Method specific Cholesky decomposition: Coulomb and exchange energies

Linus Boman,[1] Henrik Koch,[1,a)] and Alfredo Sánchez de Merás[2]
[1]*Department of Chemistry, Norwegian University of Science and Technology, N-7491 Trondheim, Norway*
[2]*Instituto de Ciencia Molecular, Universidad de Valencia, P.O. Box 22085, E-46071 Valencia, Spain*

(Received 8 July 2008; accepted 3 September 2008; published online 2 October 2008)

We present a novel approach to the calculation of the Coulomb and exchange contributions to the total electronic energy in self consistent field and density functional theory. The numerical procedure is based on the Cholesky decomposition and involves decomposition of specific Hadamard product matrices that enter the energy expression. In this way, we determine an auxiliary basis and obtain a dramatic reduction in size as compared to the resolution of identity (RI) method. Although the auxiliary basis is determined from the energy expression, we have complete control of the errors in the gradient or Fock matrix. Another important advantage of this method specific Cholesky decomposition is that the exchange energy and Fock matrix can be evaluated with a linear scaling effort contrary to the RI method or standard Cholesky decomposition of the two-electron integral matrix. The methods presented show the same scaling properties as the so-called local density fitting methods, but with full error control. © *2008 American Institute of Physics*.
[DOI: 10.1063/1.2988315]

I. INTRODUCTION

The efficient calculation of self consistent field (SCF) or density functional energies is of profound importance in modern computational chemistry. Linear scaling formalisms have been developed by many groups. However, these algorithms only work efficiently for large systems in very small basis sets.[1] When more accuracy is needed, larger basis sets are essential, and for these cases the resolution of identity (RI) method[2,3] has been used to reduce the scaling. The RI method has worked very well for the calculation of the Coulomb contribution to the Fock matrix, but the exchange contribution cannot be evaluated with the same benefits. We shall discuss these problems in more detail. Another complication using the RI method is that the accuracy of the results is not easily controlled as the approach typically uses atom-centered preoptimized auxiliary basis sets. This implies that the errors will scale with the size of the system and can be quite significant. A typical argument found in the literature—that the error due to the incompleteness of the auxiliary basis is much smaller than the basis set error—speaks in favor of the RI method.[4] However what seems forgotten is the fact that calculated size-extensive properties cannot be used to extrapolate to the basis set limit as the errors can be larger than the accuracy of the extrapolation procedures.[5–7]

In this paper, we report an approach similar to the RI method that avoids the use of preoptimized auxiliary basis sets. We simply determine the auxiliary basis using the decomposition developed by Cholesky (1875–1918) and published by Benoit[8] in 1924. The idea to apply the Cholesky decomposition to the two-electron integral matrix was first suggested by Beebe and Linderberg[9] some 30 years ago. The Cholesky decomposition is the only numerical procedure known to the authors that can remove the zero or small eigenvalues of a positive semidefinite matrix without calculating the entire matrix. This makes the procedure truly unique, and the possibilities to obtain tremendous computational savings are apparent. Just consider the two-electron integral matrix. In the limit of a complete basis, the number of integrals scales as N^4, but the number of nonzero eigenvalues scales as N in the limit of a complete basis (N is the size of the basis). Despite this, the Cholesky decomposition does not seem to have received much attention in the quantum chemistry community, and only recently has the method come into focus. There are a few notable exceptions, especially the developments by Røeggen and Wisløff-Nilssen,[10] who have used the Cholesky decomposition of the two-electron integrals in the implementation of geminal models. The use of the Cholesky decomposition in connection with the calculation of derivative integrals has been discussed by O'Neal and Simons.[11] More recently, Koch *et al.*[12] developed an implementation of the Cholesky decomposition of the two-electron integrals aiming at large scale applications. The decomposition was shown to give very large savings in the evaluation of the SCF, density functional theory (DFT), and second order Møller–Plesset perturbation theory (MP2) for large basis sets. Since this implementation in a local version of the DALTON program[13] the Cholesky decomposition has formed the basis for many computational developments and applications.[14–19] The strategy has subsequently been adopted by the group around the MOLCAS program,[20] and Aquilante *et al.*[21] recently documented the usefulness in multiconfigurational SCF calculations. Recently Røeggen and Johansen[22] reported a parallel implementation of the Cholesky decomposition that shows a practically linear scaling with the number of compute nodes. However, the future use of Cholesky decomposition based methods will depend on the abilities to evaluate molecular derivatives. This has very recently been demonstrated by Aquilante *et al.*[23] al-

a)Author to whom correspondence should be addressed. Electronic mail: henrik.koch@chem.ntnu.no.

though efficiency was not the focus in this implementation. In any case, the standard Cholesky decomposition[9,12] is still perceived to be a rather complicated procedure;[24] this view is not shared by the authors that actually find it striking in its simplicity.

The use of the RI representation of the two-electron integral matrix is sometimes referred to as density fitting.[25] This name is rather misleading as the method goes much further than fitting a density. The most correct description would probably be to denote it as an inner projection method. The concept of inner projections was introduced into quantum chemistry by Löwdin[26,27] in his landmark papers on perturbation theory from 1965 to 1971, which appeared several years before Whitten[28] addressed the issue of integral approximations. The Cholesky decomposition is a powerful method to determine the optimal basis in the inner projection. In this paper, we depart from many of the previous ideas and take inner projections to an extreme. The key ingredient, the auxiliary basis, will not be preconstructed in any form, not preoptimized like in RI methods for all molecular systems or, like we have done previously,[12] Cholesky decomposing the two-electron integral matrix for the molecular system in question. We will determine the optimal auxiliary basis when evaluating individual terms in the iterative processes. This should ensure that the minimal auxiliary basis is used at any given time. Of course this comes at a price, but we will show that using the Cholesky decomposition it is indeed possible to determine the optimal auxiliary basis very efficiently and in the same procedure calculate the target quantity. Although the method is also applicable to electron correlation models, we focus in this paper on the Coulomb and exchange contributions to the energy and Fock matrix. In the section containing our conclusions, we briefly discuss the applicability to MP2, coupled cluster, and explicitly correlated methods. We start our analysis with some general considerations regarding the Cholesky decomposition.

II. THE CHOLESKY DECOMPOSITION

The Cholesky decomposition of a real symmetric positive semidefinite matrix M is most pedagogically defined in terms of the recursive formula

$$\tilde{M}_{pq} = M_{pq} - \frac{M_{pJ}M_{Jq}}{M_{JJ}} = M_{pq} - L_p^J L_q^J, \tag{1}$$

where the Cholesky vectors are defined as

$$L_p^J = \frac{M_{pJ}}{\sqrt{M_{JJ}}}. \tag{2}$$

As the $\tilde{M}$ matrix is also positive semidefinite, we may repeat the process using $\tilde{M}$ and in this way continue until all the diagonal elements are below a predetermined threshold T. After completion we obtain an approximate representation of the matrix M given by

$$M_{pq} = \sum_J L_p^J L_q^J + \Delta_{pq}, \tag{3}$$

where J labels the auxiliary basis and we have introduced the positive semidefinite error matrix Δ with matrix elements smaller than the threshold T. We note that all rows and columns in the error matrix corresponding to the decomposed diagonals will be zero ($\Delta_{Jp}=\Delta_{pJ}=0$), as can be easily seen in Eq. (1). To perform an incomplete Cholesky decomposition *does not* require the entire M matrix to be calculated; only the diagonal elements and the relevant columns given by the auxiliary basis J are needed. This is actually the power of the Cholesky decomposition because after the inclusion of each additional Cholesky vector all the diagonal elements become smaller and the Cauchy–Schwarz inequality is improved,

$$|\tilde{M}_{pq}|^2 \leq \tilde{M}_{pp}\tilde{M}_{qq} \leq M_{pp}M_{qq}. \tag{4}$$

The Cholesky decomposition is equivalent to Löwdin's inner projection, as discussed by Beebe and Linderberg.[9] We may express this as

$$M_{pq} \approx \sum_J L_p^J L_q^J = \sum_{IJ} M_{pI} S_{IJ}^{-1} M_{Jq}, \tag{5}$$

where the metric matrix S is defined as the M matrix in the subspace spanned by the auxiliary basis $S_{IJ}=M_{IJ}$. The Cholesky vectors can be calculated by performing a Cholesky decomposition of the S matrix,

$$S = KK^T, \tag{6}$$

where we assume that S is positive definite and K is lower triangular (note that the Cholesky decomposition is unique for positive definite matrices). We then obtain the following expression for the Cholesky vectors:

$$L_q^I = \sum_J K_{IJ}^{-1} M_{Jq}. \tag{7}$$

This can easily be shown using the following explicit expressions for K and K^{-1}:

$$K_{IJ} = \frac{1}{K_{JJ}}\left(S_{IJ} - \sum_{K=1}^{J-1} K_{IK}K_{JK}\right), \tag{8}$$

$$K_{IJ}^{-1} = \frac{1}{K_{II}}\left(\delta_{IJ} - \sum_{K=J}^{I-1} K_{IK}K_{KJ}^{-1}\right). \tag{9}$$

Equation (7) establishes the formal equivalence between the Cholesky decomposition and the inner projection. As we only need to decompose the S matrix, we may formulate a direct Cholesky decomposition algorithm where the elements of the matrix M entering Eq. (7) are calculated on the fly. This will eliminate the storage of the Cholesky vectors but will require the recalculation of the M matrix elements. For application to the two-electron integral matrix, we can initially decompose the S matrix spanned by the atom-centered orbital pairs and remove the linear dependence in this set.[29] This should then be followed by including auxiliary basis functions from the entire product space. Such an implementation will make the Cholesky decomposition limited by computational time only. An implementation of this direct procedure is in progress.

In most applications, the accurate calculation of the M matrix is not the objective but rather the expressions where it enters. In many cases, the M matrix enter in the following functional form:

$$E = \sum_{pq} V_p M_{pq} V_q, \tag{10}$$

where V_p are the elements of some arbitrary vector. We may obtain large computational reductions decomposing the matrix

$$Z_{pq} = V_p M_{pq} V_q = V_p \left(\sum_J L_p^J L_q^J + \Delta_{pq} \right) V_q, \tag{11}$$

depending on the screening induced by V_p. The main objective of this paper is to explore these possibilities. We denote Z as the characteristic matrix, as this characterizes the problem at hand. In order to control the accuracy needed in the minimization of E, we need to analyze the errors in the gradient. We write the functional in Eq. (10) as

$$E = \sum_{pq} V_p \left(\sum_J L_p^J L_q^J + \Delta_{pq} \right) V_q, \tag{12}$$

and we have determined the auxiliary basis such that

$$V_p \Delta_{pp} V_p \leq T, \tag{13}$$

where T is the decomposition threshold. The gradient is now written in the following form:

$$G_p = 2 \sum_q M_{pq} V_q = 2 \sum_J L_p^J \sum_q L_q^J V_q + 2 \sum_q \Delta_{pq} V_q. \tag{14}$$

When the index p belongs to the auxiliary basis, the error terms will be zero as $\Delta_{Jp}=0$. The remaining terms will be bound by the inequality

$$|\Delta_{pq} V_q| = \frac{|V_p \Delta_{pq} V_q|}{|V_p|} \leq \frac{|V_p \Delta_{pp} V_p|^{1/2} |V_q \Delta_{qq} V_q|^{1/2}}{|V_p|} = \Delta_{pp}^{1/2} |V_q \Delta_{qq} V_q|^{1/2} \leq \Delta_{pp}^{1/2} T^{1/2}, \tag{15}$$

and we observe that the decomposition threshold is the control parameter, and the upper limit to the error in the gradient can be calculated from the diagonal elements of the error matrix Δ that is available after the completion of the Cholesky decomposition. Another way to view the calculation of the functional and gradient is to consider the following positive semidefinite matrix of double the original dimension:

$$\Omega = \begin{pmatrix} V_p M_{pq} V_q & V_p M_{ps} \\ M_{rq} V_q & M_{rs} \end{pmatrix} = \sum_J \begin{pmatrix} K_p^J \\ L_r^J \end{pmatrix} \begin{pmatrix} K_q^J \\ L_s^J \end{pmatrix}^T. \tag{16}$$

The first Cholesky vector is obtained using Eq. (1),

$$\begin{pmatrix} K_p^J \\ L_r^J \end{pmatrix} \begin{pmatrix} K_q^J \\ L_s^J \end{pmatrix}^T = \begin{pmatrix} V_p \frac{M_{pJ} M_{Jq}}{M_{JJ}} V_q & V_p \frac{M_{pJ} M_{Js}}{M_{JJ}} \\ \frac{M_{rJ} M_{Jq}}{M_{JJ}} V_q & \frac{M_{rJ} M_{Js}}{M_{JJ}} \end{pmatrix}, \tag{17}$$

and we observe that the expression for the gradient in the off-diagonal block in Eq. (17) is identical to the terms in Eq. (14), except for a factor of 2. As the gradient is obtained from the decomposition of a positive semidefinite matrix, we have shown that the gradient approximation given by Eq. (14) is robust, as defined by Dunlap *et al.*[30] The advantage with the double dimension expression is that we see how to control the decomposition when decomposing the Z matrix in Eq. (11). We may simply keep track of the two diagonals simultaneously, and we can choose the threshold in the two blocks separately if only the off-diagonal block is to be computed.

In passing, we note that during a typical optimization we update the V_p vector and need to calculate a new gradient using $(V_p + \tilde{V}_p)$,

$$\tilde{G}_p = 2 \sum_q M_{pq} V_q + 2 \sum_q M_{pq} \tilde{V}_q. \tag{18}$$

If the results from the computation of the first term is available, then the second term can be calculated decomposing the matrix $\tilde{V}_p M_{pq} \tilde{V}_q$. This may require a smaller effort, especially if the update is small. Needless to say, the errors are governed by the inequality in Eq. (15) and may easily be calculated.

III. ORBITAL LOCALIZATION

Before we begin the discussion of the Coulomb and exchange contributions, a few words on localized orbitals are needed. We denote the one-electron density in the atomic orbital (AO) basis $D_{\alpha\beta}$ and Cholesky decompose[31] it,

$$D_{\alpha\beta} = \sum_k C_\alpha^k C_\beta^k = \sum_k D_{\alpha\beta}^k, \tag{19}$$

where Greek letters label AOs. The individual orbital contributions to the density are denoted as $D_{\alpha\beta}^k = C_\alpha^k C_\beta^k$. These so-called Cholesky orbitals $\{C_\alpha^k\}$ have recently been shown to be orthogonal,[31] i.e.,

$$\sum_{\alpha\beta} C_\alpha^k \langle \varphi_\alpha | \varphi_\beta \rangle C_\beta^l = \delta_{kl}, \tag{20}$$

where $\langle \varphi_\alpha | \varphi_\beta \rangle$ is the AO overlap matrix and $\{\varphi_\alpha(\vec{r})\}$ are the AOs. Furthermore, the Cholesky orbitals are localized as the Cholesky decomposition preserves the sparsity of the AO density matrix. We shall employ these orbitals throughout this paper. In passing we note that we recently developed a modified orbital localization procedure for subsystems aiming at calculating size-intensive molecular properties.[32] The subsystem localization is obtained by restricting the decomposed diagonals to the AOs that are centered in the subsystem. This localization procedure could offer some advantages compared to the straightforward Cholesky decomposition of the density used here, mainly due to the more chemical distribution of the orbitals on atomic centers. We shall investigate this alternative localization procedure elsewhere. Other localization procedures can be used, but the advantage of the Cholesky localization is the ease of evaluation.

IV. THE COULOMB ENERGY

When evaluating the performance of algorithms, we shall investigate two limiting cases. The first is the complete

basis set limit where we keep the number of atoms fixed and increase the cardinal number[5] of the basis set. The other is the limit of a large system where we keep the basis set fixed and increase the number of atoms. The conventional direct SCF method scales around N^4 in the limit of a complete basis. However, depending on the density screening of higher angular momentum basis functions and the integral evaluation algorithm, the scaling can be much higher. In the limit of a large system, the scaling of the direct SCF is N^2 due to the Coulomb contribution, as the exchange contribution scales linearly when the density scales linearly with system size. The scaling of the standard Cholesky decomposition is N^3 in the limit of both a large system and a complete basis, not including the integral evaluation time.

Our starting point is the definitions of the Coulomb energy and Fock matrix expressed in terms of the AO density,

$$E_C = \sum_{\alpha\beta\gamma\delta} D_{\alpha\beta}(\alpha\beta|\gamma\delta)D_{\gamma\delta}, \tag{21}$$

$$F^C_{\alpha\beta} = 2\sum_{\gamma\delta} (\alpha\beta|\gamma\delta)D_{\gamma\delta}. \tag{22}$$

From the energy, we obtain the characteristic matrix

$$M^C_{\alpha\beta,\gamma\delta} = D_{\alpha\beta}(\alpha\beta|\gamma\delta)D_{\gamma\delta} = \sum_{kl} D^k_{\alpha\beta}(\alpha\beta|\gamma\delta)D^l_{\gamma\delta}, \tag{23}$$

which is written as the sum of individual orbital charge density interactions. The Cholesky decomposition of $M^C_{\alpha\beta,\gamma\delta}$ provides the optimal auxiliary basis that can fit the real space density,

$$\rho(\vec{r}) = \sum_{\alpha\beta} \varphi_\alpha(\vec{r})^*\varphi_\beta(\vec{r})D_{\alpha\beta}, \tag{24}$$

which ensures a prespecified accuracy T in the Coulomb energy E_C. Thus we have that

$$M^C_{\alpha\beta,\gamma\delta} = \sum_J L^J_{\alpha\beta}L^J_{\gamma\delta} = \sum_{IJ} D_{\alpha\beta}(\alpha\beta|I)S^{-1}_{IJ}(J|\gamma\delta)D_{\gamma\delta}, \tag{25}$$

where I and J label the product functions in the auxiliary basis and $S_{IJ}=(I|J)$ is the corresponding two-electron integral. The approximate Coulomb energy can be evaluated using the Cholesky vectors $\{L^J_{\alpha\beta}\}$ directly,

$$E_C = \sum_J \left(\sum_{\alpha\beta} L^J_{\alpha\beta}\right)^2, \tag{26}$$

and the Fock matrix is calculated from either of the expressions

$$\begin{aligned} F^C_{\alpha\beta} &= 2\sum_{IJ} (\alpha\beta|I)S^{-1}_{IJ}\sum_{\gamma\delta} (J|\gamma\delta)D_{\gamma\delta} \\ &= 2\sum_{IJ} (\alpha\beta|I)K^{-1}_{JI}\sum_{\gamma\delta} L^J_{\gamma\delta}, \end{aligned} \tag{27}$$

where we have used Eqs. (6) and (7) to express the Fock matrix in terms of the previously calculated Cholesky vectors. The errors in the Fock matrix are determined by the inequality in Eq. (15), and this approach is similar to the RI method using an auxiliary basis that has been determined using the standard Cholesky decomposition. As we shall see in Sec. VI, the size of the auxiliary basis is dramatically reduced as compared to the standard Cholesky decomposition of the two-electron integral matrix. The computational scaling of evaluating the Fock matrix in Eq. (27) is N^2 in the limit of a complete basis. This can be made linear in N if we only evaluate the Fock matrix with one occupied index. The scaling of the decomposition in Eq. (25) will become independent of the size of the basis due to the exponential convergence of SCF and DFT models with respect to the cardinal number of the basis[33] (screening of the high angular momentum functions). Consequently the number of Cholesky vectors will become constant in the limit of a complete basis. In the limit of a large system, we expect the decomposition in Eq. (25) to scale quadratically in the integral calculation part and cubically in the decomposition as the number of Cholesky vectors scale linearly. In Sec. VI we demonstrate the rapid convergence of the decomposition. We denote this approach as the *Coulomb decomposition*.

As pointed out by many authors, the applicability of the RI method is eventually limited by the calculation of the inverse matrix in Eq. (25) as this will scale cubically with the size of the molecular system. This can be avoided if we consider a different specific Cholesky decomposition as we realize that the complete Fock matrix is not needed in the optimization of the energy. To calculate the gradient, we only need the Fock matrix with one occupied index and one general AO index,

$$F^C_{\alpha k} = \sum_{\beta\gamma\delta} C^k_\beta(\alpha\beta|\gamma\delta)D_{\gamma\delta}. \tag{28}$$

This will clearly improve the screening also in the case when the basis set contains higher angular momentum functions. The matrix that enters this Fock matrix is an off-diagonal block of the positive semidefinite matrix,

$$\Omega^k = \begin{pmatrix} C^k_\beta(\alpha\beta|\gamma\delta)C^k_\delta & C^k_\beta(\alpha\beta|\kappa\lambda)D_{\kappa\lambda} \\ D_{\mu\nu}(\mu\nu|\gamma\delta)C^k_\delta & D_{\mu\nu}(\mu\nu|\kappa\lambda)D_{\kappa\lambda} \end{pmatrix}, \tag{29}$$

and thus the elements of the off-diagonal block are limited by the inequality

$$|D_{\alpha\beta}(\alpha\beta|\gamma\delta)C^k_\delta| \le (D_{\alpha\beta}(\alpha\beta|\alpha\beta)D_{\alpha\beta})^{1/2}(C^k_\delta(\gamma\delta|\gamma\delta)C^k_\delta)^{1/2}. \tag{30}$$

This suggests a Cholesky decomposition that is controlled by the largest diagonal elements in the two sub-blocks. If we denote the indices of the largest diagonal elements in the upper and lower blocks as $(\alpha\beta)$ and $(\kappa\lambda)$, respectively, then we choose the auxiliary function that gives the largest product,

$$(D_{\gamma\delta}(\gamma\delta|\gamma\delta)D_{\gamma\delta})(C^k_\delta(\gamma\delta|\gamma\delta)C^k_\delta), \quad \gamma\delta = \alpha\beta \vee \kappa\lambda, \tag{31}$$

as this choice will favor minimizing the off-diagonal elements entering the characteristic matrix in Eq. (29). However, there might be situations where different strategies would be preferred, for instance, when the sizes of the diagonal elements in the upper and lower submatrices differ significantly. In that case, the sum of the diagonal elements will be more appropriate than the product in Eq. (31).

The Fock matrix can be obtained directly from the Cholesky vectors according to the following expressions:

$$\Omega^k = \begin{pmatrix} C^k_\beta(\alpha\beta|\gamma\delta)C^k_\delta & C^k_\beta(\alpha\beta|\kappa\lambda)D_{\kappa\lambda} \\ D_{\mu\nu}(\mu\nu|\gamma\delta)C^k_\delta & D_{\mu\nu}(\mu\nu|\kappa\lambda)D_{\kappa\lambda} \end{pmatrix} = \sum_J \begin{pmatrix} L^J_{\alpha\beta} \\ K^J_{\mu\nu} \end{pmatrix} \begin{pmatrix} L^J_{\gamma\delta} \\ K^J_{\kappa\lambda} \end{pmatrix}^T, \tag{32}$$

$$F^C_{\alpha k} = \sum_{\beta\gamma\delta} C^k_\beta(\alpha\beta|\gamma\delta)D_{\gamma\delta} = \sum_J \left(\sum_\beta L^J_{\alpha\beta}\right)\left(\sum_{\gamma\delta} K^J_{\gamma\delta}\right). \tag{33}$$

We observe the computationally simple expression in terms of sums of the nonzero elements of the Cholesky vectors. We should emphasize that the Cholesky decomposition of the double dimension matrix is not needed, but only the screening protocol is based on the double dimension matrix when decomposing the two-electron integral matrix. The computational scaling of this approach is quadratic with respect to the size of the molecular system, and we have in this way bypassed the cubic scaling of the inverse matrix construction. This is easily seen as the localized orbital C^k_β will restrict the AO index, making the upper diagonal block in Eq. (29) a local quantity in the limit of a large system. Furthermore, assuming that the density scales linearly with the size of the system, we obtain an overall quadratic scaling. The method can be compared to what some authors denote local density fitting. The most important difference is that we have strict error control with the same quadratic computational scaling as the local density fitting methods described by Polly *et al.*[25] and Sodt *et al.*[34] Moreover, diffuse basis sets can be used without any computational penalties, which is not the case for local density fitting procedures. Another clear advantage is that the auxiliary basis is determined in an optimal way for a given density and orbital localization. Thus the basis will be significantly smaller than the preoptimized auxiliary basis set employed in RI methods even compared to the preoptimized Coulomb specific auxiliary basis sets. We denote this approach the Coulomb-*k* decomposition, but we have not implemented this yet since in most applications the Coulomb decomposition is not dominating.

For delocalized systems, such as metals, the scaling behavior of the density matrix is not linear with the size of the system. This will clearly have implications on the scaling properties of the above mentioned methods. When the delocalization can be associated with a small number of orbitals, we may split the density into two separate contributions, one for the insulator part and one for the delocalized part, such that

$$D_{\alpha\beta} = D^I_{\alpha\beta} + D^D_{\alpha\beta}. \tag{34}$$

We now assume that the rank of the insulator part is much higher than the delocalized part and that the insulator density scales linearly and the delocalized density scales quadratically with the size of the system. To illustrate, without too much details, consider the evaluation of the Coulomb energy for the density in Eq. (34). The method specific Cholesky decomposition may take advantage of this splitting by considering the matrix

$$\Omega = \begin{pmatrix} \Omega^{II} & \Omega^{ID} \\ \Omega^{DI} & \Omega^{DD} \end{pmatrix} = \begin{pmatrix} D^I_{\alpha\beta}(\alpha\beta|\gamma\delta)D^I_{\gamma\delta} & D^I_{\alpha\beta}(\alpha\beta|\kappa\lambda)D^D_{\kappa\lambda} \\ D^D_{\mu\nu}(\mu\nu|\gamma\delta)D^I_{\gamma\delta} & D^D_{\mu\nu}(\mu\nu|\kappa\lambda)D^D_{\kappa\lambda} \end{pmatrix}. \tag{35}$$

When the rank of Ω^{DD} is much smaller than the rank of Ω^{II}, we may evaluate the off-diagonal contributions to the energy, only decomposing the diagonal elements in the lower diagonal block Ω^{DD}. The Ω^{II} contribution to the energy can be obtained considering a similar characteristic matrix as in Eq. (29),

$$\Omega^k = \begin{pmatrix} D^k_{\alpha\beta}(\alpha\beta|\gamma\delta)D^k_{\gamma\delta} & D^k_{\alpha\beta}(\alpha\beta|\kappa\lambda)D^I_{\kappa\lambda} \\ D^I_{\mu\nu}(\mu\nu|\gamma\delta)D^k_{\gamma\delta} & D^I_{\mu\nu}(\mu\nu|\kappa\lambda)D^I_{\kappa\lambda} \end{pmatrix}, \tag{36}$$

where the orbitals in $D^k_{\alpha\beta}$ are restricted to the localized insulator part and thus a quadratic procedure is obtained. This implies that the delocalized part of the density will only affect the computational prefactor and a quadratic scaling will be maintained.

Although the above discussed algorithms are not linear scaling procedures, the computational prefactors have been significantly improved, as we shall see in Sec. VI. We stipulate that a combination of the continuous fast multipole method[35] (CFMM) for the long range interactions and the Coulomb decomposition for the short range interactions could lead to an efficient linear scaling procedure. Complications in CFMM arising from diffuse functions will require a separate procedure.[36] In the next section, we discuss the exchange contributions and show that linear scaling is indeed possible using the method specific Cholesky decomposition.

V. THE EXCHANGE ENERGY

The exchange terms are notoriously complicated in connection with the RI and Cholesky decomposition methods as the direct usage of the Cholesky vectors determined from the standard Cholesky decomposition of the two-electron integral matrix,[12]

$$(\alpha\beta|\gamma\delta) = \sum_J L^J_{\alpha\beta}L^J_{\gamma\delta}, \tag{37}$$

implies an N^4 scaling when evaluating the exchange contribution to the Fock matrix in the limit of a complete orbital basis. This should be compared to the direct SCF that also displays an N^4 scaling in the contraction of the density with the integrals, and thus no apparent saving is obtained. In fact, the largest saving obtained using the Cholesky decomposition is in the evaluation of the two-electron integrals. This can be quite significant as the scaling of the integral evaluation increases strongly with the cardinal number, and the decomposition only needs to be done once. However, the usefulness of this method is limited as many quantum chemical applications are performed using medium sized basis sets, and then the advantages compared to direct SCF are much smaller. Recently, Aquilante *et al.*[37] developed a quadratic scaling procedure for contracting density and Cholesky vectors using a screening protocol. The method, however, relies on the standard Cholesky decomposition in

Eq. (37) and thus displays an overall cubic scaling with the size of the system. As pointed out by the authors, the direct SCF evaluation of the Fock matrix is faster than the Cholesky decomposition for medium sized basis sets, and thus savings first start to appear when the number of Fock matrix evaluations is high. This is the case when correlated wave functions or molecular properties are to be evaluated.

The method specific Cholesky decomposition offers different ways of evaluating the exchange contributions to the Fock matrix. We shall here discuss three algorithms for evaluating the exchange part of the Fock matrix where two of them display a linear scaling behavior with respect to the size of the system. We start our discussion with the large basis set case.

The exchange energy and Fock matrix are defined as (with opposite signs)

$$E_X = \sum_{\alpha\beta\gamma\delta} D_{\alpha\gamma}(\alpha\beta|\gamma\delta)D_{\beta\delta}, \tag{38}$$

$$F^X_{\alpha\beta} = \sum_{\gamma\delta} (\alpha\gamma|\beta\delta)D_{\gamma\delta}, \tag{39}$$

and the energy expression leads to the following definition of the characteristic matrix:

$$M^X_{\alpha\beta,\gamma\delta} = D_{\alpha\gamma}(\alpha\beta|\gamma\delta)D_{\beta\delta} = \sum_{kl} C^k_\alpha C^l_\beta (\alpha\beta|\gamma\delta) C^k_\gamma C^l_\delta. \tag{40}$$

As the rank of the two-electron density matrix,

$$W_{\alpha\beta,\gamma\delta} = D_{\alpha\gamma}D_{\beta\delta} = \sum_{kl} (C^k_\alpha C^l_\beta)(C^k_\gamma C^l_\delta), \tag{41}$$

scales as the number of kl-pairs, a decomposition of the matrix in Eq. (40) will not lead to any reductions in rank compared to the standard Cholesky decomposition. We may, however, in analogy with the Coulomb decomposition, define an exchange density matrix $D^X_{\alpha\beta}$ as the sum of the transition densities that enter the expression above. We introduce the following notation:

$$D^{kl}_{\alpha\beta} = C^k_\alpha C^l_\beta, \tag{42}$$

$$D^X_{\alpha\beta} = \frac{2}{N_e}\sum_{kl} C^k_\alpha C^l_\beta = Y_\alpha Y_\beta, \tag{43}$$

$$Y_\alpha = \sqrt{\frac{2}{N_e}}\sum_k C^k_\alpha, \tag{44}$$

and note that $D^X_{\alpha\beta}$ has rank 1. The exchange density has been normalized with respect to the number of electrons N_e such that

$$\sum_{\alpha\beta} \langle\varphi_\alpha|\varphi_\beta\rangle D^X_{\alpha\beta} = 1, \tag{45}$$

where we have used the orthogonality of the Cholesky orbitals in Eq. (20). The construction of $D^X_{\alpha\beta}$ is motivated by considering the functional

$$Q^X = \sum_{\alpha\beta\gamma\delta}\sum_{klk'l'} D^{kl}_{\alpha\beta}(\alpha\beta|\gamma\delta)D^{k'l'}_{\gamma\delta}, \tag{46}$$

where the characteristic matrix is given as

$$\Omega = \begin{pmatrix} D^{kl}_{\alpha\beta}(\alpha\beta|\gamma\delta)D^{kl}_{\gamma\delta} & \cdots & D^{kl}_{\alpha\beta}(\alpha\beta|\kappa\lambda)D^{k'l'}_{\kappa\lambda} \\ \vdots & & \vdots \\ D^{k'l'}_{\mu\nu}(\mu\nu|\gamma\delta)D^{kl}_{\gamma\delta} & \cdots & D^{k'l'}_{\mu\nu}(\mu\nu|\kappa\lambda)D^{k'l'}_{\kappa\lambda} \end{pmatrix}. \tag{47}$$

The diagonal kl-blocks enter the expression for the exchange energy in Eq. (38). When determining the auxiliary basis for an exchange energy calculation, we may choose to decompose this characteristic matrix as the accurate representation of this matrix will ensure the accuracy in the diagonal blocks and thus the exchange energy. Therefore we recommend decomposing the matrix

$$M_{\alpha\beta,\gamma\delta} = D^X_{\alpha\beta}(\alpha\beta|\gamma\delta)D^X_{\gamma\delta} \tag{48}$$

to determine the optimal auxiliary basis to be used in the calculation of the exchange energy. The number of nonzero terms in the characteristic matrix will depend on the type of system we study. In compact systems, the matrix contains $(N_e/2)^4$ terms of the same size and the normalization in Eq. (43) is appropriate as $(N_e/2)^2$ terms enter the exchange energy. In the limit of a large system (and a linear scaling AO density) the exchange energy contains $N_e/2$ terms and the characteristic matrix contains $(N_e/2)^2$ terms. In this case a more appropriate normalization would be $\sqrt{N_e/2}$. We must emphasize that changing the normalization corresponds to changing the threshold of the decomposition. In practical applications, a calibration needs to be performed to determine the threshold that is adequate. After the auxiliary basis has been determined, we evaluate the Fock matrix using the inner projection expression. The scaling of the method is the same as for the standard Cholesky decomposition but with the number of auxiliary functions dramatically reduced. We shall, in brief, call this approach *exchange decomposition*, and the performance is discussed in Sec. VI. An alternative[36] to decompose the matrix in Eq. (48) is to use a screening protocol based on the characteristic matrix in Eq. (47).

Before we describe the other algorithms, we would like to point out an interesting property of the functional in Eq. (46). In the limit where the exchange contributions are zero, the remaining elements in the functional sum up to give the Coulomb energy. As the density in Eq. (43) has rank 1, this implies that the Coulomb energy can be evaluated as a single orbital self-energy,

$$E_C = \left(\frac{N_e}{2}\right)^2 (YY|YY), \tag{49}$$

where the orbital is defined in Eq. (44). This could have some computational implications for the evaluation of long range Coulomb interactions when the orbitals do not overlap and the exchange contributions are zero.

We denote the next algorithm as the *Exchange-k* decomposition and start the analysis from the exchange Fock matrix

$$F^X_{\alpha\gamma} = \sum_{\beta\delta k} C^k_\beta (\alpha\beta|\gamma\delta) C^k_\delta, \tag{50}$$

considering the characteristic matrix

$$M^k_{\alpha\beta,\gamma\delta} = C^k_\beta (\alpha\beta|\gamma\delta) C^k_\delta = \sum_J L^J_{\alpha\beta} L^J_{\gamma\delta}. \tag{51}$$

The decomposition of this matrix will provide a specific auxiliary basis for each localized orbital C^k_α, making the matrix local in the limit of a large system, and thus the auxiliary basis is independent of the system size. However, since only two orbital indices are screened, large basis sets will add to the computational prefactor while still maintaining the linear scaling. The individual contributions to the Fock matrix are calculated directly from the Cholesky vectors in the following form:

$$F^k_{\alpha\gamma} = \sum_J H^J_\alpha H^J_\gamma, \tag{52}$$

$$H^J_\alpha = \sum_\beta L^J_{\alpha\beta}. \tag{53}$$

Since the error in the SCF energy is typically quadratic in the norm of the gradient and we directly decompose the matrices entering the Fock matrix, we may choose a higher threshold. When the iterative process is completed, we calculate the energy using a lower threshold if needed. Furthermore, the expressions in Eqs. (51)–(53) suggest a simple parallel implementation where the contribution for each orbital is calculated on a separate compute node with memory requirements scaling linearly with the size of the system.

In most cases some of the localized orbitals have a significant overlap, and a common auxiliary basis is more efficient. In such cases we can consider characteristic matrices of the form

$$\Omega = \begin{pmatrix} C^k_\beta(\alpha\beta|\gamma\delta)C^k_\delta & \cdots & C^k_\beta(\alpha\beta|\kappa\lambda)C^l_\lambda \\ \vdots & & \vdots \\ C^l_\nu(\mu\nu|\gamma\delta)C^k_\delta & \cdots & C^l_\nu(\mu\nu|\kappa\lambda)C^l_\lambda \end{pmatrix} \tag{54}$$

and determine the decomposition sequence from the diagonal elements of this matrix. In practice it is sufficient to decompose the two-electron integral matrix, but the screening of the diagonals should be carried out based on the matrix above. We will not elaborate on this screening protocol here[36] but just mention that decomposing the sum of the diagonal block matrices will, in general, not lead to a favorable Cholesky decomposition. This is simply due to the fact that the decomposition of Hadamard product matrices gives an overall rank equal to the product of the ranks of the individual matrices.

As mentioned above, only two indices are screened in the Exchange-*k* decomposition, and we shall now consider an algorithm, denoted as the *Exchange-kl* decomposition, where all orbital indices are screened. Using the expression for the exchange energy in Eq. (38), we analyze the characteristic matrix

$$M^{kl}_{\alpha\beta,\gamma\delta} = D^{kl}_{\alpha\beta}(\alpha\beta|\gamma\delta)D^{kl}_{\gamma\delta} = \sum_J L^J_{\alpha\beta} L^J_{\gamma\delta}. \tag{55}$$

The decomposition of this matrix gives an auxiliary basis for each *kl*-pair individually. The exchange energy is obtained directly from the Cholesky vectors as in Eq. (26) for each *kl*-pair. The Fock matrix with one occupied index can be evaluated from the sum of contributions

$$F^X_{\alpha k} = \sum_l F^l_{\alpha k}, \tag{56}$$

$$F^l_{\alpha k} = \sum_{\beta\gamma\delta} C^l_\beta (\alpha\beta|\gamma\delta) D^{kl}_{\gamma\delta}. \tag{57}$$

These are calculated in terms of the auxiliary basis in the following form:

$$\begin{aligned} F^l_{\alpha k} &= \sum_{IJ\beta} C^l_\beta (\alpha\beta|I) S^{-1}_{IJ} \sum_{\gamma\delta} (J|\gamma\delta) D^{kl}_{\gamma\delta} \\ &= \sum_{IJ\beta} C^l_\beta (\alpha\beta|I) K^{-1}_{JI} \sum_{\gamma\delta} L^J_{\gamma\delta}, \end{aligned} \tag{58}$$

where I and J label the auxiliary basis determined in Eq. (55) and $S=KK^T$. The errors are controlled by the inequality in Eq. (15). Clearly the efficiency of this decomposition is determined by the locality of the orbitals. However, in the limit of a large basis, many *kl*-pairs will have a significant overlap, and in this situation it might be more advantageous to find a common auxiliary basis for some of these pairs. This can be achieved by considering a matrix similar to Eq. (54) to control the selection of the decomposition sequence.

In order to avoid the inverse matrix in Eq. (58) and to obtain a more direct control of the errors in the Fock matrix contributions $F^l_{\alpha k}$, we consider the matrix

$$\begin{aligned} \Omega^{kl} &= \begin{pmatrix} C^l_\beta(\alpha\beta|\gamma\delta)C^l_\delta & C^l_\beta(\alpha\beta|\kappa\lambda)D^{kl}_{\kappa\lambda} \\ D^{kl}_{\mu\nu}(\mu\nu|\gamma\delta)C^l_\delta & D^{kl}_{\mu\nu}(\mu\nu|\kappa\lambda)D^{kl}_{\kappa\lambda} \end{pmatrix} \\ &= \sum_J \begin{pmatrix} L^J_{\alpha\beta} \\ K^J_{\mu\nu} \end{pmatrix} \begin{pmatrix} L^J_{\gamma\delta} \\ K^J_{\kappa\lambda} \end{pmatrix}^T, \end{aligned} \tag{59}$$

where $F^l_{\alpha k}$ is the off-diagonal matrix and where we have suppressed the k and l indices in the Cholesky vectors. The Fock matrix contribution is then obtained from an expression similar to Eq. (33),

$$F^l_{\alpha k} = \sum_J \left(\sum_\beta L^J_{\alpha\beta} \right) \left(\sum_{\gamma\delta} K^J_{\gamma\delta} \right). \tag{60}$$

As we have shown in the above analysis of the exchange contribution, there are many different ways to introduce the method specific Cholesky decompositions. In Sec. VI we will gain insight into the applicability of the methods presented.

VI. RESULTS AND DISCUSSION

In order to show the behavior of the decompositions outlined in the previous sections, we present results for four different systems: an alpha helix glycine structure with up to 30 glycine units, a water molecule, a benzene molecule, and a helium crystal. The crystal has two layers with a quadratic

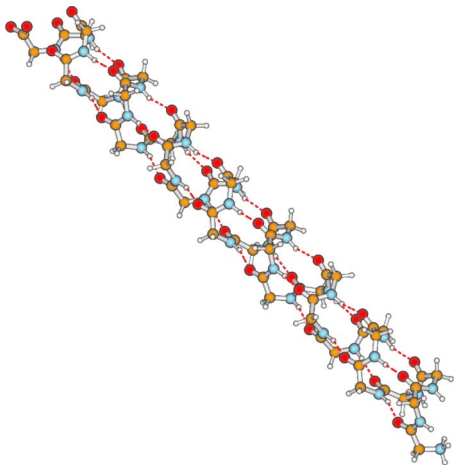

FIG. 1. (Color online) An alpha helix structure containing 30 glycine molecules where the number of atoms is 213 of which 92 are hydrogen atoms and the largest interatomic distance is 47 Å.

shape, and the nearest neighbor distance is 4.2 Å, as found experimentally by Schuch and Mills.[38] The glycine structure is shown in Fig. 1, with the geometry obtained from the MOLDEN program.[39] All calculations were performed with a local version of the DALTON program[13] on a Xeon 2.66 GHz quad-core processor. Timings are compared to the direct SCF as implemented in DALTON. Even though the direct SCF of DALTON is perhaps not the fastest available code, we still find it appropriate for comparison since this allows us to use the same integral evaluation code everywhere. The timings for the direct SCF refer to a complete Fock build (with a screening threshold of 10^{-7}), and the timings for all decompositions have been obtained using the density from a Hückel initial guess. Throughout we have used the correlation consistent basis sets cc-pVXZ augmented with diffuse functions where $X=D,T,Q,5,6,7$. Including diffuse functions is important for the evaluation of many molecular properties, and they are also notoriously challenging for reduced scaling methods and are thus suitable for showing the strengths and possibilities of Cholesky decomposition based methods compared to the alternatives. The well behaved convergence toward the basis set limit with respect to the cardinal number X also makes the correlation consistent basis sets excellent to investigate the scaling behavior of our methods. Point group symmetry has not been used in any of the reported calculations.

We start analyzing the results for the Coulomb decomposition. As we discussed in Sec. IV, the number of Cholesky vectors will be constant in the limit of a complete basis set. In Table I, we show the number of auxiliary functions obtained for water and benzene using different basis sets. We have used two different densities, a converged SCF density and a Hückel density. For the converged density, we observe a practically constant number of Cholesky vectors, as anticipated. This is not the case for the Hückel density that shows a small linear increase with respect to the cardinal number, and this suggests that the initial density should be obtained from a converged density using a smaller basis set. Comparing the number of auxiliary functions required in the Coulomb decomposition and the standard Cholesky decomposition, the differences are impressive. For aug-cc-pVQZ, the differences represent a factor of 20, and for aug-cc-pV7Z, the factor is 30–60 depending on the density used. This means that for the Hückel density, the integral time will be 30 times shorter in the Coulomb decomposition, but in fact it is 75 times faster due to favorable screening. The decomposition time is reduced by a much larger factor due to the cubic scaling and is for these cases negligible. Compared to RI auxiliary basis sets that are normally around three times the number of basis functions, the reduction is about a factor of 7 for aug-cc-pVQZ. This is a very significant difference for such a small system and a high price to pay when error control is also lost. For benzene the number of auxiliary functions increases almost linearly with the number of electrons compared to water. For the aug-cc-pVQZ basis set, the standard Cholesky decomposition and RI methods require nine and seven times as many auxiliary functions as needed for the Coulomb decomposition. We now turn our attention to the fairly sparse helium crystal using an aug-cc-pVQZ basis that is essential in describing the dispersion interaction

TABLE I. Number of Cholesky vectors for water and benzene for Coulomb, Exchange(3), and standard Cholesky decomposition for different basis sets, where N denotes the number of basis functions. The normalization [see Eq. (43)] used in the Exchange(3) decomposition is $2/N_e$. The number of vectors is presented both using the converged SCF density and the density from the Hückel guess. The decomposition threshold is 10^{-8}.

		Coulomb		Exchange(3)		
	N	**Converged**	**Hückel**	**Converged**	**Hückel**	SCD
			Water			
aug-cc-pVDZ	41	66	45	94	73	414
aug-cc-pVTZ	92	67	47	130	77	984
aug-cc-pVQZ	172	77	62	134	108	1750
aug-cc-pV5Z	287	71	89	135	201	2839
aug-cc-pV6Z	443	70	110	144	342	4268
aug-cc-pV7Z	643	89	197	259	849	5779
			Benzene			
aug-cc-pVDZ	192	324	242	326	264	2000
aug-cc-pVTZ	414	305	220	442	394	4108
aug-cc-pVQZ	756	299	283	577	454	6867

TABLE II. Timings in seconds for a two-layered helium fcc crystal with the nearest neighbor distance of 4.2 Å using an aug-cc-pVQZ basis set. Timings for integral evaluation and decomposition are given separately with the decomposition times in parentheses. The number of atoms in the crystal is denoted as # helium, and N is the number of basis functions. The decomposition threshold is 10^{-8}. There is no normalization [see Eq. (43)] in the Exchange(1) decomposition.

# helium	N	Coulomb	Exchange(1)	Exchange-kl	SCD	Direct SCF
9	414	2 (1)	8 (2)	3 (7)	289	138
16	736	3 (2)	21 (4)	5 (14)	1218	933
25	1150	5 (2)	34 (5)	10 (23)	4192	4205
36	1656	9 (3)	51 (7)	15 (37)	22 782	14 789
49	2245	15 (5)	72 (8)	21 (53)	61 514	44 010
64	2944	20 (5)	114 (12)	32 (78)	146 244	116 430
121	5566	63 (10)	257 (19)	87 (224)	···	1 153 040 [a]

[a]Estimated using scaling between 49 and 64 helium atoms.

in this system. In Table II we report the number of Cholesky vectors needed to obtain an accuracy of 10^{-8}. The small number of auxiliary functions needed cannot be matched by any other method, and we observe the complete failure of the standard Cholesky decomposition in this case. The number of auxiliary functions is only 10% of the size of the basis set. This clearly demonstrates the power of the Coulomb decomposition. For the alpha helix glycine chain, which is an electron rich and compact system, we observe a relative increase in the number of auxiliary functions compared to the other systems. From Table III we observe that the number of Cholesky vectors is twice the number of basis functions for an accuracy of 10^{-8}. In production applications, most quantum chemists will be satisfied with less accuracy. Incidentally, the Coulomb decomposition for five glycine molecules with a threshold of 10^{-4} gives 223 auxiliary functions compared to 1132 for 10^{-8}. This should be compared to the 1908 auxiliary functions typical of the RI method. Detailed timings for the Coulomb decomposition are reported in Tables III–V. Interestingly, the helium crystal shows sublinear scaling both in integral and decomposition times; this could be attributed to the crystal symmetry. For the glycine alpha helix, we observe a dramatic increase in decomposition time due to the cubical scaling, and for 25 glycine molecules it dominates over the integral calculation. This problem can be resolved using the Coulomb-k decomposition that only scales quadratically. We will be reporting on this in a forthcoming paper. Needless to say, compared to the timings of one Fock matrix construction in the direct SCF, we observe a many orders of magnitude difference in favor of the Coulomb decomposition.

The evaluation of the exchange contributions will, in general, require a larger auxiliary basis than Coulomb contributions. The Exchange decomposition has been designed for the case where the occupied orbitals share many AOs as in compact systems using large basis sets. For medium size basis sets and systems such as benzene, the number of auxiliary functions is about the same as the number of basis functions, using a $2/N_e$ normalization in Eq. (43) of the Q^X functional. In the case of the helium crystal, we report numbers without any normalization as the system is very sparse. The exchange basis is twice the size of the Coulomb basis and could actually be used to calculate both terms. However, as seen from Table II, Coulomb is faster than Exchange(1), and to use a common basis will lead to a longer execution time. For the glycine chain, we report in Tables III and V two

TABLE III. Number of Cholesky vectors (auxiliary functions) for different decompositions for an alpha helix glycine chain using an aug-cc-pVDZ basis set. For Coulomb and Exchange(1–2) decompositions, the threshold is 10^{-8}, and for Exchange-k the threshold is 10^{-4}. The normalization values [see Eq. (43)] used in the Exchange(1) and Exchange(2) decompositions are 1 and $\sqrt{2/N_e}$, respectively. The number of basis functions is denoted as N.

# glycine	N	Coulomb	Exchange(1)	Exchange(2)	Exchange-k	SCD[a]
1	160	258	736	534	1524	1792
2	279	469	1213	771	3034	3071
3	398	688	1850	1145	4562	4345
5	636	1132	3033	1752	8054	6899
10	1231	2264	6262	3361	18 730	13 195
14	1707	3177	8606	4311	27 589	···
20	2421	4545	12 601	5994	40 804	···
25	3016	5665	15 954	7369	51 687	···
30	3611	6815	···	8346	62 690	···

[a]Timings for SCD in seconds: 56 (35), 182 (139), 453 (435), 1572 (3055), and 7750 (43487) for 1, 2, 3, 5, and 10 glycine molecules, respectively. For comparison, 30 glycine molecules using 3-21G (1273 basis functions) [Coulomb: 1197 (842), Exchange(1): 2632 (4258), Exchange-k: 2662 (1216), SCD: 3467 (13725), and direct SCF: 5779]. Timings for integral evaluation and decomposition are given separately with the decomposition times in parentheses.

TABLE IV. Timings in seconds for water and benzene. The times for integral evaluation and the decomposition are given separately with the decomposition times in parentheses. The normalization [see Eq. (43)] used in the Exchange(3) decomposition is $2/N_e$. The decomposition threshold is 10^{-8}, except for Exchange-*k* where the threshold is 10^{-4}. The numbers of basis functions are denoted by *N*.

	N	Coulomb	Exchange(3)	Exchange-*kl*	Exchange-*k*	SCD	Direct SCF
				Water			
aug-cc-pVDZ[a]	41	2 (2)	4 (2)	23 (15)	10 (6)	11	0.05
aug-cc-pVTZ[a]	92	3 (2)	6 (2)	34 (18)	15 (8)	25	0.68
aug-cc-pVQZ	172	8 (3)	14 (3)	71 (24)	32 (12)	80	10
aug-cc-pV5Z	287	16 (3)	48 (4)	164 (31)	77 (17)	400	124
aug-cc-pV6Z	443	34 (4)	260 (8)	404 (38)	220 (29)	2603	972
aug-cc-pV7Z	646	156 (9)	1885 (26)	1324 (61)	730 (67)	14 531	6784
				Benzene			
aug-cc-pVDZ	192	31 (7)	24 (6)	2915 (649)	250 (62)	130	25
aug-cc-pVTZ	414	64 (7)	58 (9)	5649 (673)	561 (95)	789	305
aug-cc-pVQZ	756	197 (13)	266 (11)	12 988 (931)	1645 (220)	8727	4009

[a]The direct SCF timings are faster as the decomposition code writes the Cholesky vectors to disk; for larger calculations this becomes negligible.

normalizations (that effectively correspond to a change in threshold): one with no normalization denoted as Exchange(1) and one with $\sqrt{2/N_e}$ denoted as Exchange(2). To determine the auxiliary basis using Exchange(1) will give a basis that will be too large, and Exchange(2) is more appropriate due to the compactness of the system. The size of the auxiliary basis is 1.5 times the size of the corresponding Coulomb basis and is still smaller than that for the RI method. However, in production calculations the threshold or normalization should be chosen to give the desired accuracy in the energy. Regarding the execution times for Exchange(2), we observe that for 30 glycine molecules the decomposition time dominates, and in this case Exchange-*k* might be a better algorithm. Another way to bypass the almost cubic scaling is to study matrices of the form

$$\Omega^k = \begin{pmatrix} C^k_\alpha(\alpha Y|\beta Y)C^k_\beta & C^l_\alpha(\alpha Y|\delta Y) \\ (\gamma Y|\beta Y)C^k_\beta & (\gamma Y|\delta Y) \end{pmatrix} \tag{61}$$

using an obvious notation from the previous sections. However, we shall not analyze this any further here.

The Exchange-*k* decomposition for the glycine chain is carried out with a threshold of 10^{-4} as we may evaluate the Fock matrix directly from the Cholesky vectors, and the error in the energy is quadratic in the accuracy of the Fock matrix. If the energy is calculated using this Fock matrix, then the accuracy will be the same as in the Fock matrix. The characteristic matrix for the Exchange-*k* decomposition is that of Eq. (51), where we screen two of the orbital indices. The Exchange-*k* scaling for glycine is almost linear, having an exponent of 1.35 when going from 25 to 30 glycine molecules and an exponent of 1.75 when going from 10 to 14 units. This means that the onset point for the linear scaling is reached at approximately 30 glycine units. As seen in Figs. 2 and 3, the integral time starts to scale linearly before the decomposition time. For those readers who might be interested in using very small basis sets, we report in the footnote of Table III timings for 30 glycine molecules using the 3-21G basis set. In this case the Coulomb decomposition is about three times faster than direct SCF, and Exchange-*k* is about the same. The standard Cholesky decomposition is not the method of choice for this system, as we have already explained.

The glycine chain is an electron rich and compact system, and the Exchange-*kl* decomposition will therefore not

TABLE V. Timings in seconds for an alpha helix glycine chain using aug-cc-pVDZ basis set. For Coulomb and Exchange(1–2) decompositions the threshold is 10^{-8}, and for Exchange-*k* the threshold is 10^{-4}. The normalization values [see Eq. (43)] used in the Exchange(1) and Exchange(2) decompositions are 1 and $\sqrt{2/N_e}$, respectively. The number of basis functions is denoted as *N*, and the timings for integral evaluation and decomposition are given separately with the decomposition times in parentheses.

# glycine	*N*	Coulomb	Exchange(1)	Exchange(2)	Exchange-*k*	Direct SCF
1	160	24 (7)	50 (16)	37 (11)	154 (44)	18
2	279	64 (13)	134 (34)	89 (20)	471 (98)	171
3	398	142 (24)	297 (82)	174 (39)	893 (170)	709
5	636	428 (69)	887 (326)	458 (105)	2803 (419)	4608
10	1231	1889 (442)	4018 (4626)	1799 (644)	10 480 (1643)	60 302
14	1707	3810 (1140)	7909 (13 890)	3227 (1138)	18 381 (3077)	210 279
20	2421	7991 (5104)	16 889 (46 061)	6319 (5714)	30 681 (5844)	797 406
25	3016	12 997 (15 758)	29 082 (151 187)	9609 (10 996)	41 618 (8758)	⋯
30	3611	18 834 (24 615)	⋯	13 343 (17 200)	52 059 (12 220)	⋯

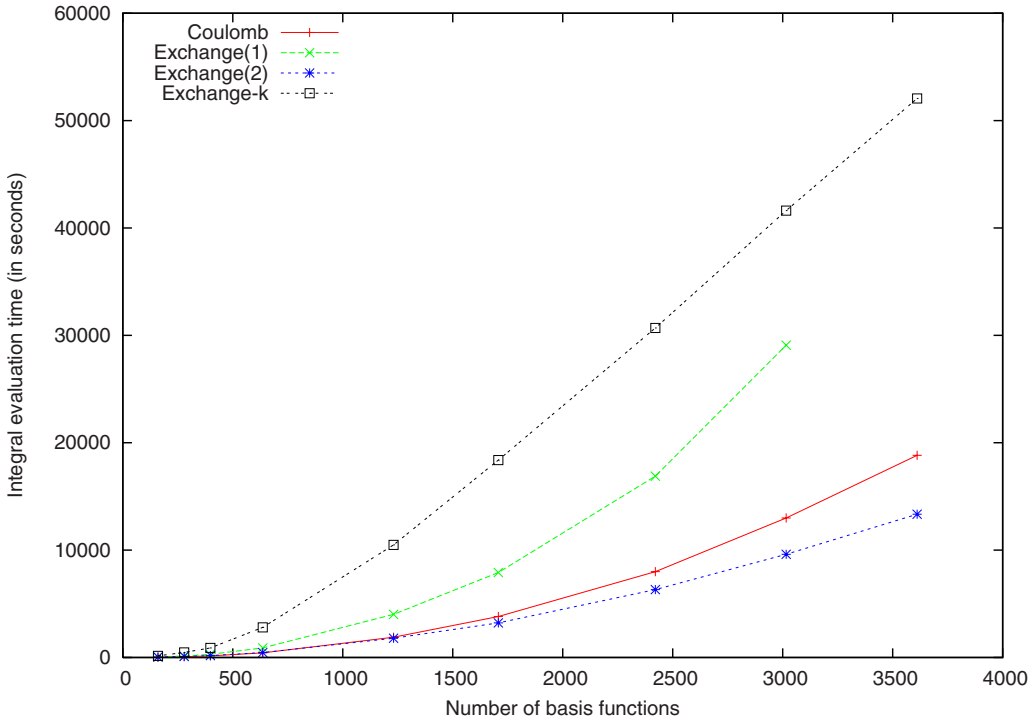

FIG. 2. (Color online) Integral evaluation times for the glycine alpha helix using different decomposition algorithms. In particular, we note the linear scaling of the Exchange-*k* algorithm.

be the method of choice for this system. The number of *kl*-pairs scales quadratically with the number of electrons, and the same number of characteristic matrices in Eq. (55) needs to be decomposed in the Exchange-*kl* decomposition. Prescreening of the diagonal elements reduces the number, but in electron rich systems, it may still be expensive due to overlapping auxiliary bases among the *kl*-pairs. The strength of the method lies in the very local nature of the characteristic matrix. In Table II this strength becomes clear. The helium crystal has both of the properties for making Exchange-*kl* the preferred choice: it has few electrons and is spatially extended. The total number of vectors for the helium crystal, as presented in Table VI and Fig. 4, is not representative since each vector is much less computation-

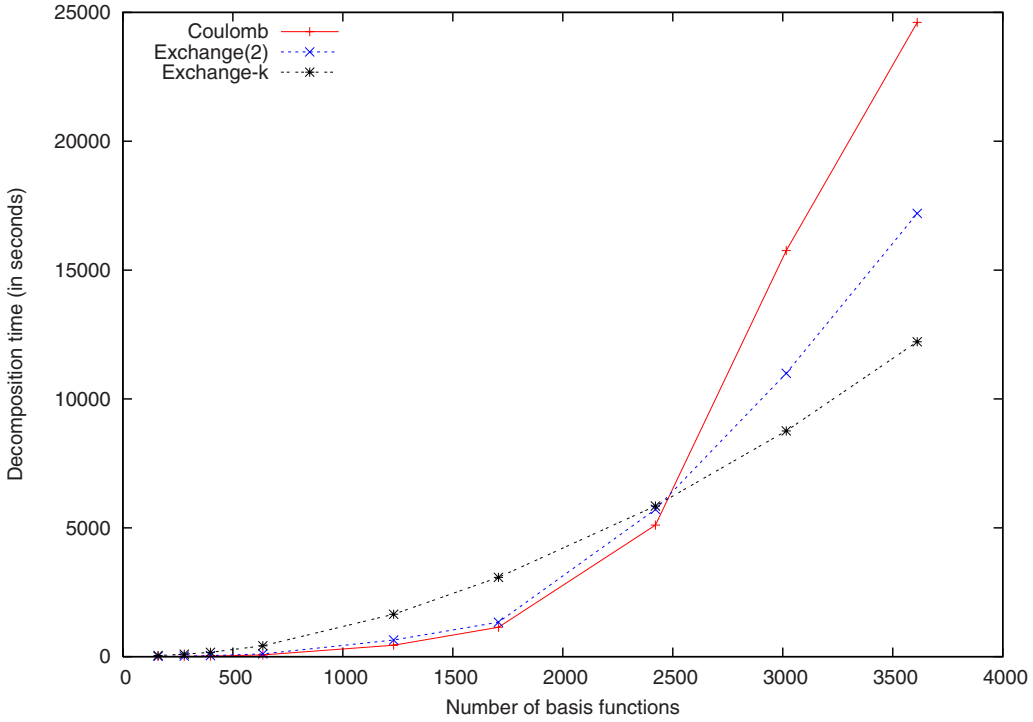

FIG. 3. (Color online) The decomposition time is displayed for the glycine alpha helix using different decomposition algorithms. In particular, we note the near linear scaling of the Exchange-*k* algorithm (see text) and the cubic scaling of the Coulomb decomposition.

TABLE VI. Number of Cholesky vectors (auxiliary functions) for different decompositions for a helium crystal in an aug-cc-pVQZ basis set. The decomposition threshold is 10^{-8}, and N is the number of basis functions. There is no normalization [see Eq. (43)] in the Exchange(1) decomposition.

# helium	N	Coulomb	Exchange(1)	Exchange-kl	SCD
9	414	27	83	76	4153
16	736	48	141	145	7483
25	1150	75	207	238	11 760
36	1656	108	288	361	17 040
49	2245	147	340	505	23 277
64	2944	192	477	673	30 432
121	5566	363	716	1321	⋯

ally demanding to handle than that for Exchange(1). However it still demonstrates the local nature of the kl-pairs since it is linear with the size of the system. The total cost for Coulomb and Exchange-kl decompositions is many orders of magnitude smaller than that for the direct SCF.

Finally, we will address the accuracy of the decomposition algorithms discussed previously. Besides the reductions in computational effort documented above, the advantage using the Cholesky decomposition to determine auxiliary bases is that we can control the accuracy using a single threshold. In Table VII we report energy optimizations of the water molecule using Coulomb and Exchange-k decompositions. For Coulomb we exclude the exchange term, and for Exchange-k we exclude the Coulomb term. The converged energies are compared to corresponding results obtained using the standard Cholesky decomposition. The total SCF energy is obtained combining the two approximations, and errors are reported compared with standard SCF energies. As expected, lowering the threshold an order of magnitude reduces the error in the Coulomb and the exchange contributions with an order of magnitude. Needless to say, the largest error between Coulomb and Exchange-k will determine the overall error in the total SCF energy. In Table VIII we address the errors in the Exchange decomposition in Eq. (48) using different normalization factors of the density in Eq. (43). We observe that increasing the size of the basis set improves the accuracy of the energy, and the normalization or the threshold can actually be increased to give the same accuracy in the energies. When the density is converged, we may increase the integral accuracy and calculate the energy more accurately. The errors reported demonstrate how exceedingly stable the Cholesky decomposition really is.

VII. CONCLUSIONS

We have presented a detailed discussion of the Coulomb and exchange contributions in the framework of the method specific Cholesky decomposition. The flexibility of the method is illustrated by the variety of algorithms discussed.

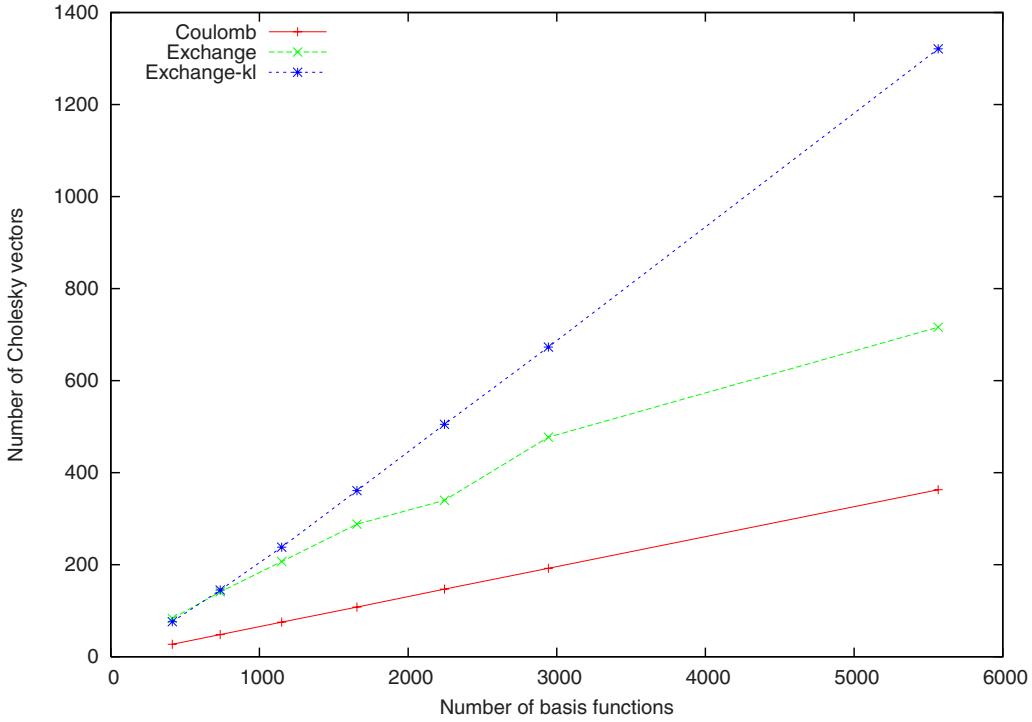

FIG. 4. (Color online) The number of Cholesky vectors for the helium crystal (see text). We note that the number of vectors is not directly related to the computational performance.

TABLE VII. Errors in energies in hartree (see text) are reported for water using different basis sets and decomposition thresholds. The number in parentheses indicates the order of magnitude of each error.

Basis set	Threshold	Coulomb	Exchange-k	Total
aug-cc-pVDZ	10^{-7}	1.43(−5)	−6.18(−6)	1.56(−5)
	10^{-8}	1.04(−6)	−2.06(−7)	1.17(−6)
	10^{-9}	7.86(−8)	−2.89(−8)	9.46(−8)
aug-cc-pVTZ	10^{-7}	1.42(−5)	4.05(−6)	1.36(−5)
	10^{-8}	1.92(−6)	−1.67(−7)	1.69(−6)
	10^{-9}	2.30(−7)	−6.00(−8)	8.45(−8)

For the Coulomb contributions, we have shown a very large reduction in the computational prefactor compared to the standard Cholesky decomposition. In the large basis set limit, we obtain auxiliary bases that are smaller than the total number of orbitals—eventually this becomes constant in the basis set limit. For large systems the Coulomb decomposition scales cubically with system size and the integral part quadratically, but the computational prefactor is dramatically reduced compared to the standard Cholesky decomposition, which has a similar scaling. We outlined a quadratic scaling algorithm that has a similar scaling as local density fitting procedures, but with full error control. For the exchange contributions, we have introduced an exchange density that can be used to determine auxiliary bases for large basis set cases. We show a clear reduction of the computational prefactor as compared to the standard Cholesky decomposition. We have discussed two linear scaling algorithms for large systems. The Exchange-kl is optimal for sparse systems, and the Exchange-k is optimal for intermediate size systems. The latter should not be used if high angular momentum functions are present in the basis set as they are not screened.

The actual algorithmic implementation has not been discussed in detail. However, the purpose of this paper is to describe the theoretical foundation, and the implementation of the screening algorithms used will be the subject of a forthcoming paper.[36] It should be emphasized that the implementations of the screening protocols are essential to obtain the reductions in computational effort reported.

As promised in Sec. I, we now address the question of applicability in electron correlation models. For the simple MP2 model, the answer is affirmative as the integrals with two occupied indices, $(\alpha k|\beta l)$, can be evaluated using the characteristic matrix given in Eq. (54). We also anticipate that coupled cluster models will benefit from an implementation of the method specific Cholesky decomposition. Using the notation of Koch *et al.*,[40] we consider the computational intensive B-term in the coupled cluster singles and doubles model,

$$\Omega_{ij}^{\alpha\beta} = \sum_{\gamma\delta} (\alpha\gamma|\beta\delta)(\lambda_{\gamma i}\lambda_{\delta j} + t_{ij}^{\gamma\delta}). \tag{62}$$

We now fix the ij indices and perform a sparse LU-factorization

$$(\lambda_{\gamma i}\lambda_{\delta j} + t_{ij}^{\gamma\delta}) = \sum_{\mu} L_{\gamma\mu}U_{\mu\delta}, \tag{63}$$

where we have suppressed the ij indices on the right-hand side. Thus the B-term can be evaluated in the form

$$\Omega_{ij}^{\alpha\beta} = \sum_{\gamma\delta\mu} L_{\gamma\mu}(\alpha\gamma|\beta\delta)U_{\mu\delta}, \tag{64}$$

which can be evaluated using the algorithms outlined in this paper. The efficiency of such an approach will be determined by the locality of the orbitals. For large basis sets, many of the ij-pairs will have overlapping auxiliary bases, and a simultaneous treatment will be necessary. Numerical experiments should be conducted to obtain more insight into the problem. In explicitly correlated methods, the main obstacle is the evaluation of complicated integrals,[41] and we may apply the method specific Cholesky decomposition also in this case. We can illustrate this considering two real commuting operators G and H, where GH need not be positive definite. We construct the positive semidefinite characteristic matrix using the state vector $|\Psi\rangle$,

$$\Omega = \begin{pmatrix} \langle\Psi|\Psi\rangle & \langle\Psi|G|\Psi\rangle & \langle\Psi|H|\Psi\rangle \\ \langle\Psi|G|\Psi\rangle & \langle\Psi|G^2|\Psi\rangle & \langle\Psi|GH|\Psi\rangle \\ \langle\Psi|H|\Psi\rangle & \langle\Psi|GH|\Psi\rangle & \langle\Psi|H^2|\Psi\rangle \end{pmatrix}. \tag{65}$$

We obtain from the decomposition of this matrix the following inequality:

$$\left(\langle\Psi|GH|\Psi\rangle - \frac{\langle\Psi|G|\Psi\rangle\langle\Psi|H|\Psi\rangle}{\langle\Psi|\Psi\rangle}\right)^2 \leq \left(\langle\Psi|G^2|\Psi\rangle - \frac{\langle\Psi|G|\Psi\rangle^2}{\langle\Psi|\Psi\rangle}\right)\left(\langle\Psi|H^2|\Psi\rangle - \frac{\langle\Psi|H|\Psi\rangle^2}{\langle\Psi|\Psi\rangle}\right), \tag{66}$$

and we observe a clear improvement of the straightforward Cauchy–Schwarz inequality,

TABLE VIII. Errors in energies in hartree (see text) are reported for water using different basis sets. We employ the Coulomb and exchange decomposition with a threshold of 10^{-8}. Results are reported for different normalizations. The numbers in square parentheses are the errors obtained using converged densities and the standard Cholesky decomposition to evaluate the energy. The number in parentheses indicates the order of magnitude of each error.

Basis set	No normalization	$\sqrt{2/N_e}$	$2/N_e$
aug-cc-pVDZ	3.13(−5)[−3.85(−8)]	9.44(−5)[−1.08(−6)]	1.28(−4)[−5.51(−6)]
aug-cc-pVTZ	3.40(−6)[−1.71(−8)]	1.12(−5)[1.76(−9)]	1.38(−4)[−5.61(−6)]
aug-cc-pVQZ	3.81(−6)[−4.41(−8)]	1.91(−5)[−1.62(−9)]	1.97(−6)[−1.91(−6)]

$$\langle\Psi|GH|\Psi\rangle^2 \leq \langle\Psi|G^2|\Psi\rangle\langle\Psi|H^2|\Psi\rangle. \quad (67)$$

If the integrals $\langle\Psi|GH|\Psi\rangle$ are to be calculated in the explicitly correlated model, we may define a positive semidefinite characteristic matrix from where we may obtain integral approximations or evaluate the expressions. No further Cholesky decompositions will be considered here.

ACKNOWLEDGMENTS

We would like to thank T. B. Pedersen for commenting on the manuscript. The present work is supported by the Norwegian Research Council through the Strategic University Program in Quantum Chemistry (Grant No. 154011/420), CREST, Japan Science and Technology Agency (JST), 4-1-8 Honcho Kawaguchi, Saitama 332-0012, Japan, and the Spanish FEDER+MEC Project No. CTQ2007-67143-C02-01/BQU. We are also grateful to the Center for Advanced Study at the Academy of Science and Letters in Oslo for financial support. We further acknowledge NOTUR for computational resources.

[1] E. Rudberg, E. H. Rubensson, and P. Sałek, J. Chem. Phys. **128**, 184106 (2008).
[2] O. Vahtras, J. E. Almlöf, and M. W. Feyereisen, Chem. Phys. Lett. **213**, 514 (1993).
[3] F. Weigend and M. Häser, Theor. Chim. Acta **97**, 331 (1997).
[4] F. Weigend, A. Köhn, and C. Hättig, J. Chem. Phys. **116**, 3175 (2002).
[5] T. Helgaker, W. Klopper, H. Koch, and J. Noga, J. Chem. Phys. **106**, 9639 (1997).
[6] A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, H. Koch, J. Olsen, and A. K. Wilson, Chem. Phys. Lett. **286**, 243 (1998).
[7] D. Yamaki, H. Koch, and S. Ten-no, J. Chem. Phys. **127**, 144104 (2007).
[8] C. E. Benoit, Bull. Geod. **2**, 67 (1924).
[9] N. H. F. Beebe and J. Linderberg, Int. J. Quantum Chem. **12**, 683 (1977).
[10] I. Røeggen and E. Wisløff-Nilssen, Chem. Phys. Lett. **132**, 154 (1986).
[11] D. W. O'Neal and J. Simons, Int. J. Quantum Chem. **36**, 673 (1989).
[12] H. Koch, A. Sánchez de Merás, and T. B. Pedersen, J. Chem. Phys. **118**, 9481 (2003).
[13] DALTON, Release 2.0, a molecular electronic structure program, 2005 (see http://www.kjemi.uio.no/software/dalton/dalton.html).
[14] T. B. Pedersen, A. Sánchez de Merás, and H. Koch, J. Chem. Phys. **120**, 8887 (2004).
[15] I. Garcìa Cuesta, T. B. Pedersen, H. Koch, and A. Sánchez de Merás, Chem. Phys. Lett. **390**, 170 (2004).
[16] T. B. Pedersen, H. Koch, L. Boman, and A. Sánchez de Merás, Chem. Phys. Lett. **393**, 319 (2004).
[17] T. D. Crawford, L. S. Owens, M. C. Tam, P. R. Schreiner, and H. Koch, J. Am. Chem. Soc. **127**, 1368 (2005).
[18] I. G. Cuesta, T. B. Pedersen, H. Koch, and A. Sánchez de Merás, ChemPhysChem **7**, 2503 (2006).
[19] I. Garcìa Cuesta, J. Sánchez Marín, T. B. Pedersen, H. Koch, and A. M. J. Sánchez de Merás, Phys. Chem. Chem. Phys. **10**, 361 (2008).
[20] MOLCAS 6, University of Lund, Sweden, 2005 (see http://www.teokem.lu.se/molcas/).
[21] F. Aquilante, T. B. Pedersen, R. Lindh, B. O. Roos, A. Sánchez de Merás, and H. Koch, J. Chem. Phys. **129**, 024113 (2008).
[22] I. Røeggen and T. Johansen, J. Chem. Phys. **128**, 194107 (2008).
[23] F. Aquilante, R. Lindh, and T. B. Pedersen, J. Chem. Phys. **129**, 034106 (2008).
[24] A. G. Taube and R. J. Bartlett, Collect. Czech. Chem. Commun. **70**, 837 (2005).
[25] R. Polly, H. J. Werner, F. R. Manby, and P. J. Knowles, Mol. Phys. **102**, 2311 (2004).
[26] P. O. Löwdin, J. Chem. Phys. **43**, S175 (1965).
[27] P. O. Löwdin, Int. J. Quantum Chem. **S4**, 231 (1971).
[28] J. L. Whitten, J. Chem. Phys. **58**, 4496 (1973).
[29] F. Aquilante, R. Lindh, and T. B. Pedersen, J. Chem. Phys. **127**, 114107 (2007).
[30] B. I. Dunlap, J. W. D. Connolly, and J. R. Sabin, J. Chem. Phys. **71**, 3396 (1979).
[31] F. Aquilante, T. B. Pedersen, A. Sánchez de Merás, and H. Koch, J. Chem. Phys. **125**, 174101 (2006).
[32] H. Koch, A. Sánchez de Merás, L. Boman, and I. G. Cuesta, J. Chem. Phys. (to be published).
[33] A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, and J. Olsen, Chem. Phys. Lett. **302**, 437 (1999).
[34] A. Sodt, J. E. Subotnik, and M. Head-Gordon, J. Chem. Phys. **125**, 194109 (2006).
[35] C. A. White, B. G. Johnson, P. M. W. Gill, and M. Head-Gordon, Chem. Phys. Lett. **253**, 268 (1996).
[36] L. Boman, H. Koch, and A. Sánchez de Merás (to be published).
[37] F. Aquilante, T. B. Pedersen, and R. Lindh, J. Chem. Phys. **126**, 194106 (2007).
[38] A. F. Schuch and R. L. Mills, Phys. Rev. Lett. **6**, 596 (1961).
[39] G. Schaftenaar and J. H. Noordik, J. Comput.-Aided Mol. Des. **14**, 123 (2000).
[40] H. Koch, A. Sánchez de Merás, T. Helgaker, and O. Christiansen, J. Chem. Phys. **104**, 4157 (1996).
[41] S. Ten-no, Chem. Phys. Lett. **398**, 56 (2004).