Chapter 13

Cholesky decomposition techniques in electronic structure theory

Francesco Aquilante, Linus Boman, Jonas Boström, Henrik Koch, Roland Lindh, Alfredo Sánchez de Merás and Thomas Bondo Pedersen

Abstract We review recently developed methods to efficiently utilize the Cholesky decomposition technique in electronic structure calculations. The review starts with a brief introduction to the basics of the Cholesky decomposition technique. Subsequently, examples of applications of the technique to *ab initio* procedures are presented. The technique is demonstrated to be a special type of a resolution-of-identity or density-fitting scheme. This is followed by explicit examples of the Cholesky techniques used in orbital localization, computation of the exchange contribution to the Fock matrix, in MP2, gradient calculations, and so-called method specific

Francesco Aquilante
Department of Physical Chemistry, Sciences II, University of Geneva, Quai E. Ansermet 30, 1211 Geneva 4, Switzerland e-mail: francesco.aquilante@gmail.com

Linus Boman
Department of Chemistry, Norwegian University of Science and Technology, N-7491 Trondheim, Norway e-mail: linus.boman@chem.ntnu.no

Jonas Boström
Department of Theoretical Chemistry, Chemical Center, University of Lund, P.O. Box 124 S-221 00 Lund, Sweden e-mail: Jonas.Bostrom@teokem.lu.se

Henrik Koch
Department of Chemistry, Norwegian University of Science and Technology, N-7491 Trondheim, Norway e-mail: henrik.koch@chem.ntnu.no

Roland Lindh
Department of Quantum Chemistry, Uppsala University, P.O. Box 518, SE-751 20, Uppsala, Sweden. e-mail: roland.lindh@kvac.uu.se

Alfredo Sánchez de Merás
Instituto de Ciencia Molecular, Universitat de València, P.O. Box 22085, ES-46071 Valencia, Spain e-mail: sanchez@uv.es

Thomas Bondo Pedersen
Centre for Theoretical and Computational Chemistry, Department of Chemistry, University of Oslo, P.O. Box 1033, Blindern, N-0315 Oslo, Norway e-mail: t.b.pedersen@kjemi.uio.no

Cholesky decomposition. Subsequently, examples of calibration of the method with respect to computed total energies, excitation energies, and auxiliary basis set pruning are presented. In particular, it is demonstrated that the Cholesky method is an unbiased method to derive auxiliary basis sets. Furthermore, details of the implementational considerations are put forward and examples from a parallel Cholesky decomposition scheme is presented. Finally, an outlook and perspectives are presented, followed by a summary and conclusions section. We are of the opinion that the Cholesky decomposition method is a technique that has been overlooked for too long. We have just recently started to understand how to efficiently incorporate the method in existing *ab initio* programs. The full potential of the Cholesky technique has not yet been fully explored.

13.1 Introduction

The application of electronic structure methods to large molecular systems using accurate basis sets is limited due to the steep computational scaling many of the methods encounter. The evaluation of two-electron integrals has turned out to be a major bottleneck in most of these methods. Although many different schemes have been devised for calculating two-electron integrals, for instance McMurchie-Davidson, Rys quadrature and Obara-Saika schemes, the algorithms are still demanding, especially for two-electron integrals that contain orbitals with high angular momentum. The altogether best approach to reduce the computational time used for calculating two-electron integrals is to reduce the number of integrals that need to be calculated. The first step in these developments is to avoid calculating integrals that are zero or small and give small contributions to the properties we are calculating. This has led to the development of the so-called linear scaling methods and they have proven highly efficient for small basis sets. Linear scaling formalisms have been developed by many groups. However, these algorithms currently only work efficiently for large systems in very small basis sets. When more accuracy is needed larger basis sets are essential and for these cases the resolution of identity (RI) method [1–8] has been used to reduce the computational effort. The RI method has worked very well for the calculation of the Coulomb contribution to the Fock matrix [9–18], but the exchange contribution is much more difficult to compute with the same benefits [15; 19–22]. 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 pre-optimized auxiliary basis sets. The argument found in the literature is that the error due to the incompleteness of the auxiliary basis is smaller than the basis set error speaks in favor of the RI method. However, care should be exercised as size-extensive properties cannot be used to extrapolate to the basis set limit if the errors are larger than the accuracy of the extrapolation procedures.

In this review we report on different approaches similar to the RI method that avoids the use of pre-optimized auxiliary basis sets. We simply determine the

auxiliary basis using the decomposition developed by Commandant André-Louis Cholesky (1875-1918) and published by Commandant Benoît in 1924 [23]. The idea to apply the Cholesky decomposition (CD) to the two-electron integral matrix was first suggested by Beebe and Linderberg [24] more than thirty years ago. The CD is the only numerical procedure known to the authors that can remove the zero or small eigenvalues of a positive semi-definite 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 will scale as N^4 , but the number of non-zero eigenvalues scales as N in the limit of a complete basis (N is the size of the basis). Despite this, the CD does not seem to have received much attention in the quantum chemistry community. There are notable exceptions, especially the developments by Røeggen and co-workers who have used the Cholesky decomposition of the two-electron integrals in the implementation of geminal models [25–30]. The use of the CD in connection with the calculation of derivative integrals has been discussed by O’Neal and Simons [31]. More recently, Koch, Sánchez de Merás and Pedersen [32; 33] have developed an implementation of the CD of the two-electron integrals aiming at large scale applications. The decomposition was shown to give large savings for large basis sets with a variety of theoretical methods: Hartree-Fock (HF), density functional theory (DFT), second order Møller-Plesset perturbation theory (MP2), and the second-order approximate coupled cluster model (CC2) [34]. These implementations in the DALTON program has formed the basis for many computational developments and applications [35–43]. The CD strategy has subsequently been implemented in the MOLCAS program [44] for multiconfigurational methods (multiconfigurational self-consistent field (MCSCF) [45] and multiconfigurational second-order perturbation theory (CASPT2) [46]) as well as scaled opposite spin MP2 [47] and coupled cluster (CC) methods [44]. The MOLCAS implementation has been crucial for a number of further developments and applications [21; 48–69]. Røeggen and Johansen [30] were the first to report a parallel implementation of the CD that shows a practically linear scaling with the number of compute nodes. The future use of CD based methods depends on the ability to evaluate geometrical derivatives of the molecular energy. In contrast to the numerical approach by O’Neal and Simons [31], analytic energy derivatives based on CD have been discussed recently by Aquilante et al. [52]. The CD is still perceived to be a rather complicated procedure. The aim of this chapter is to demonstrate that CD is a very rich tool and finds many practical applications in electronic structure theory.

The use of the RI representation of the two-electron integral matrix is sometimes referred to as density fitting (DF). Both these acronyms, RI and DF, are somewhat misleading. The density fitting community often anchors the method in the 1973 paper by Whitten [1] on integral approximations, whereas the RI terminology is mostly a descendent of the work by Almlöf and Feyereisen [7; 8]. The most correct description would probably be to denote them as inner projection methods. The concept of inner projections was introduced into quantum chemistry by Löwdin in his 1967-1971 landmark papers on perturbation theory [70; 71]. The CD is a powerful method to determine the optimal basis in the inner projection.

We start our review with some general mathematical considerations regarding the CD. This is then followed by section 13.3 discussing several of our recent applications of the method in electronic structure theory. Section 13.4 is devoted to accuracy calibration and implementation aspects are discussed in section 13.5. The chapter ends with our outlooks and conclusions.

13.2 Mathematical Background

In April 1924 a paper [23] by Commandant Benoît was published summarizing some handwritten notes dated in Vincennes on December 2nd 1910 by André-Louis Cholesky. Commandant Cholesky had developed a method to solve the systems of linear equations appearing in the compensation of geodesic networks, using a particular lower-upper (LU) decomposition [72] of a symmetric positive definite matrix $\mathbf{M}$

$$\mathbf{M}\mathbf{x} = \mathbf{L}\mathbf{L}^T\mathbf{x} = \mathbf{b} \quad (13.1)$$

where the matrix $\mathbf{M}$ has been decomposed in the product of a lower triangular matrix with strictly positive diagonals, $\mathbf{L}$, and its transpose. With such decomposition, the solution of the linear system is trivial by Gauss elimination in a two-step procedure: Solve first $\mathbf{L}\mathbf{y} = \mathbf{b}$ for $\mathbf{y}$, and then $\mathbf{L}^T\mathbf{x} = \mathbf{y}$ for $\mathbf{x}$.

Cholesky decomposition is easily understood as an iterative procedure where in each iteration a Cholesky vector L_p^J is being subtracted from the updated matrix,

$$\mathbf{M}^{(n+1)} = \mathbf{M}^{(n)} - \mathbf{L}^{J_n}(\mathbf{L}^{J_n})^T \quad (13.2)$$

which remains positive definite at all steps in the process. Note that $\mathbf{M}^{(0)} = \mathbf{M}$. Convergence is achieved when the residuals in the diagonal elements are smaller than a predefined threshold, τ . The original matrix can be approximated to almost arbitrary accuracy as

$$\mathbf{M} \approx \sum_{n=1}^K \mathbf{L}^{J_n}(\mathbf{L}^{J_n})^T \quad (13.3)$$

with Cholesky vectors defined according to

$$L_i^{J_n} = \frac{M_{iJ_n}^{(n)}}{\sqrt{M_{J_n J_n}^{(n)}}} \quad (13.4)$$

The Cholesky vectors constitute the columns of the $\mathbf{L}$ matrix and, therefore, in order to obtain the required lower triangular form, one needs to carry out the decomposition in such a way that the selected $M_{J_n J_n}^{(n)}$ is always the first non-zero diagonal element in the matrix. For the sake of numerical stability of the algorithm, it is often preferable to choose as pivoting element the maximum diagonal element in each step of the procedure

$$M_{J_n J_n}^{(n)} = \max_i \left(M_{ii}^{(n)} \right) \quad (13.5)$$

Consequently, while the CD of a positive definite matrix is unique, there can be different sets of Cholesky vectors depending on the ordering in which the diagonal elements are selected. At any rate, this does not cause any additional difficulty, as the different vector sets span the same subspace. Moreover, the super-matrix of two-electron integrals $M_{\alpha\beta, \gamma\delta} = (\alpha\beta | \gamma\delta)$ is positive semi-definite and the CD exists, but is in general not unique.

The reason why CD is a valuable tool in electronic structure calculations lies in the fact that it provides an efficient way of removing linear dependencies in a given matrix. As a matter of fact, CD can eliminate the zero or almost zero eigenvalues of a matrix without computing the whole matrix, but only calculating the complete diagonal as well as the columns of the decomposed diagonal elements. For a strictly positive matrix, the number K of Cholesky vectors equals the dimension of the matrix N . However, for a positive semi-definite matrix, K is smaller than N , $N - K$ being the number of linear dependencies. As a simple example, let us consider the following matrix

$$\mathbf{A} = \begin{pmatrix} a & c & \alpha a + \beta c \\ c & b & \alpha c + \beta b \\ \alpha a + \beta c & \alpha c + \beta b & \alpha^2 a + \beta^2 b + 2\alpha\beta c \end{pmatrix} \quad (13.6)$$

in which, obviously, the third column is a linear combination of the first two. When $\mathbf{A}$ is decomposed, we obtain only two Cholesky vectors

$$\begin{aligned} \mathbf{A} &= \frac{1}{\sqrt{a}} \begin{pmatrix} a \\ c \\ \alpha a + \beta c \end{pmatrix} \begin{pmatrix} a & c & \alpha a + \beta c \end{pmatrix} \frac{1}{\sqrt{a}} \\ &+ \begin{pmatrix} 0 \\ \sqrt{b - \frac{c^2}{a}} \\ \beta \sqrt{b - \frac{c^2}{a}} \end{pmatrix} \begin{pmatrix} 0 & \sqrt{b - \frac{c^2}{a}} & \beta \sqrt{b - \frac{c^2}{a}} \end{pmatrix} \end{aligned} \quad (13.7)$$

At this point, it is important to recall that the updated matrix remains positive definite only when exact arithmetic is used, but this property can be lost if round-off errors are significant. That is, round-off errors might cause matrices, which formally should be positive semi-definite, to have a slightly negative definite part. This is, for instance, the case for the two-electron integral matrix [32], which in practice limits the accuracy of the CD to about 10^{-12} ; still, this is more than enough for practical applications. For matrices that are better conditioned, such as the orbital energy denominators appearing in many-body perturbation theory (MBPT) calculations, it is possible to reach machine precision (i.e. 10^{-16}) in a straightforward manner [73].

To further illustrate the removal of linear or near-linear dependence by Cholesky decomposition, consider a positive definite operator $\hat{M}(\mathbf{r}_1, \mathbf{r}_2)$ with the following matrix representation in a real basis $\{\phi_i(\mathbf{r})\}$,

$$M_{ij} = \int \int \phi_i(\mathbf{r}_1) \hat{M}(\mathbf{r}_1, \mathbf{r}_2) \phi_j(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \equiv (\phi_i | \phi_j) \quad (13.8)$$

The positive definite operator is here assumed to be a two-electron operator such as the Coulomb operator $\hat{M}(\mathbf{r}_1, \mathbf{r}_2) = |\mathbf{r}_1 - \mathbf{r}_2|^{-1}$, or the attenuated Coulomb operator $\hat{M}(\mathbf{r}_1, \mathbf{r}_2) = A(|\mathbf{r}_1 - \mathbf{r}_2|) |\mathbf{r}_1 - \mathbf{r}_2|^{-1}$ (A is a strictly positive function; e.g., a Gaussian function or complementary error function). It could also be the Dirac delta function $\hat{M}(\mathbf{r}_1, \mathbf{r}_2) = \delta(\mathbf{r}_1 - \mathbf{r}_2)$. Now, even if the operator is positive definite the matrix representation may be positive semi-definite or nearly so. This occurs whenever the basis is linearly dependent in the metric defined by the operator $\hat{M}(\mathbf{r}_1, \mathbf{r}_2)$. As discussed above, CD of a positive semi-definite matrix leads to a number of Cholesky vectors which is smaller than the dimension of the matrix. The pivoting procedure defined by Eq. (13.5) leads to the concept of the Cholesky basis, which is defined as the subset of the basis set for which the corresponding diagonal elements give rise to Cholesky vectors. We use h_J to denote members of the Cholesky basis. Performing a modified Gram-Schmidt orthonormalization in the $\hat{M}$ -metric of the Cholesky basis leads to the orthonormal Cholesky basis

$$Q_J = N_J \left(h_J - \sum_{K=1}^{J-1} Q_K (h_J | Q_K) \right) \quad (13.9)$$

where the normalization constant is given by

$$N_J = \left[(h_J | h_J) - \sum_{K=1}^{J-1} (h_J | Q_K)^2 \right]^{-\frac{1}{2}} \quad (13.10)$$

and the matrix elements $(\cdot | \cdot)$ are defined in Eq. (13.8). We can now show that the Cholesky vectors can be expressed in terms of the orthonormal Cholesky basis as [49; 52; 58]

$$L_i^J = (\phi_i | Q_J) \quad (13.11)$$

This implies that CD can be used to remove linear dependence in a given basis using any positive definite metric. The original matrix can then be expressed as an inner projection in two equivalent ways

$$\begin{aligned} (\phi_i | \phi_j) &= \sum_J (\phi_i | Q_J) (Q_J | \phi_j) \\ &= \sum_{IJ} (\phi_i | h_I) \tilde{M}_{IJ}^{-1} (h_J | \phi_j) \end{aligned} \quad (13.12)$$

where $\tilde{M}_{IJ}^{-1} = (\tilde{\mathbf{M}}^{-1})_{IJ}$ and

$$\tilde{M}_{IJ} = (h_I | h_J) \quad (13.13)$$

Alternatively, we may expand the linearly dependent basis in the Cholesky basis

$$\phi_i = \sum_J C_i^J h_J \quad (13.14)$$

and obtain yet another equivalent expression

$$(\phi_i | \phi_j) = \sum_{IJ} C_i^I \tilde{M}_{IJ} C_j^J \quad (13.15)$$

The expansion coefficients are determined by a least-squares fitting (in the $\hat{M}$ -metric), leading to the equations

$$\sum_I C_j^I \tilde{M}_{IJ} = (\phi_j | h_J) \quad (13.16)$$

from which the coefficients can be computed. Using that the matrix $\tilde{\mathbf{M}}$ is the sub-block of $\mathbf{M}$ which is represented exactly by the CD, these fitting equations can be recast as [58]

$$\sum_I C_j^I L_I^J = L_j^J \quad (13.17)$$

providing a relation between fitting coefficients and Cholesky vectors.

We close this section by showing an important property of the CD that is rarely discussed in standard textbooks on linear algebra. The decomposition of a projection operator gives orthonormal Cholesky vectors [48], such that an incomplete decomposition is again a projection operator. Consider a projector P and the projection on the complement $Q = 1 - P$, such that $P^2 = P$ and $Q^2 = Q$. In the complete basis $\{|n\rangle\}$ of finite or infinite dimension we have the resolution of identity

$$\sum_n |n\rangle\langle n| = 1 \quad (13.18)$$

and the matrix representations that we assume real

$$\begin{aligned} P_{nm} &= \langle n | P | m \rangle = P_{mn} \\ Q_{nm} &= \langle n | Q | m \rangle = Q_{mn} . \end{aligned} \quad (13.19)$$

The first Cholesky vector

$$L_n^1 = \langle n | P | 1 \rangle \langle 1 | P | 1 \rangle^{-1/2} = P_{n1} P_{11}^{-1/2} \quad (13.20)$$

is easily seen to be normalized. The second Cholesky vector

$$L_n^2 = \left(P_{n2} - \frac{P_{12}^2}{P_{11}} \right) \left(P_{22} - \frac{P_{12}^2}{P_{11}} \right)^{-1/2} \quad (13.21)$$

is also normalized

$$\sum_n (L_n^2)^2 = \left(P_{22} - \frac{P_{12}^2}{P_{11}} \right)^{-1} \sum_n \left(P_{2n} - \frac{P_{1n} P_{12}}{P_{11}} \right) \left(P_{n2} - \frac{P_{n1} P_{12}}{P_{11}} \right) = 1 \quad (13.22)$$

and orthogonal to the first Cholesky vector

$$\begin{aligned}
\sum_n L_n^1 L_n^2 &= P_{11}^{-1/2} \left(P_{22} - \frac{P_{12}^2}{P_{11}} \right)^{-1/2} \sum_n P_{1n} \left(P_{n2} - \frac{P_{n1} P_{12}}{P_{11}} \right) \\
&= P_{11}^{-1/2} \left(P_{22} - \frac{P_{12}^2}{P_{11}} \right)^{-1/2} \left(P_{12} - \frac{P_{11} P_{12}}{P_{11}} \right) = 0 \quad (13.23)
\end{aligned}$$

An induction proof is now straightforwardly completed. As P and Q are orthogonal projectors, i.e. $PQ = 0$, any incomplete decomposition of P and Q give Cholesky vectors that are orthonormal. We shall apply this property in section 13.3.3 to localize occupied and virtual orbitals.

13.3 Applications

In this section we give a number of examples of the use of the CD technique in *ab initio* methods. These examples include the approximate representation of two-electron integrals, of course, but also other aspects such as orbital localization and quartic-scaling MP2.

13.3.1 Connection between density fitting and Cholesky decomposition

Although Beebe and Linderberg [24] noted that the CD technique was an inner projection approximation and Vahtras and co-workers [8] noted that the so-called V approximation in their integral approximation scheme is “an inner projection similar to the Beebe-Linderberg approximation” this has never been explicitly demonstrated and exploited. In this section we will clearly demonstrate that the CD method implicitly generates an auxiliary basis, the orthonormal Cholesky auxiliary basis, which in some respects is no different from those used in standard DF methods. The connection between DF and CD is discussed in detail in the recent review by Pedersen, Aquilante, and Lindh [58].

As discussed in the previous section, the CD procedure is nothing but a modified Gram-Schmidt orthonormalization procedure applied to the positive definite molecular super-matrix, $(\alpha\beta | \gamma\delta)$ and includes in a pivoting algorithm the original atomic orbital (AO) product functions $\gamma\delta$ one by one into the orthonormalization procedure. Let us use this order index, J , to denote a particular AO product as h_J , the parent Cholesky auxiliary basis. The CD procedure contains the two elements — orthogonalization and normalization. The orthonormalized Cholesky auxiliary basis, Q_J , is defined by Eq. (13.9) with the normalization constant given by Eq. (13.10).

We note that the parent Cholesky auxiliary basis $\{h_J\}$ and the orthonormal Cholesky auxiliary basis $\{Q_J\}$ span the same space and that for the latter, by virtue of design, the molecular super-matrix is trivially expressed as $(Q_J | Q_K) = \delta_{JK}$. This

equivalence now allows us to express the approximated two-electron integrals using the DF formalism in two equivalent ways (the first two lines of the equation),

$$\begin{aligned}
 (\alpha\beta | \gamma\delta) &= \sum_{JK} (\alpha\beta | h_J) G_{JK}^{-1} (h_K | \gamma\delta) \\
 &= \sum_{JK} (\alpha\beta | Q_J) \tilde{G}_{JK}^{-1} (Q_K | \gamma\delta) \\
 &= \sum_J (\alpha\beta | Q_J) (Q_J | \gamma\delta)
 \end{aligned} \tag{13.24}$$

where $G_{JK} = (h_J | h_K)$ and $\tilde{G}_{JK} = (Q_J | Q_K) = \delta_{JK}$. The last step is a trivial consequence of the orthonormal Cholesky auxiliary basis and leads us to the identification of the Cholesky vectors as

$$L_{\alpha\beta}^J = (\alpha\beta | Q_J) \tag{13.25}$$

which, apart from a slightly different notation, is identical to Eq. (13.11). We have now established that the CD approach indirectly form a particular auxiliary basis to be used within the DF formalism. This auxiliary basis is similar in many respects to the standard available DF auxiliary basis sets. However, there are three major differences. First, the orthonormal Cholesky auxiliary basis is in principle a two-center type of auxiliary basis. This allows us to, with a single parameter, control the error of the approximated integrals to an arbitrary accuracy, while it also presents some challenges in the calculation of analytic gradients. This problem has been resolved with the atomic CD procedure [49; 62], in which accurate analytic gradients can be expressed with a reduction of accuracy of the approximated integrals that is of no practical consequence [52]. Second, the Cholesky auxiliary basis set is always a subset of the full product space. Third, the auxiliary basis set is a set of (product) basis functions ordered with respect to their significance in the reproduction of the two-electron integrals, i.e. the set can be further truncated on-the-fly in a screening scheme with complete control of the accuracy. That is, it is a compact and efficient auxiliary basis set. Normally, DF auxiliary basis sets are improved by *ad hoc* contraction or addition of primitive Gaussians. This procedure does not control that the additional functions are completely within the space of the original product basis, and can in this way be wasteful. Rather, the quality of the extended DF auxiliary basis sets are checked in subsequent test calculations based on energy criteria (see for example Refs. [74–76]). To summarize, there are good arguments why the CD auxiliary basis sets, in the parent or orthonormal form, are the optimal general auxiliary basis sets. Benchmark results that support this view are presented in section 13.4.

13.3.2 One-center CD auxiliary basis sets

In contrast to conventional RI or DF schemes, the CD procedure generates auxiliary basis sets which in general are of two-center type. This allows the standard

CD technique applied to the two-electron integral, which we denote Full-CD, to approximate the conventional result to any degree of accuracy with the adjustment of a single threshold parameter. However, the two-center character of the Full-CD auxiliary basis set makes it difficult to formulate a general analytic expression for derivatives of the two-electron integrals with respect to the nuclear coordinates. As the geometry changes, the set of Full-CD auxiliary basis functions changes as well which may introduce discontinuities on the potential energy surface of the order of magnitude of the decomposition threshold. A brute force way to control this would of course be to reduce the threshold to the extent that this feature is of no practical significance. However, for larger CD thresholds the problem remains. Having established the connection between CD and RI/DF techniques in section 13.3.1 we can solve this problem in a pragmatic way. For RI/DF techniques analytic expressions for gradients have already been derived and is a consequence of the strict one-center type of auxiliary basis sets used in these techniques – the auxiliary basis set is not a function of the molecular geometry. Still with these limitations the RI/DF approach has been shown to provide an approximation with reasonable accuracy. The question which follows at once is "Can a CD procedure be used to generate meaningful one-center type auxiliary basis sets?".

In a series of papers, methods for one-center type of CD auxiliary basis sets have been described [49; 58; 62]. The first paper developed the one-center CD (1C-CD) and the atomic CD (aCD) auxiliary basis sets. Both of these methods are straightforward modifications of the Full-CD procedure. In the 1C-CD approach only one-center product functions are allowed to enter the Cholesky basis [49]. Thus, a one-center auxiliary basis set is constructed which for all practical purposes, though not strictly, is invariant to the molecular geometry. This is, however, strictly the case for the aCD approach [49] in which the auxiliary basis set is generated from a decomposition of the separate atomic blocks of the molecular two-electron integral matrix. For these two types of auxiliary basis set one remaining deficiency in common with Full-CD and in difference to standard RI/DF auxiliary basis sets remains: a rather large set of primitive Gaussians is used. The so-called atomic compact CD (acCD) auxiliary basis sets [62] offer a significant reduction of these primitive sets without significant loss of accuracy. The hallmark of the acCD procedure is that in the atomic CD procedure the set of primitive Gaussian products itself is subject to a Cholesky decomposition and that the remaining functions are fitted to produce an accurate representation of the parent aCD auxiliary basis set. This compact representation of the aCD auxiliary basis set introduces significant reduction of the primitive space and speeds up the computation of two- and three-center two-electron integrals when the acCD auxiliary basis sets are used in the RI/DF procedure (see Figure 13.1 for a representation of the reduction of the primitive product space in the case of an all-electron atomic natural orbital valence basis set for technetium). In addition, it should be noted that these new one-center auxiliary basis sets are constructed on the fly. That is, with these new procedures, the need for auxiliary basis set libraries is a feature of the past and that RI/DF calculations now can be performed directly for any existing or future valence basis set. The accuracy of the one-center CD auxiliary basis sets is discussed in section 13.4.

Fig. 13.1 Progression of the primitive set of technetium in a) the valence ANO-RCC s-shell and corresponding b) aCD-4 and c) acCD-4 auxiliary basis sets. The decimal logarithm of the Gaussian exponent is reported on the horizontal axes. (Reproduced with the permission of AIP)

Finally, we point out a unique feature of the aCD and acCD auxiliary basis sets. Since the auxiliary functions span the same space (within the decomposition threshold) as the one-center atomic orbital product functions that are to be fitted, only auxiliary functions on the same center as the product functions are needed to produce an accurate fit [62]. That is, the fitting of one-center atomic orbital products becomes increasingly local as the threshold is decreased. This feature of the atomic CD sets implies that local density fitting can be performed without the use of less accurate short-range metrics such as the overlap or attenuated Coulomb metrics discussed by Jung *et al.* [77]. The situation is different for two-center atomic orbital products where two-center auxiliary functions are needed to obtain an inherently local fit with the accurate long-range Coulomb metric [58]. As discussed in [58], it is possible to get an accurate local two-center fit with a minimum amount of two-center auxiliary functions, namely exactly those that are needed to make the auxiliary set complete (within the decomposition threshold) in the subspace of interest. These observations pave the way for truly local density fitting where each approximate atomic orbital product function is represented in exactly the same way in any molecule: the fit becomes transferable and thus free from (auxiliary) basis set superposition errors. By the same token, this approach eliminates the size-extensivity errors uncovered by Werner and Manby [78] with the explicitly correlated DF-MP2-R12 method.

To illustrate the locality of the atomic CD sets, we follow Jung *et al.* [77] in considering a simple model problem. Suppose that we wish to fit a spherical unit-charge density s_T using two identical auxiliary basis functions s_1 and s_2 , which are also spherical unit-charge densities. While s_1 is centered at the same point as s_T , s_2 is a distance r away. We now seek the coefficients c_1 and c_2 such that the approximation $|s_T) \approx c_1|s_1) + c_2|s_2)$ is the best possible in a least-squares sense. This is achieved by solving the fitting equations

$$\begin{bmatrix} c_1 & c_2 \end{bmatrix} \begin{bmatrix} A & B \\ B & A \end{bmatrix} = \begin{bmatrix} C & D \end{bmatrix} \quad (13.26)$$

where $A = (s_1|s_1) = (s_2|s_2)$, $B = (s_1|s_2) = (s_2|s_1)$, $C = (s_1|s_T)$, and $D = (s_2|s_T)$. The solution is given by

$$c_1 = \frac{C}{A} - \frac{B}{A} \frac{AD - BC}{A^2 - B^2} \quad (13.27)$$

$$c_2 = \frac{AD - BC}{A^2 - B^2} \quad (13.28)$$

showing that, in general, there is a non-vanishing contribution to the fit from the distant auxiliary function s_2 . The central question is whether this contribution decays sufficiently fast to allow a local fit involving the function s_1 only. For large distances r , we have [77]

$$B \rightarrow r^{-1} \quad (13.29)$$

$$D \rightarrow r^{-1} \quad (13.30)$$

and therefore

$$c_1 \rightarrow \frac{C}{A} \quad (13.31)$$

$$c_2 \rightarrow \frac{A - C}{A^2} \frac{1}{r} \quad (13.32)$$

This shows that c_2 is a slowly decaying function of the distance, making the fitting procedure inherently nonlocal. However, using the Cholesky approach to auxiliary basis functions for this simple model system, we have $s_1 = s_T$ and therefore $C = A$ and $D = B$. Hence, regardless of the distance r , Eqs. (13.27) and (13.28) become

$$c_1 = 1 \quad (13.33)$$

$$c_2 = 0 \quad (13.34)$$

showing that the fit is manifestly local. Fitting a single function using itself as auxiliary basis is silly, of course, but the example underlines the importance of choosing the auxiliary basis such that it spans the same space, or very nearly so, as the functions to be fitted. Not only will the fit be more accurate, it will render the long-range decay of the fitting coefficients irrelevant [62]. More general discussions, theoretic-

cal as well as through numerical examples, of this inherent locality can be found in [58; 62].

13.3.3 Orbital localization using Cholesky decomposition

A key ingredient needed for state of the art electronic structure methods to be efficient is the sparsity of the molecular orbital (MO) coefficient representation of the density matrix. Sparsity is in this context the property of an array to have relatively few, and usually scattered, large elements. Due to the invariance of HF theory with respect to unitary transformations among the occupied (or virtual) orbital space, an infinite number of such representations are possible. Of course, sparsity of the MO coefficients does not necessarily imply that the resulting orbitals are localized. Locality carries the notion of spatial extent of an object, which in this case indicates that the orbitals are confined to few neighboring atoms. Orbital locality implies a degree of sparsity in the MO coefficients but not *vice versa*. Several localization schemes have been developed for choosing a unitary transformation of the molecular orbitals. Among the most common are the Boys [79], Edmiston-Ruedenberg [80], and Pipek-Mezey [81] procedures, which are all formulated as an optimization problem where a localization functional is maximized with respect to rotations among the orbitals. The orbital localization thus becomes an iterative procedure. An alternative approach to obtain localized orbitals is provided by the Cholesky decomposition of the AO density matrix [48]. This Cholesky localization has several computational advantages over the standard schemes and also some unique features. First, Cholesky decomposition is a numerically stable and fast algorithm that can be made linear scaling for matrices with linear scaling number of non-zero elements [82]. Second, being a non-iterative procedure, complicated optimization techniques are not needed. Third, as no initial orbitals need to be given, the procedure is particularly well suited for determining local MOs directly from density matrix-based HF theory. Fourth, it is the natural choice for obtaining localized orbitals to be used in connection with the local-K exchange screening discussed in section 13.3.4.

In addition to local occupied orbitals, Cholesky decomposition of *ad hoc* defined density-type matrices can produce local MOs spanning the virtual space, as required by linear-scaling wave function-based electron correlation models. In particular, the corresponding Cholesky MOs are reasonably localized even though they are orthonormal by construction (since the density is a projector, see also section 13.2). The convenience of using orthonormal orbitals for the virtual space instead of the usual (redundant) projected AOs could increase in the future the popularity of Cholesky localization techniques — probably the only orbital localization scheme that works for both occupied and virtual orbitals, with and without symmetry, and produces very quickly localized orbitals directly from the parent density matrix.

As is well known, the Hartree-Fock energy is invariant under any unitary transformation that preserves the mutual orthogonality between the occupied and virtual

subspaces. The underlying reason is that the density matrix

$$D_{\alpha\beta} = \sum_i^{\text{occ}} C_{\alpha i} C_{\beta i} \quad (13.35)$$

where the molecular orbital coefficients are collected in the matrix $\mathbf{C}$ and i denotes an occupied MO, is indeed invariant under rotations of the occupied orbitals. Similarly, the virtual density-like matrix

$$D_{\alpha\beta}^V = \left(\mathbf{S}^{-1} - \mathbf{D} \right)_{\alpha\beta} = \sum_a^{\text{virt}} C_{\alpha a} C_{\beta a} \quad (13.36)$$

where $\mathbf{S}$ is the overlap matrix in the AO basis and a denotes a virtual MO, is invariant under rotations of the virtual orbitals. Both $\mathbf{D}$ and $\mathbf{D}^V$ matrices are positive semi-definite, their ranks being the number of, respectively, occupied and virtual orbitals. It is possible to Cholesky decompose each of these matrices selecting only those diagonals that correspond to atomic orbitals centered on atoms belonging to a predefined set of active centers defining a subsystem [43]. Once the decomposition is done in the reduced subspace, the residual matrix is still positive semi-definite and can, therefore, be decomposed without any restriction. In this way, we obtain an active set of orbitals, occupied and virtual, orthogonal to an inactive set that is eliminated, i.e. kept frozen, in the subsequent correlation treatment. Alternatively, one can also decompose completely the two density matrices on an atom-by-atom basis, thus obtaining a set of orthonormal MOs that span the same subspace as the canonical Hartree-Fock set. When the decomposition is finished, the active orbital space is selected from the atomic centers forming the subsystem. Either way, we obtain a localized MO basis spanning a reduced space located on the subsystem. If orbital energies are required, the Fock operator may be diagonalized in the active space.

The great advantage of using a small subsystem space is to concentrate the computational effort where it is in fact required. Indeed, many molecular properties are essentially local in character and, thus, it is reasonable to assume that only the atoms in the neighborhood of a specific site, the subsystem, significantly contribute to the considered property. In Figure 13.2 we represent the orbital spread [83] of the Cholesky orbitals in C_{10}H_2 compared to that of the canonical Hartree-Fock orbitals. It is easily verified that the procedure is able to localize the molecular orbitals, although augmentation (depicted in the right part) introduce a larger degree of delocalization.

For the calculation of intermolecular interaction energies, we start by noting that the canonical Hartree-Fock molecular orbitals are not well suited for this purpose (see Figure 13.3). In fact, they are spread out along the entire organic unit, but have very small components in the vicinity of the interaction area. Thus, the largest amplitudes will give virtually zero contribution to the calculated interaction. On the other hand, the Cholesky orbitals are concentrated in the area of interest. The interaction energy has been calculated for several active spaces and using double zeta correla-

Fig. 13.2 Histogram of orbital spreads of canonical (above) and Cholesky (below) molecular orbitals for $C_{10}H_2$ with cc-pVDZ (left) and aug-cc-pVDZ (right) basis sets. The abscissa denotes the orbital spread in atomic units and the ordinate the number of molecular orbitals (both occupied and virtual) containing the spread for a step of 0.25 a.u.

tion consistent basis sets [84]. The smallest subsystem is formed by the neon atom, the $CH_2=CH-$ group in front of it and a $[3s3p2d1f1g]$ set of bond functions [85] located at the center of the line connecting the neon atom and the $C=C$ double bond. In each successive space an additional CH_2 unit is attached. Our results are depicted in Figure 13.4. This is a very complicated problem, as the interaction is completely described by weak dispersion forces, and augmented basis sets are required. As discussed above, augmentation introduces some delocalization, implying that rather large subsystems are needed to get meaningful results. However, if the interaction is calculated according to

$$\Delta E_{MP2/CCSD(T)} = \Delta E_{CCSD(T)}^{act} - \Delta E_{MP2}^{act} + \Delta E_{MP2}^{exact} \quad (13.37)$$

the convergence towards to the exact value is fast and almost uniform, making it feasible to get accurate results using very reduced spaces.

13.3.4 The LK algorithm

The use of the Cholesky representation of the two-electron integrals, as proposed by Beebe and Linderberg [24] and extended by Koch and co-workers, [32], does not

Fig. 13.3 Complex octene-neon: HOMO (left) and LUMO (right) in the canonical (above) and Cholesky (below) molecular orbital basis.

Fig. 13.4 Interaction energy (in cm^{-1}) for the neon-octene complex using cc-pVDZ (left) and aug-cc-pVDZ (right) supplemented with midbond functions. The horizontal axis represents the number of $\text{CH}_{(n)}$ groups in the active space.

always yield a satisfactory performance for SCF calculations involving the evaluation of HF-exchange. For large molecules (>100 atoms) and compact basis sets, standard integral-direct algorithms are by far the fastest. The quartic scaling of the evaluation of the exchange Fock matrix $\mathbf{K}$ quickly downgrades the capacity of the straightforward Cholesky and DF SCF implementations [19; 32] for electron-rich (many occupied orbitals) systems and large basis set. Particularly unpleasant is the dependence of the computational costs on the number of occupied orbitals, as this is not the case for integral-based SCF algorithms. Polly et al. [20] were the first to propose an alternative algorithm for computing the exchange Fock matrix in DF-based SCF with reduced scaling. Although asymptotically linear scaling, this procedure

does not yield bounded errors — therefore the accuracy of the result cannot be controlled solely on the basis of energy thresholds. As discussed previously, it is exactly this error control that makes the Cholesky approximation extremely appealing. Consequently, an alternative solution to this “*exchange problem*” that could maintain the prime characteristic of the method was proposed by some of the present authors [21]. This approach, called *Local K* or simply *LK* algorithm is general in the sense that it does not involve *a priori* assumptions about the size of the system or the nature of its electronic structure. Only the short-range character of the exchange interactions is required for the screening to be effective.

In order to implement an efficient and also error bounded screening on the contribution of a molecular orbital i to the matrix elements of K , the *LK* algorithm makes use of the following chain of inequalities

$$\begin{aligned} |K_{\alpha\beta}^i| &= |(\alpha i | \beta i)| \leq \sum_{\gamma\delta} |C_{\gamma i}| |(\alpha \gamma | \beta \delta)| |C_{\delta i}| \\ &\leq \sum_{\gamma\delta} |C_{\gamma i}| |D_{\alpha\gamma}^{1/2} D_{\beta\delta}^{1/2}| |C_{\delta i}| = Y_{\alpha}^i Y_{\beta}^i \end{aligned} \quad (13.38)$$

where $D_{\alpha\beta} = (\alpha\beta | \alpha\beta)$ are the (exact) diagonal elements of the two-electron matrix in AO basis and the i -th vector $\mathbf{Y}^i$ has elements $Y_{\alpha}^i = \sum_{\gamma} D_{\alpha\gamma}^{1/2} |C_{\gamma i}|$. Clearly, the vector $\mathbf{Y}^i$ will be sparse whenever the corresponding MO coefficient vector $\mathbf{C}^i$ represents a localized orbital. Experience shows that the use of Cholesky orbitals, described in section 13.3.3, mediates perfectly between the need for a fast localization (to be performed at each SCF cycle) with that of a sufficient sparsity in the resulting $\mathbf{Y}$ vectors.

It is important to notice that in deriving the inequalities (13.38) we have only used the fact that the two-electron integral matrix in AO basis is positive definite and therefore satisfies the Schwarz inequality. Whenever the contribution to the exchange Fock matrix from a certain number m of Cholesky vectors has been computed, the inequalities can be applied in the very same way to the remainder matrix. In practice this means that the Y vectors can be computed using updated integral diagonals, namely

$$\tilde{D}_{\alpha\beta} = D_{\alpha\beta} - \sum_{j=1}^m (L_{\alpha\beta}^j)^2 \quad (13.39)$$

Due to the inner projection nature of the Cholesky decomposition, these updated integral diagonals are guaranteed to remain nonnegative. Most importantly, the number of significant elements $\tilde{D}_{\alpha\beta}$ will decrease, since more and more Cholesky vectors have already contributed to the exchange Fock matrix. The details of the *LK* screening are presented and extensively discussed in the original paper [21]. A sketch of the shell-driven *LK* algorithm is shown in Figure 13.5. (A shell is defined as a set of atomic orbital basis functions on a given center with the same angular momentum.) The need to perform an MO half-transformation of the Cholesky vectors, formally as expensive as the build of $\mathbf{K}$, is tackled by employing estimates based on the Frobenius norm of the resulting matrices. The exploitation of the per-

```

Loop over Cholesky vectors
Loop over occupied orbitals ( $i$ )

  (1) Pre-selection and pre-ordering of the eligible shells
  Compute the  $Y^{(i)}$  screening vectors from (updated) integral diagonals
  Loop over all  $\alpha$ -shells
    If (  $MY^{(i)} \cdot \max(Y_\alpha^{(i)}) \geq \tau$  ) then
      Add the  $\alpha$ -shell to a list  $ML^{(i)}$ 
    EndIf
  End Loop
  Sort the  $\alpha$ -shells in  $ML^{(i)}$  by the value  $\max(Y_\alpha^{(i)})$ 

  (2) Frobenius-norm screened MO transformation
  Loop over  $\alpha$ -shells in the list  $ML^{(i)}$ 
  Loop over all  $\beta$ -shells in the shell-pairs ( $\alpha\beta$ ) of the AO Cholesky vectors
    If (  $\|L_{\alpha\beta}^J\|_F \cdot \|C_\beta^{(i)}\|_F \geq \tau_F$  ) then
      Perform the MO half-transformation  $L_{\alpha(i)}^J = \sum_\beta L_{\alpha\beta}^J C_\beta^{(i)}$ 
    EndIf
  End Loop
  Remove from  $ML^{(i)}$  the  $\alpha$ -shells that did not qualify for the MO transformation
  End Loop

  (3) Evaluation of the exchange Fock matrix  $K_{\alpha\beta}^{(i)}$ 
  Loop over  $\alpha$ -shells in  $ML^{(i)}$ 
  Loop over the  $\beta$ -shells ( $\beta \leq \alpha$ ) in  $ML^{(i)}$ 
    If (  $\max(Y_\alpha^{(i)}) \cdot \max(Y_\beta^{(i)}) \geq \tau$  ) then
      Compute  $K_{\alpha\beta}^{(i)} = \sum_J L_{\alpha(i)}^J L_{\beta(i)}^J$ 
    Else
      Leave  $\beta$ -loop
    EndIf
  End Loop
End Loop

  (4) Update integral diagonals
End Loop

```

Fig. 13.5 A sketch of the LK algorithm. For a given occupied orbital i , $MY^{(i)}$ is defined as the maximum element of the vector $Y^{(i)}$ defined in Eq. (13.38), and $ML^{(i)}$ is a list of contributing shells of basis functions.

mutational symmetry of the two-electron integrals, as well as the shell pre-ordering, represent some of the key features for the efficiency of the screening. By the nature of the method, the accuracy of the LK screening is absolutely reliable and can be controlled by the choice of the screening thresholds. Moreover, the LK screening is indeed capable of reducing the scaling of the evaluation of the exchange Fock matrix from quartic to quadratic. In Figure 13.6, we can see an example of what that means in terms of performances in Cholesky SCF calculations: with LK it is now possible to perform HF wavefunction optimizations faster than with integral-direct algorithms. Currently, the LK screening is used in MOLCAS for any type of Fock ma-

Fig. 13.6 Timings for the construction of the exchange Fock matrix for linear alkanes using standard Cholesky, LK-Cholesky and integral-direct. Decomposition threshold $\delta = 10^{-4}$ and cc-pVTZ basis set.

trix build based on Cholesky vectors. We also stress the importance of the speedup achieved for CASSCF calculations, as documented in the original implementation paper [45]. Here, as an example, we will only report the performances of the LK-CASSCF algorithm for some of the systems described in Ref. [59]. These systems are important intermediates in the reaction of O_2 with a Cu(I)- α -Ketocarboxylate, and the accurate evaluation of the singlet-triplet splitting in each species is essential to the understanding of the mechanism of activation of molecular oxygen by copper coordination complexes [60]. With atomic natural orbital (ANO) basis sets of double- ζ quality, we are in a range of about 280 to 450 contracted Gaussian basis functions, depending on the system (point group symmetry not employed). In this range, the time spent to generate the DF-vectors is only a small fraction of the total time for any choice of the Cholesky basis (Full-CD, aCD, etc.), which also has no significant effect on the time required by the subsequent steps. The following timings refer always to wall-time observed on an architecture of the type Intel(R) Xeon(TM) 3.20GHz with 8GB RAM and employing the MOLCAS software. For a typical calculation with 300 basis functions, about 2.5 hours are required for computing and storing the conventional two-electron integrals. The subsequent CASSCF wavefunction optimization of the singlet ground state takes about 38 minutes per iteration, for an active space of (12-in-12). Using acCD-4, the vector generation is performed in only 3 minutes, and 3.4 minutes per iteration are required in the LK-CASSCF step. Considering that with the present CASSCF optimizer in MOLCAS

an average of 100 iterations are needed for converging the wavefunction, the observed speed-up for the overall calculation is nearly a factor 12. Moving towards 450 basis functions, the conventional calculations can hardly be afforded due to the large disk-space requirements. With 437 basis functions, the acCD-4 vector generation goes up to 8 minutes, whereas the LK-CASSCF time per iteration stays within 4 minutes. Already with this relatively small basis set, the overall speed-up compared to the conventional calculation is nearly of two orders of magnitude. Speed-up for larger basis sets is difficult to measure, as the conventional calculation becomes impossible.

We close this section with a final remark. There is not a single notion in the LK algorithm that could distinguish between a Cholesky or any other DF representation of the integrals. In other words, the LK screening is a simple, accurate and general solution to the exchange problem.

13.3.5 Quartic-scaling MP2

In MP2 theory, the need for the set of integrals $(ai | bj)$, where i, j and a, b label occupied and virtual orbitals, respectively, makes the conventional calculation a fifth-order process. Using Cholesky or DF, we can compute the same set of integrals as follows

$$(ai | bj) = \sum_J L_{ai}^J L_{bj}^J \quad (13.40)$$

where L_{ai}^J are the MO-transformed Cholesky or DF vectors. The MO transformation of these vectors is not the bottleneck of the calculation. With O and V denoting, respectively, the number of occupied and virtual orbitals and M the number of Cholesky vectors, it requires $\sim ON^2M$ operations, while the evaluation of Eq. (13.40) has a computational cost of $\sim O^2V^2M$. Compared to the conventional $\sim ON^4$ computational requirement, the smaller prefactor allows substantial speed-ups. The evaluation of the integrals from Eq. (13.40) is particularly well suited for canonical MP2 calculations, since in this case they can be computed on-the-fly in a batched loop, used to evaluate the energy contribution and never stored on disk [32]. Comparatively, in the Cholesky-based CASPT2 method [46] there is a possibility to avoid the evaluation of the corresponding $(ai | bj)$ -type integrals and reformulate the method directly in terms of Cholesky vectors. However, the existing MOLCAS implementation of CASPT2 still goes through the evaluation of such integrals as in Eq. (13.40) and stores them on disk. Nonetheless, the present implementation makes possible CASPT2 calculations otherwise beyond the capabilities of the conventional implementation [51; 53–55; 57; 59; 60]. This is achieved because it bypasses completely the AO integral storage bottleneck and also because it produces the needed MO integrals at reduced computational cost and input-output overheads.

Let us analyze in more detail the formal scaling of MP2. The question is whether or not the approach adopted so far in Cholesky and DF (closed-shell) MP2 is the best possible. We shall demonstrate that it is not, and that the scaling of the method

can be reduced from fifth to fourth order, with an algorithm that is also perfectly parallelizable. This is achieved by performing the Cholesky decomposition of the MP2 amplitude matrix (with minus sign), namely

$$-t_{ai,bj} = \frac{(ai | bj)}{\epsilon_a - \epsilon_i + \epsilon_b - \epsilon_j} = \sum_{K=1}^m R_{ai}^K R_{bj}^K. \quad (13.41)$$

The Cholesky decomposition algorithm applied to this matrix requires $\sim OVm^2$ operations plus the evaluation of m columns of the matrix $\mathbf{t}$. The necessary two-electron integrals can be computed using Cholesky or DF representations as in Eq. (13.40). It is crucial to understand how the value of m scales with system size. Although the matrix $\mathbf{t}$ has a quadratic-scaling dimension (OV), its effective rank m scales only linearly with the size of the system. In fact, it is easy to realize that m is bounded by a linear scaling quantity, $n_r M$, where n_r is the number of Cholesky vectors needed for an exact decomposition of the orbital energy denominator matrix. It is known that this number is very small and, more importantly, independent of the size of the molecule [73]. Therefore the effective rank of $\mathbf{t}$ scales linearly with the system size.

We can now write the expression for the closed-shell canonical MP2 energy as follows

$$\begin{aligned} E_2 &= \sum_{K=1}^m \sum_{aibj} R_{ai}^K R_{bj}^K [2(ai | bj) - (aj | bi)] \\ &= \sum_{\alpha\beta\gamma\delta} \Theta_{\alpha\beta,\gamma\delta} [2(\alpha\beta | \gamma\delta) - (\alpha\delta | \gamma\beta)] \end{aligned} \quad (13.42)$$

where

$$\Theta_{\alpha\beta,\gamma\delta} = \sum_{K=1}^m R_{\alpha\beta}^K R_{\gamma\delta}^K \quad (13.43)$$

and $R_{\alpha\beta}^K$ are the elements of the back-transformed Cholesky vectors of the amplitudes. It is possible to show that Θ is a sparse matrix [86]. Together with the sparsity of the AO two-electron integrals, this implies that for large systems an efficient screening is possible in order to reduce the costs for the evaluation of E_2 . In fact, by applying the Schwarz inequality to both Θ and the integrals, we can show that the exchange-type term can be computed with an effort that is asymptotically linear. However, due to the fact that Θ needs to be computed from its Cholesky representation, most likely this step becomes quadratic scaling with a low prefactor. In the same way, the Coulomb-type term will require a cubic scaling effort.

In order to efficiently implement Eq. (13.42), the exact AO two-electron integrals need to be computed in a direct fashion and not from their Cholesky or DF representation. The reason for that is the possibility to achieve efficient parallelization of the code. The Cholesky decomposition of $\mathbf{t}$ can be performed separately on each node without any communication. The amplitude matrix $\mathbf{t}$ will be different on each node and corresponds to a partial contribution to the $(ai | bj)$ integrals, which is

computed from MO transformed Cholesky vectors distributed among the nodes. On the other hand, since the evaluation of the AO two-electron integrals is at most a quadratic step, it can be performed by each node without jeopardizing the overall efficiency of the calculation. Indeed, since the matrix on each node is only a partial contribution to the total matrix, the application of the Schwarz inequality will result in the evaluation of an even smaller (and different) set of AO integrals on each node. Aquilante and Pedersen [47] have shown an application of a simplified form of this algorithm for the evaluation of the Coulomb-type term only (namely, scaled opposite spin (SOS) MP2).

13.3.6 Calculation of molecular gradients

The calculation of the forces acting on the atomic nuclei of molecules is a bread-and-butter task of quantum chemistry. Knowledge of the atomic forces makes it possible to study molecular geometries, such as equilibrium structures and transition states, and is also needed for molecular dynamics simulations. Atomic forces are defined as (minus) the first derivatives of the total electronic energy with respect to nuclear positions. Higher-order energy derivatives with respect to nuclear positions are needed for the calculation of harmonic force constants and vibrational frequencies (second derivatives), anharmonic force constants (third derivatives), and so on. In order to compute atomic forces and higher-order derivatives we must be able to calculate derivatives of the two-electron integrals. Since CD of the two-electron integrals is a numerical procedure, defining analytic derivatives of Cholesky vectors is non-trivial.

O’Neal and Simons [31] have proposed a procedure in which the undifferentiated and differentiated integrals are considered as elements of an extended positive semi-definite matrix, which is Cholesky decomposed. Owing to the fact that most of the first derivative atomic orbital product functions belong to the space spanned by the undifferentiated products, only a modest increase in the number of Cholesky vectors is observed in comparison to the undifferentiated case. Combining this approach with the method specific CD technique (section 13.3.7) could provide a viable path to reduced scaling evaluation of atomic forces.

An alternative analytic approach was recently proposed by Aquilante, Lindh, and Pedersen [52]. It is based on the connection between CD and DF discussed in section 13.3.1. The first derivative of Eq. (13.24) can be written as

$$(\alpha\beta | \gamma\delta)^{(1)} = \sum_J C_{\alpha\beta}^J (\gamma\delta | h_J)^{(1)} + \sum_J (\alpha\beta | h_J)^{(1)} C_{\gamma\delta}^J - \sum_{JK} C_{\alpha\beta}^J G_{JK}^{(1)} C_{\gamma\delta}^K \quad (13.44)$$

where

$$C_{\alpha\beta}^J = \sum_K G_{JK}^{-1} (\alpha\beta | h_K) \quad (13.45)$$

and all derivatives are to be evaluated at the nuclear geometry of interest. This expression solely involves derivative integrals that can be evaluated analytically. While Eq. (13.44) takes into account the explicit geometry-dependence of the auxiliary basis functions, any implicit geometry-dependence is neglected. This means that Eq. (13.44) is exact for any auxiliary basis set whose composition is independent of molecular geometry. This is a significant difference as compared to the approach suggested by O’Neal and Simons which has an accuracy depending on the CD threshold.

Predefined auxiliary basis sets are composed of atom-centered functions, which are used regardless of the molecular geometry. Hence, predefined auxiliary basis sets only contain explicit geometry-dependence, making Eq. (13.44) exact for this case [74; 87–91]. Cholesky auxiliary basis sets are generally more complicated. The molecular two-electron integral matrix is a function of geometry. Consequently, its decomposition varies with molecular geometry. In particular, this means that the composition of the Cholesky auxiliary basis set, i.e. the set of linearly independent product functions singled out by the decomposition, varies with geometry. It is, however, reasonable to assume that the Cholesky auxiliary basis set composition is invariant under infinitesimal changes in molecular geometry, making Eq. (13.44) a good approximation. The test calculations reported by Aquilante, Lindh, and Pedersen [52] have not revealed problems associated with the use of Eq. (13.44). On the contrary, it was found that the accuracy (relative to conventional calculations) of equilibrium structures can be controlled by adjusting the decomposition threshold [52].

13.3.7 Method specific Cholesky decomposition

Although the CD of the two-electron integral matrix has proven to be very useful in the determination of molecular properties of small to medium-sized molecular systems, there are still several limitations. The major drawback of standard CD comes from the fact that integrals actually not needed are nonetheless calculated. In most cases the goal is not the exact determination of the two-electron integral matrix, but the calculation of some expression such as

$$E = \sum_{\alpha\beta\gamma\delta} V_{\alpha\beta}(\alpha\beta | \gamma\delta) V_{\gamma\delta} = \sum_{pq} V_p M_{pq} V_q = \sum_{pq} Z_{pq} \quad (13.46)$$

where a characteristic matrix $\mathbf{Z}$ is implicitly defined. Instead of directly decomposing the matrix $\mathbf{M}$

$$M_{pq} = \sum_J L_p^J L_q^J + \Delta_{pq} \quad (13.47)$$

such that all elements of the residual matrix Δ are smaller than a predetermined threshold τ , it is normally more efficient to decompose the characteristic matrix $\mathbf{Z}$, as in this way the screening introduced by $\mathbf{V}$ is taken into account. Alternatively, we

can understand this decomposition considering a positive semidefinite matrix whose dimension is twice that of $\mathbf{M}$

$$\begin{aligned}
 \begin{pmatrix} V_p M_{pq} V_q & V_p M_{ps} \\ M_{rq} V_q & M_{rs} \end{pmatrix} &= \sum_J \begin{pmatrix} K_p^J \\ L_r^J \end{pmatrix} \begin{pmatrix} K_q^J \\ L_s^J \end{pmatrix}^T \\
 &= \sum_J \begin{pmatrix} V_p \frac{M_{pJ} M_{Jq}}{M_{JJ}} V_q & V_p \frac{M_{pJ} M_{Js}}{M_{JJ}} \\ \frac{M_{rJ} M_{Jq}}{M_{JJ}} V_q & \frac{M_{rJ} M_{Js}}{M_{JJ}} \end{pmatrix} \\
 &= \sum_J \begin{pmatrix} V_p L_p^J L_q^J V_q & V_p L_p^J L_s^J \\ L_r^J L_q^J V_q & L_r^J L_s^J \end{pmatrix} \quad (13.48)
 \end{aligned}$$

The off-diagonal block of the matrix in Eq. (13.48) enters in the expression of the gradient of the energy E

$$G_p = 2 \sum_q M_{pq} V_q = 2 \sum_J L_p^J \sum_q L_q^J V_q + 2 \sum_q \Delta_{pq} V_q \quad (13.49)$$

where the error terms are either zero ($\Delta_{pq} = 0$) or bound by the inequality

$$\begin{aligned}
 |\Delta_{pq} V_p| &= \frac{|V_p \Delta_{pq} V_q|}{|V_p|} \leq \frac{|V_p \Delta_{pp} V_p|^{1/2} |V_q \Delta_{qq} V_q|^{1/2}}{|V_p|} \\
 &= \Delta_{pp}^{1/2} |V_q \Delta_{qq} V_q|^{1/2} \leq \Delta_{pp}^{1/2} T^{1/2} \quad (13.50)
 \end{aligned}$$

Therefore, the threshold of the decomposition also controls the error in the gradient. Moreover, the double dimension matrix illustrates that if we are only interested in the off-diagonal block, we can keep track of the two diagonals in Eq. (13.48) and choose different thresholds for each block.

The usefulness of the previous discussion becomes apparent when taking into account that the functional in Eq. (13.46) has the form of a variety of matrix contractions that appear in electronic structure computations. In particular, it is related to the calculation of the Coulomb and the exchange terms in Hartree-Fock and DFT calculations [40]. We start by what we denominate *Coulomb decomposition*. The calculation of the Coulomb term in conventional direct SCF methods scales as N^4 in the limit of a complete basis and as N^2 in the limit of a large system if the density scales linearly with the system size. In contrast, the standard CD scales as N^3 in both limits (due the scaling of the decomposition itself).

The Coulomb energy is

$$E_C = \sum_{\alpha\beta\gamma\delta} D_{\alpha\beta}(\alpha\beta | \gamma\delta) D_{\gamma\delta} \quad (13.51)$$

and thus, according to our previous discussion, the characteristic matrix is

$$M_{\alpha\beta, \gamma\delta}^C = D_{\alpha\beta}(\alpha\beta | \gamma\delta) D_{\gamma\delta} \quad (13.52)$$

which can be Cholesky decomposed to give

$$M_{\alpha\beta,\gamma\delta}^C = \sum_J L_{\alpha\beta}^J L_{\gamma\delta}^J = \sum_{IJ} D_{\alpha\beta}(\alpha\beta | I) S_{IJ}^{-1} (J | \gamma\delta) D_{\gamma\delta} \quad (13.53)$$

where we have expressed the CD in an inner product [70; 71] form to emphasize its relationship with the RI method, as discussed in sections 13.2 and 13.3.1. The matrix $\mathbf{S}$ has elements $S_{IJ} = (I | J)$ where I and J denote functions of the Cholesky basis. The decomposition might eventually be carried out in a reduced space — say, the atomic orbitals centered on a particular atom — to get a smaller Cholesky basis. Once the Cholesky vectors have been computed, the Coulomb energy is evaluated very easily

$$E_C = \sum_J \left(\sum_{\alpha\beta} L_{\alpha\beta}^J \right)^2 \quad (13.54)$$

The Coulomb Fock operator can be calculated in terms of the auxiliary basis or in terms of the CD of $S = KK^T$

$$F_{\alpha\beta}^C = 2 \sum_{IJ} (\alpha\beta | I) S_{IJ}^{-1} (J | \gamma\delta) D_{\gamma\delta} = 2 \sum_{IJ} (\alpha\beta | I) K_{JI}^{-1} \sum_{\gamma\delta} L_{\gamma\delta}^J \quad (13.55)$$

Due to the screening of high angular momentum functions in SCF and DFT methods, the number of Cholesky vectors required to compute the energy given by Eq. (13.54) becomes constant in the limit of a complete basis and therefore the global scaling is N^2 . The scaling of the Fock operator is also N^2 in the limit of a complete basis, but the scaling can become linear if we only calculate the elements with one occupied index and one general, as required for optimizing the energy. In the limit of a large system, the decomposition of the characteristic Coulomb matrix should scale quadratically in the integral calculation and cubically in the decomposition since the number of Cholesky vectors increases linearly.

The exchange contribution is very difficult to calculate with the RI and CD methods [19; 21] since using the vectors from standard CD of the two-electron integral matrix formally shows quartic scaling. Local density fitting [15; 20; 22] can reduce the scaling but pays the price of losing strict error control. Recently, Aquilante et al. [21] showed that it is possible to contract the density with the Cholesky vectors with only quadratic scaling, although the global procedure still scales as N^3 because of the scaling of the CD itself. As a matter of fact, the savings derived from the use of the standard CD are a consequence of the fact that the decomposition needs to be done only once, while a direct SCF method needs several calculations of the two-electron integrals. Therefore, it is very convenient to apply a specific decomposition-based technique to calculate the exchange. With opposite signs, the exchange energy and the Fock matrix read

$$E_X = \sum_{\alpha\beta\gamma\delta} D_{\alpha\gamma}(\alpha\beta | \gamma\delta) D_{\beta\delta} \quad (13.56)$$

$$F_{\alpha\beta}^X = \sum_{\gamma\delta} (\alpha\gamma | \beta\delta) D_{\gamma\delta} \quad (13.57)$$

The product of densities entering in Eq. (13.56) is an element of the Kronecker product of the density matrix times itself and therefore its rank is the number of kl -pairs, i.e., $O(O+1)/2$, O being the number of occupied orbitals. Therefore, the direct decomposition of a matrix with elements $D_{\alpha\gamma}(\alpha\beta | \gamma\delta)D_{\beta\delta}$ does not imply any saving as the number of kl -pairs easily becomes larger than the numerical rank of the two-electron matrix. We turn our attention, then, to the exchange Fock operator, which — after Cholesky decomposing the one-electron density matrix — is written as

$$F_{\alpha\beta}^X = \sum_{\gamma\delta k} C_{\gamma}^k(\alpha\gamma | \beta\delta) C_{\delta}^k \quad (13.58)$$

where, as previously shown by us [48], the orbitals C^k are localized. The former expression defines a characteristic matrix for the so-called *Exchange-k* algorithm

$$M_{\alpha\beta,\gamma\delta}^k = C_{\beta}^k(\alpha\beta | \gamma\delta) C_{\delta}^k = \sum_J {}^kL_{\alpha\beta}^J {}^kL_{\gamma\delta}^J \quad (13.59)$$

Let us recall that the *Exchange-k* algorithm builds up a localized auxiliary basis for each occupied orbital and therefore O decompositions of this kind are required for each calculation of the Fock operator in the SCF process. However, in many cases the localized orbitals have a significant overlap and, thus, it is more efficient to use a common auxiliary basis obtained from the decomposition of a composed characteristic matrix

$$\Omega = \begin{pmatrix} C_{\beta}^k(\alpha\beta | \gamma\delta) C_{\delta}^k \cdots C_{\beta}^k(\alpha\beta | \kappa\lambda) C_{\lambda}^l \\ \vdots \qquad \qquad \qquad \vdots \\ C_{\nu}^l(\mu\nu | \gamma\delta) C_{\delta}^k \cdots C_{\nu}^l(\mu\nu | \kappa\lambda) C_{\lambda}^l \end{pmatrix} \quad (13.60)$$

Since the orbitals C^k are localized, they make the characteristic matrix local in the limit of a large system and, consequently, the auxiliary basis becomes independent of the system size and the scaling becomes linear. On the other hand, as only two indices are screened, large basis sets will make the pre-factor larger. As the error in the energy is quadratic in the norm of the gradient, i.e. the Fock operator, the decomposition threshold needs not be very small if we are only interested in the energy.

Once the decomposition is carried out, the contribution of each localized orbital to the exchange Fock operator may be evaluated according to

$$F_{\alpha\gamma}^k = \sum_J {}^kH_{\alpha}^J {}^kH_{\gamma}^J = \sum_J \left(\sum_{\beta} {}^kL_{\alpha\beta}^J \right) \left(\sum_{\delta} {}^kL_{\gamma\delta}^J \right) \quad (13.61)$$

where we have implicitly defined the ${}^k\mathbf{H}^J$ vector. In practice, it is sufficient to decompose simply the two-electron integral matrix, but doing the screening on the diagonal elements on the basis of the characteristic matrices above.

As we have seen, in the *Exchange-k* decomposition only two indices are screened. It is also possible to screen the four indices of the two-electron integral while keep-

ing the linear scaling in the computation of the exchange term in the Fock operator, but at the price of increasing the number of decompositions. This is the basic idea of the *Exchange-kl* algorithm, which can be derived from the expression of the exchange energy in Eq. (13.56) after decomposing the density matrices:

$$D_{\alpha\gamma}D_{\beta\delta} = \sum_k C_{\alpha}^k C_{\gamma}^k \sum_l C_{\beta}^l C_{\delta}^l = \sum_{kl} D_{\alpha\beta}^{kl} D_{\gamma\delta}^{kl} \quad (13.62)$$

This resorting of the density elements motivates the new characteristic matrix $M_{\alpha\beta,\gamma\delta}^{kl}$

$$E_X = \sum_{kl} \sum_{\alpha\beta,\gamma\delta} M_{\alpha\beta,\gamma\delta}^{kl} = \sum_{kl} \sum_{\alpha\beta,\gamma\delta} D_{\alpha\beta}^{kl} (\alpha\beta | \gamma\delta) D_{\gamma\delta}^{kl} \quad (13.63)$$

The decomposition of this matrix gives a localized auxiliary basis for each kl -pair of occupied orbitals. Therefore, the linearity with the molecular size of the scaling is kept, but the number of decompositions scales in a quadratic manner. Therefore, the method is in general not appropriate for compact systems with a high number of electrons. Parallel to the *Exchange-k* algorithm, in many cases the overlap among the different orbital pairs makes it more efficient to consider a composed characteristic matrix similar to that in Eq. (13.60). While the exchange energy may be directly evaluated in terms of the Cholesky vectors from the decomposition of $M_{\alpha\beta,\gamma\delta}^{kl}$, the Fock matrix with one occupied index can be calculated by adding all the pair contributions in terms of the auxiliary basis denoted by I and J

$$\begin{aligned} F_{\alpha k}^X &= \sum_l F_{\alpha k}^l = \sum_{\beta\gamma\delta l} C_{\beta}^l (\alpha\beta | \gamma\delta) D_{\gamma\delta}^{kl} \\ &= \sum_{IJ\beta l} C_{\beta}^l (\alpha\beta | I) S_{IJ}^{-1} \sum_{\gamma\delta} (J | \gamma\delta) D_{\gamma\delta}^{kl} \end{aligned} \quad (13.64)$$

The two approaches to compute the exchange discussed so far suffer from the fact that several decompositions are required. To eliminate this disadvantage, but loosing the linear scaling with the system size, it is possible to add the transition exchange densities into a single exchange density matrix

$$D_{\alpha\beta}^X = N \sum_{kl} D_{\alpha\beta}^{kl} = N \sum_{kl} C_{\alpha}^k C_{\beta}^l \quad (13.65)$$

where N is a normalization constant that can be chosen in several ways, which actually correspond to different thresholds in the decomposition. For instance, the normalization with respect to the number of electrons N_e , such that

$$\sum_{\alpha\beta} \langle \phi_{\alpha} | \phi_{\beta} \rangle D_{\alpha\beta}^X = 1 \quad (13.66)$$

is achieved by choosing N as $2/N_e$. This is a convenient normalization constant for compact systems since $(N_e/2)^2$ terms enter the exchange energy, but in the large system limit only $(N_e/2)$ terms enter and, thus, taking N as $\sqrt{2/N_e}$ is more appropriate. It is even possible to take N simply as unity. In the *Exchange(n)* algorithms

(where n is an integer indicating the different normalizations according to $n = 1$: $N = 1$, $n = 2$: $N = \sqrt{2/N_e}$, $n = 3$: $N = 2/N_e$) the characteristic matrix can be found by considering the matrix

$$\Omega = \begin{pmatrix} D_{\alpha\beta}^{kl}(\alpha\beta | \gamma\delta) D_{\gamma\delta}^{kl} & \cdots & D_{\alpha\beta}^{kl}(\alpha\beta | \kappa\lambda) D_{\kappa\lambda}^{k'l'} \\ \vdots & & \vdots \\ D_{\mu\nu}^{k'l'}(\mu\nu | \gamma\delta) D_{\gamma\delta}^{kl} & \cdots & D_{\mu\nu}^{k'l'}(\mu\nu | \kappa\lambda) D_{\kappa\lambda}^{k'l'} \end{pmatrix} \quad (13.67)$$

and we observe that the sum of the diagonal elements is precisely the exchange energy. Accordingly, we suggest to decompose the characteristic matrix

$$M_{\alpha\beta,\gamma\delta} = D_{\alpha\beta}^X(\alpha\beta | \gamma\delta) D_{\gamma\delta}^X, \quad (13.68)$$

to get the auxiliary basis. The scaling of such decomposition is the same as the standard CD, but the pre-factor is significantly smaller.

To illustrate the performance of the method, we present in Tables 13.1 and 13.2 the number of Cholesky vectors, i.e. the number of auxiliary basis functions, required to fit the Coulomb and the Exchange contributions for some model systems in the Hartree-Fock method. For the sake of comparison, recall that preoptimized standard RI auxiliary basis sets normally contain around 2-4 times the number of basis functions.

Table 13.1 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 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} .

Basis	N	Coulomb	Exchange(3)		CD	
		Converged	Hückel	Converged		Hückel
Water						
aug-cc-pVDZ	41	66	45	94	73	414
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 13.2 Number of Cholesky vectors for different decompositions for an alpha helix glycine chain using aug-cc-pVDZ basis set, where N denotes the number of basis functions. The normalization used in the *Exchange(1)* and *Exchange(2)* decompositions are 1 and $\sqrt{2/N_e}$, respectively. The decomposition threshold is 10^{-8} for *Coulomb* and *Exchange(1-2)* decompositions and 10^{-4} for *Exchange-k* decomposition.

Nr. glycines	N	Coulomb	Exchange(1)	Exchange(2)	Exchange-k	CD
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	18730	13195
14	1707	3177	8606	4311	27589	-
20	2421	4545	12601	5994	40804	-
25	3016	5665	15954	7369	51687	-
30	3611	6815	-	8346	62690	-

13.4 Calibration of accuracy

Several sets of benchmark calculations have been performed in order to establish the accuracy of the Cholesky auxiliary basis sets. In these benchmark papers the accuracy has been assessed as a function of the CD threshold, the AO basis set quality, the wave function model, variations of the CD auxiliary basis set generation and the impact of auxiliary basis set pruning. In the first investigation Aquilante et al. [49] compared Full-CD, 1C-CD, and aCD auxiliary basis sets with pre-optimized auxiliary basis sets using the test set of Baker and Chan [92]. Results with respect to total energies and activation energies at the HF, DFT and MP2 level of approximation were analyzed. In a second study the 118 closed shell molecules of the G2/97 test set [93] and a small set of 7 transition metal containing elements of the MLBE21/05 database [94] were used to investigate the accuracy of the CD auxiliary basis sets [61]. In this benchmark, accuracies for total energies and dipole moments were presented in association with the HF, DFT and MP2 methods using several AO basis sets. This study also included assessments of the acCD auxiliary basis sets. Finally, in the third study, Boström et al. [69] assessed the accuracy with respect to the CASSCF/CASPT2 vertical excitation energies of 196 valence states of the test suite by Schreiber et al.[95] and 72 Rydberg states of 3 small organic systems. Below we give a brief summary of the findings.

13.4.1 Accuracy of total energies

The accuracy of total energies are presented in all of the above mentioned benchmark articles. Aquilante et al. [49] compared on-the-fly CD auxiliary basis sets with the pre-optimized RI-J [76] and RI-C [75] auxiliary basis sets, claiming that the for-

mer in contrast to the latter does not show a bias towards a specific quantum chemical method. The additional approximation introduced in the 1C-CD and aCD were also claimed to be negligible. The subsequent benchmark study by Boström and co-workers [61] confirmed these findings. Figure 13.7 shows Full-CD accuracy assess-

Fig. 13.7 Mean absolute errors, μ , in kcal/(mol-electron) of the total energy associated with the Full-CD auxiliary basis set as a function of the Cholesky threshold, τ , in au for different AO basis sets. Each panel shows the result for a specific quantum chemical method.

ments in combination with the HF, MP2, DFT(BLYP) and DFT(B3LYP) models. Similar accuracies were observed independently of the quantum chemical model, supporting the claim that the CD auxiliary basis sets are unbiased. The accuracies were also seen to improve with tighter CD thresholds. For Full-CD the improvement was almost linear in these log-log-plots. A saturation effect could be observed for the 1C-CD and atomic CD methods (shown in Figure 13.8) as the CD threshold was reduced. This was especially noted for AO basis sets of lower quality. It was generally found that the higher the quality of the AO basis set, the better the accuracy of the auxiliary basis set, this is particularly true for the aCD and acCD auxiliary basis sets. It was also observed that there is no significant difference between the aCD and acCD auxiliary basis sets. In general, an error of 0.01 kcal/(mol-electron) was found to be associated with a CD threshold of 10^{-4} au. Studies on a few transition metal complexes and additional properties as the dipole moments, and results presented for total CASPT2-energies of ground and excited states further support these findings.

13.4.2 Accuracy of vertical transition energies

In the study by Aquilante *et al.* [49], the accuracy of activation energies were reported. A favorable cancellation of errors was observed in the computed activation energies corresponding to a reduction by a factor of 2-3 as compared to total energies. In a subsequent benchmark study [69] on CASPT2 vertical excitation energies,

Fig. 13.8 Mean absolute errors, μ , in kcal/(mol-electron) of the total energy associated with the HF wave function model vs the Cholesky threshold, τ , in au for different AO basis sets. Each panel shows the result for a specific CD auxiliary basis set.

optimal with respect to error cancellation, a reduction of the error of roughly one order of magnitude was observed (see Figure 13.9). Just as for total energies a depen-

Fig. 13.9 Mean absolute errors, μ , in eV of the vertical transition energies for the ANO-RCC-VTZP AO basis set at the CASSCF (left panel) and CASPT2 (center panel) level of theory and in total CASPT2 energies (right panel) as a function of the Cholesky threshold, τ , in au, the CD auxiliary basis set, and auxiliary basis set pruning.

dence upon the AO basis set is present, favoring CD auxiliary basis sets generated in connection with higher quality AO basis sets. In Figure 13.9 it is demonstrated that the errors at the CASSCF and CASPT2 levels of theory for all practical purposes are similar. Just as for the G2/97 test it was observed that the Full-CD error decays almost exponentially whereas the gains in moving to smaller thresholds are less pronounced for the acCD and aCD auxiliary basis set. It was also noticed that the difference between the aCD and acCD approximations is insignificant and that the auxiliary basis set pruning gives an acceptable additional error. Computing excitations to Rydberg states can be challenging and special primitive functions are

often added to the AO basis sets to correctly describe their diffuse character. To accommodate this, pre-optimized auxiliary basis sets are normally augmented with primitive auxiliary basis functions to maintain the accuracy established for valence excited states. In the CD auxiliary basis set, however, no special care is needed as long as the parent AO basis set used to generate it is properly augmented for the purpose. The trend observed for valence states was shown to hold for Rydberg states as well.

Qualitative results from benchmarking vertical transition energies showed that a CD threshold of 10^{-3} yields average errors smaller than 0.01 eV in all but some of the Full-CD cases and a CD threshold of 10^{-4} , as recommended for absolute energies above, will lower this to around 0.001 eV and even better in some cases.

13.4.3 Auxiliary basis set pruning

In the study by Boström and co-workers[61] it was noted that the accuracy associated with the atomic CD auxiliary basis sets was better than necessary and that additional reductions of the basis set could be possible without too severe reduction in the accuracy. A method to prune the auxiliary basis set by skipping some higher angular momentum functions (denoted SHAC), a concept originally introduced by Eichkorn et al. [74], has been presented together with the loss of accuracy in utilizing SHAC on a small subset of 24 G2/97 molecules. The results using pruned auxiliary basis sets were also presented alongside the full aCD and acCD results in the CASSCF/CASPT2 benchmark paper. We note that the method of pruning in the study of Boström and co-workers is different from and slightly more sophisticated than the one used in the study by Aquilante et al. [49] In the latter the pruning is done after the atomic CD, while in the former the pruning it is done before the atomic CD.

A rather uniform loss of about one order of magnitude in accuracy (see Figure 13.10) was observed while using the pruned auxiliary basis set to calculate relative energies. This can generally be afforded while still maintaining accuracies almost as good as, or in some cases better, than with Full-CD. For total energies the trend is not as clear, but pruning might still be a viable alternative, at least for high quality AO basis sets.

13.5 Implementational aspects

Cholesky decomposition of a positive semi-definite matrix $\mathbf{M}$ is often based on the assumption that the matrix can be stored in memory. This is a reasonable assumption for a number of matrices such as the atomic orbital overlap matrix for which the number of nonzero elements scales linearly with system size (number of atoms). In such cases, the Cholesky decomposition can be performed in linear-

Fig. 13.10 Mean absolute errors, μ , of the total MP2 energy in kcal/(mol-electron) evaluated with the aid of an aCD auxiliary basis set as a function of the Cholesky threshold, τ , in au, the AO basis set and auxiliary basis set pruning. The dashed lines represent results using auxiliary basis set pruning.

scaling time [82]. In many cases, however, the matrix can not be stored in memory. For example, the entire molecular two-electron integral matrix can be stored in core only for the smallest basis sets and molecules. In addition, the semi-definite nature implies that only a fraction of the matrix columns is needed to decompose the matrix to a prescribed accuracy specified by the decomposition threshold. An out-of-core (matrix-direct) algorithm is clearly needed. Using τ to denote the decomposition threshold, an out-of-core algorithm for the decomposition of a positive semi-definite matrix $\mathbf{M}$ can be summarized as follows:

1. Get diagonal elements of the matrix:

$$D_p = M_{pp}$$

$$D_{\max}^{(1)} = \max_p D_p$$

2. Compute reduced set of significant diagonal elements:

$$\mathcal{L}^{(1)} = \left\{ p \mid d \sqrt{D_{\max}^{(1)} D_p} > \tau, d \geq 1 \right\}$$

3. Initialize vector counter: $\mathcal{N} = 0$.

4. Initialize counter: $i = 0$

5. While $D_{\max}^{(i+1)} > \tau$:

- a. Update counter: $i = i + 1$
- b. Compute smallest diagonal that may be treated:

$$D_{\min}^{(i)} = \max \left(s D_{\max}^{(i)}, \tau \right), \quad s \leq 1$$

- c. Compute set of qualified diagonals:

$$\mathcal{Q}^{(i)} = \left\{ q \mid q \in \mathcal{L}^{(i)}, D_q > D_{\min}^{(i)} \right\}$$

- d. Get matrix columns corresponding to qualified diagonals:

$$M_{pq}, \quad p \in \mathcal{L}^{(i)}, q \in \mathcal{Q}^{(i)}$$

- e. Subtract contributions from previous vectors:

$$\Delta_{pq} = M_{pq} - \sum_{J=1}^{\mathcal{N}} L_p^J L_q^J, \quad p \in \mathcal{L}^{(i)}, q \in \mathcal{Q}^{(i)}$$

- f. Compute largest diagonal among qualified:

$$Q_{\max} = \max_{q \in \mathcal{Q}^{(i)}} D_q^{(i)}$$

- g. Initialize counter: $j = 0$

- h. While $j < \dim(\mathcal{Q}^{(i)})$ and $Q_{\max} > D_{\min}^{(i)}$:

- i. Update counters: $j = j + 1, J = \mathcal{N} + j$
- ii. Assign Cholesky basis function h_J to $[q]_J$, the index corresponding to $Q_{\max}$
- iii. Calculate Cholesky vector:

$$L_p^J = Q_{\max}^{-1/2} \Delta_{p,[q]_J}, \quad p \in \mathcal{L}^{(i)}$$

- iv. Update:

$$\begin{aligned} \Delta_{pq} &\leftarrow \Delta_{pq} - L_p^J L_q^J, \quad p \in \mathcal{L}^{(i)}, q \in \mathcal{Q}^{(i)} \\ D_p &\leftarrow D_p - (L_p^J)^2, \quad p \in \mathcal{L}^{(i)} \\ Q_{\max} &= \max_{q \in \mathcal{Q}^{(i)}} D_q \end{aligned}$$

- i. Update vector counter: $\mathcal{N} = \mathcal{N} + j$

- j. Compute largest diagonal:

$$D_{\max}^{(i+1)} = \max_{p \in \mathcal{L}^{(i)}} D_p$$

- k. Compute reduced set of significant diagonal elements:

$$\mathcal{L}^{(i+1)} = \left\{ p \mid d \sqrt{D_{\max}^{(i+1)} D_p} > \tau, d \geq 1 \right\}$$

Due to the subtraction step (point 5e), the computational effort is cubic with respect to the dimension of the reduced set (point 2). This implies that the decomposition of the entire molecular two-electron integral matrix scales cubically with the size of the system. As discussed above, the decomposition of the characteris-

tic matrices appearing in method-specific Cholesky decomposition can be done in quadratic or linear time, depending on the characteristic matrix.

In addition to the decomposition threshold τ , two input parameters are used in this algorithm: the screening damping $d \geq 1$ (points 2 and 5k) and the span factor $s \leq 1$ (point 5b). The screening damping ensures that matrix elements with values below the decomposition threshold have a nonzero Cholesky representation. For decomposition of two-electron integrals, we choose the damping according to $d \approx 10^9 \tau$ for thresholds above 10^{-8} and $d = 1.0$ for lower thresholds. The span factor modifies the pivoting by allowing diagonals to be treated even if the value of the diagonal element is not the largest. It has been introduced in order to minimize the overhead that may be involved in fetching the columns (point 5d) as well as to minimize the I/O operation necessary for the subtraction (point 5e) in cases where the Cholesky vectors can not be stored in core. This is normally the case when decomposing two-electron integrals. For two-electron integrals we choose the span factor as $s = 0.01$, i.e. diagonals that are at most 100 times smaller than the largest diagonal may be treated. Since two-electron integrals in the atomic orbital basis are usually computed in shell quadruples, i.e. a subblock of the integral matrix where each of the four atomic orbital indices belongs to a given atomic shell, the span factor allows us to make use of more computed integrals than just the column corresponding to the largest diagonal. This, in turn, minimizes the potentially huge number of integral recalculations that would result from strict pivoting. In order to keep the dimension of the set of qualified diagonals (point 5c) at a reasonable level, it is further restricted to contain at most 100 diagonals. For two-electron integral decompositions, these are chosen from the shell pairs containing the largest diagonals.

The decomposition time is dominated by the subtraction step (point 5e) in most cases, although the fetching of integral columns may be equally time-consuming. The latter is the case, e.g., for two-electron integral evaluations involving basis sets with a very large number of primitive Gaussian functions. The main bottleneck of the subtraction step is reading the previous vectors from disk, which is needed when available memory is insufficient to store the vectors in core. Two approaches have been implemented to circumvent the I/O bottleneck: a distributed parallel algorithm and a two-step algorithm.

In a parallel execution, the elements of the the reduced sets (points 2 and 5k) are distributed among the processes. Each process fetches the corresponding rows of the matrix (point 5d) and performs the subtraction of previous Cholesky vectors (point 5e). This requires that the Cholesky vector elements corresponding to the set of qualified diagonals (L_q^j in point 5e) are broadcast from the process holding them. This design ensures that the memory requirement per process is minimized and that Cholesky vector I/O can be avoided by increasing the number of computational nodes (thus increasing available memory).

The two-step algorithm, as the name indicates, proceeds in two steps. In the first step, the algorithm outlined above is used to obtain the Cholesky basis (assuming that the matrix is a finite basis set representation of an operator) such that

$$M_{IJ} = \sum_K L_I^K L_J^K$$

Note that this is exact regardless of the chosen decomposition threshold. The only modification of the algorithm is that reduced sets (points 2 and 5k) are computed according to

$$\mathcal{L}^{(i)} = \left\{ p \mid D_p > \tau \right\}$$

I.e., only those diagonals that may give rise to Cholesky vectors are retained. In the second step, complete Cholesky vectors are computed according to

$$L_p^J = (L_J^J)^{-1/2} \left[M_{pJ} - \sum_{K=1}^{J-1} L_p^K L_J^K \right]$$

where M_{pJ} denotes element p of the column corresponding to Cholesky vector J . By driving the calculation of full Cholesky vectors with a loop over (blocks of) row indices p , I/O is avoided if the vector components corresponding to the Cholesky basis, L_J^K , can be stored in core. If they can not be stored in core, the I/O bottleneck is not removed but significantly reduced. The two-step algorithm can fairly easily be parallelized. The parallel algorithm discussed above can be utilized for the first step, while the loop over row indices in the second step can be parallelized.

In order to illustrate the performance of these algorithms, as implemented in the MOLCAS program package [44], we consider the Cholesky decomposition of the molecular two-electron integral matrix for the three units of 3-butylthiophene shown in Figure 13.11. The atomic orbital basis sets cc-pVTZ and cc-pVQZ correspond to a total of 1270 and 2457 basis functions, respectively, for this system. The two-electron integral matrix thus contains approximately 326×10^9 and 4559×10^9 elements. Using a single processor on a 2.66 GHz Intel Xeon X5355 Quad Core with 1900 Mb memory, the Cholesky decomposition of the cc-pVTZ integral matrix with a threshold of 10^{-4} au requires 10.5 h wall time. The Cholesky vectors require 12.4 Gb storage and the CPU time of 1.8 h clearly indicates the I/O bottleneck alluded to above. Using the two-step algorithm, the total decomposition time reduces to 1.2 h wall time and 0.8 h CPU time. Though not entirely removed, the I/O bottleneck is significantly reduced. Using the parallel one-step algorithm for the cc-pVQZ basis set with the same decomposition threshold (10^{-4} au), the speedup (in terms of wall time) is given in Figure 13.12. The speedup is measured for 8, 16, 32, 64, 128, and 256 processors relative to the calculation on 4 processors. All processors are of the above mentioned type connected with Infiniband and 1900 Mb memory is used on each processor. The Cholesky vectors require 73.5 Gb storage. The calculation on 4 processors requires 70.2 h wall time/32.1 h CPU time, whereas the calculation on 256 processors requires 1.1 h wall time/1.0 h CPU time. These timings show that a major advantage of parallel execution is the reduction of I/O. Running the same decomposition with the serial two-step algorithm takes 29.8 h wall time/9.3 h CPU time with 1900 Mb memory and 8.3 h wall time/6.4 h CPU time with 15200 Mb memory. Comparing to the timings of the parallel one-step algorithm, it is clear that optimal performance would be achieved by parallelizing the two-step algorithm. This is an on-going project.

Fig. 13.11 System consisting of 3 units of 3-butylthiophene

Finally, it is worth mentioning that the first step in the two-step algorithm can be regarded as computation of the Cholesky basis to be used with the density fitting approach. The timings for this step are therefore of interest. With the cc-pVTZ basis set, generation of the Cholesky basis requires 9.5 min wall time/7.1 min CPU time, whereas the cc-pVQZ basis set requires 4.7 h wall time/1.3 h CPU time with 1900 Mb memory and 1.2 h wall time/1.1 h CPU time with 15200 Mb memory on a single processor.

The calculation of integrals in the CD procedure is in practice not different from that of an ordinary two-electron integral driver. The only significant difference is in the respect that the CD algorithm is expressed in terms of specific basis functions,

Fig. 13.12 Parallel speedup for the cc-pVQZ decomposition

whereas the most efficient two-electron integrals implementations and algorithms derive their efficiency in not treating the basis functions one at the time but rather to compound all basis functions of a center and with the same total angular momentum into a single entity – a shell. Hence, efficient modern computer codes compile integrals in batches of shell quadruples. This is in conflict with the CD procedure and would, if not handled or considered, make the program compute a significant number of integrals in vain. The effect is of particular significance for generally contracted basis functions (so-called ANO basis set) as compared to the segmented basis sets. The so-called span factor has been introduced to reduce this loss and is of particular importance when dealing with basis sets of the ANO type. In common to a traditional two-electron code a CD implementation requires the standard set of pre-computed intermediates to be computed. After that the CD procedure requires the two electron integral code to be able to compute a shell quadruple on demand and in any order. This is ensured with some minor modifications to any existing integral code to introduce the required modularity of the code.

13.6 Outlook and perspectives

As we have demonstrated in the previous sections, CD is a valuable tool to control the accuracy and associated computational effort with only a single threshold parameter. The advantage using a decomposed form of the two-electron integrals is not apparent in all cases. For instance, using this integral representation in coupled cluster models does not lead directly to lower scaling expressions.

An essential solution to this problem is to perform method specific decomposition (Section 13.3.7), such that the auxiliary basis sets are tailor-made for a given model or even a particular term. As was shown in the previous section, the actual construction of the basis is the smallest part of the calculation. The scaling of the subsequent contractions will benefit from the reduced size as compared to any pre-optimized procedure. In some cases the basis becomes practically constant (e.g. for the Coulomb energy) and these reductions are almost impossible to obtain using pre-optimized auxiliary basis sets.

Another solution is to exploit the inherent locality of the Cholesky basis (Section 13.3.2) to design a trivially linear-scaling density fitting algorithm which retains the unbiased (i.e., non-method specific) nature and complete error control of full CD [58; 62]. Compared to method specific CD, the local approach will require more auxiliary basis functions (since it is unbiased) but it will be the same auxiliary functions that are used for all contributions (Coulomb, exchange, dynamical and static correlation).

In the limit of large basis, CD is clearly favored as one may save the Cholesky vectors in compact form and in this way save expensive recalculation of two-electron integrals. One should remember that, in the complete basis set limit, the number of two-electron integrals scales as N^4 and to treat 10,000 atomic orbitals will result in roughly 10,000 TByte of two-electron integrals to be processed. This calls for even more compact treatment of the two-electron integrals beyond the $3N^3$ to $8N^3$ scaling attainable by the full CD. The ultimate goal is an N^2 scaling in the limit of a complete basis, and we have already made significant progress in this direction [96].

The development of correlated methods such as coupled cluster and CASPT2 for energy, gradient and response properties will be needed in order to take full advantage of the developments presented in this chapter. Subsystem based methods are becoming very important in quantum chemistry as the need for accurate electronic structure methods for larger systems is increasing. Techniques based on CD will play a central role in these developments.

13.7 Summary and conclusions

In this review we have presented the original scheme behind Cholesky decomposition [23] of positive semidefinite matrices, which can be considered a special type of Gram-Schmidt orthonormalization or as a special case of LU factorization. In particular, a point is made that the iterative procedure, controlled with a single parameter, completely removes linear dependencies by eliminating zero or near-zero eigenvectors in a controlled manner. Hence, the CD procedure can be considered an effective way to represent large matrices in a compact manner. The CD procedure has in the past been considered a purely numerical procedure used mainly, if not exclusively, for solving positive definite linear equations. In 1977, however, the

CD procedure was introduced as a tool for simplifying the handling of two-electron integrals in quantum chemistry by Beebe and Linderberg [24].

For a number of cases, we demonstrate how the CD procedure may be applied to the methods of electronic structure theory. We show that the CD procedure is a special type of resolution-of-the identity or density-fitting scheme. This connection is of particular importance, since it provides an analytic view of CD. From this point of view, it becomes possible to derive, with some limitations, analytic expression for gradients and higher-order derivatives with respect to nuclear geometry. We furthermore demonstrate that the CD procedure forms a reliable and accurate tool for constructing one-center auxiliary basis sets to be used in the RI/DF formalism. The CD procedure can also be applied to the one-particle occupied density and the virtual pseudo-density matrix to obtain localized MOs (Cholesky MOs) in an effective non-iterative fashion. The Cholesky MOs are instrumental for an effective and fast computation of the exchange contribution to the Fock matrix, the LK algorithm. Special versions of the CD MO localization scheme allow the generation of active orbitals to be used in correlation treatments. The use of these orbitals has demonstrated considerable reduction of the computational effort in the evaluation of, for example, intermolecular interaction energies at the MP2 and CCSD(T) levels of theory. Yet another use of the CD procedure is demonstrated in quartic-scaling MP2. Here the CD procedure is applied directly to the MP2 amplitude matrix. Combined with efficient prescreening, this yields a canonical MP2 approach scaling as N^4 . Finally, we provide yet another case in which the CD procedure is applied to give dramatic reduction as compared to a conventional implementation. In method specific CD, the two-electron integrals are not treated alone but are combined with density matrices such that the resulting Cholesky basis is tailored to accurately represent the contribution (Coulomb or exchange) of interest. This leads to what we have coined *Coulomb decomposition* and three different *Exchange decompositions*. It is demonstrated that the method specific CD procedure leads to dramatic savings as compared to the standard CD procedure applied to the two-electron integrals.

We present results from calibration calculations in which the statistical accuracy of the CD procedure is studied. These assessments are investigated for total energies for a range of wave function models and density functional theory, as well as for vertical excitation energies at the CASSCF and CASPT2 levels of theory. The benchmark calculations were performed for various sets of valence and CD auxiliary basis sets. Accuracies for total energies associated with CD thresholds of 10^{-4} au were found to be of 0.01 kcal/(mol · electron) and for excitation energies the corresponding accuracy was found to be 0.001 eV on average. Furthermore, these studies verify that CD (full as well as atomic) is an unbiased, highly accurate alternative to pre-optimized auxiliary basis sets, albeit at the price of a slightly larger number of auxiliary basis functions. For Full-CD basis sets the Cholesky threshold gives complete control of the accuracy. A good compromise of speed and accuracy is obtained using a threshold of 10^{-4} in connection with absolute energies. For relative energies a reduction in error due to favorable cancellation of errors have been observed and for vertical excitation energies this might even allow for a threshold of 10^{-3} to be used in most cases. For high quality basis sets, pruning gives a significant

reduction in auxiliary basis set size with an affordable loss of accuracy. Moreover, the special type of atomic CD auxiliary basis sets, the so-called atomic compact CD (acCD), comes with significant reduction of the number of primitive functions in the auxiliary basis set at virtually no loss of accuracy.

Towards the end of this review, details of the implementational aspects of the CD method are presented. Techniques are presented which take into consideration the fact that the computer representation of numbers is not infinitely accurate. In addition, the fact that effective two-electron integral computation is done in shell-quadruples — a procedure which is not perfect for a straightforward CD implementation — is considered and compromises are suggested. Finally, a scheme of efficient implementation of a parallel CD procedure is presented together with benchmark results which demonstrate encouraging speed-up with up to 256 CPUs.

Finally an outlook and perspective of the CD method in *ab initio* quantum chemistry is presented. The authors take the view that the CD procedure offers several advantages in reducing the scaling of well-known quantum chemical methods. To fully exploit CD techniques, however, further work has to be done with respect to implementation.

To conclude, it is our opinion that the CD technique is a much overlooked method that holds a great potential in the field of quantum chemistry. We hope that the current flux of papers on CD technique has changed this general attitude — to our knowledge to this date (February 2010) there has been published in total 22 papers with respect to the CD technique in quantum chemistry, 16 of these over the last three years. We are convinced that the CD procedure has much more to offer if we look more carefully for opportunities.

Acknowledgements ASM acknowledges financial support from Spanish FEDER+MEC project CTQ2007-67143-C02-01/BQU. RL and JB would like to thank the Swedish Research Council directly and through the Linnaeus Center of Excellence on Organizing Molecular Matter at Lund University, Sweden, for financial support. TBP acknowledges financial support from the CoE Centre for Theoretical and Computational Chemistry (179568/V30). FA would like to acknowledge funding from the Swiss National Science Foundation (SNF), Grant No. 200020-120007. HK and LB would like to thank the Norwegian Research Council for support (Grant No. 154011/420). The authors thank Eirik Hjertenæs at the Norwegian University of Science and Technology for proof-reading the manuscript.

References

- [1] J.L. Whitten, J. Chem. Phys. **58**, 4496 (1973)
- [2] E.J. Baerends, D.E. Ellis, P. Ros, Chem. Phys. **2**, 41 (1973)
- [3] J.A. Jafri, J.L. Whitten, J. Chem. Phys. **61**, 2116 (1974)
- [4] H. Sambe, R.H. Felton, J. Chem. Phys. **62**, 1122 (1975)
- [5] B.I. Dunlap, J.W.D. Connolly, J.R. Sabin, J. Chem. Phys. **71**, 3396 (1979)
- [6] B.I. Dunlap, J.W.D. Connolly, J.R. Sabin, J. Chem. Phys. **71**, 4993 (1979)

- [7] M. Feyereisen, G. Fitzgerald, A. Komornicki, *Chem. Phys. Lett.* **208**, 359 (1993)
- [8] O. Vahtras, J. Almlöf, M. Feyereisen, *Chem. Phys. Lett.* **213**, 514 (1993)
- [9] R.T. Gallant, A. St-Amant, *Chem. Phys. Lett.* **256**, 569 (1996)
- [10] C. Fonseca Guerra, J.G. Snijders, G. te Velde, E.J. Baerends, *Theor. Chem. Acc.* **99**, 391 (1998)
- [11] B.I. Dunlap, *J. Mol. Struct. Theochem* **501-502**, 221 (2000)
- [12] M. Sierka, A. Hogekamp, R. Ahlrichs, *J. Chem. Phys.* **118**, 9136 (2003)
- [13] A. Sodt, J.E. Subotnik, M. Head-Gordon, *J. Chem. Phys.* **125**, 194109 (2006)
- [14] P. Sałek, S. Høst, L. Thøgersen, P. Jørgensen, P. Manninen, J. Olsen, B. Jansík, S. Reine, F. Pawłowski, E. Tellgren, T. Helgaker, S. Coriani, *J. Chem. Phys.* **126**, Art.Nr. 114110 (2007)
- [15] S. Reine, E. Tellgren, A. Krapp, T. Kjærgaard, T. Helgaker, B. Jansík, S. Høst, P. Sałek, *J. Chem. Phys.* **129**, Art.Nr. 104101 (2008)
- [16] A.M. Köster, J.M. del Campo, F. Janetzko, B. Zuniga-Gutierrez, *J. Chem. Phys.* **130**, Art. Nr. 114106 (2009)
- [17] V.D. Domínguez-Soria, G. Geudtner, J.L. Morales, P. Calaminici, A.M. Köster, *J. Chem. Phys.* **131**, Art.Nr. 124102 (2009)
- [18] A.M. Burow, M. Sierka, F. Mohamed, *J. Chem. Phys.* **131**, Art.Nr. 214101 (2009)
- [19] F. Weigend, *Phys. Chem. Chem. Phys.* **4**, 4285 (2002)
- [20] R. Polly, H.J. Werner, F.R. Manby, P. Knowles, *Mol. Phys.* **102**, 2311 (2004)
- [21] F. Aquilante, T.B. Pedersen, R. Lindh, *J. Chem. Phys.* **126**, Art.Nr. 194106 (2007)
- [22] A. Sodt, M. Head-Gordon, *J. Chem. Phys.* **128**, Art.Nr. 104106 (2008)
- [23] Cdt. Benoît, *Bulletin Géodésique* **2**, 67 (1924)
- [24] N.H.F. Beebe, J. Linderberg, *Int. J. Quantum Chem.* **12**, 683 (1977)
- [25] I. Røeggen, E. Wisløff-Nilssen, *Chem. Phys. Lett.* **132**, 154 (1986)
- [26] I. Røeggen, L. Veseth, *Int. J. Quantum Chem.* **101**, 201 (2005)
- [27] I. Røeggen, *J. Chem. Phys.* **124**, Art. Nr.184502 (2006)
- [28] I. Røeggen, *Theor. Chem. Acc.* **116**, 683 (2006)
- [29] I. Røeggen, *J. Chem. Phys.* **126**, Art.Nr. 204303 (2007)
- [30] I. Røeggen, T. Johansen, *J. Chem. Phys.* **128**, Art.Nr. 194107 (2008)
- [31] D.W. O'Neal, J. Simons, *Int. J. Quantum Chem.* **36**, 673 (1989)
- [32] H. Koch, A. Sánchez de Merás, T.B. Pedersen, *J. Chem. Phys.* **118**, 9481 (2003)
- [33] T.B. Pedersen, A. Sánchez de Merás, H. Koch, *J. Chem. Phys.* **120**, 8887 (2004)
- [34] O. Christiansen, H. Koch, P. Jørgensen, *Chem. Phys. Lett.* **243**, 409 (1995)
- [35] I. García Cuesta, T.B. Pedersen, H. Koch, A.M.J. Sánchez de Merás, *Chem. Phys. Lett.* **390**, 170 (2004)
- [36] T.B. Pedersen, H. Koch, L. Boman, A.M.J. Sánchez de Merás, *Chem. Phys. Lett.* **393**, 319 (2004)
- [37] I. García Cuesta, T.B. Pedersen, H. Koch, A. Sánchez de Merás, *ChemPhysChem* **7**, 2503 (2006)

- [38] B. Fernández, T.B. Pedersen, A. Sánchez de Merás, H. Koch, *Chem. Phys. Lett.* **441**, 332 (2007)
- [39] I. García Cuesta, J. Sánchez Marín, T.B. Pedersen, H. Koch, A. Sánchez de Merás, *Phys. Chem. Chem. Phys.* **10**, 361 (2008)
- [40] L. Boman, H. Koch, A. Sánchez de Merás, *J. Chem. Phys.* **129**, Art.Nr. 134107 (2008)
- [41] T.B. Pedersen, J. Kongsted, T.D. Crawford, K. Ruud, *J. Chem. Phys.* **130**, Art.Nr. 034310 (2009)
- [42] T.B. Pedersen, J. Kongsted, T.D. Crawford, *Chirality* **21**, S68 (2009)
- [43] A. Sánchez de Merás, H. Koch, I. García Cuesta, L. Boman, *J. Chem. Phys.* **132**, Art.Nr. 204105 (2010)
- [44] F. Aquilante, L. De Vico, N. Ferré, G. Ghigo, P.Å. Malmqvist, P. Neogrády, T.B. Pedersen, M. Pitoňák, M. Reiher, B.O. Roos, L. Serrano-Andrés, M. Urban, V. Veryazov, R. Lindh, *J. Comput. Chem.* **31**, 224 (2010)
- [45] F. Aquilante, T.B. Pedersen, B.O. Roos, A. Sánchez de Merás, H. Koch, *J. Chem. Phys.* **129**, Art.Nr. 24113 (2008)
- [46] F. Aquilante, P.Å. Malmqvist, T.B. Pedersen, A. Ghosh, B.O. Roos, *J. Chem. Theory Comput.* **4**, 694 (2008)
- [47] F. Aquilante, T.B. Pedersen, *Chem. Phys. Lett.* **449**, 354 (2007)
- [48] F. Aquilante, T.B. Pedersen, A. Sánchez de Merás, H. Koch, *J. Chem. Phys.* **125**, Art.Nr. 174101 (2006)
- [49] F. Aquilante, R. Lindh, T.B. Pedersen, *J. Chem. Phys.* **127**, Art.Nr. 114107 (2007)
- [50] A. Öhrn, F. Aquilante, *Phys. Chem. Chem. Phys.* **9**, 470 (2007)
- [51] B.O. Roos, V. Veryazov, J. Conradie, P.R. Taylor, A. Ghosh, *J. Phys. Chem. B* **112**, 14099 (2008)
- [52] F. Aquilante, R. Lindh, T.B. Pedersen, *J. Chem. Phys.* **129**, Art.Nr. 034106 (2008)
- [53] K. Pierloot, S. Vancoillie, *J. Chem. Phys.* **128**, Art.Nr. 034104 (2008)
- [54] M. Radoń, K. Pierloot, *J. Phys. Chem. A* **112**, 11824 (2008)
- [55] F. Bozoglian, S. Romain, M.Z. Ertem, T.K. Todorova, C. Sens, J. Mola, R. Montserrat, I. Romero, J. Benet-Buchholz, X. Fontrodona, C.J. Cramer, L. Gagliardi, A. Llobet, *J. Am. Chem. Soc.* **131**, 15176 (2009)
- [56] P. Söderhjelm, F. Aquilante, U. Ryde, *J. Phys. Chem. B* **113**, 11085 (2009)
- [57] M. Srnec, F. Aquilante, U. Ryde, L. Rulíšek, *J. Phys. Chem. B* **113**, 6074 (2009)
- [58] T.B. Pedersen, F. Aquilante, R. Lindh, *Theor. Chem. Acc.* **124**, 1 (2009)
- [59] S.M. Huber, A.R.M. Shahi, F. Aquilante, C.J. Cramer, L. Gagliardi, *J. Chem. Theory Comput.* **5**, 2967 (2009)
- [60] S.M. Huber, M.Z. Ertem, F. Aquilante, L. Gagliardi, W.B. Tolman, C.J. Cramer, *Chem. Eur. J.* **15**, 4886 (2009)
- [61] J. Boström, F. Aquilante, T.B. Pedersen, R. Lindh, *J. Chem. Theory Comput.* **5**, 1545 (2009)
- [62] F. Aquilante, L. Gagliardi, T.B. Pedersen, R. Lindh, *J. Chem. Phys.* **130**, Art.Nr. 154107 (2009)

- [63] K.E. Riley, M. Pitoňák, J. Černý, P. Hobza, *J. Chem. Theory Comput.* **6**, 66 (2010)
- [64] M. Pitoňák, A. Heßelmann, *J. Chem. Theory Comput.* **6**, 168 (2010)
- [65] M. Pitoňák, P. Neogrády, P. Hobza, *Phys. Chem. Chem. Phys.* **12**, 1369 (2010)
- [66] V.P. Vysotskiy, L.S. Cederbaum, *J. Chem. Phys.* **132**, Art. Nr.044110 (2010)
- [67] T.S. Chwee, E.A. Carter, *J. Chem. Phys.* **132**, 074104 (2010)
- [68] Z. Tabookht, X. López, C. de Graaf, *J. Phys. Chem. A* **114**, 2028 (2010)
- [69] J. Boström, M.G. Delcey, F. Aquilante, L. Serrano-Andrés, T.B. Pedersen, R. Lindh, *J. Chem. Theory Comput.* **6**, 747 (2010)
- [70] P.O. Löwdin, *J. Chem. Phys.* **43**, S175 (1965)
- [71] P.O. Löwdin, *Int. J. Quantum Chem.* **S4**, 231 (1970)
- [72] G.H. Golub, C.F. Van Loan, *Matrix Computations*, 3rd edn. (Johns Hopkins University Press, 1996)
- [73] H. Koch, A. Sánchez de Merás, *J. Chem. Phys.* **113**, 508 (2000)
- [74] K. Eichkorn, O. Treutler, H. Öhm, M. Häser, R. Ahlrichs, *Chem. Phys. Lett.* **240**, 283 (1995)
- [75] F. Weigend, A. Köhn, C. Hättig, *J. Chem. Phys.* **116**, 3175 (2002)
- [76] F. Weigend, *Phys. Chem. Chem. Phys.* **8**, 1057 (2006)
- [77] Y. Jung, A. Sodt, P.M.W. Gill, M. Head-Gordon, *Proc. Natl. Acad. Sci. USA* **102**, 6692 (2005)
- [78] H.J. Werner, F.R. Manby, *J. Chem. Phys.* **124**, Art.Nr. 054114 (2006)
- [79] S.F. Boys, *Rev. Mod. Phys.* **32**, 296 (1960)
- [80] C. Edmiston, K. Ruedenberg, *Rev. Mod. Phys.* **35**, 457 (1963)
- [81] J. Pipek, P.G. Mezey, *J. Chem. Phys.* **90**, 4916 (1989)
- [82] S. Schweizer, J. Kussmann, B. Doser, C. Ochsenfeld, *J. Comput. Chem.* **29**, 1004 (2008)
- [83] M. Ziolkowski, B. Jansík, P. Jørgensen, J. Olsen, *J. Chem. Phys.* **131**, Art.Nr. 124112 (2009)
- [84] T.H. Dunning Jr., *J. Chem. Phys.* **90**, 1007 (1989)
- [85] B. Fernández, H. Koch, J. Makarewicz, *J. Chem. Phys.* **111**, 5922 (1999)
- [86] G.E. Scuseria, P.Y. Ayala, *J. Chem. Phys.* **111**, 8330 (1999)
- [87] F. Weigend, M. Häser, *Theor. Chem. Acc.* **97**, 331 (1997)
- [88] C. Hättig, *J. Chem. Phys.* **118**, 7751 (2003)
- [89] A. Köhn, C. Hättig, *J. Chem. Phys.* **119**, 5021 (2003)
- [90] P. Deglmann, K. May, F. Furche, R. Ahlrichs, *Chem. Phys. Lett.* **384**, 103 (2004)
- [91] D. Rappoport, F. Furche, *J. Chem. Phys.* **122**, Art.Nr. 064105 (2005)
- [92] J. Baker, F. Chan, *J. Comput. Chem.* **17**, 888 (1996)
- [93] L.A. Curtiss, K. Raghavachari, P.C. Redfern, J.A. Pople, *J. Chem. Phys.* **106**, 1063 (1997)
- [94] N.E. Schultz, Y. Zhao, D.G. Truhlar, *J. Phys. Chem. A* **109**, 11127 (2005)
- [95] M. Schreiber, M. Silva-Junior, S.P.A. Sauer, W. Thiel, *J. Chem. Phys.* **128**, Art.Nr. 134110 (2008)
- [96] H. Koch, *J. Chem. Phys.* To be submitted