.[10] Such approaches naturally lend themselves to reduced- and linear-scaling schemes, given that the bulk of the computational effort goes into the calculation of the ERI tensor, which can be made very sparse via the TC operator.

Mixed basis set approaches for the evaluation of ERIs within the context of exchange energy calculations in a Gaussian basis have also been presented. For example, in recent work each product of contracted Gaussian basis functions is expanded in a plane wave basis set (which plays the role of an auxiliary basis set) and its electrostatic potential is numerically integrated with the product of the Gaussians for the second electron in the ERI.[11] Related approaches have been developed in codes where the $\{\varphi_\alpha\}$ are numerical atomic orbitals (NAOs)[12] using fitting functions (resolution of the identity (RI) technique)[13–15] and specially developed NAO auxiliary (fitting) basis sets.

The use of fitting exclusively within a Gaussian function context (both the "main" and the auxiliary basis sets being Gaussian functions) is perhaps the most widely explored approach for efficient calculation of exchange energies with several important developments over the years. Notable developments in this area include the approach to achieving reduced scaling and smaller prefactors by using "half-transformed" ERIs by Polly *et al.*[16] and the recently developed linear-scaling technique by Merlot *et al.*,[17] which is based on local fitting of either the bra or the ket side of the ERI that are also combined with fitting of both sides via the "robust fitting" formula of Dunlap[18] that eliminates first order errors. Such approaches have also been extended to wavefunction methods beyond Hartree-Fock such as the method by Lorenz *et al.*[19] who have developed an approach for configuration interaction singles (CIS) calculations on crystalline solids based on the calculation of exchange matrix-vector product tensor in terms of Wannier functions expressed in a Gaussian basis set.

The non-local nature of the exchange operator makes it even more challenging to calculate the Hartree-Fock exchange energy when extended basis sets such as plane waves are used. Methods for such calculations have nevertheless been presented in codes involving plane wave basis sets,[20,21] particularly in the context of calculations on periodic crystalline solids. A related approach has been presented by Wu *et al.*,[22] which even though it has been implemented in a plane wave basis set, is expected to be asymptotically linear-scaling in computational cost as it employs Wannier functions, which are highly localised in systems with a band gap, thus expression (5) is evaluated entirely in terms of Wannier functions. ERIs are not explicitly evaluated, rather the action of the exchange potential due to products of Wannier functions is evaluated using a real-space solution of the Poisson equation in regions of space which are smaller and independent of the size of the simulation cell.[23]

One of the major advantages of being able to calculate Hartree-Fock exchange energy is the ability to use it in hybrid exchange-correlation functionals. These provide superior accuracy, justified in part by their derivation in terms of the formally exact adiabatic connection formula and the fact that they remove (at least part of) the self-interaction error which is inherent in the available exchange density functionals. In practice, hybrid functionals comprise of a mixture of Hartree-Fock exchange and density functionals which are based on the local density approximation (LDA) and the generalised gradient approximation (GGA). One of the most popular hybrid functionals is the B3LYP functional,[24] which combines GGA functionals with Hartree-Fock exchange using an expression with 3 adjustable parameters. More recent attempts to improve the accuracy of such hybrid functionals use a screened Hartree-Fock exchange[25] term, which decays rapidly at long range.

In this paper, we present our theoretical developments in order to implement the calculation of the Hartree-Fock exchange energy in the ONETEP program for first principles quantum chemistry calculations with linear-scaling cost. ONETEP belongs to a new generation of linear-scaling methods, where the localised functions $\varphi_\alpha(\mathbf{r})$ are not fixed basis functions but rather are optimised *in situ* as dictated by their chemical environment, in order to achieve the accuracy typical of very large atomic orbital or plane wave basis sets. In Sec. II, we present an outline of the theory on which the ONETEP code is based and the approach which we have developed for the calculation of ERIs with a computational cost which is independent of system size by employing an

auxiliary basis set of spherical waves. The way the ERIs are combined to build up the exchange energy with linear-scaling cost is also discussed as well as the calculation of the gradient of the energy with respect to the localised functions $\{\varphi_\alpha\}$, which is needed for their *in situ* optimisation. In Sec. III, we present extensive testing and validation of our method. The accuracy of the method and the factors that control it are investigated and compared against calculations with other methods for the calculation of Hartree-Fock exchange. We also compare with the DFT+U implementation in ONETEP,[26,27] which is a method radically different from Hartree-Fock exchange but it also aims to reduce the self-interaction error. Finally, we demonstrate the linear-scaling computational cost of our approach and investigate the effect of employing an exchange cutoff on polyethylene and polyacetylene chains of increasing length. We finish the paper with conclusions and thoughts about future work in this area.

II. THEORY

A. ONETEP

ONETEP[28] is based on a reformulation of Kohn-Sham DFT with norm-conserving pseudopotentials in terms of the single-particle density matrix, $\rho(\mathbf{r}, \mathbf{r}')$. The density matrix is represented as

$$\rho(\mathbf{r}, \mathbf{r}') = \varphi_\alpha(\mathbf{r}) K^{\alpha\beta} \varphi^*_\beta(\mathbf{r}'), \tag{7}$$

where the $\{\varphi_\alpha\}$ are Non-Orthogonal Generalised Wannier Functions (NGWFs).[29] The elements of the density kernel $K^{\alpha\beta}$ are nonzero only if $|\mathbf{r}_\alpha - \mathbf{r}_\beta| < r_K$, with $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ being the coordinates of the centres of the NGWFs α and β, and r_K a real-space cutoff length. Each NGWF is centred on a nuclear coordinate and is strictly localised within a sphere of radius R_α. Their overlap matrix is

$$S_{\alpha\beta} = \int \varphi^*_\alpha(\mathbf{r}) \varphi_\beta(\mathbf{r})\, d\mathbf{r}\,. \tag{8}$$

The NGWFs are expanded as a linear combination of psinc functions,[30] $D_m(\mathbf{r}) = D(\mathbf{r} - \mathbf{r}_m)$, as

$$\varphi_\alpha(\mathbf{r}) = \sum_m^{m \in LR(\alpha)} D(\mathbf{r} - \mathbf{r}_m) c_{m\alpha}, \tag{9}$$

where the index m runs over the points of the real-space Cartesian grid $\mathbf{r}_m$, which are the centres of the psinc functions, inside the localization region of φ_α, $LR(\alpha)$. The psinc functions form an orthogonal basis set of bandwidth-limited delta functions related to plane-waves by a unitary transformation, and hence they share many of the desirable properties of these, notably the independence on the nuclear coordinates and the ability of the basis set to be systematically improved by increasing a single parameter: the kinetic energy cutoff. The total energy is minimised self-consistently with respect to $K^{\alpha\beta}$ and $c_{m\alpha}$ in two nested loops,[29,31] so that the converged solution satisfies

$$\frac{\partial E}{\partial K^{\alpha\beta}} = 0 \;\forall\; \alpha, \beta, \tag{10}$$

and

$$\frac{\partial E}{\partial c_{m\alpha}} = 0 \;\forall\; m, \alpha. \tag{11}$$

Therefore, the NGWFs are optimised *in situ* by finding the set of coefficients $c_{m\alpha}$ that minimise the total energy under the constraints of idempotency of the density matrix and conservation of the number of electrons N_e. The condition in Eq. (11) refers to the stationarity of the energy with respect to the NGWFs expressed on the grid.

B. Two-electron integral "engine" for non-orthogonal generalised Wannier functions

We aim to develop an "integral engine" for the generation of batches of ERIs. A batch of ERIs in our case consists of all the ERIs in a quartet of atoms. For example, for atoms A, B, C, and D, a batch consists of all the ERIs $(\varphi_\alpha\varphi_\delta|\varphi_\beta\varphi_\gamma)$, where the index α runs over all the NGWFs on atom A, the index β over all the NGWFs on atom B, etc., as shown in Figure 1. We follow a "direct-SCF" approach[32] so that no ERIs are stored on disk but rather they are contracted "on the fly," as generated, with the corresponding elements of the density kernel in order to build the exchange matrix, which is stored in memory and used to construct the Kohn-Sham Hamiltonian matrix.

C. Straightforward calculation of ERIs

A relatively straightforward approach for the calculation of the ERIs can be developed by modifying and using the existing machinery of ONETEP for molecular integrals with NGWFs based on Fast Fourier Transforms (FFTs) and the cardinality property of the psinc basis set, which ensures exact calculation of integrals as discrete sums of values on the real space grid.[33] For example, to obtain $(\varphi_\alpha\varphi_\delta|\varphi_\beta\varphi_\gamma)$, the product $\varphi_\beta(\mathbf{r})\varphi^*_\gamma(\mathbf{r})$ is first evaluated on the real space grid. Then its

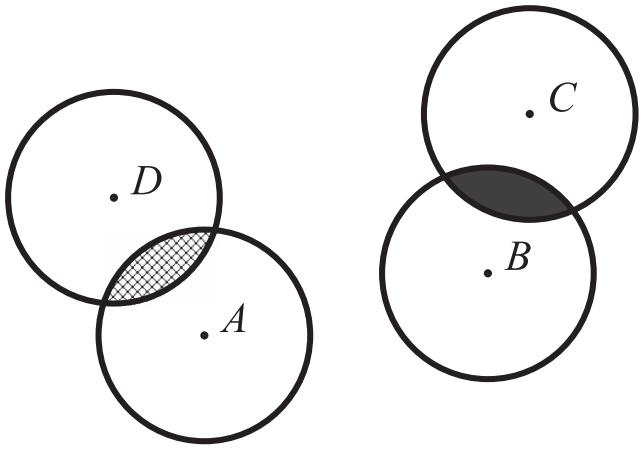

FIG. 1. Atoms A, B, C, and D whose respective NGWFs ($\varphi_\alpha, \varphi_\beta, \varphi_\gamma, \varphi_\delta$, not shown) feature in each term of (5). The densities interacting via exchange are indicated – the potential of the shaded density acts on the cross-hatched density. Terms where the localisation sphere of A is disjoint from the localisation sphere of D vanish. Terms where the localisation sphere of B is disjoint from the localisation sphere of C vanish. The non-local nature of Hartree-Fock exchange is reflected in the fact that terms where the localisation sphere of A is disjoint from the localisation sphere of B do *not* necessarily vanish.

electrostatic (Coulomb) potential is evaluated using FFTs, employing the spherical cutoff-Coulomb operator[34,35] in order to avoid introducing artificial periodicity. Finally, the overlap of this electrostatic potential with the product of functions $\varphi_\alpha^*(\mathbf{r})\varphi_\delta(\mathbf{r})$ is computed in the real space psinc grid as a dot product between the values of these two quantities on the grid points. Due to the long-range nature of the Coulomb potential this approach is computationally very costly as FFTs over the entire simulation cell are required, in general. Thus, contrary to the FFT box technique[23] that is used in other parts of ONETEP to ensure calculation of integrals with cost which is small and independent of system size, in this case the cost of single ERI scales proportionally with the system size (the total number of atoms). We have therefore implemented this approach not as a method that is viable for applications, but simply as the "exact" benchmark against which we evaluate the accuracy of our linear-scaling approach for the exchange energy, which we present in this paper.

D. Resolution of identity calculation of ERIs

A set of (in general) non-orthogonal functions $\{f_a(\mathbf{r})\}_{a=1}^{N_\mathrm{f}}$ which are linearly independent can be used to define a Hilbert space, provided there is also an inner product defined (with an associated metric) which should be positive definite. Choices for the metric that are commonly used in quantum chemistry are the overlap integral,

$$O_{pq} = \int f_p^*(\mathbf{r}) f_q(\mathbf{r}) d\mathbf{r} = \langle f_p | f_q \rangle \tag{12}$$

and the electrostatic integral,

$$V_{pq} = \iint f_p^*(\mathbf{r}) \frac{1}{|\mathbf{r}-\mathbf{r}'|} f_q(\mathbf{r}')\, d\mathbf{r}\, d\mathbf{r}' = (f_p | f_q). \tag{13}$$

We can define the following RI operators based on the above metrics:

$$\hat{I}_O = |f_p\rangle O^{pq} \langle f_q| \tag{14}$$

and

$$\hat{I}_V = |f_p) V^{pq} (f_q|, \tag{15}$$

where we assume implicit summation over the repeated p and q indices. The O^{pq} and V^{pq} are elements of the inverse metric matrices $\mathbf{O}^{-1}$ and $\mathbf{V}^{-1}$, respectively. Typically, such formulas are used in DFT approaches in quantum chemistry to fit the electronic density[13–15,36] in order to speed up the calculation of the Hartree (Coulomb) energy by computing only 3-centre, rather than 4-centre ERIs.

Here, we propose a method for the calculation of four-centre ERIs of NGWFs via RI approaches in such a way that the computational cost per ERI is small and independent of the number of atoms. We will derive formulas for both the overlap and electrostatic metrics. In the case of the overlap metric, by including identity operators for each product of NGWFs, we obtain the following for the exchange energy:

$$\begin{aligned} E_{\mathrm{HFx},O} &= -K^{\beta\alpha}(\varphi_\alpha\varphi_\delta \hat{I}_O | \hat{I}_O \varphi_\beta\varphi_\gamma) K^{\delta\gamma} \\ &= -K^{\beta\alpha}(\langle\varphi_\alpha\varphi_\delta|f_p\rangle O^{pq} f_q | f_r O^{rt} \langle f_t|\varphi_\beta\varphi_\gamma\rangle) K^{\delta\gamma} \\ &= -K^{\beta\alpha}\langle\varphi_\alpha\varphi_\delta|f_p\rangle O^{pq} V_{qr} O^{rt} \langle f_t|\varphi_\beta\varphi_\gamma\rangle K^{\delta\gamma} \\ &= -K^{\beta\alpha}\langle\varphi_\alpha\varphi_\delta|f^q\rangle V_{qr} \langle f^r|\varphi_\beta\varphi_\gamma\rangle K^{\delta\gamma}. \end{aligned} \tag{16}$$

In a similar way, if we use the electrostatic metric, we have

$$\begin{aligned} E_{\mathrm{HFx},V} &= -K^{\beta\alpha}(\varphi_\alpha\varphi_\delta \hat{I}_V | \hat{I}_V \varphi_\beta\varphi_\gamma) K^{\delta\gamma} \\ &= -K^{\beta\alpha}((\varphi_\alpha\varphi_\delta|f_p) V^{pq} f_q | f_r V^{rt} (f_t|\varphi_\beta\varphi_\gamma)) K^{\delta\gamma} \\ &= -K^{\beta\alpha}(\varphi_\alpha\varphi_\delta|f_p) V^{pq} V_{qr} V^{rt} (f_t|\varphi_\beta\varphi_\gamma) K^{\delta\gamma} \\ &= -K^{\beta\alpha}(\varphi_\alpha\varphi_\delta|f_p) V^{pq} (f_q|\varphi_\beta\varphi_\gamma) K^{\delta\gamma}. \end{aligned} \tag{17}$$

If we compare the above equations, we observe that Eq. (16) has certain numerical advantages, as the three centre overlap integrals $\langle f_t|\varphi_\beta\varphi_\gamma\rangle$ are non-zero only when f_t overlaps with the product $\varphi_\beta\varphi_\gamma^*$, so only a small subset of the auxiliary functions $\{f_a(\mathbf{r})\}_{a=1}^{N_\mathrm{f}}$ participate in each three-centre overlap integral, provided they are localised in space to a similar degree as the NGWFs. In contrast, in each three-centre electrostatic integral $(f_q|\varphi_\beta\varphi_\gamma)$ the entire set of auxiliary functions $\{f_a(\mathbf{r})\}_{a=1}^{N_\mathrm{f}}$ participates. In both cases, the entire matrix $\mathbf{V}$ needs to be computed.

Our auxiliary basis needs to satisfy two important requirements: (i) to retain the large basis set accuracy of plane waves, as embodied in the psinc basis set, and (ii) to avoid the costly FFTs required to compute the Coulomb potential of products of NGWFs in the straightforward calculation approach. To satisfy these requirements, we have selected truncated spherical waves as our set of auxiliary functions $\{f_a(\mathbf{r})\}_{a=1}^{N_f}$. Truncated spherical waves are solutions to the Helmholtz equation with boundary conditions enforcing localisation in a sphere of radius a (i.e., solutions of the Schrödinger equation for the particle in a sphere) and they are given by

$$f(\mathbf{r}) = \begin{cases} j_l(qr)\, Z_{lm}(\hat{\mathbf{r}}) & r < a, \\ 0 & r \geq a, \end{cases} \tag{18}$$

where $j_l(qr)$ is a spherical Bessel function and $Z_{lm}(\hat{\mathbf{r}})$ is a real spherical harmonic. The value of q is chosen so that the truncation does not introduce a discontinuity, i.e., $j_l(qa) = 0$. The spherical waves are eigenfunctions of the kinetic energy operator within the localisation region, with eigenvalue $E = \frac{1}{2}q^2$. Therefore, the same kinetic energy cutoff that determines the plane wave basis can be used to restrict the values of q and l in the spherical wave basis. Since plane waves and spherical waves are solutions to the same equation with different boundary conditions, a set of spherical waves can be expected to be a suitable basis set to expand a quantity expressed in plane waves, and already localised in spherical regions,[37] thus satisfying our first requirement. Our second requirement is also satisfied as analytical expressions for the potential of truncated spherical waves, $u_q(\mathbf{r}) = \int \frac{f_q(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}'$, can be derived.[38]

From now on we shall assume that the functions $\{f_p\}$ are truncated spherical waves and N_SW will be used in place of N_f to denote the total number of auxiliary functions used in

the expansion. Our truncated spherical waves, as defined in Eq. (18), include only real spherical harmonics and are therefore real functions.

1. Calculation of the metric matrices

A prerequisite for calculating the ERIs with the RI formalism is the construction of the electrostatic metric matrix **V** (13), and, if in the overlap approach, we also need to construct the overlap metric matrix **O** (12). The on-site (i.e., same-centre) elements of both matrices can be calculated analytically.[38] While there are analytical expressions for the off-site elements of **O**,[39] no suitable expressions exist for the off-site elements of **V**. In our implementation, the off-site elements of both **V** and **O** are calculated numerically.

Each off-site matrix element is a three-dimensional integral of a product of two truncated spherical waves (in the case of the **O** matrix) or a product of a truncated spherical wave and the Coulomb potential due to another truncated spherical wave (in the case of the **V** matrix), i.e.,

$$M_{Ap,Bq} = \int_{-R_{\mathrm{loc}A}}^{R_{\mathrm{loc}A}} \int_{-\sqrt{R_{\mathrm{loc}A}^2-z^2}}^{\sqrt{R_{\mathrm{loc}A}^2-z^2}} \int_{-\sqrt{R_{\mathrm{loc}A}^2-z^2-y^2}}^{\sqrt{R_{\mathrm{loc}A}^2-z^2-y^2}} f_p(\mathbf{r}) \times g_q(\mathbf{r}-\mathbf{R}_{AB})\,dx\,dy\,dz, \tag{19}$$

where

$$g_q(\mathbf{r}) = \begin{cases} f_q(\mathbf{r}) & \text{for } M = O, \\ u_q(\mathbf{r}) & \text{for } M = V, \end{cases} \tag{20}$$

and $p = 1, \ldots, n_{\mathrm{SW}}$, $q = 1, \ldots, n_{\mathrm{SW}}$ (n_{SW} is the number of spherical waves per atomic centre).

The first class of products vanishes when the two centres do not overlap (thus making **O** sparse), the second class of products never vanishes, making **V** dense in principle.

However, when certain pairs of auxiliary functions can be guaranteed never to participate together in the same expansion, **V** can be made sparse. For example, if every expansion only involves spherical waves centred on atoms whose NGWFs overlap, **V** will have the same sparsity as the NGWF overlap matrix **S** and the spherical wave overlap matrix **O**.[40] This is not to say that exchange between atoms whose NGWFs do not overlap will vanish, only that certain elements of the **V** matrix need never be calculated. Section II D 3 will describe in detail the choice of expansion centres proposed in our approach.

The integrand is highly oscillatory, making direct integration on a Cartesian grid impractical. Numerical integration on a radial grid converges with fewer integration steps, but still remains very costly.[38] Instead of these approaches, we propose to expand each truncated spherical wave and the potential thereof into Chebyshev polynomials, whose products, by virtue of being polynomials themselves, are analytically integrable.

First, we divide each localisation sphere of A into N_{i} segments (disks) parallel to the xy plane, each with a thickness of $\Delta_{\mathrm{z}} = 2R_{\mathrm{loc}A}/N_{\mathrm{i}}$ (cf. Fig. 2(a)). N_{o} Chebyshev nodes are then placed along the height of the interval Δ_{z} (cf. Fig. 2(b)). In this way, the localisation sphere is sampled with $N_{\mathrm{i}}N_{\mathrm{o}}$ circular cross-sections (disks) in total. Each of these is similarly subdivided into segments of thickness $\Delta_{\mathrm{y}} = 2\sqrt{R_{\mathrm{loc}A}^2 - z^2}/N_{\mathrm{i}}$, with Chebyshev nodes positioned accordingly (cf. Fig. 2(c)). In this way, we obtain a sampling with $N_{\mathrm{i}}^2 N_{\mathrm{o}}^2$ line segments. These are once again subdivided into N_{i} intervals, each containing N_{o} Chebyshev nodes (cf. Fig. 2(d)).

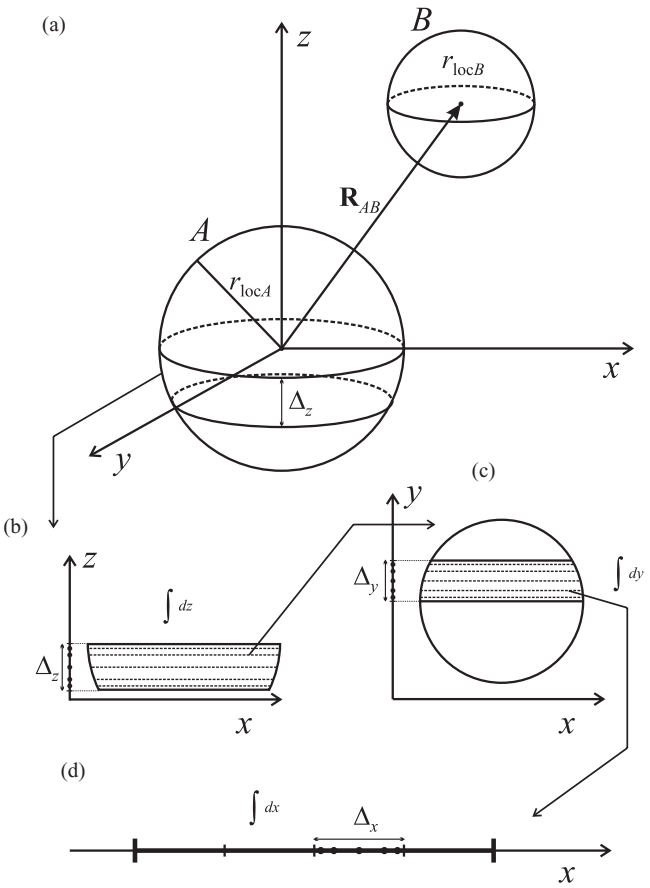

FIG. 2. Diagram showing how a single element of the metric matrix is computed, here using four intervals ($N_{\mathrm{i}} = 4$) and fifth-order Chebyshev polynomials ($N_{\mathrm{o}} = 5$). (a) The localisation sphere of atom A is divided into spherical segments of height Δ_{z}. (b) Each segment is cross-sected at Chebyshev nodes, yielding a set of circles. (c) Each circle is divided into segments of height Δ_{y}, and each segment is cross-sected at Chebyshev nodes, yielding a set of line segments. (d) Each line segment is divided into segments of length Δ_{x}, on each of these the functions making up the integrand are sampled at Chebyshev nodes, yielding their expansion into Chebyshev polynomials. The product (integrand) is thus interpolated with a product of polynomials and an integral over each segment is trivially obtained analytically.

In each of the resultant $N_{\mathrm{i}}^3 N_{\mathrm{o}}^2$ intervals, the functions $f_p(\mathbf{r})$ and $g_q(\mathbf{r} - \mathbf{R}_{AB})$ are sampled at N_{o} Chebyshev nodes, yielding the coefficients of expansion of each function into Chebyshev polynomials. The obtained interpolations, $\tilde{f}_p(\mathbf{r})$ and $\tilde{g}_q(\mathbf{r} - \mathbf{R}_{AB})$, are polynomials in x, and thus their product (itself a polynomial) is trivial to integrate analytically over x in each interval. Each of the cross-sections along x in Fig. 2(d) is piecewise-analytically integrated over in this fashion, yielding a single value, obtained at a y-Chebyshev node, for subsequent integration over y (cf. Fig. 2(c)). The integral over z is obtained in the same manner (cf. Fig. 2(b)). Only the innermost integration (over x) involves products of functions.

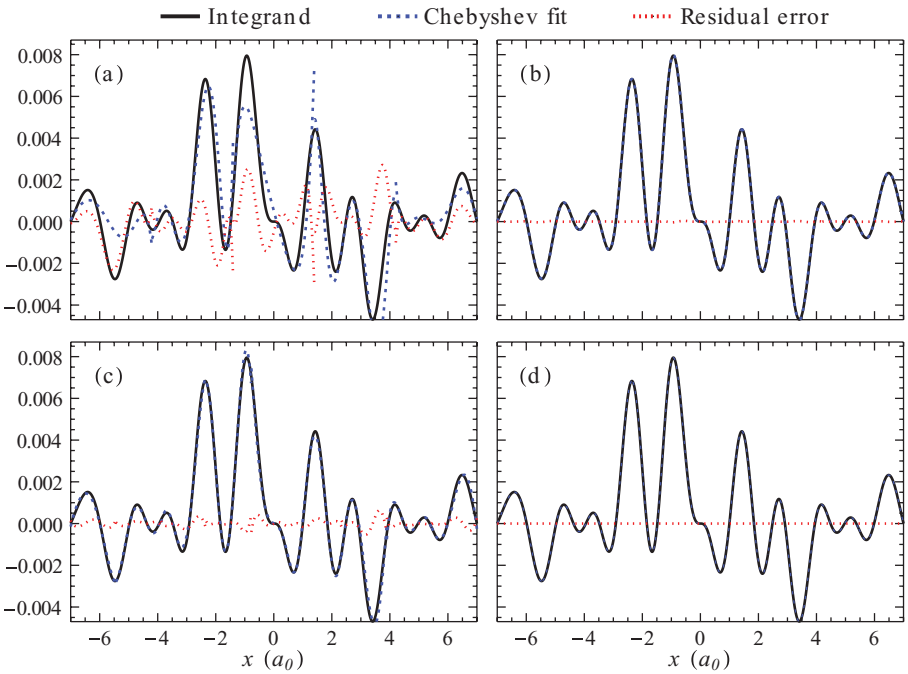

FIG. 3. Chebyshev fits of an example cross-section through the diameter of a localisation sphere. The integrand fitted here is the product of a truncated spherical wave with $l = 2$, $m = 0$, and $q = 3.122\, a_0^{-1}$ and a potential of a truncated spherical wave with $l = 3$, $m = 2$, and $q = 2.418\, a_0^{-1}$, originating on a centre displaced by $[0.5, -0.5, 2.0]\, a_0$. The panels demonstrate the convergence of the integrand when Chebyshev-fitted with: (a) 5 intervals, 4th order polynomials; (b) 5 intervals, 8th order polynomials; (c) 9 intervals, 4th order polynomials; (d) 9 intervals, 8th order polynomials.

The resultant procedure, although cubic in N_i and N_o, offers superior performance compared to sampling on a Cartesian or radial grid, and is easily parallelisable. We note that the expansions $\tilde{f}_p(\mathbf{r})$ do not depend on B and only need to be evaluated once for every atom A. Further improvement in computational efficiency is obtained by exploiting the fact that for every matrix block where A and B are constant and $p = 1, \ldots, n_\mathrm{SW}$, $q = 1, \ldots, n_\mathrm{SW}$ only n_SW evaluations of $f_p(\mathbf{r})$ and corresponding expansions must be performed. With a similar cost for $g_q(\mathbf{r} - \mathbf{R}_{AB})$, we obtain a procedure where the calculation of a matrix block with n_SW^2 elements only involves $2n_\mathrm{SW}$ expensive operations (evaluations, expansions) and n_SW^2 extremely cheap multiplications of Chebyshev coefficients with pre-calculated integrals of products of Chebyshev polynomials.

We find excellent convergence of the fit with increasing number of intervals N_i and polynomial order (cf. Fig. 3). Interpolation with 12th order Chebyshev polynomials over 12 intervals is already sufficiently accurate, thus $(12 \times 12)^3$, or about 3×10^6, coefficients are needed in practice for every representation. For fixed ionic positions, the metric matrices only need to be calculated once, as they do not depend on the electronic degrees of freedom. If all pairs of centres A and B are considered, this calculation is quadratically scaling. However, if the $\mathbf{V}$ matrix is made sparse, as described above, the computational effort of calculating this matrix scales linearly.

2. Calculation of exchange energy

Here, we explain how the calculation of the exchange energy according to the RI formulas (16) and (17) is carried out. We first note that in the case of the overlap metric, the final expression in (16) is not directly used, since it involves the duals $\{f^t\}$ of the spherical waves, and these are not localised. Rather, we use the penultimate expression of (16), which involves the spherical waves $\{f_t\}$ themselves.

The Chebyshev interpolation that was used in constructing the metric matrices is not used in this stage. The products $\varphi_\beta(\mathbf{r})\varphi_\gamma^*(\mathbf{r})$ that appear in the kets of (16) and (17) are instead evaluated on the Cartesian grid on which ONETEP represents NGWFs, i.e., they are expanded in the psinc basis set. The spherical waves $\{f_t\}$ (in the case of the overlap metric) or the potentials thereof (in the case of the electrostatic metric), for both of which analytical expressions are available, are also represented in the psinc basis set. Having expanded all quantities $\varphi_\beta(\mathbf{r})\varphi_\gamma^*(\mathbf{r})$, $f_t(\mathbf{r})$, and $\int f_t(\mathbf{r}')/|\mathbf{r} - \mathbf{r}'|d\mathbf{r}'$ in terms of the psinc basis, the desired integrals $\langle f_t|\varphi_\beta\varphi_\gamma\rangle$ and $(f_q|\varphi_\beta\varphi_\gamma)$, are obtained straightforwardly as dot products over the psinc grid points, and apart from the fact that the integrands need obviously to be bandwidth limited up to the same plane wave kinetic energy cutoffs as the psinc functions, they are calculated exactly using the properties of the psinc basis set.[29,41] The integration regions are localised on NGWF localisation spheres, which makes the computational effort of evaluating a single integral independent of the number of atoms.

For the next stage, we need to compute the right-hand sides of Eqs. (16) and (17). To perform this calculation more efficiently, we can recognise that, for example, in the case of the electrostatic metric the quantity obtained so far, contracted with the relevant blocks of the $\mathbf{V}$ matrix, $\{V^{pq}(f_q|\varphi_\beta\varphi_\gamma)\}_{p=1\ldots N_\mathrm{SW}}$ is none other but a set of expansion coefficients $\{c^p_{\beta\gamma}\}_{p=1\ldots N_\mathrm{SW}}$ for the product $\varphi_\alpha^*(\mathbf{r})\varphi_\delta(\mathbf{r})$ in terms of the truncated spherical waves. The $N_\mathrm{SW} = N_\mathrm{c} n_\mathrm{SW}$

is the number of spherical waves in the expansion (N_c is the number of centres generating the spherical waves). Therefore, we obtain these expansion coefficients by solving the system of linear equations $\{V_{qp}\, c^p_{\beta\gamma} = (f_q|\varphi_\beta\varphi_\gamma)\}_{q=1\ldots N_{SW}}$ in order to avoid inverting the **V** matrix blocks.

When the overlap metric is used, we instead invert blocks of the **O** matrix and perform two matrix multiplications to obtain a matrix block $W^{pt} = O^{pq} V_{qr} O^{rt}$, which we then multiply by a vector of integrals $\langle f_t|\varphi_\beta\varphi_\gamma\rangle$ to obtain the coefficients of expansion. We found that in typical calculations where n_{SW} is in the order of 200, the matrix inversions are somewhat ill-conditioned (with condition numbers in the order of 2×10^{10}), but this did not pose numerical problems as typical matrix inversion library routines can easily cope with such cases.

Subsequently, the expanded potential, contracted for efficiency with the density kernel over the index γ, with the following form for the electrostatic metric:

$$\left(\int \frac{f_p(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}'\right) c^p_{\beta\gamma} K^{\delta\gamma} \tag{21}$$

and this form for the overlap metric

$$f_p(\mathbf{r})\, c^p_{\beta\gamma} K^{\delta\gamma} \tag{22}$$

is acted on the product $\varphi^*_\alpha(\mathbf{r})\varphi_\delta(\mathbf{r})$ that appears in the bras of (16) and (17). The final step, for both metrics, consists in multiplying this result by $-K^{\beta\alpha}$, yielding a component of the exchange energy due to the density $\varphi_\beta\varphi^*_\gamma$ acting on $\varphi^*_\alpha\varphi_\delta$.

The cost of the calculation described above is asymptotically constant (independent of system size) for a single pair (α, β). Thus, the proposed approach is linear-scaling, if and only if the number of pairs (α, β) increases linearly with the size of the system. For this to be possible, the exchange interaction needs to be truncated somehow, which is standard practice in linear-scaling methods.[42] In this approach, we propose to use a distance-based cutoff, r_X, where the contributions from pairs (α, β) whose centres that are further away than a prescribed distance (e.g., $r_X = 20\, a_0$) are neglected, effectively making **X** sparse. This is similar to the density kernel truncation that can be employed for calculating the remaining energy terms in ONETEP with a linear-scaling cost.

3. Choice of centres for the auxiliary basis set

The auxiliary basis set does not need to be complete in the sense of spanning the full space of square integrable functions on R^3. Clearly, it is sufficient that the NGWF products $\varphi_\beta\varphi^*_\gamma$ and $\varphi^*_\alpha\varphi_\delta$ and the auxiliary basis span the same subspace. Our auxiliary set (18) converges to a complete basis set within its spherical localisation region in the limit of including all its (infinite) functions, which are eigenfunctions of a Hermitian operator. We thus employ a local "two-centre" (2c) fitting approach for the products of NGWFs, where we either fit the bra or the ket side of each ERI.[17] For example, for the electrostatic metric based approach, the formula that we use to compute the exchange energy is the following:

$$E_{HFx,V,2c} = -\frac{1}{2} K^{\beta\alpha}[(\varphi_\alpha\varphi_\delta \hat{I}_{V,\alpha\delta}|\varphi_\beta\varphi_\gamma) + (\varphi_\alpha\varphi_\delta|\hat{I}_{V,\beta\gamma}\varphi_\beta\varphi_\gamma)]K^{\delta\gamma} \tag{23}$$

$$= -\frac{1}{2} K^{\beta\alpha}[(\varphi_\alpha\varphi_\delta|f_{p,\alpha\delta})V^{pq}(f_{q,\alpha\delta}|\varphi_\beta\varphi_\gamma) + (\varphi_\alpha\varphi_\delta|f_{p,\beta\gamma})V^{pq}(f_{q,\beta\gamma}|\varphi_\beta\varphi_\gamma)]K^{\delta\gamma}, \tag{24}$$

where $\hat{I}_{V,\alpha\delta}$ is the projection operator consisting only of spherical waves centred on centres of NGWFs ϕ_α and ϕ_δ, and accordingly $f_{p,\alpha\delta}$ denotes spherical waves centred on the same centres.

An important advantage of this two-centre approach is that the expansion always includes spherical waves centred only on atoms whose NGWFs overlap, thus making the **V** matrix sparse. This is a necessary, but not sufficient, condition for the approach to be linear-scaling.

If spherical wave fitting functions on all four centres were used, the exchange energy would be evaluated by the following four-centre formula:

$$E_{HFx,V,4c} = -K^{\beta\alpha}(\varphi_\alpha\varphi_\delta|f_{p,\alpha\delta\beta\gamma})V^{pq}(f_{q,\alpha\delta\beta\gamma}|\varphi_\beta\varphi_\gamma)K^{\delta\gamma}, \tag{25}$$

but this choice is computationally very inefficient, and thus we have only used it as a means of numerically benchmarking the 2-centre approach.

It is worth noting that, in contrast to codes with, e.g., Gaussian atomic orbitals, where the auxiliary basis set fitting is used to reduce the total number of functions, in our case the number of functions is significantly increased when the auxiliary basis is introduced. For example, a carbon atom is typically described by 4 NGWFs, but for the same atom our spherical wave fitting basis set may have l of up to 4 and with 10 distinct values for q for each l (for which we shall subsequently use the notation $l_{max} = 4$, $q_{max} = 10$), for a total of $(1 + 3 + 5 + 7 + 9) \times 10 = 250$ spherical waves. Thus, for the two-centre expansion of Eq. (24) we may typically use $250 \times 2 = 500$ spherical waves to fit each of the $4 \times 4 = 16$ NGWF products.

4. Gradients for energy optimisation

As we are using direct energy minimisation approaches based on the conjugate gradients technique in order to optimise the energy and reach self-consistency, we need the derivatives of the exchange energy with respect to the density kernel and the NGWFs. By differentiation with respect to the kernel, we obtain

$$\frac{\partial E_{HFx,O,2c}}{\partial K^{\eta\theta}} = -K^{\beta\alpha}\left[\langle\varphi_\alpha\varphi_\eta|f_{p,\alpha\eta}\rangle\, O^{pq} V_{qr} O^{rt} \langle f_{t,\alpha\eta}|\varphi_\beta\varphi_\theta\rangle + \langle\varphi_\alpha\varphi_\eta|f_{p,\beta\theta}\rangle\, O^{pq}\, V_{qr} O^{rt} \langle f_{t,\beta\theta}|\varphi_\beta\varphi_\theta\rangle\right] = -(X_{\theta\eta} + X_{\eta\theta}) \tag{26}$$

for the overlap metric and

$$\frac{\partial E_{\mathrm{HFx,V,2c}}}{\partial K^{\eta\theta}} = -K^{\beta\alpha}[(\varphi_\alpha\varphi_\eta|f_{p,\alpha\eta})V^{pq}(f_{q,\alpha\eta}|\varphi_\beta\varphi_\theta) + (\varphi_\alpha\varphi_\eta|f_{p,\beta\theta})V^{pq}(f_{q,\beta\theta}|\varphi_\beta\varphi_\theta)] = -(X_{\theta\eta} + X_{\eta\theta}) \quad (27)$$

for the electrostatic metric. It is understood that the exchange matrices in the above expressions are obtained with different approximations (expansion in an auxiliary basis set with the overlap and electrostatic metric, respectively). We point out that the values of $X_{\theta\eta}$ and $X_{\eta\theta}$ are strictly identical only in the limit of an infinite auxiliary basis set that spans the same subspace as the NGWF product it is used to fit.

In a typical ONETEP calculation, the NGWFs are also optimised[29] and thus the gradient of the exchange energy with respect to NGWFs needs to be calculated. Formally, the gradient G_ε of the exchange energy with respect to an NGWF φ_ε is

$$\begin{aligned} G_\varepsilon(\mathbf{r}) &= \frac{\delta}{\delta\varphi_\varepsilon^*}[-K^{\beta\alpha}(\varphi_\alpha\varphi_\delta|\varphi_\beta\varphi_\gamma)K^{\delta\gamma}] \\ &= \frac{\delta}{\delta\varphi_\varepsilon^*}[-K^{\beta\alpha}X_{\alpha\beta}] \\ &= -2K^{\beta\varepsilon}\varphi_\delta|\varphi_\beta\varphi_\gamma)K^{\delta\gamma} \\ &= -2K^{\beta\varepsilon}\varphi_\delta(\mathbf{r})\left(\int\frac{\varphi_\beta(\mathbf{r}')\varphi_\gamma^*(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}'\right)K^{\delta\gamma}, \end{aligned} \quad (28)$$

where a functional derivative has been used since we differentiate with respect to a function. Under the density fitting approximation (assuming the electrostatic metric is used), this becomes

$$\begin{aligned} G_\varepsilon &= \frac{\delta}{\delta\varphi_\varepsilon^*}\left[-\frac{1}{2}K^{\beta\alpha}[(\varphi_\alpha\varphi_\delta|f_{p,\alpha\delta})V^{pq}(f_{q,\alpha\delta}|\varphi_\beta\varphi_\gamma) + (\varphi_\alpha\varphi_\delta|f_{p,\beta\gamma})V^{pq}(f_{q,\beta\gamma}|\varphi_\beta\varphi_\gamma)]K^{\delta\gamma}\right] \\ &= \frac{\delta}{\delta\varphi_\varepsilon^*}\left[-\frac{1}{2}K^{\beta\alpha}\left(X_{\alpha\beta}+X_{\beta\alpha}\right)\right], \end{aligned} \quad (29)$$

where we have omitted the dependence of the NGWFs and the gradient on $\mathbf{r}$ for the sake of brevity.

However, as explained in Sec. II D 3, in the 2-centre approach we fit *either* the bra or the ket side of each ERI (but never both), and, consequently, for finite auxiliary basis sets, the above is not strictly equal to

$$G_\varepsilon \neq -2K^{\beta\varepsilon}[\varphi_\delta(\mathbf{r})|f_{p,\beta\gamma})V^{pq}(f_{q,\beta\gamma}|\varphi_\beta\varphi_\gamma)]K^{\delta\gamma}, \quad (30)$$

but rather

$$\begin{aligned} G_\varepsilon &= -[K^{\beta\varepsilon}\varphi_\delta(\mathbf{r})|f_{p,\beta\gamma})V^{pq}(f_{q,\beta\gamma}|\varphi_\beta\varphi_\gamma)K^{\delta\gamma} + K^{\beta\alpha}(\varphi_\alpha\varphi_\delta|f_{p,\beta\varepsilon})V^{pq}(f_{q,\beta\varepsilon}|\varphi_\beta(\mathbf{r})K^{\delta\varepsilon}] \\ &= -[K^{\beta\varepsilon}\varphi_\delta(\mathbf{r})|f_{p,\beta\gamma})V^{pq}(f_{q,\beta\gamma}|\varphi_\beta\varphi_\gamma)K^{\delta\gamma} + K^{\delta\varepsilon}\varphi_\beta(\mathbf{r})|f_{q,\beta\varepsilon})V^{qp}(f_{p,\beta\varepsilon}|\varphi_\delta\varphi_\alpha)K^{\beta\alpha}] \\ &= -[K^{\beta\varepsilon}\varphi_\delta(\mathbf{r})|f_{p,\beta\gamma})V^{pq}(f_{q,\beta\gamma}|\varphi_\beta\varphi_\gamma)K^{\delta\gamma} + K^{\beta\varepsilon}\varphi_\delta(\mathbf{r})|f_{q,\delta\varepsilon})V^{qp}(f_{p,\delta\varepsilon}|\varphi_\beta\varphi_\alpha)K^{\delta\alpha}] \\ &= -[K^{\beta\varepsilon}[\varphi_\delta(\mathbf{r})|f_{p,\beta\gamma})V^{pq}(f_{q,\beta\gamma}|\varphi_\beta\varphi_\gamma) + \varphi_\delta(\mathbf{r})|f_{q,\delta\varepsilon})V^{qp}(f_{p,\delta\varepsilon}|\varphi_\beta\varphi_\gamma)]K^{\delta\gamma}, \end{aligned} \quad (31)$$

which is a direct consequence of using the 2-centre fitting.

An analogous expression can be derived for the expansion using the overlap metric. The first term in the final square bracket corresponds to the potential due to the density of the product of φ_β and φ_γ^*, expanded in spherical waves centred on β and γ, acting on φ_δ, which must overlap with the NGWF with respect to which we differentiate, φ_ε^*. Here, the implicit summation involves all NGWFs φ_β within a cut-off distance of φ_ε^*, all NGWFs φ_γ^* that overlap with φ_β and all NGWFs φ_δ that overlap with φ_ε^*. The second term is the potential due to the same density $\varphi_\beta\varphi_\gamma^*$, acting on the same φ_δ, but now the expansion basis involves only the spherical waves centred on NGWFs φ_δ and φ_ε^*.

While the first term is straightforward to compute, evaluating the second term in the form above is computationally intensive, since the product $\varphi_\beta\varphi_\gamma^*$ needs to be re-expanded every time δ or ε changes. Below we show how a simple reordering of summations and of the expansion operation can be used to work around this problem.

First, let us denote with $\hat{P}_{\kappa\lambda}$ the projection operator that expands a potential $|\varphi_\eta\varphi_\theta)$ due to a density $\varphi_\eta\varphi_\theta^*$ into spherical waves centred on atoms to which NGWFs κ and λ belong, i.e.,

$$\hat{P}_{\kappa\lambda} = |f_{q,\kappa\lambda})V^{qp}(f_{p,\kappa\lambda}|. \quad (32)$$

Using this notation we can rewrite (31) as

$$G_\varepsilon = K^{\beta\varepsilon}\varphi_\delta(\mathbf{r})[\hat{P}_{\beta\gamma}|\varphi_\beta\varphi_\gamma) + \hat{P}_{\varepsilon\delta}|\varphi_\beta\varphi_\gamma)]K^{\delta\gamma}. \quad (33)$$

We note that $\hat{P}_{\beta\gamma}|\varphi_\beta\varphi_\gamma)$ and $\hat{P}_{\varepsilon\delta}|\varphi_\beta\varphi_\gamma)$ both tend to $|\varphi_\beta\varphi_\gamma)$ as the quality of the auxiliary basis set is improved, but for a finite auxiliary basis these three quantities are not strictly interchangeable, which is the reason that the two terms in the final square bracket of (31) are not identical.

We now rewrite (31) again using real-valued NGWFs, and explicitly denoting the summations to indicate the order in which they can be performed efficiently

$$G_\varepsilon = 2\sum_\delta\varphi_\delta(\mathbf{r})\sum_\beta K^{\beta\varepsilon}\sum_\gamma K^{\delta\gamma}\hat{P}_{\beta\gamma}|\varphi_\beta\varphi_\gamma) + 2\sum_\delta\varphi_\delta(\mathbf{r})\hat{P}_{\varepsilon\delta}\left(\sum_\beta K^{\beta\varepsilon}\sum_\gamma K^{\delta\gamma}|\varphi_\beta\varphi_\gamma)\right), \quad (34)$$

where the factor of 2 is due to the fact that the NGWFs are now real-valued.

For the calculation of the first term in (33), each potential $|\varphi_\beta\varphi_\gamma)$ only needs to have its generating density expanded in terms of spherical waves originating on the same centres that generate the density. For the calculation of the second term, we observe that the quantities summed over β and γ are all connected with the same ε and δ. We can therefore sum them over β and γ and then expand them in one go. Furthermore,

the sums over γ do not depend on ε, which makes it possible to store and re-use their values, further improving efficiency.

III. RESULTS AND DISCUSSION

A. Validation and tests

1. Accuracy of the metric matrix

Our first validation test focused on assessing how the accuracy of the metric matrix **V** obtained through Chebyshev interpolation (cf. Sec. II D 1) depends on the number of intervals and the order of the Chebyshev polynomials used. We chose an isolated chlorosilane molecule in a $(45\,a_0)^3$ cubic box as our test case. The psinc kinetic energy cutoff was 827 eV. The quality of the auxiliary basis set was kept at $l_{\max} = 3$, $q_{\max} = 10$ for a total of 160 spherical waves per centre, yielding a dense 1280×1280 matrix **V** (cf. Fig. 4). 13.2% of the elements were zero, owing to the orthogonality of same-site spherical waves with different angular momenta. Out of the remaining elements, 96.7% were larger in magnitude than 0.0001 and 22.8% were larger in magnitude than 0.1. We found that obtaining the elements of **V** to an accuracy of the sixth decimal was sufficient for stable calculations.

Figure 5 shows the resulting accuracy of the matrix elements as a function of the order of the Chebyshev polynomials and the number of intervals used. Since the exact values were not available for comparison, we used an extremely accurate calculation (N_o = 16th-order polynomials on N_i = 16 intervals) as reference. The results demonstrate that, e.g., interpolation with 10th-order polynomials over 10 intervals is already sufficiently accurate. We propose $N_o = 12$, $N_i = 12$ as the default setting.

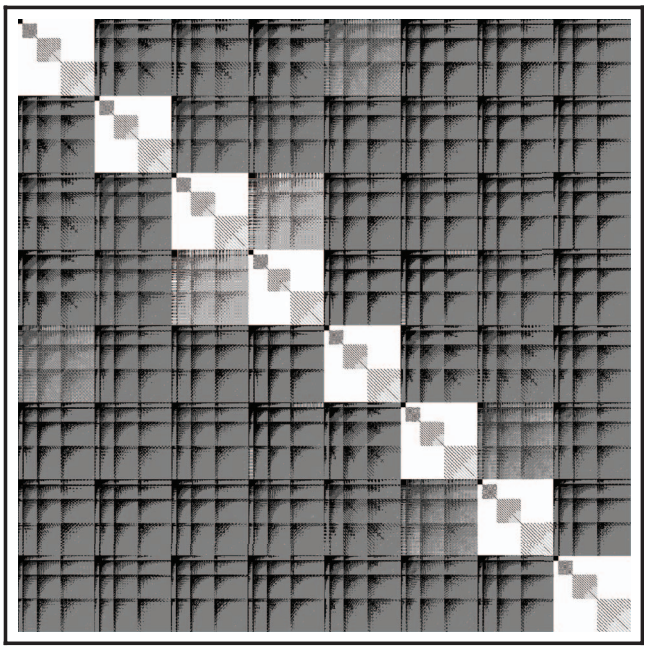

FIG. 4. Structure of the metric matrix **V** for chlorosilane. Values larger than 0.1 are drawn in black, values between 0.0001 and 0.1 are drawn in dark grey, non-zero values below 0.0001 are drawn in light grey, values of exactly zero are shown in white.

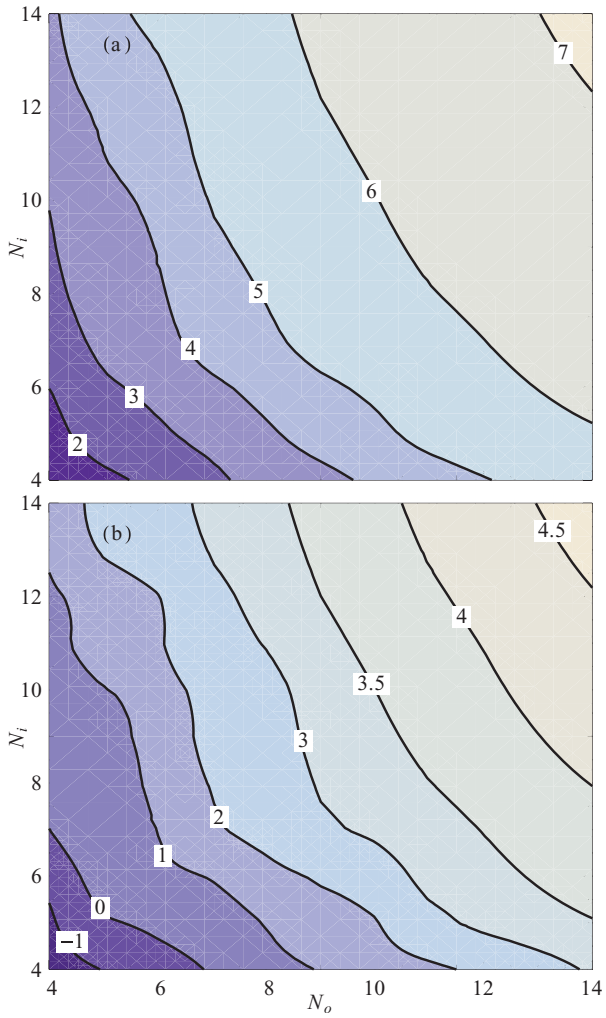

FIG. 5. Number of correct decimals in the **V** matrix: (a) average, (b) for the element with the maximum error, for various values of the interpolation order (N_o) and number of intervals (N_i).

2. Stability, accuracy of the exchange matrix and exchange energy

Our next test was aimed at assessing the minimum quality of the auxiliary basis set sufficient for obtaining chemically accurate results. Furthermore, we wanted to ensure that the two-centre expansion remains stable when auxiliary basis sets of reasonable quality are used. Our initial tests indicated that, for instance, with extremely inaccurate auxiliary basis sets, e.g., with only 24 spherical waves per centre ($l_{\max} = 1$, $q_{\max} = 6$, yielding $N_{\mathrm{SW}} = 48$), the obtained exchange matrix was accurate to only 3.2 decimals (we explain below how this was calculated), which was not enough to recover the broken bra-ket symmetry by simple symmetrisation. Thus, unlike with the four-centre expansion, which remains stable even for extremely poor qualities of the auxiliary basis set (only becoming inaccurate), here it becomes crucial to demonstrate that the calculation remains stable at least for moderately accurate auxiliary basis sets. Again, we chose an isolated chlorosilane molecule as a test case. A cubic box $(42.0\,a_0)^3$ in size was used, with the psinc kinetic energy cutoff set at 1292 eV. This slightly higher value

(corresponding to a psinc spacing of $0.4\,a_0$) was chosen, because it allows spherical waves of up to $q_{\max} = 20$ (for $l_{\max} \leq 4$) to be represented on the psinc grid without aliasing. The metric matrix was calculated to a higher-than-default accuracy ($N_o = 14$, $N_i = 14$) to ensure that the Chebyshev interpolation would not introduce any significant further error.

Our results confirm that the calculation becomes unstable only when the quality of the auxiliary basis set is very poor, i.e., when either the maximum angular momentum $l_{\max}$ or the number of Bessels functions $q_{\max}$ used is insufficient to adequately represent the fitted charge densities. This instability manifests itself as inaccuracies in the NGWF gradient, causing the NGWF optimisation to stall about half an order of magnitude above the default threshold. For example, with only $q_{\max} = 4$ Bessel functions it was not possible to converge the calculation regardless of how big $l_{\max}$ was. As more Bessel functions were added, the noise in the gradient quickly diminished to a level that allowed convergence to default thresholds. Fig. 6 demonstrates this for $l_{\max} = 3$. Similarly, when $l_{\max}$ was too small, e.g., when spherical waves with only s and p symmetry ($l_{\max} = 1$) were used, convergence was not reached regardless of the number of Bessel functions used in the expansion.

Fig. 7 makes it clear that the underlying reason for impaired convergence is that the exchange matrix $\mathbf{X}$ becomes less accurate as the quality of the auxiliary basis set is made worse. We find that the average number of correct decimals in $\mathbf{X}$, calculated as $c(\mathbf{X}) = \frac{1}{N_{\text{SW}}^2} \sum_{i=1}^{N_{\text{SW}}} \sum_{j=1}^{N_{\text{SW}}} \log_{10} |(\mathbf{X} - \mathbf{X}^T)_{ij}|$ is a useful metric of how adequate the quality of the auxiliary basis is, and an excellent predictor of the stability of the calculations. None of the calculations with $c(\mathbf{X}) < 3.8$ converged, while all calculations with $c(\mathbf{X}) > 3.8$ did. The above gives us confidence that the number of spherical waves needed for a stable operation of the proposed approach is not excessive and is easily achievable on standard psinc grids (with corresponding kinetic energy cutoffs in the range of 800–1200 eV).

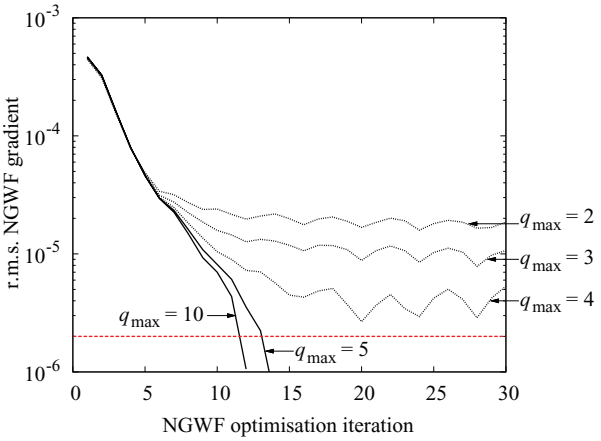

FIG. 6. Convergence of NGWF optimisation for the example case of $l_{\max} = 3$. Calculations using excessively low-quality auxiliary basis sets (4 and fewer Bessel functions), shown with dotted lines, did not converge to the default gradient threshold (denoted with the dashed red line). Using 5 Bessel functions was enough to achieve stable convergence.

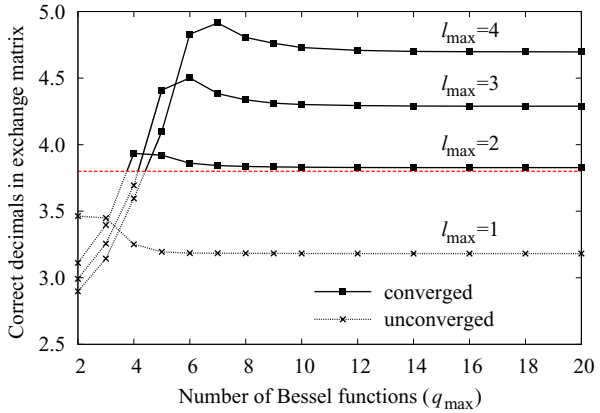

FIG. 7. Average number of correct decimals, $c(\mathbf{X})$, in the exchange matrix, depending on the quality of the auxiliary basis set. The points represent results of calculations, the lines serve as guides for the eye. Calculations that failed to converge are denoted with crosses and are shown for completeness. The dashed red line at $c(\mathbf{X}) = 3.8$ roughly delineates the quality of the auxiliary basis set that was necessary to achieve stable convergence (for chlorosilane).

Having ensured that our calculations are stable, we now show how the accuracy of the calculated energies depends on the quality of the auxiliary basis set. Our metric of accuracy is the fraction of exchange energy correctly recovered (with the direct $O(N^2)$ approach as reference) and we show the results in Fig. 8. For our test case, the auxiliary basis sets that were barely sufficient for stable operation already recover 96% of exchange energy, while using the high-quality basis sets ($l_{\max} = 4$, $q_{\max} \geq 12$) allowed us to recover more than 99.8%.

Subsequently, we set out to demonstrate that for reasonable qualities of the auxiliary basis set, the 2-centre expansion that we propose does not introduce significant errors compared to the more straightforward (but impractically expensive) 4-centre expansion, both when the exchange-interacting NGWFs overlap and when they do not (in the latter case the approaches are strictly equivalent). Our test case, which necessarily needed to be small, in order to permit calculations with the 4-centre approach, was a hydrogen-bonded water dimer, on which we ran calculations for varying lengths of the

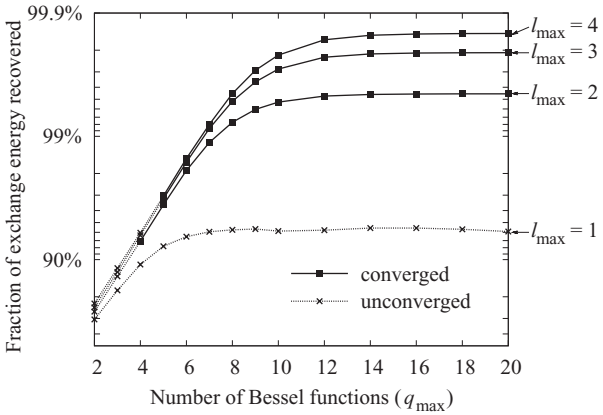

FIG. 8. Fraction of the exchange energy recovered, depending on the quality of the auxiliary basis set. The points represent results of calculations, the lines serve as guides for the eye. Calculations that failed to converge are denoted with crosses and are shown for completeness.

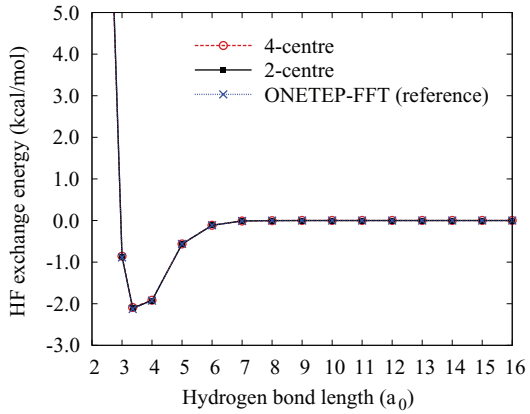

FIG. 9. Inter-molecular exchange energy between two hydrogen-bonded water molecules, as a function of hydrogen bond length.

hydrogen bond. In contrast to the remaining validation tests, we chose an auxiliary basis set of only moderate quality, by setting $l_{\max} = 3$, $q_{\max} = 10$. In order to make our presentation clearer, we only show the exchange interaction *between* the H_2O molecules, i.e., we subtract the intra-molecular exchange of each H_2O molecule with itself in each case. Fig. 9 shows that the energies obtained with the 2-centre expansion are extremely close to those obtained with the 4-centre expansion and to the results of the reference FFT-based approach. As the inter-molecular exchange decays very rapidly, to gain better insight into the magnitude of the error, we show the relative error of the two approaches, with the FFT-based approach as reference, with a separate plot, Fig. 10. This plot clearly demonstrates that the inaccuracy of both approaches (which is an expected consequence of density fitting) is very similar and does not exceed 6% even when the inter-molecular exchange energy decays to as little as 10^{-6} kcal/mol. The error in the *total* exchange energy was in the order of 0.7% for both approaches, regardless of the hydrogen bond length, which we believe is very good for this quality of the auxiliary basis set.

Finally, we demonstrate how the accuracy of the 2-centre expansion is similar to that of the 4-centre expansion regardless of the quality of the auxiliary basis set. Since the latter

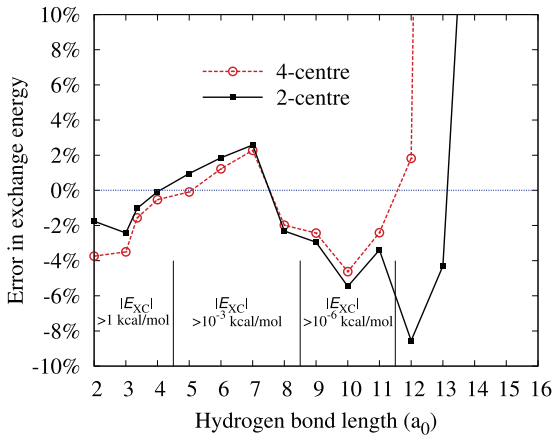

FIG. 10. Relative error in the inter-molecular exchange energy between two hydrogen-bonded water molecules, as a function of hydrogen bond length.

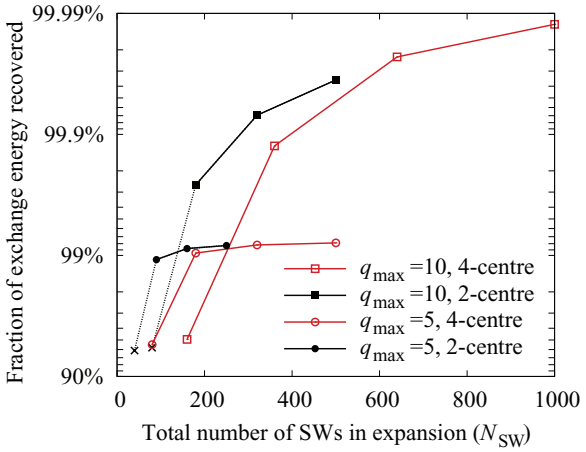

FIG. 11. Fraction of exchange energy recovered with the 2-centre approach proposed in this work (filled black symbols) and the 4-centre approach (empty red symbols). Squares and circles denote calculations with 10 and 5 Bessel functions per centre, respectively. The lines are only meant as guides for the eye. The quadratically scaling FFT approach (cf. Sec. II C) was used as reference. The points on the plot correspond to $l_{\max}$ of 1, 2, 3, and 4. The corresponding total numbers of spherical waves (N_{SW}) differ between the approaches, because they use 2 and 4 centres, respectively. Calculations with $l_{\max} = 0$ did not converge and are not shown. Calculations with $l_{\max} = 1$ did not converge with the 2-centre approach and their results (denoted with crosses) are only shown for completeness.

approach is very computationally expensive, we chose a small system (chlorosilane) as our test case. In Fig. 11, we show the percentage of exchange energy recovered by the two approaches, with the $O(N^2)$ FFT approach as reference. Even with only 5 Bessel functions per centre, both approaches easily recover over 99% of exchange energy. With the default setting of 10 Bessel functions per centre, both approaches are able to recover over 99.9% of exchange energy. For a fixed total number of spherical waves in the expansion, the 2-centre expansion performs better, although, as expected, for poor qualities of the auxiliary basis set (at $l_{\max} \leq 1$) it is not accurate enough for the calculation to converge. Neither approach converges when only s spherical waves are used (i.e., for $l_{\max} = 0$).

The accuracy of relative energies is demonstrated in Fig. 12, on the example of a bond-stretch of the same chlorosilane molecule. With a high-quality auxiliary basis set ($l_{\max} = 4$, $q_{\max} = 10$), the bond-stretch energy of 6.309 kcal/mol was calculated with an error < 0.02 kcal/mol with both the proposed 2-centre approach and the 4-centre approach.

B. Bond-stretch curve for ethene

We subsequently compared the bond-stretch curves obtained with ONETEP with those obtained from CASTEP and NWCHEM for three functionals: PBE, HF, and B3LYP. For the non-local functionals, the ONETEP calculations were performed with the approach put forward in this work and with the direct $O(N^2)$ approach as reference. A kinetic energy cutoff of 827 eV was used in ONETEP and CASTEP, and NWCHEM used a cc-pVQZ basis set. The agreement (cf. Fig. 13) between ONETEP and CASTEP is excellent (given that the same norm-conserving pseudopotentials were used),

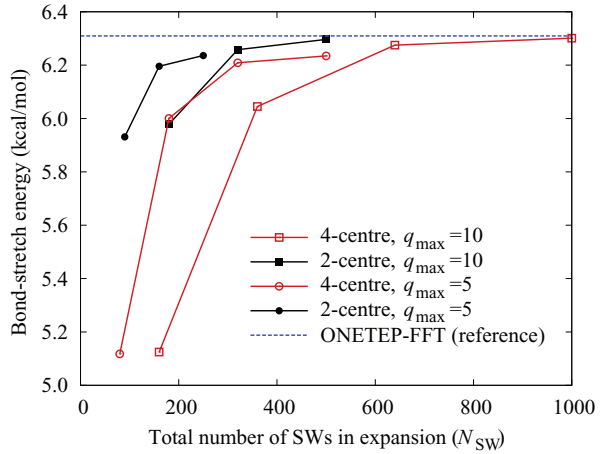

FIG. 12. The bond-stretch energy for a chlorosilane molecule, calculated with the 2-centre approach proposed in this work (filled black symbols) and the 4-centre approach (empty red symbols). Squares and circles denote calculations with 10 and 5 Bessel functions per centre, respectively. The lines are only meant as guides for the eye. The quadratically scaling FFT approach (dashed blue line) (cf. Sec. II C) was used as reference. The points on the plot correspond to l_{max} of 1, 2, 3, and 4. The corresponding total numbers of spherical waves (N_{SW}) differ between the approaches, because they use 2 and 4 centres, respectively. Results of calculations with extremely poor quality auxiliary basis sets that did not converge ($l_{max} = 0$ for the 4-centre approach, $l_{max} \leq 1$ for the 2-centre approach) are not shown to retain clarity.

and these results agree remarkably well with NWCHEM, considering that the first two approaches use plane waves and pseudopotentials, while NWCHEM uses Gaussian basis sets and performs an all-electron calculation.

Similar bond-stretch curves were produced for four additional qualities of the auxiliary basis set. As it would be difficult to present them on a single plot, we instead show (cf. Fig. 14) how the predicted equilibrium bondlength varied depending on the quality of the auxiliary basis set, for the calculation with a pure Hartree-Fock approach, which was the most sensitive. We find good convergence of the equilibrium bondlength with increasing number of spherical waves used in the expansion. Already with 250 spherical waves the result was no further than 0.003 a_0 from the reference ONETEP result and the CASTEP result, and no further than 0.007 a_0 from what NWCHEM predicted.

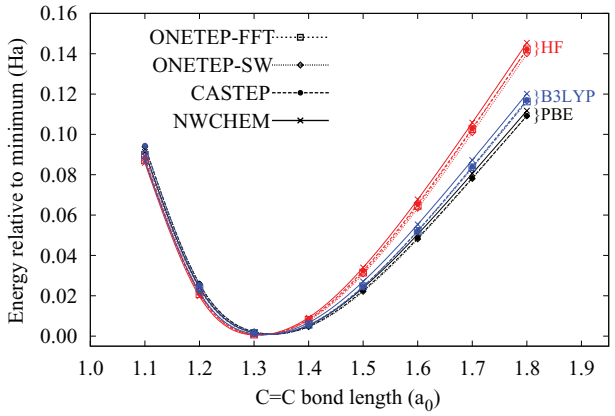

FIG. 13. Bond stretch curves for ethene obtained with Hartree-Fock exchange, B3LYP, and PBE using ONETEP, CASTEP, and NWCHEM. ONETEP-FFT denotes the direct $O(N^2)$ approach that uses simulation-cell FFTs to obtain the potential when calculating exchange (cf. Sec. II C). ONETEP-SW denotes the $O(N)$ approach that uses spherical waves as auxiliary basis, put forward in this work. The auxiliary basis set used $n_{SW} = 250$ spherical waves ($l_{max} = 4$, $q_{max} = 10$) on each centre, for a total of $N_{SW} = 500$. Points on the plot correspond to calculated values, the lines represent a spline fit to data. The curves have been shifted so that their minima are at zero. The position of each minimum was determined from a parabolic fit to its three closest points.

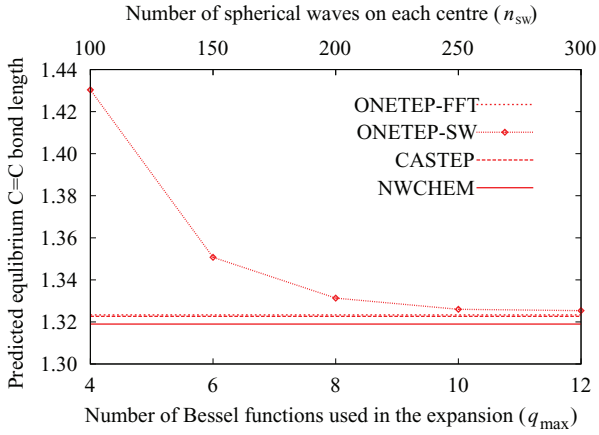

FIG. 14. Convergence of the predicted equilibrium C=C bondlength in ethene with the increasing quality of the auxiliary basis set. Calculations were performed with the Hartree-Fock approach. Spherical waves up to $l_{max} = 4$ were used in the expansion. The total number of spherical waves used in the expansion, N_{SW}, is twice the value shown on the upper x axis, as two-centre expansion was performed.

C. Demonstration of accuracy: Geometry optimisation of small molecules

Here, we demonstrate the accuracy of the proposed approach by using its implementation in ONETEP to perform geometry optimisation for eight molecules from the T-96R test set[43] using the B3LYP functional. Two local functionals (LDA, PBE) were used for comparison. Results (equilibrium bondlengths obtained from full geometry optimisation) were compared against ONETEP's direct $O(N^2)$ reference implementation and against CASTEP. Each molecule was geometry-optimised in a $(30\,a_0)^3$ cubic box, with a plane-wave kinetic energy cutoff of 1292 eV. The cutoff-Coulomb technique[34] was used to ensure that the periodic images of the molecules did not interact. A high-quality auxiliary basis set was used ($l_{max} = 4$, $q_{max} = 12$), for a total of $N_{SW} = 600$ spherical waves used in the expansion.

Fig. 15 shows the deviation from the experimental[44] bondlengths, which are given at the top of each panel. Our aim was not to compare how well each functional reproduces experimental bondlengths, but rather to demonstrate that our predictions with B3LYP agree with CASTEP to a similar degree as for PBE or LDA calculations. Table I summarises the findings, which show that this is indeed the case.

D. Comparison against DFT+U

DFT+U is a computationally inexpensive and well-established approach for including a self-interaction correction in DFT calculations, at least for certain atomic sites

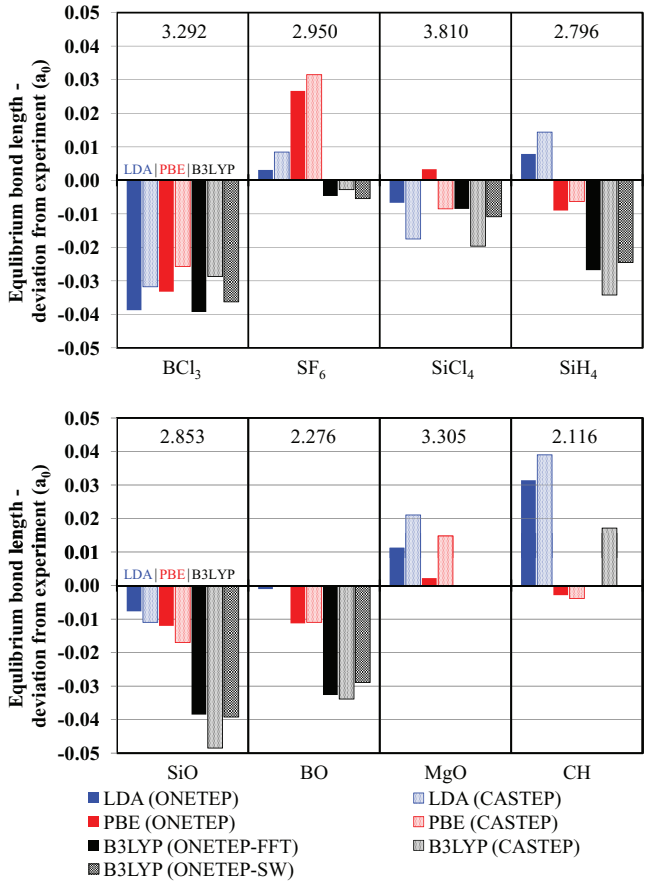

FIG. 15. Equilibrium bondlengths of BCl_3 (B–Cl), SF_6 (S–F), $SiCl_4$ (Si–Cl), SiH_4 (Si–H), SiO, BO, MgO, and CH compared with experiment, calculated with ONETEP and CASTEP using different functionals. The first two bars for each molecules correspond to LDA, the next two – to PBE, the last three bars correspond to B3LYP. The experimental equilibrium bondlength is given at the top, the bars show the signed deviation from experiment.

(such as transition metals) for which this correction is crucial. Here, we aim to compare results obtained with standard DFT calculations with the B3LYP hybrid exchange-correlation functional (employing the approach put forward in this work) against those of the DFT+U implementation[26,27] available in ONETEP. Given that the two approaches set out to include the same physical effect, albeit in a radically different manner, the results they produce should be consistent, at least qualitatively. By using the two techniques as implemented within the same code (ONETEP), we minimise the potential

TABLE I. Mean absolute deviation (MAD) between the equilibrium bondlength obtained with ONETEP with respect to CASTEP. SW denotes the approach put forward in this work, FFT denotes the reference $O(N^2)$ approach. The MAD between the SW and FFT approaches was 0.0016 a_0.

XC functional	Mean absolute deviation (a_0)
LDA	0.0064
PBE	0.0057
B3LYP-SW	0.0075
B3LYP-FFT	0.0074

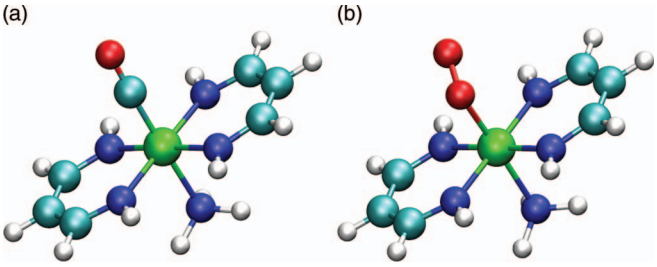

FIG. 16. The truncated myoglobin models used in our calculations (a) with CO bound, and (b) with O_2 bound.

sources of differences related to the code and the underlying numerical framework (such as differences due to basis sets or pseudopotentials).

We have used the two methods to investigate the relative binding energies of CO and O_2 to a truncated myoglobin model, similar to the model of Oláh and Harvey.[45] Larger models, which also included portions of the protein, were used previously by Cole *et al.*[46] to study ligand (CO and O_2) discrimination in myoglobin with the DFT+U approach in ONETEP. The truncation we applied was dictated by reasons of computational efficiency – as our implementation of Hartree-Fock exchange is not yet parallel-ready, calculations on the larger models would take an impractically long time. The chosen models are shown in Fig. 16. The psinc kinetic energy cutoff was 827 eV and the NGWF localisation radius was taken as 9 a_0.

The choice of the correct ground state for the unligated system was not obvious, as the triplet and quintet states lie very close in energy. Pure DFT calculations with the PBE exchange-correlation functional predicted the triplet to be more favourable (by 8.8 kcal/mol), while the result of DFT+U varied depending on the choice of U. As U was increased, the triplet ground state became progressively less favourable and at higher values of U it was the quintet that was favoured, with a cross-over at about $U = 2.5$ eV. With B3LYP the triplet state was always preferred, regardless of the quality of the auxiliary basis set (although the energy difference varied by as much as 1.5 kcal/mol). The ground state of the complex with O_2 was also carefully chosen, and with all approaches we found that an open shell singlet (a singlet diradical) is more favourable than a closed-shell singlet, in agreement with the observations by Cole *et al.*[46] on the larger myoglobin models. Magnetic symmetry was artificially broken to obtain the singlet diradical state through the application of effective magnetic fields of opposite sign to the Fe 3*d* and O_2 manifolds, following the approach outlined in Ref. 46.

Fig. 17 shows the binding energy of both ligands. CO is seen to be always preferred, at least for the range of values of U studied, and both ligands are predicted to be binders, except at the highest U. The binding energies from the B3LYP calculations are very close to the DFT+U values when $U \approx 2$ eV.

Fig. 18 shows the relative binding energy of the ligands (with positive values indicating preference for CO). We compare the results obtained with DFT+U for our truncated model, the DFT+U results of Cole *et al.*[46] obtained for larger

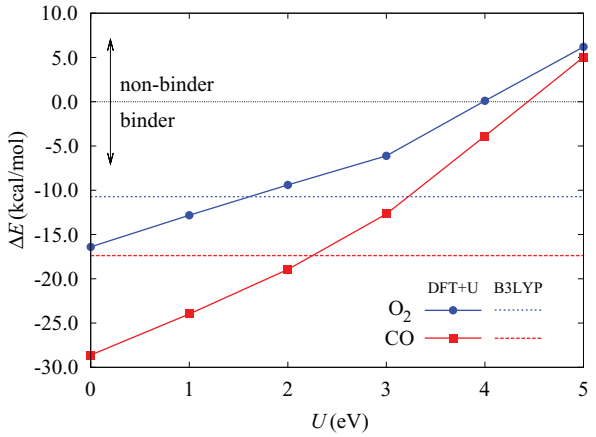

FIG. 17. Binding energy of CO and O_2 to the truncated myoglobin model. Negative values indicate binding is favoured. The curves with points show DFT+U (PBE) results for varying U, the dashed lines correspond to DFT (B3LYP) calculations with the approach presented in this work.

models, and the results obtained with B3LYP, employing the technique proposed in this work (for the truncated model). The figure demonstrates that as the size of the auxiliary basis set is increased, the results converge rapidly. Our B3LYP relative binding energy coincides with the DFT+U value obtained when U is set to about 3 eV. The similar shape of the dependence of DFT+U predictions on U between our truncated model and the three models of Cole *et al.*[46] demonstrate that our model, although drastically truncated, remains physically sound.

It is pleasing to be able to demonstrate that the two approaches give near-quantitative agreement for certain values of U, as would be expected.

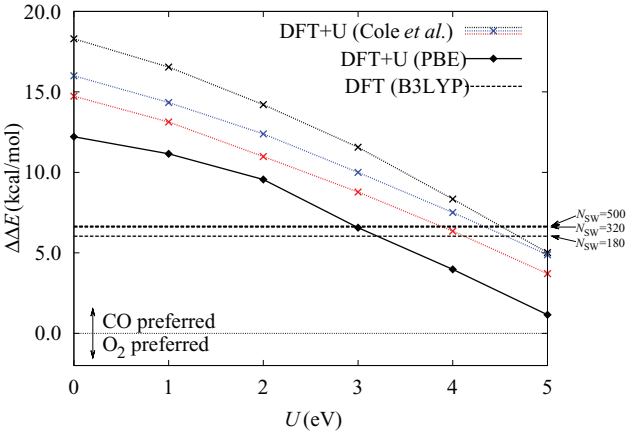

FIG. 18. Relative binding energy to the truncated myoglobin model of O_2 with respect to CO. Positive values indicate preference for CO. The curve with diamonds shows DFT+U (PBE) results for varying U, the dashed lines correspond to DFT (B3LYP) calculations with the approach presented in this work. Results for three qualities of auxiliary basis set are shown: $l_{max} = 2$, $q_{max} = 10$, $N_{SW} = 180$; $l_{max} = 3$, $q_{max} = 10$, $N_{SW} = 320$, and $l_{max} = 4$, $q_{max} = 10$, $N_{SW} = 500$. The dotted lines with crosses are the results of Cole *et al.*[46] obtained through DFT+U for different haemoglobin models. The colour-coding follows that of Ref. 46 (black: model with 1 residue, blue: model with 3 residues, red: model with 53 residues).

E. Linear scaling and effect of truncation

Finally, to demonstrate that the proposed approach is indeed linear-scaling, we benchmarked the code on two hydrocarbon polymers of increasing length: a chain of polyethylene, which is an example of an insulating material, and a chain of polyacetylene, which is an example of a small-bandgap system ("model conductor"). The calculations employed an NGWF radius of $7\,a_0$. Exchange interactions were truncated beyond $r_X = 15\,a_0$, $20\,a_0$, $25\,a_0$, and infinity (no truncation). Spherical waves up to $l_{max} = 4$, $q_{max} = 10$ were used in the expansion. 10th-order polynomials over 10 intervals were used in the Chebyshev interpolation to calculate the metric matrix, $\mathbf{V}$. The density kernel was not truncated. The quadratically scaling FFT-based approach (cf. Sec. II C) was used as reference.

In Fig. 19, we show the relative error in exchange energy, with respect to the reference calculation. We note that the total error for both systems remains below 0.2% and is due to the finite size of the auxiliary basis set. The additional error incurred by truncating exchange interactions beyond a distance-based cutoff is very small in comparison – below 0.1% for the model conductor and below 10^{-5}% for the model insulator, which makes the curves for the latter overlap so that they are indistinguishable on the plot.

The limited system sizes tractable with the current serial implementation make it difficult to determine the asymptotic behaviour of the error. Only in the case of the conducting system and only for the crudest approximation ($r_X = 15\,a_0$), the relative error increases monotonically, and even then it is expected to be modest, even for systems with thousands of atoms, to the best that can be inferred with a limited number of data points. Further investigation is needed to demonstrate with certainty that the error committed by truncating exchange at reasonable cutoffs (say, $r_X = 20\,a_0$) is well-behaved for the largest systems of interest in DFT calculations. We

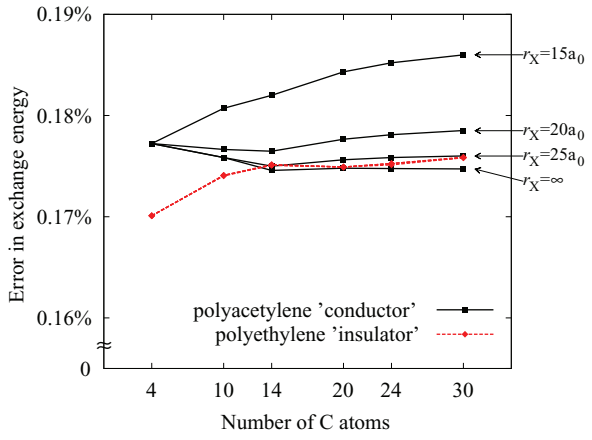

FIG. 19. Relative error in exchange energy obtained with the proposed approach, with the quadratically scaling approach as reference. The bulk of the error, which is below 0.2%, is due to the finite size of the auxiliary basis set. For the model conductor, the additional error incurred by truncating exchange is seen to be below 0.01% even when the cut-off radius is very short, regardless of system size. For the model insulator, the error incurred by truncating exchange is negligible ($<10^{-5}$%), regardless of system size.

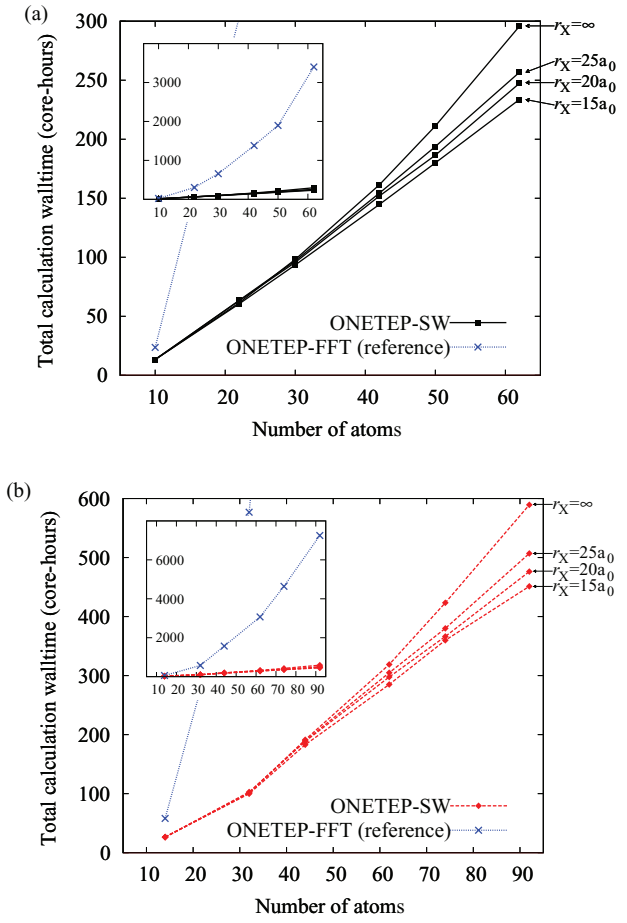

FIG. 20. Total computational effort (core-hours) for a fully converged calculation on polyethylene chains (a) and polyacetylene chains (b) of increasing length. The reference approach (dotted blue line) is quadratically scaling. The proposed approach (solid black lines – polyethylene, dashed red lines – polyacetylene) is linear-scaling when a finite cutoff for exchange is assumed, and is already faster than the reference calculation for the smallest system studied, C_4H_6.

plan to demonstrate this once the implementation of our approach is parallel-ready.

Fig. 20 shows how the total time of the calculation depends on system size and exchange truncation cutoff. As expected, the proposed approach is linear-scaling when a finite exchange interaction cutoff is employed. Calculations with no exchange truncation scale quadratically; however, in this size regime the quadratic-scaling term does not yet dominate the calculation time.

IV. CONCLUSIONS

We have presented a method for the calculation of four-centre two-electron repulsion integrals in terms of NGWFs. We have developed this method within the ONETEP program where the NGWFs are expressed in terms of a basis set of psinc functions, which is equivalent to a plane wave basis set and systematically improvable. We have used our ERI method to implement the calculation of Hartree-Fock exchange energy in the context of pure Hartree-Fock calculations and DFT calculations with hybrid exchange-correlation functionals. During these calculations, the NGWFs are optimised *in situ*, as influenced by their chemical environment. This approach allows calculations to be performed with the accuracy of large, high-quality basis sets, as we have confirmed by tests on a wide variety of molecules in which we have compared with results obtained with codes which use plane-wave and Gaussian basis sets. In these tests, we obtain excellent agreement with the plane wave calculations or the Gaussian basis set calculations, provided a large, high-quality Gaussian basis set has been used.

We have also investigated in depth the dependence of the accuracy of our ERI algorithm on numerical calculation parameters, related to the spherical waves which are used as an auxiliary basis set, in order to choose default values suitable for high-accuracy calculations.

The DFT+U approach, which is also available in the ONETEP code, aims to introduce the same physical effects as hybrid functionals but via a completely different methodology. We have performed calculations on small myoglobin models using both our approach (B3LYP functional) and DFT+U, with identical calculation parameters, and have confirmed that the two methods have strong qualitative agreement. Quantitative agreement can be obtained for the appropriate value of the Hubbard U parameter, or in other words our implementation of hybrid exchange-correlation functionals could be used to parameterise the U on small model systems in order to obtain the most physically suitable U values for calculations on larger systems.

Even though our ERI code is so far serial, we were also able to perform calculations on large enough numbers of atoms to demonstrate linear-scaling of the computational effort with respect to the number of atoms, taking advantage of the strict localisation of the NGWFs. We have performed single-point energy calculations on polyethylene and polyacetylene chains of increasing length, up to $C_{30}H_{62}$ and $C_{30}H_{32}$. With the future parallelisation and optimisation of the ERI code, we expect to be able to run Hartree-Fock and hybrid DFT calculations on several thousand atoms.

The methods we have presented in this paper are the foundation for many important future developments within the ONETEP program which depend on the availability of ERIs. These will include methods for ground-state properties such as the more recent screened-exchange functionals, wavefunction-based approaches such as Møller-Plesset and Random Phase Approximation (RPA) perturbation theories, as well as methods for electron addition or removal energies such as GW perturbation theory and methods for excited states such as various levels of configuration interaction.

ACKNOWLEDGMENTS

We would like to thank Dr. David D. O'Regan and Dr. Daniel J. Cole for help with DFT+U calculations on myoglobin models. We would also like to thank Professor Peter D. Haynes and Dr. Gilberto Teobaldi for useful discussions with regard to spherical waves and truncation of Hartree-Fock exchange, respectively. J.D. acknowledges the support of the Engineering and Physical Sciences Research Council (EPSRC) UK (EPSRC Grant Nos. EP/G055882/1 and EP/J015059/1) and of the Polish National Science Centre and

the Ministry of Science and Higher Education (Grant Nos. N N519 577838 and IP2012 043972). Q.H. would like to thank the EPSRC for research studentship funding. C.-K. S. would like to thank the Royal Society for a University Research Fellowship. The calculations in this work were carried out on the Iridis3 supercomputer of the University of Southampton and on the Galera supercomputer at the TASK Computer Centre (Gdansk, Poland).

[1]P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[2]W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[3]J. Tao, J. P. Perdew, V. N. Staroverov, and G. E. Scuseria, Phys. Rev. Lett. **91**, 146401 (2003).
[4]E. Artacho and L. Miláns del Bosch, Phys. Rev. A **43**, 5770 (1991).
[5]E. Schwegler, M. Challacombe, and M. Head-Gordon, J. Chem. Phys. **106**, 9708 (1997).
[6]C. Ochsenfeld, C. A. White, and M. Head-Gordon, J. Chem. Phys. **109**, 1663 (1998).
[7]S. A. Maurer, D. S. Lambrecht, D. Flaig, and C. Ochsenfeld, J. Chem. Phys. **136**, 144107 (2012).
[8]E. Rudberg, E. H. Rubensson, and P. Salek, J. Chem. Theory Comput. **7**, 340 (2011).
[9]J. Kussmann and C. Ochsenfeld, J. Chem. Phys. **138**, 134114 (2013).
[10]M. Guidon, J. Hutter, and J. VandeVondele, J. Chem. Theory Comput. **5**, 3010 (2009).
[11]M. Del Ben, J. Hutter, and J. VandeVondele, J. Chem. Theory Comput. **8**, 4177 (2012).
[12]X. Ren, P. Rinke, V. Blum, J. Wieferink, A. Tkatchenko, A. Sanfilippo, K. Reuter, and M. Scheffler, New J. Phys. **14**, 053020 (2012).
[13]B. I. Dunlap, J. W. D. Connolly, and J. R. Sabin, J. Chem. Phys. **71**, 3396 (1979).
[14]B. I. Dunlap, J. W. D. Connolly, and J. R. Sabin, J. Chem. Phys. **71**, 4993 (1979).
[15]C.-K. Skylaris, L. Gagliardi, N. C. Handy, A. G. Ioannou, S. Spencer, and A. Willetts, J. Mol. Struct.: THEOCHEM **501–502**, 229 (2002).
[16]R. Polly, H.-J. Werner, F. Manby, and P. J. Knowles, Mol. Phys. **102**, 2311 (2004).
[17]P. Merlot, T. Kjaergaard, T. Helgaker, R. Lindh, F. Aquilante, S. Reine, and T. B. Pedersen, J. Comput. Chem. **34**, 1486 (2013).
[18]B. I. Dunlap, J. Mol. Struct.: THEOCHEM **529**, 37 (2000).
[19]M. Lorenz, L. Maschio, M. Schütz, and D. Usvyat, J. Chem. Phys. **137**, 204119 (2012).
[20]M. Gibson, S. Brandt, and S. Clark, Phys. Rev. B **73**, 125120 (2006).
[21]M. Marsman, J. Paier, A. Stroppa, and G. Kresse, J. Phys.: Condens. Matter **20**, 064201 (2008).
[22]X. Wu, A. Selloni, and R. Car, Phys. Rev. B **79**, 085102 (2009).
[23]C.-K. Skylaris, A. A. Mostofi, P. D. Haynes, C. J. Pickard, and M. C. Payne, Comput. Phys. Commun. **140**, 315 (2001).
[24]P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, J. Phys. Chem. **98**, 11623 (1994).
[25]R. Peverati and D. G. Truhlar, Phys. Chem. Chem. Phys. **14**, 16187 (2012).
[26]D. D. O'Regan, N. D. M. Hine, M. C. Payne, and A. A. Mostofi, Phys. Rev. B **85**, 085107 (2012).
[27]D. D. O'Regan, N. D. M. Hine, M. C. Payne, and A. A. Mostofi, Phys. Rev. B **82**, 081102 (2010).
[28]C.-K. Skylaris, P. D. Haynes, A. A. Mostofi, and M. C. Payne, J. Chem. Phys. **122**, 084119 (2005).
[29]C.-K. Skylaris, A. A. Mostofi, P. D. Haynes, O. Diéguez, and M. C. Payne, Phys. Rev. B **66**, 035119 (2002).
[30]A. A. Mostofi, P. D. Haynes, C.-K. Skylaris, and M. C. Payne, J. Chem. Phys. **119**, 8842 (2003).
[31]P. D. Haynes, C.-K. Skylaris, A. A. Mostofi, and M. C. Payne, J. Physics: Condens. Mat. **20**, 294207 (2008).
[32]J. Almlöf, K. Korsel, and K. Faegri, Jr., J. Comput. Chem. **3**, 385 (1982).
[33]C.-K. Skylaris, P. D. Haynes, A. A. Mostofi, and M. C. Payne, Phys. Status Solidi B **243**, 973 (2006).
[34]M. R. Jarvis, I. D. White, R. W. Godby, and M. C. Payne, Phys. Rev. B **56**, 14972 (1997).
[35]N. D. M. Hine, J. Dziedzic, P. D. Haynes, and C.-K. Skylaris, J. Chem. Phys. **135**, 204103 (2011).
[36]O. Vahtras, J. Almlöf, and M. W. Feyereisen, Chem. Phys. Lett. **213**, 514 (1993).
[37]P. D. Haynes and M. C. Payne, Comput. Phys. Commun. **102**, 17 (1997).
[38]Q. Hill, "Development of more accurate computational methods within linear-scaling density functional theory," Ph.D. thesis (University of Southampton, 2010).
[39]B. Monserrat and P. D. Haynes, J. Phys. A: Math. Theor. **43**, 465205 (2010).
[40]N. D. M. Hine, P. D. Haynes, A. A. Mostofi, C.-K. Skylaris, and M. C. Payne, Comput. Phys. Commun. **180**, 1041 (2009).
[41]A. A. Mostofi, C.-K. Skylaris, P. D. Haynes, and M. C. Payne, Comput. Phys. Commun. **147**, 788 (2002).
[42]S. Goedecker and G. Scuseria, Comput. Sci. Eng. **5**, 14 (2003).
[43]V. N. Staroverov, G. E. Scuseria, J. Tao, and J. P. Perdew, J. Chem. Phys. **119**, 12129 (2003).
[44]*CRC Handbook of Chemistry and Physics*, 3rd ed., edited by D. R. Lide (CRC, Boca Raton, FL, 2002).
[45]J. Oláh and J. Harvey, J. Phys. Chem. A **113**, 7338 (2009).
[46]D. J. Cole, D. D. O'Regan, and M. C. Payne, J. Phys. Chem. Lett. **3**, 1448 (2012).