Linear scaling computation of the Hartree-Fock exchange matrix [1]

Eric Raymond Schwegler

October 30, 1998

[1] Cite this work as LA-UR-98-4964

Abstract

A new method is introduced for fast construction of the Cartesian Gaussian-based exchange matrix that is central to Hartree-Fock and hybrid Hartree-Fock/density functional theories. By employing a novel loop structure and loop skip-out statements, linear scaling construction of the exchange matrix is demonstrated for insulating systems. For metallic systems, the method naturally reverts to a quadratic scaling. Importantly, numerical equivalence to standard direct SCF is maintained, and no spatial restrictions are placed on locality of the density matrix. The importance of exploiting permutational symmetry of electron repulsion integrals in linear scaling methods is also examined.

Accurate and computationally inexpensive estimates of multipole expansion errors are crucial to the success of a number of fast electronic structure methods. A new non-empirical multipole acceptability criterion is developed that is directly applicable to expansions of high order moments. Several model calculations typical of electronic structure theory are used to demonstrate improvements of up to five orders of magnitude over an empirical criterion. A multipole based method for electron repulsion integral evaluation that uses the new multipole acceptability criterion is developed and shown to be effective for avoiding the high cost of basis set contraction.

The analytical gradient of the Hartree-Fock energy with respect to nuclear displacement is reviewed. The exchange component of the gradient is identified and efficient thresholding criteria for its computation are given. The linear scaling methods for formation of the Hartree-Fock exchange matrix are extended to computation of the exchange component of the gradient and linear scaling is demonstrated.

Acknowledgements

I would like to express my gratitude to the following:

Dr. Matt Challacombe – for being my advisor and friend. I am deeply indebted to him for unselfishly sharing his knowledge with me and for always putting my interests first.

The ASCI program at the Los Alamos National Laboratory – for providing support during my last year as a graduate student.

Professor Martin Head-Gordon – for not hesitating to add a Minnesota branch to his research group during a difficult time for me.

Professors Donald Truhlar and Jim Chelikowsky – for their advice and help during my stay in Minnesota.

The Minnesota Supercomputer Institute – for providing access to unmatched computational resources, a terrific place to do research, and a warm place to park my car.

Mom, Dad, Ma, Pa, Gretchen, Heidi, Brigetta, Kevin, Kevin, Connor, Ryan, Saryna, Lada, Michael, Van, Alex, Stephany, Bryan, and Marysa – I couldn't ask for a better family!

Finally, I thank my wife Yaravi. Only through her constant love, support, and encouragement has any of this been possible.

Contents

Introduction

The Hartree-Fock (HF) approximation is a central model in first-principles electronic structure methods. In molecular orbital theory, the HF approximation provides a reference for methods that include the effect of dynamical electron correlation[1]. Within the density functional theory, HF exchange is a necessary ingredient in the most accurate functionals[2]. Unmodified, HF theory models the essential physics in systems where the effects of dynamical correlation are small, and generally yields qualitatively correct results[3].

Solution of the HF equations by the linear combination of atomic orbitals (LCAO) method [4, 5] leads to an algorithm that requires work that grows formally as N^4, where N is proportional to the system size. This steep scaling is due to the computation and handling of the notorious two-electron repulsion integrals (ERIs),

$$(\phi_a \phi_b | \phi_c \phi_d) = \int \int d\mathbf{r} d\mathbf{r}' \phi_a^* (\mathbf{r}) \, \phi_b (\mathbf{r}) \frac{1}{|\mathbf{r} - \mathbf{r}'|} \phi_c^* (\mathbf{r}') \, \phi_d (\mathbf{r}') \, . \tag{1}$$

The evaluation of ERIs is needed for the construction of the Fock matrix

$$\mathbf{F} = \mathbf{h} + \mathbf{J} - \frac{1}{2}\mathbf{K} \tag{2}$$

where $\mathbf{h}$ is the core Hamiltonian and

$$J_{ab} = \sum_{cd} D_{cd} (\phi_a \phi_b | \phi_c \phi_d), \tag{3}$$

and

$$K_{ab} = \sum_{cd} D_{cd} (\phi_a \phi_c | \phi_b \phi_d), \tag{4}$$

are the Coulomb and exchange matrices. The density matrix

$$D_{cd} = 2 \sum_{\mu} C_{d\mu} C_{c\mu}^*, \tag{5}$$

is obtained by contraction of molecular orbital coefficients $C_{d\mu}$[4, 5], or by direct minimization[6, 7, 8, 9].

Although the number of ERIs grow formally as N^4, it has been known for some time that in large systems a sizable number of ERIs are small, and that avoiding their calculation leads to algorithms that have $\mathcal{O}(N^2 \log N)$ or even $\mathcal{O}(N^2)$ complexity[10, 11]. However, because ERIs are contracted with elements of $\mathbf{D}$, the size of an ERI does not always correlate well with its overall significance to the Fock matrix. With the ubiquitous direct SCF methods introduced by Almlöf and co-workers[12], errors in the Fock matrix are controlled by thresholding with density weighted ERIs. The sophisticated integral prescreening used in the direct SCF method is effective as even large ERIs can be eliminated when their contribution to the Fock matrix is predicted to be small.

The use of density weighted integral estimates[12, 13, 14] along with the continued development of efficient ERI evaluation schemes[15, 16, 17, 18, 19, 20] has enabled direct SCF calculations on molecular systems large enough to realize an $\mathcal{O}(N^2)$ computational complexity in construction of the Fock matrix[21]. While other steps in the standard Roothaan-Hall procedure do scale faster than $\mathcal{O}(N^2)$, such as matrix diagonalization, methods for direct minimization of $\mathbf{D}$ can be formulated that require only $\mathcal{O}(N)$ work[9]. Therefore, large-scale direct SCF calculations are limited by calculation of the $\mathcal{O}(N^2)$ ERIs that are needed to form the Coulomb and exchange matrices.

Recently, several new methods based on hierarchical multipole expansions have appeared for construction of the Coulomb matrix that only require $\mathcal{O}(N)$ work[22, 23, 24]. Unfortunately, these fast methods are not directly applicable to computation of the exchange matrix, and a different approach must be taken.

The focus of this volume is on the development and implementation of new computational methods that facilitate formation of the HF exchange matrix. The chapters are organized as follows: In Chapter 1 we introduce ONX, which is the first ever method for linear scaling computation of the Hartree-Fock exchange matrix. Calculations on water clusters, large proteins, and graphitic sheets are used to demonstrate its computational efficiency and rigorous control of error. Much of the work presented in this chapter appears in:

> "Linear scaling computation of the Fock matrix. II. Rigorous bounds on exchange integrals and incremental Fock builds," E. Schwegler, M. Challacombe, and M. Head-Gordon, *J. Chem. Phys.*, 107(23):9703–17, 1996.

In Chapter 2, the use of ERI permutational symmetry in linear scaling computations of the Hartree-Fock exchange matrix is discussed. Several key factors are pointed out that are shown to have a marked effect on the observed computational benefits of permutational symmetry. Accurate and computationally inexpensive estimates of multipole expansion errors are examined in Chapter 3, and a new nonempirical multipole acceptability criterion (MAC) is described that is appropriate for use in electronic structure theory. This work appears in:

> "A multipole acceptability criterion for electronic structure theory," E. Schwegler, M. Challacombe, and M. Head-Gordon, *J. Chem. Phys.*, 1998. accepted.

In Chapter 4, use of the MAC is demonstrated in the evaluation of contracted ERIs within the multipole approximation. The combination of this multipole based method with ONX is shown to be a highly efficient way to form the HF exchange matrix. Finally, in Chapter 5, calculation of the analytical first derivative of the Hartree-Fock energy is reviewed and linear scaling computation of the exchange energy component of the derivative is demonstrated for the first time.

Chapter 1

Order N exchange

1.1 Introduction

The applicability of modern quantum chemistry is rapidly expanding due to the success and continued evolution of fast electronic structure methods. These fast methods may be loosely classified as those that involve hierarchical methods and those that do not. Hierarchical methods are able to reduce the computational complexity of building the Coulomb matrix $\mathbf{J}$ from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$, and include continuous versions of the Fast Multipole Method [25, 22, 26, 27, 28, 23], the quantum chemical treecode [24], and FFT based KWIK technologies [29]. In the latter class are methods that exploit or enforce the locality of short-range quantum interactions to achieve a reduced computational complexity. In particular, a wide variety of methods have recently been introduced that enforce locality of the density matrix $\mathbf{D}$ [8, 30, 31, 32, 33, 34, 35, 36, 37, 38] or the molecular orbital matrix $\mathbf{C}$ [39, 40, 41, 42, 43, 44, 45, 46, 47] to achieve linear scaling in various aspects of SCF theory, as for example computation of the Hartree-Fock (HF) exchange matrix $\mathbf{K}$ [37, 38]. However, methods that enforce locality have not been able to simultaneously achieve the speed and accuracy associated with hierarchical methods. This is relevant to application of the increasingly useful HF and hybrid HF/DFT theories [48, 2, 49, 50, 51, 52], which are currently limited by the computation of $\mathbf{J}$ and $\mathbf{K}$ (and in the latter case, the exchange-correlation matrix).

Recently, we introduced a linear scaling method for computation of the HF Fock matrix $\mathbf{F} = \mathbf{h} + \mathbf{J} - 1/2\mathbf{K}$, where $\mathbf{h}$ is the core Hamiltonian,

$$J_{ab} = \sum_{cd} D_{cd} \left(\phi_a \phi_b | \phi_c \phi_d\right) , \tag{1.1}$$

$$K_{ab} = \sum_{cd} D_{cd} \left(\phi_a \phi_c | \phi_b \phi_d\right) \ , \tag{1.2}$$

and the four-index quantities

$$(\phi_a \phi_b | \phi_c \phi_d) = \int d\mathbf{r} \int d\mathbf{r}' \ \phi_a^*(\mathbf{r}) \phi_b(\mathbf{r}) \left|\mathbf{r} - \mathbf{r}'\right|^{-1} \phi_c^*(\mathbf{r}') \phi_d(\mathbf{r}') \tag{1.3}$$

are electron repulsion integrals (ERIs) involving Cartesian Gaussian functions (CGTFs)

$$\phi_a(\mathbf{r}) = (x - A_x)^l (y - A_y)^m (z - A_z)^n \exp\left[-\zeta_a (\mathbf{r} - \mathbf{A})^2\right]. \tag{1.4}$$

In Eqs. (1.1) and (1.2) $\mathbf{D}$ is the density matrix

$$D_{cd} = 2\sum_{\mu} C_{c\mu} C^*_{d\mu} \tag{1.5}$$

obtained by contraction of molecular orbital coefficients $C_{c\mu}$[4, 5], or found directly[6, 7, 8]. To achieve linear scaling, we employed the quantum chemical treecode QCTC [24] to compute $\mathbf{J}$, and enforced the locality of exchange interactions with ONX [37], hearafter referred to as ONX_A. With this method, we found ONX_A to be 3-7 times slower than QCTC. Perhaps a more significant disadvantage of ONX_A is the reliance on system dependent parameters to balance performance with accuracy. Nevertheless, this approach allowed us to carry out HF calculations of unprecedented size, and has provided much of the motivation and insight for the present study.

Here, a new method for computation of the Hartree-Fock exchange matrix is described, which may be the first algorithm to achieve linear scaling without enforced locality. Also, linear scaling incremental Fock builds are demonstrated for the first time, and useful benchmarks are obtained.

This chapter is organized as follows: In the next Section, locality of the density matrix and implications for fast electronic structure methods are discussed. Existing methods for computing the HF exchange matrix are reviewed in Section 1.3 and various issues relevant to computation of the exchange matrix are considered in Section 1.4. In Section 1.5, a new implementation of subroutine ONX is outlined. Then, in Section 1.6 the use of ONX and QCTC in incremental Fock builds is described. Benchmark RHF/3-21G and RHF/6-31G calculations on sequences of water clusters, polyglycine α-helices, and single layer graphic sheets are presented in Section 1.7, along with calculations on the P53 tetramerization monomer. These results are discussed in Section 1.8, and in Section 1.9 we present our conclusions.

1.2 Locality of the density matrix

The long-range behavior of the density matrix,

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{cd} D_{cd} \phi_c(\mathbf{r}) \phi_d(\mathbf{r}') \tag{1.6}$$

is a subject of fundamental and practical importance. In Hartree-Fock theory, the density matrix describes the entire physical system [53]. Knowledge of its long-range behavior is essential to achieve linear scaling in a variety of electronic structure theories [24, 8, 30, 31, 32, 33, 34, 35, 36]. Also, locality of the density matrix is intimately related to orbital localization [39, 54, 55], and is therefore relevant to local-MO implementations of various SCF theories [39, 40, 41, 42, 43, 44, 45, 46, 47].

While the density matrix is dependent on both model chemistry and the physical system, generalizations of its asymptotic decay have been made for infinite systems [56, 57, 58, 59, 60]. For insulators that can be represented by localized Wannier functions, the density matrix decays as

$$\rho(\mathbf{r},\mathbf{r}') \sim e^{-\sqrt{G}|\mathbf{r}-\mathbf{r}'|} \tag{1.7}$$

where G is proportional to the band gap of the system [61, 62]. For conductors, which do not have a band gap, the decay rate is slower and a power law dependence

$$\rho(\mathbf{r},\mathbf{r}') \sim |\mathbf{r}-\mathbf{r}'|^{-d} \tag{1.8}$$

is expected [58, 59, 63], where d is related to the dimensionality of the system.

For finite systems, especially disordered ones, the situation is less clear cut [64]. Kohn has conjectured that the density matrix behaves as

$$\rho(\mathbf{r},\mathbf{r}') \sim e^{-|\mathbf{r}-\mathbf{r}'|/\ell} \quad (|\mathbf{r}-\mathbf{r}'| \to \infty) \ , \tag{1.9}$$

where ℓ is a system-specific length-scale parameter[63]. In practice, we find that for sys-

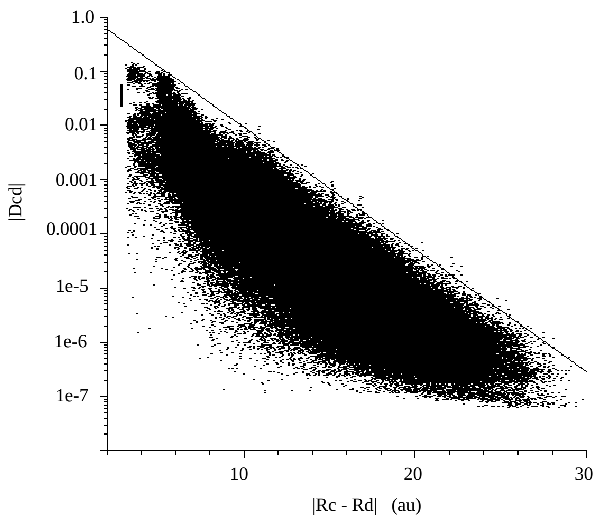

Figure 1.1: Magnitudes of RHF/3-21G density matrix elements with separation of basis function centers for the $(H_2O)_{110}$ water cluster. The density matrix has been converged to within 10^{-7} au. The line is $f(x) = ae^{-bx}$ with a=1.7 and b=0.52 au^{-1}.

tems such as proteins and water clusters, elements of the density matrix may be bounded (approximately) by an exponential, which decays with basis function separation. For example, see Figure 1.1. On the other hand, elements of the density matrix corresponding to molecular systems with a small HOMO-LUMO gap exhibit a decay that is much slower, consistent with the algebraic decay of conductors. For example, see Figure 1.2.

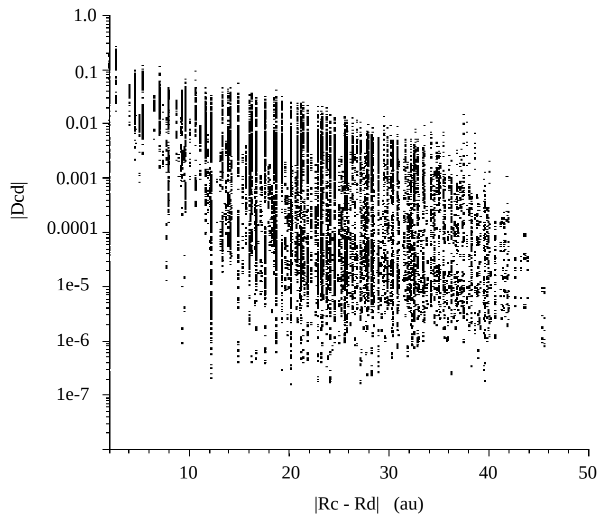

Figure 1.2: Magnitudes of RHF/3-21G density matrix elements with separation of basis function centers for the graphitic sheet $C_{150}H_{30}$. The density matrix has been converged to within 10^{-7} au. The stratification of density matrix elementsis due to molecular symmetry.

1.3 Methods for K builds

The first specialized methods for computing **K** were developed to efficiently truncate the potentially infinite summations that arise in the HF theory of periodic systems [56, 58, 59, 60]. These methods, implemented in the periodic HF program CRYSTAL, are based on truncating density and exchange matrices with distance dependent cutoffs.

Locality of the density matrix for insulating systems has been used to argue the existence of reduced scaling methods for computing exchange interactions in HF theory [26, 37, 65, 66, 67, 68, 69]. The most straightforward approach to exploiting this locality is to use the exchange component of the density weighted integral estimates commonly employed in direct SCF methods [70, 14];

$$\begin{aligned}\text{Max}\left[|D_{cd}|\,,|D_{ab}|\,,|D_{cb}|\,,|D_{ad}|\right]\times\\ \left|(\rho_{ac}|\rho_{ac})\right|^{1/2}\left|(\rho_{bd}|\rho_{bd})\right|^{1/2}\leq\ \text{thresh}\end{aligned} \tag{1.10}$$

where

$$\rho_{ac}(\mathbf{r})=\phi_a(\mathbf{r})\phi_c(\mathbf{r}) \tag{1.11}$$

is a distribution. However, this approach results in an algorithm with N^2 complexity, due at least to logical testing.

The first linear scaling method for computing $\mathbf{K}$ was developed in Ref. [37]. This method, which we refer to as ONX_A, is based on the assumed exponential decay of the density matrix, and the corresponding estimate

$$\widetilde{K}_{ab} = \int d\mathbf{r} \int d\mathbf{r}' \; \phi_a(\mathbf{r})\, \phi_b(\mathbf{r}')\, |\mathbf{r}-\mathbf{r}'|^{-1}\, e^{-|\mathbf{r}-\mathbf{r}'|/\ell} \tag{1.12}$$

for thresholding insignificant elements of $\mathbf{K}$.

More recently, Burant *et al.* have presented an alternative method, near-field exchange (NFX), for reducing the cost of $\mathbf{K}$ builds [38]. NFX limits exchange interactions to domains that are analogous to the near-field regions of the Gaussian very Fast Multipole Method (GvFMM) [23]. In the lowest order approximation $R+0$, the region of exchange interaction is identical to the GvFMM near-field, and the cost of constructing the near-field $\mathbf{J}$ can be amortized in the $\mathbf{K}$ build.

1.4 Exchange reconsidered

The aforementioned methods for computing $\mathbf{K}$ [37, 38] have demonstrated reduced computational scalings while maintaining errors at a reasonable level. However, the success of these methods relies on an empirically determined knowledge of density matrix asymptotics to achieve a balance between speed and accuracy. Implicit is the assumption that exchange interactions become negligible when basis function separations exceed certain distance dependent cutoff criteria. A subtle difference between ONX_A and NFX is that when the centers A and B are close, ONX_A may employ the entire density matrix, while NFX will not compute elements of $\mathbf{K}$ when C and D are well separated, which is equivalent to employing a sparse density matrix. In any case, as can be clearly seen in Fig. 1.1, simply bounding these interactions with a cutoff is necessarily inefficient as elements of $\mathbf{D}$ corresponding to a given separation can vary dramatically in size, in this example by four orders of magnitude.

If an effective zero is considered 10^{-8} au, numerical studies show that well converged density matrices remain dense, even for very extended systems with large HOMO-LUMO gaps [55, 71]. See for example Fig. 1.1. On careful examination, we find that thresholding itself easily dominates the computation of ERIs. However, retaining a dense density matrix does not necessarily lead to an N^2-scaling computation of ERIs, as for insulating systems small but numerically significant elements of the density matrix generally incur negligible work (see Section 1.8 for an illustration of this principle).

1.5 Order N exchange

A new implementation of ONX is described that employs the density weighted integral estimate

$$|D_{cd}|\,|(\rho_{ac}|\rho_{ac})|^{1/2}\,|(\rho_{bd}|\rho_{bd})|^{1/2} \leq \texttt{thresh}\,, \tag{1.13}$$

which is equivalent to standard direct methods. As demonstrated in Section 1.7.1, this approach achieves linear scaling for systems with a large HOMO-LUMO gap, and reduces naturally to a quadratic dependence in systems where the HOMO-LUMO gap is small.

```
do c =1,N
   do a =1,N
      compute |(ρ_ac|ρ_ac)|^{1/2}
      if(|(ρ_ac|ρ_ac)|^{1/2} > thresh /[D_cd^max|(ρ_bd^max|ρ_bd^max)|^{1/2}]) then
         add ρ_ac to {ρ_ac}
      endif
   enddo a
   for current c, sort {ρ_ac} by decreasing |(ρ_ac|ρ_ac)|^{1/2}
   store pointers first(c) and last(c)
enddo c
```

Figure 1.3: Ordering the list of significant distributions.

The N^2 complexity of thresholding can be avoided by foregoing permutational symmetry and employing a structured list of distributions $\{\rho_{ac}\}$, which for each basis function c are ordered by decreasing value of the integral estimate $|(\rho_{ac}|\rho_{ac})|^{1/2}$. To avoid an N^2 list, only those distributions that satisfy

$$|(\rho_{ac}|\rho_{ac})|^{1/2} \leq \frac{\texttt{thresh}}{D_{cd}^{\max}\,|(\rho_{bd}^{\max}|\rho_{bd}^{\max})|^{1/2}} \tag{1.14}$$

are considered, where $D_{cd}^{\max}$ is the largest element of the density matrix and $|(\rho_{bd}^{\max}|\rho_{bd}^{\max})|$ is the largest integral estimate. The procedure for structuring this list of distributions is outlined in Fig. 1.3.

ONX $\mathbf{K}$ builds begin with an outer loop over density matrix elements, as shown in Fig. 1.4. An intermediate list of distributions significant to basis function c is then formed using the criterion

$$|D_{cd}||(\rho_{ac}|\rho_{ac})|^{1/2}\,|(\rho_{b_{\mathfrak{f}}d}|\rho_{b_{\mathfrak{f}}d})|^{1/2} \leq \texttt{thresh}\,, \tag{1.15}$$

where $\rho_{b_{\mathfrak{f}}d}$ corresponds to the largest (first) integral estimate involving basis function d. An intermediate list of distributions significant to d are likewise formed.

Next, a two-fold loop over the intermediate lists of ρ_{ac} and ρ_{bd} is begun in which ERIs are computed and contracted with the density matrix to form elements of $\mathbf{K}$. For each ρ_{ac}, satisfaction of Eq. (1.15) validates `goto 501`, while the inner-most loop over ρ_{bd} can be skipped out of when Eq. (1.13) is met.

```
do c = 1,N
   do d = 1,N
      a_f = first(c); a_l = last(c)
      b_f = first(d); b_l = last(d)
      do  a = a_f,a_l
         if(|D_cd| * |(ρ_ac|ρ_ac)|^{1/2} * |(ρ_{b_f d}|ρ_{b_f d})|^{1/2} < thresh) goto 201
         buffer intermediate values for ρ_ac
      enddo a
201   continue
      do b = b_f,b_l
         if(|D_cd| * |(ρ_{a_f c}|ρ_{a_f c})|^{1/2} * |(ρ_bd|ρ_bd)|^{1/2}< thresh) goto 301
         buffer intermediate values for ρ_bd
      enddo b
301   continue
      do  a = a_f,a_l
         if(|D_cd| * |(ρ_ac|ρ_ac)|^{1/2} * |(ρ_{b_f d}|ρ_{b_f d})|^{1/2}< thresh) goto 501
         do  b = b_f,b_l
            if(|D_cd| * |(ρ_ac|ρ_ac)|^{1/2} * |(ρ_bd|ρ_bd)|^{1/2}< thresh) goto 601
            compute (ρ_ac|ρ_bd)
            K_ab = K_ab + D_cd * (ρ_ac|ρ_bd)
         enddo b
601      continue
      enddo a
501   continue
   endddo d
enddo c
```

Figure 1.4: The ONX loops. Skipping entirely out of loops (goto 201, 301, 501, and 601) is possible due to the ordering of distribution and integral estimate lists.

1.6 Recursive F builds

A unique feature of both ONX and QCTC is that they employ density based error estimates, and can therefore employ incremental $\mathbf{F}$ builds [70, 14, 13]. In the incremental approach, $\mathbf{F}$ is computed by recursion as

$$\mathbf{F}^m = \mathbf{F}^{m-1} + \Delta\mathbf{J}^{m-1} - \frac{1}{2}\Delta\mathbf{K}^{m-1} \,, \tag{1.16}$$

where m is the SCF cycle, and $\Delta\mathbf{J}^m$ and $\Delta\mathbf{K}^m$ are computed using the difference density matrix $\Delta\mathbf{D}^m = \mathbf{D}^m - \mathbf{D}^{m-1}$. For $m = 1$, the core Hamiltonian matrix and a guess density matrix are used.

With recursion, the computation of $\Delta\mathbf{J}$ and $\Delta\mathbf{K}$ is greatly accelerated as convergence is neared. However, the most recent full $\mathbf{F}$ build determines the maximum precision achievable, and errors that grow like $\sqrt{m}$ propagate during recursion [14]. To minimize this effect, a full Fock build can be carried out periodically, and/or the thresholding parameter can be continuously tightened.

We have implemented a simple algorithm for recursive Fock builds. Minimized density differences [14] and other acceleration techniques are applicable to our approach and are likely to substantially improve performance, but have not been implemented at this time. Subroutines ONX and QCTC [24] have been interfaced with a modified version of the *ab initio* electronic structure program HONDO 95.3[72] (MONDO). A single thresholding parameter `thresh`, described above for ONX and in Ref. [24] for QCTC, is used to control errors in $\mathbf{F}$. With this arrangement, the errors in $\mathbf{J}$ are smaller than those in $\mathbf{K}$, as `thresh` has a somewhat different function in QCTC than it does in `onx`. As $\mathbf{J}$ builds are still more competitive than $\mathbf{K}$ builds, balancing the errors from $\mathbf{J}$ and $\mathbf{K}$ builds is presently unwarranted.

In the following, two levels of precision are considered; those achieved with `thresh` $= 10^{-7}$ and those with `thresh` $= 10^{-9}$. All calculations are started from an extended Hückel guess. The first three SCF cycles are then performed without recursion using `thresh` $= 10^{-5}$. A full Fock build with `thresh` $= 10^{-7}$ is then performed, followed by incremental updates with decreasing values of `thresh` to avoid error propagation. Convergence is achieved when the criteria $\Delta E < 10^{-5}$ and $\max[\Delta\mathbf{D}] < 10^{-4}$ are satisfied. This protocol is illustrated in Table 1.1. To reach an error commensurate with `thresh` $= 10^{-9}$, an identical calculation is carried out, with a full `thresh` $= 10^{-9}$ Fock build after the aforementioned criteria are met. This is followed with recursive $\mathbf{F}$ builds involving decreasing values of `thresh`, reaching convergence when $\Delta E < 10^{-7}$ and $\max[\Delta\mathbf{D}] < 10^{-5}$, typically within three to four additional cycles.

1.7 Benchmarks

1.7.1 CPU Timings

To demonstrate the scaling properties of our new approach, MONDO RHF calculations with the 3-21G and 6-31G basis sets and `thresh` $= 10^{-7}$ and `thresh` $= 10^{-9}$ have been performed on a sequence of three-dimensional water clusters, polyglycine α-helices, and graphitic sheets. All calculations have been performed on a single processor IBM RS/6000 model

Table 1.1: Typical output from a RHF/3-21G calculation of a 150 water molecule cluster with `thresh` $= 10^{-7}$. Note that after SCF cycle 4, `thresh` is tightened to avoid the propagation of errors. For the more accurate calculations, SCF cycle 13 would be a full Fock build with `thresh` $= 10^{-9}$, followed by recursion.

SCF cycle	Energy (hartree)	ΔE	max[ΔD]	`thresh`	CPU time (sec) ONX	CPU time (sec) QCTC
1[a]	-11239.101 259	1.1e04	5.5e-01	1.0e-05	377.2	305.3
2[a]	-11335.567 964	9.6e01	1.3e-01	1.0e-05	271.6	299.4
3[a]	-11340.406 789	4.8e00	4.1e-02	1.0e-05	257.7	296.2
4[a]	-11341.005 010	6.0e-01	2.3e-02	1.0e-07	632.8	621.7
5	-11341.104 223	9.9e-02	7.4e-03	8.1e-08	280.9	320.3
6	-11341.122 070	1.8e-02	3.1e-03	3.6e-08	223.6	295.8
7	-11341.126 944	4.9e-03	1.6e-03	1.7e-08	270.6	293.3
8	-11341.127 859	9.1e-04	7.1e-04	9.4e-09	246.3	271.4
9	-11341.128 046	1.9e-04	3.5e-04	4.5e-09	246.0	248.2
10	-11341.128 086	4.0e-05	1.6e-04	2.2e-09	231.9	235.7
11	-11341.128 097	1.1e-05	7.5e-05	1.1e-09	226.9	216.2
12	-11341.128 101	3.5e-06	3.8e-05	5.9e-10	204.1	202.5

[a] Full Fock matrix computed in this SCF cycle.

590. These results are shown in Figures 1.5-1.14, where ONX(avg) and QCTC(avg) refer to the central processing unit (CPU) time per formation of the exchange and Coulomb matrix averaged over the entire calculation. MONDO(avg) is simply the sum of ONX(avg) and QCTC(avg) times. MONDO(max) is the most expensive Fock matrix construction time incurred in the calculation. This corresponds to the fourth SCF cycle when `thresh` is set to 10^{-7}, or the full Fock build with `thresh` $= 10^{-9}$.

The coordinates for the water clusters are the same as in Refs. [24, 37, 38]. Note that the coordinates for the polyglycine calculations are not the same as those employed by Burant *et. al.* [38]; their glycine coordinates correspond to an extended linear chain, while ours is in the α-helix conformation [37]. A carbon-carbon bond length of 1.42Å has been used to produce the coordinates of graphitic sheets with stoichiometry $C_{6n^2}H_{6n}$, for n=2 to 7[73].

To illustrate the potential N^2 scaling of thresholding, we have also included CPU timings labeled No Skipout in Figure 1.5. These timings correspond to full Fock builds in which the ONX loop skipouts have been disabled. In particular, `601 continue` and `501 continue` have been moved within their respective loops (see Figure 1.4). The CPU times for this method have been fit to a quadratic and are also shown in Figure 1.5 as No Skipout(fit).

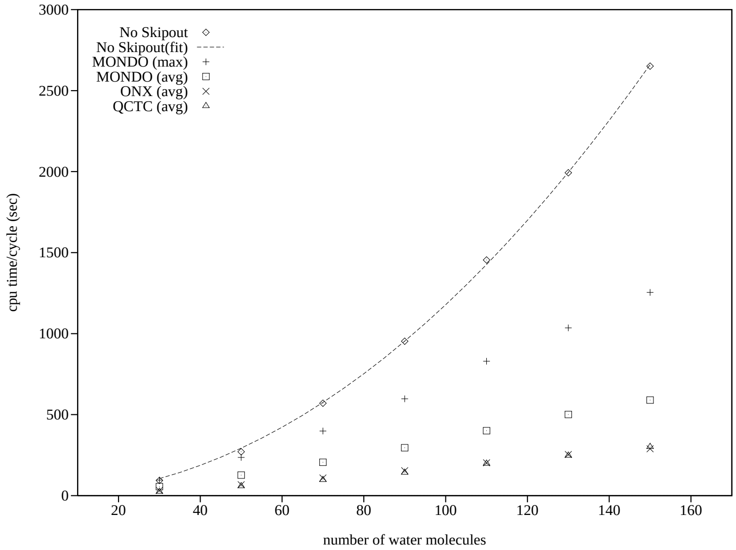

Figure 1.5: RHF/3-21G water cluster timings with $\mathtt{thresh} = 10^{-7}$.

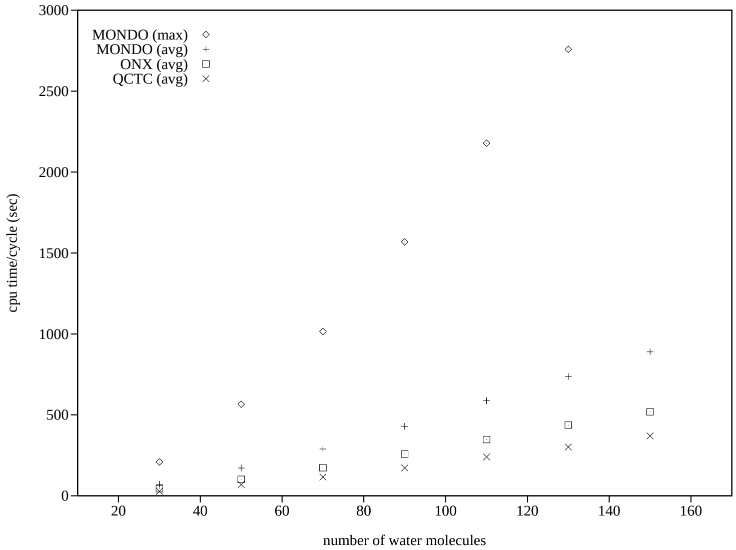

Figure 1.6: RHF/3-21G water cluster timings with $\mathtt{thresh} = 10^{-9}$.

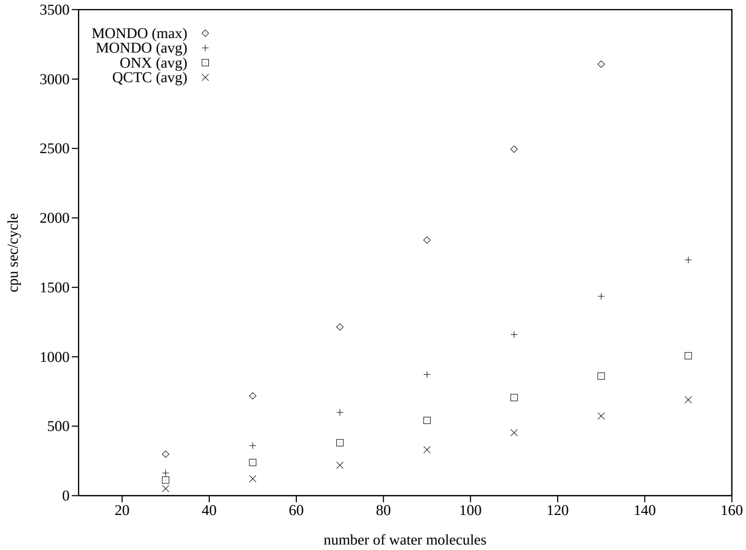

Figure 1.7: RHF/6-31G water cluster timings with $\texttt{thresh} = 10^{-7}$.

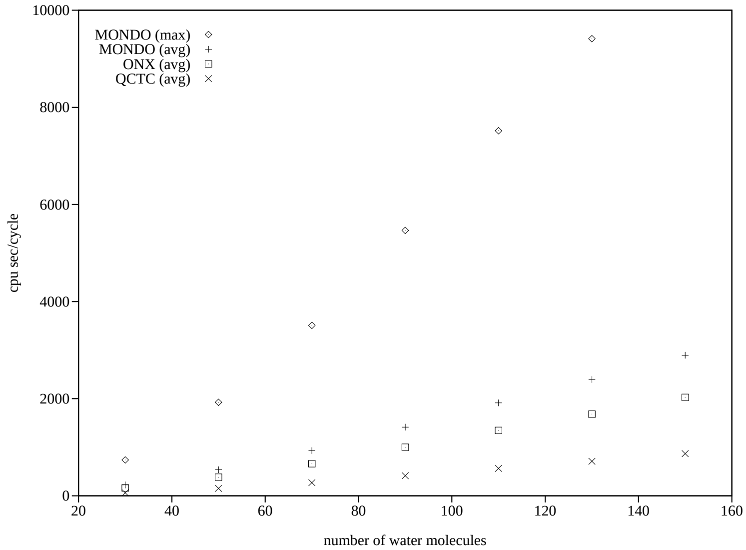

Figure 1.8: RHF/6-31G water cluster timings with $\texttt{thresh} = 10^{-9}$.

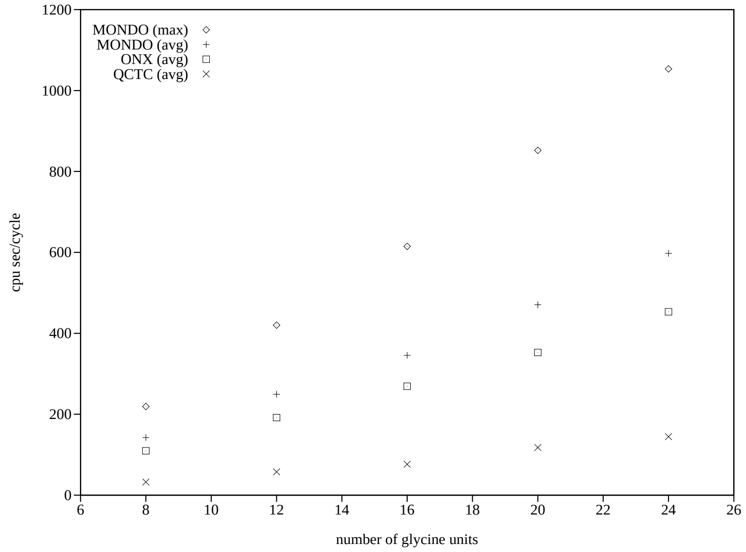

Figure 1.9: RHF/3-21G polyglycine α-helix timings with $\texttt{thresh} = 10^{-7}$.

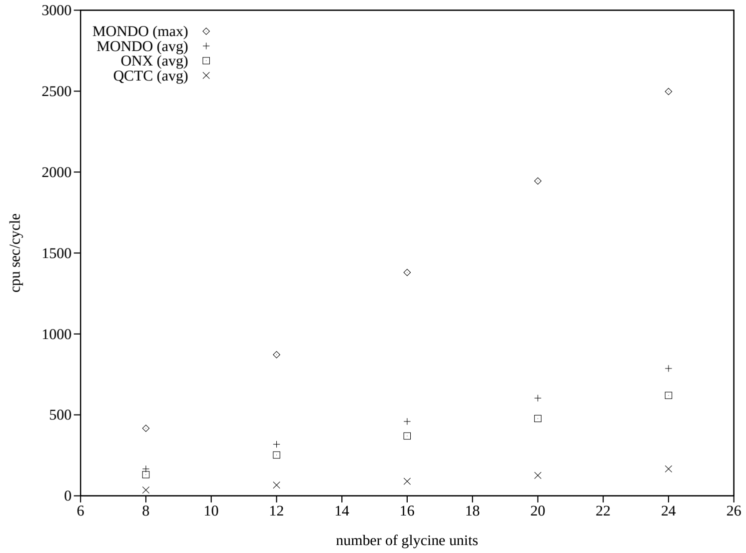

Figure 1.10: RHF/3-21G polyglycine α-helix timings with $\texttt{thresh} = 10^{-9}$.

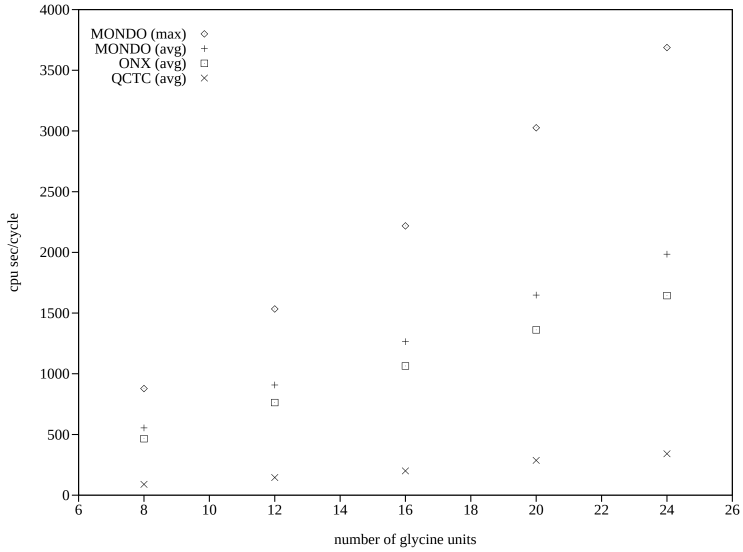

Figure 1.11: RHF/6-31G polyglycine α-helix timings with $\texttt{thresh} = 10^{-7}$.

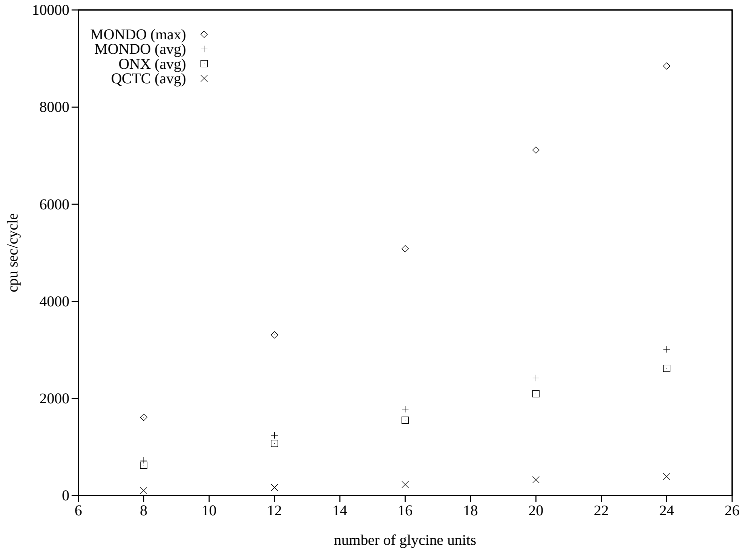

Figure 1.12: RHF/6-31G polyglycine α-helix timings with $\texttt{thresh} = 10^{-9}$.

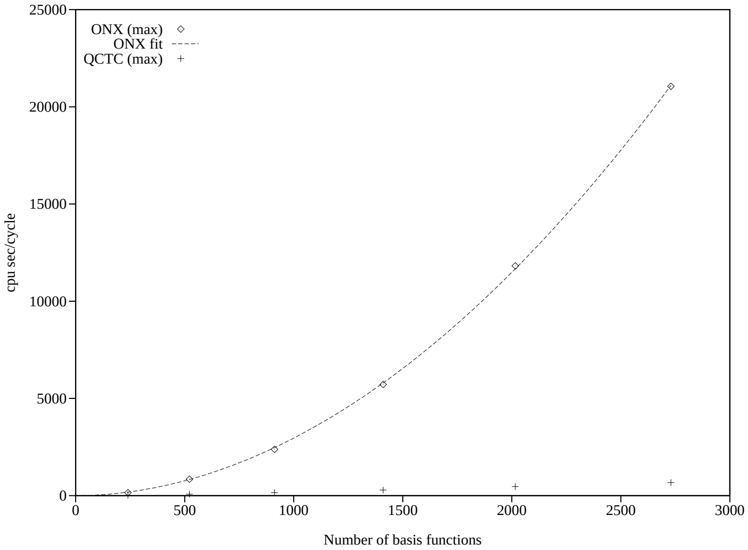

Figure 1.13: RHF/3-21G graphitic sheet timings with **thresh** $= 10^{-7}$.

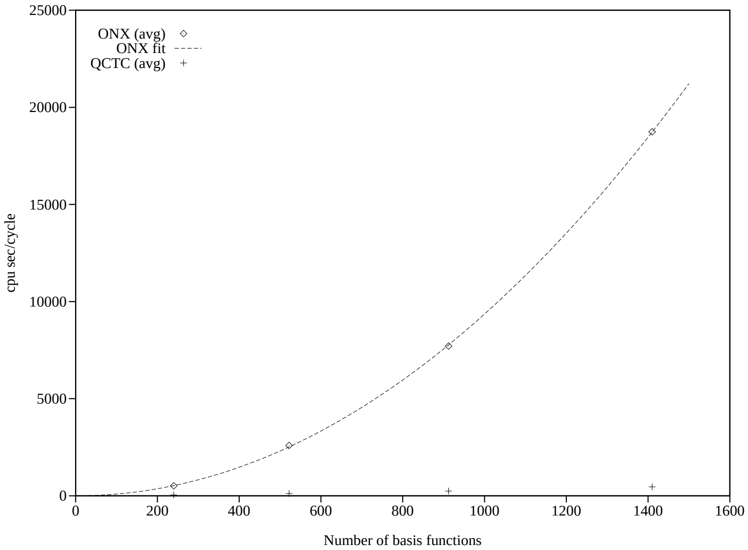

Figure 1.14: RHF/6-31G graphitic sheet timings with **thresh** $= 10^{-7}$.

Table 1.2: The convergence of the RHF/3-21G total energy for different molecular systems and values of `thresh`. For each calculation `thresh` was held constant and the incremental Fock build technique was not used.

		Total energy (hartree)		
`thresh`	max[$\Delta\mathbf{D}$]	$(H_2O)_{110}$	$(Gly)_{24}$	$C_{54}H_{18}$
10e-6	1.0e-5	-8316.844 140	-4936.961 534	-2044.145 492
10e-7	1.0e-6	-8316.845 692	-4936.961 942	-2044.147 993
10e-8	1.0e-6	-8316.845 757	-4936.962 076	-2044.147 809
10e-9	1.0e-6	-8316.845 776	-4936.962 087	-2044.147 818
10e-11	1.0e-7	-8316.845 776	-4936.962 087	-2044.147 819

Table 1.3: The RHF/3-21G error per atom for different molecular systems and values of `thresh`. Within each calculation `thresh` was held constant and the incremental Fock build technique was not used.

	Error per atom (μhartree)		
`thresh`	$(H_2O)_{110}$	$(Gly)_{24}$	$C_{54}H_{18}$
10e-6	5.0	3.3	32.
10e-7	0.25	0.86	2.4
10e-8	0.057	0.071	0.13
10e-9	0.00069	0.0080	0.020

1.7.2 Errors and convergence

Absolute errors in the converged total energies for each of the molecular sequences and basis sets are shown in Figures 1.15 to 1.19. In each case, the errors have been computed by comparison to MONDO with `thresh` $= 10^{-11}$. In Table 1.2, we have also listed three representative systems from each sequence along with corresponding converged total energies for decreasing values of `thresh`. The column labeled max[$\Delta\mathbf{D}$] in Table 1.2 is the convergence criterion used. The error per atom for the the same systems is listed in Table 1.3.

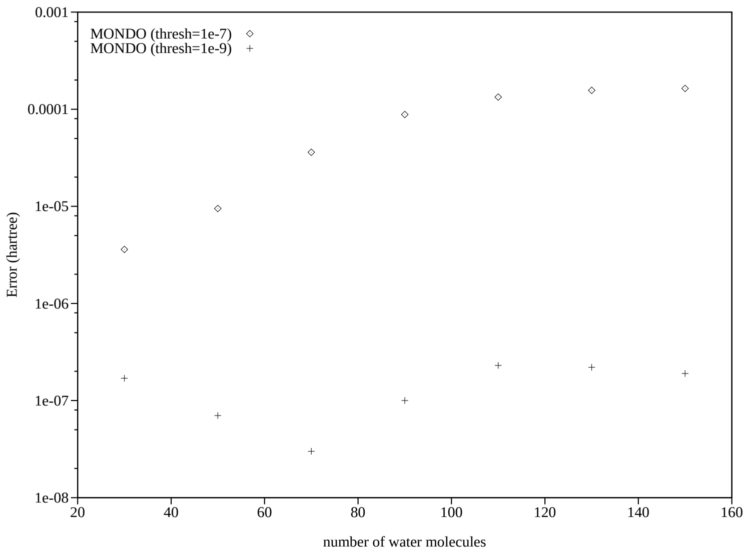

Figure 1.15: RHF/3-21G water cluster errors in converged total energy.

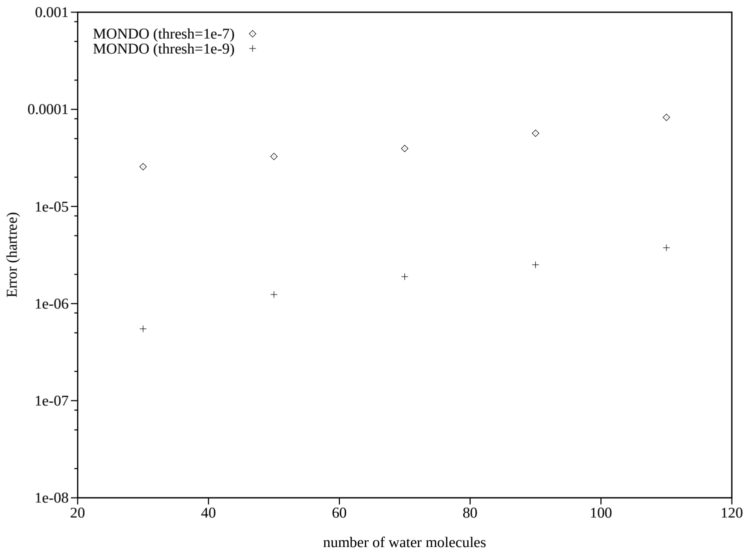

Figure 1.16: RHF/6-31G water cluster errors in converged total energy.

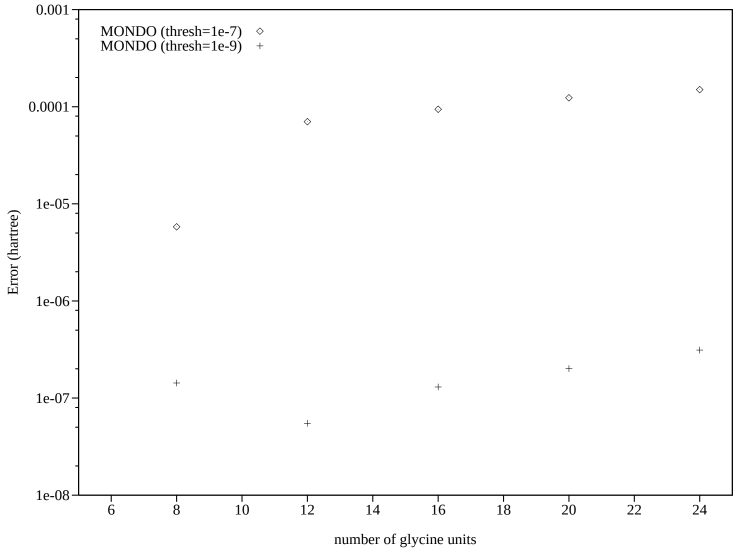

Figure 1.17: RHF/3-21G glycine α-helix errors in converged total energy.

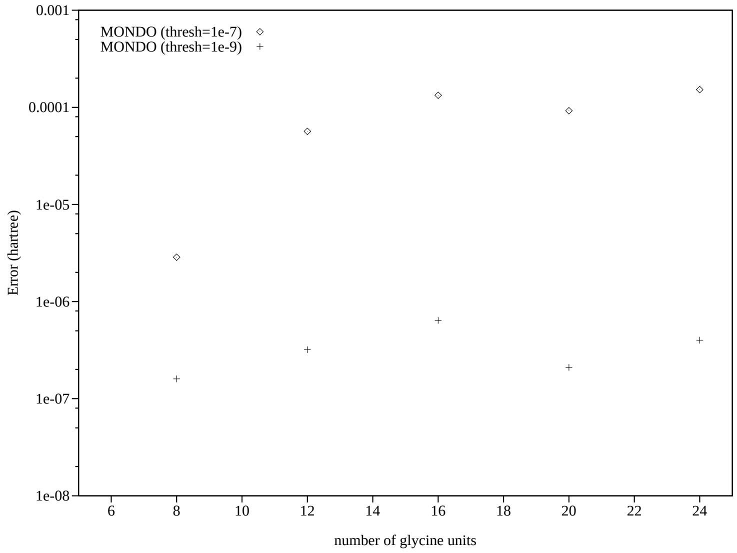

Figure 1.18: RHF/6-31G glycine α-helix errors in converged total energy.

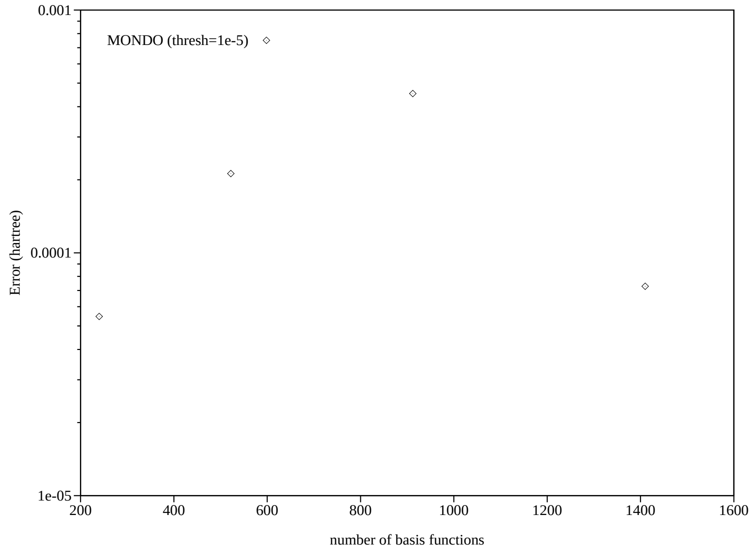

Figure 1.19: RHF/3-21G graphitic sheet errors in converged total energy.

Table 1.4: Convergence of the P53 RHF/3-21G total energy and corresponding CPU times per SCF cycle for decreasing values of `thresh`. Energies and timings correspond to full Fock builds and a constant value of `thresh`.

			CPU time (hours)	
`thresh`	max[$\Delta\mathbf{D}$]	E_{RHF} (hartree)	ONX	QCTC
10e-6	1.0e-5	-17115.306 141	0.67	0.36
10e-7	1.0e-6	-17115.305 875	1.36	0.49
10e-8	1.0e-6	-17115.305 227	2.31	0.66
10e-9	1.0e-6	-17115.305 248	3.74	0.87
10e-11	1.0e-7	-17115.305 249	8.35	1.46

1.7.3 P53

In Ref. [24], we reported a RHF/3-21G calculation on the tetramerization monomer of P53, in what we believe to be the largest non-empirical HF calculation to date. In Table 1.4, the convergence of this calculation with decreasing values of `thresh` is reported, along with CPU times for ONX and QCTC.

1.8 Discussion

The new implementation of ONX is a marked improvement over previous linear scaling methods for computing the HF exchange matrix. For full Fock builds, ONX is approximately 3-4 times faster than ONX_A, and is numerically equivalent to the direct SCF methods now standard in modern quantum chemistry. Perhaps equally dramatic, an unsophisticated approach to incremental Fock builds has been found to increase this margin by another factor of 2, with similar gains for QCTC. It is likely that a more sophisticated approach to incremental Fock builds will improve this factor.

For the insulating systems studied, the computation of $\mathbf{K}$ with ONX scales linearly with system size. For graphitic sheets which have small HOMO-LUMO gaps tending to zero with N, a quadratic scaling is observed. It is satisfying that the transition from linear to quadratic is entirely controlled by the structure of $\mathbf{D}$.

Calculations with `thresh` $= 10^{-9}$ have demonstrated linear scaling and μhartree accuracy in converged total energies for the 3-21G and 6-31G basis sets. We reconcile this level of accuracy, the use of an effectively dense density matrix, and linear scaling by noting that even though $\mathbf{D}$ is an N^2 quantity, ONX is dominated by the evaluation of ERIs, which is $\mathcal{O}(N)$ for water clusters and α-helices. This surprising result is highlighted by Figures 1.20-1.23, which show elements of $\mathbf{D}$ versus basis function separation, binned by the number of associated ERIs $M_{\rm int}$, for the RHF/3-21G $(H_2O)_{150}$ calculation of $\mathbf{K}$. These results are contrasted by Figures 1.24-1.27, which show the corresponding values for the $C_{150}H_{30}$ calculation of $\mathbf{K}$. For the water cluster, matrix elements involving large numbers of ERIs correspond to small separations, with a definite correlation between larger separations and fewer ERIs. For the graphitic sheet, density matrix elements corresponding to large $M_{\rm int}$ persist even for separations of up to 40 atomic units. Note that the percentages quoted in Figures 1.20-1.23 correspond to a much smaller ERI total than do those in Figures 1.24-1.27.

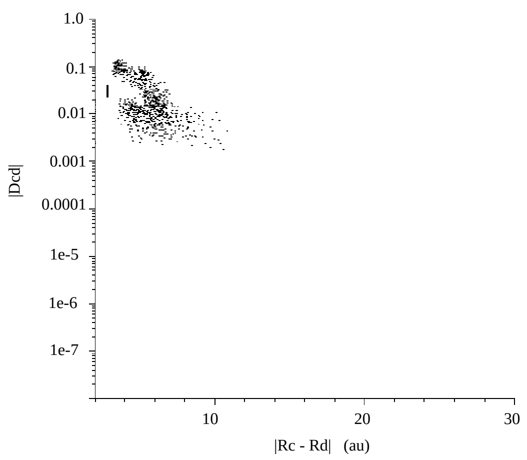

Figure 1.20: $|D_{cd}|$ versus $|C - D|$ for the $(H_2O)_{110}$ cluster for which $M_{\text{int}} > 1000$. This plot represents 6.5% of the integrals and 0.4% of $\mathbf{D}$.

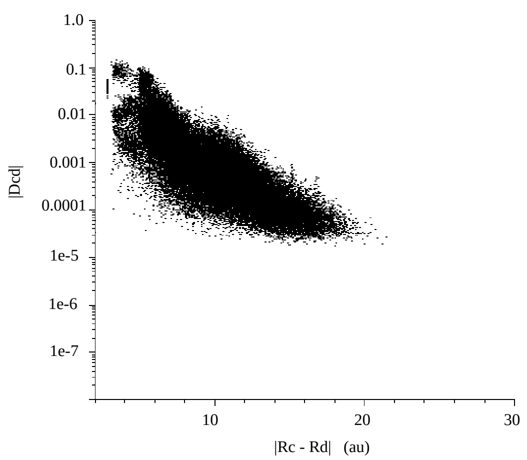

Figure 1.21: $|D_{cd}|$ versus $|C - D|$ for the $(H_2O)_{110}$ cluster for which $1000 > M_{\text{int}} > 100$. This plot represents 66.4% of the integrals and 20.8% of $\mathbf{D}$.

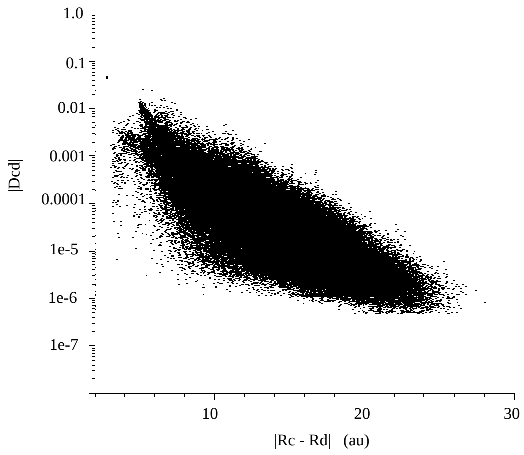

Figure 1.22: $|D_{cd}|$ versus $|C - D|$ for the $(H_2O)_{110}$ cluster for which $100 > M_{\text{int}} > 10$. This plot represents 25.6% of the integrals and 53.2% of $\mathbf{D}$.

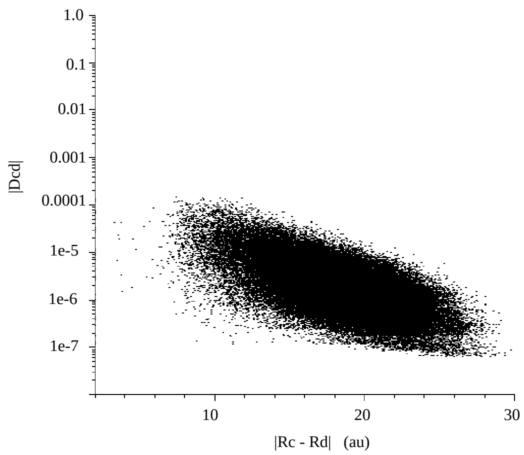

Figure 1.23: $|D_{cd}|$ versus $|C - D|$ for the $(H_2O)_{110}$ cluster for which $10 > M_{\text{int}}$. This plot represents 1.5% of the integrals and 25.5% of $\mathbf{D}$.

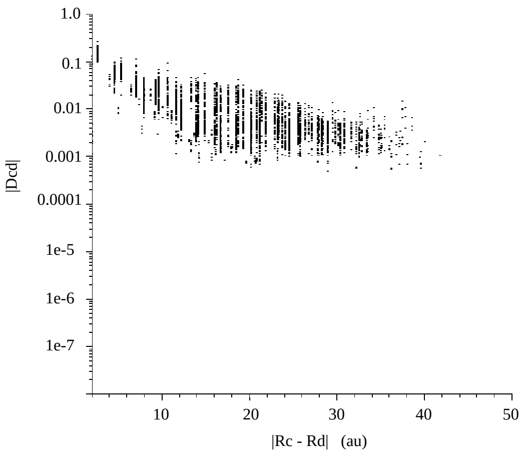

Figure 1.24: $|D_{cd}|$ versus $|C - D|$ for $C_{150}H_{30}$ for which $M_{\mathrm{int}} > 1000$. This plot represents 58.6% of the integrals and 17.9% of **D**.

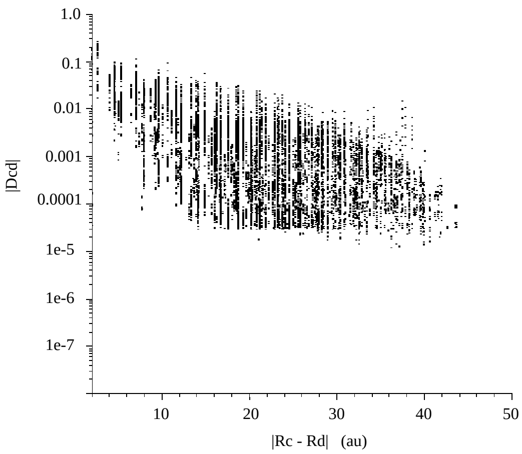

Figure 1.25: $|D_{cd}|$ versus $|C - D|$ for $C_{150}H_{30}$ for which $1000 > M_{\mathrm{int}} > 100$. This plot represents 39.4% of the integrals and 52.0% of **D**.

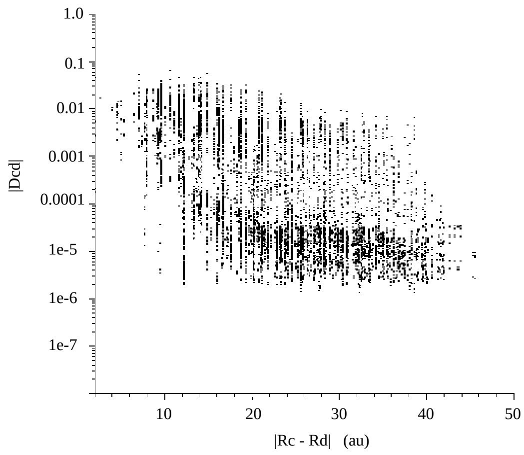

Figure 1.26: $|D_{cd}|$ versus $|C - D|$ for $C_{150}H_{30}$ for which $100 > M_{\mathrm{int}} > 10$. This plot represents 1.9% of the integrals and 23.6% of $\mathbf{D}$.

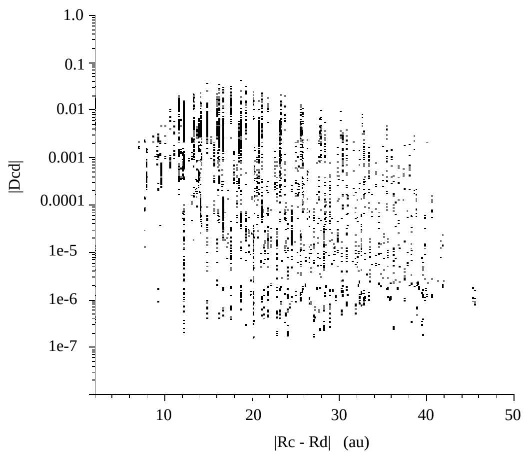

Figure 1.27: $|D_{cd}|$ versus $|C - D|$ for $C_{150}H_{30}$ for which $10 > M_{\mathrm{int}}$. This plot represents 0.04% of the integrals and 6.4% of $\mathbf{D}$.

1.9 Conclusions

We have shown that locality of the Gaussian-based density matrix is directly exploitable without sacrificing accuracy, and without enforcing locality. However, this requires the careful minimization of costs associated with data handling and manipulation, such as the thresholding of integrals. One effective strategy for achieving this is to employ data structures that permit loop skip-outs. Note that this approach is also employed by QCTC in thresholding near-field Coulomb integrals [24]. Related strategies may also prove useful in reducing the complexity of MO-based theories of electron-correlation.

Achieving a linear scaling SCF method for all N will ultimately require the use of sparse matrices to circumvent the N^{2-3} CPU time and N^2 memory associated with dense MO and/or density matrices. However, this work demonstrates the independence of achieving linear scaling Fock builds and the use of sparse matrices.

In conclusion, a new method has been presented for computing the Hartree-Fock exchange matrix, and linear scaling incremental Fock builds have been demonstrated for the first time. This new approach retains the reliability associated with direct SCF, and is able to achieve μhartree accuracies and linear scaling for insulating systems with both the 3-21G and the 6-31G basis set. Significantly, a factor of 8 improvement over our previous method [24] has been achieved. For systems with small HOMO-LUMO gaps, such as graphitic sheets, the approach naturally reverts to an $\mathcal{O}(N^2)$ complexity.

MONDO RHF/3-21G calculations on the P53 tetramerization monomer, involving 698 atoms and 3836 basis functions, have been carried out to a precision of 1 μhartree on the commonly available IBM RS/6000 model 590, and may provide a useful benchmark for the development of new algorithms [73].

Chapter 2

Permutational symmetry

2.1 Introduction

It is common in modern electronic structure methods to use basis functions formed by contractions of primitive functions,

$$\phi_a(\mathbf{r}) = \sum_i^{K_a} d_{ai}\varphi_{ai}(\mathbf{r}), \tag{2.1}$$

where d_{ai} are contraction coefficients, φ_{ai} are primitive basis functions, and K_a is the contraction length. For molecular computations, the primitive functions of choice are the Cartesian Gaussian type functions (CGTFs),

$$\begin{aligned} \varphi_{ai}(\mathbf{r}) = (x - A_x)^l (y - A_y)^m (z - A_z)^n \\ \times \exp\left[-\zeta_{ai}(\mathbf{r} - \mathbf{A})^2\right], \end{aligned} \tag{2.2}$$

first proposed by Boys[74]. Primitive CGTFs are preferred because, as stated by the Gaussian product theorem (GPT), a product of CGTFs can be expressed exactly as a one-center finite sum of CGTFs. For example, the product of two s-type ($l = m = n = 0$) primitive CGTFs is

$$\begin{aligned} \phi_a(\mathbf{r})\phi_b(\mathbf{r}) &= \exp\left[-\zeta_a(\mathbf{r} - \mathbf{A})^2\right] \times \exp\left[-\zeta_b(\mathbf{r} - \mathbf{B})^2\right] \\ &= \exp\left[-\zeta_a\zeta_b/(\zeta_a + \zeta_b)(\mathbf{A} - \mathbf{B})^2\right] \times \exp\left[-\zeta_p(\mathbf{r} - \mathbf{P})^2\right], \end{aligned} \tag{2.3}$$

where $\zeta_p = \zeta_a + \zeta_b$ and $\mathbf{P} = (\zeta_a\mathbf{A} + \zeta_b\mathbf{B})/\zeta_p$. Because of the GPT, it follows that the four-center electron repulsion integrals (ERIs),

$$(\varphi_a\varphi_b|\varphi_c\varphi_d) = \int\int d\mathbf{r}d\mathbf{r}'\varphi_a^*(\mathbf{r})\varphi_b(\mathbf{r})\frac{1}{|\mathbf{r} - \mathbf{r}'|}\varphi_c^*(\mathbf{r}')\varphi_d(\mathbf{r}'), \tag{2.4}$$

can be reduced to two-center integrals, which are readily evaluated.

In addition, since CGTFs are real, there is an eight-fold permutational symmetry between the indices a, b, c, and d in Eq. (2.4), and the following integrals are equal

$$\begin{aligned} (\varphi_a\varphi_b|\varphi_c\varphi_d) &= (\varphi_b\varphi_a|\varphi_c\varphi_d) = (\varphi_a\varphi_b|\varphi_d\varphi_c) = (\varphi_b\varphi_a|\varphi_d\varphi_c) \\ &= (\varphi_c\varphi_d|\varphi_a\varphi_b) = (\varphi_d\varphi_c|\varphi_a\varphi_b) = (\varphi_c\varphi_d|\varphi_b\varphi_a) \\ &= (\varphi_d\varphi_c|\varphi_b\varphi_a). \end{aligned} \tag{2.5}$$

Equation (2.5) is also true for ERIs over contracted basis functions. By taking into account the equalities in Eq. (2.5), and by considering how the ERIs are combined with elements of the density matrix, the HF exchange matrix for a closed-shell system may be calculated with

$$\begin{aligned} K_{ac} &= K_{ac} - D_{bd}\ (\phi_a\phi_b|\phi_c\phi_d)\,, \\ K_{ad} &= K_{ad} - D_{bc}\ (\phi_a\phi_b|\phi_c\phi_d)\,, \\ K_{bc} &= K_{bc} - D_{ad}\ (\phi_a\phi_b|\phi_c\phi_d)\,, \quad \text{and} \\ K_{bd} &= K_{bd} - D_{ac}\ (\phi_a\phi_b|\phi_c\phi_d)\,, \end{aligned} \tag{2.6}$$

instead of Eq. (1.2). Because the calculation of ERIs is often the most time consuming step, the use of Eq. (2.6) offers the possibility of saving a factor of four in the computation of **K**.

In the formulation of ONX presented in Chapter 1, the permutational symmetry of ERI indices was foregone in favor of a novel loop structure that enables loop skip-out statements. These loop skip-out statements are crucial to avoiding quadratically scaling prescreening of small ERIs and enables linear scaling computation of **K** for insulating systems. However, the recently developed LinK algorithm[75] demonstrates that it is possible to use the loop skip-out statements of ONX while at the same time taking advantage of the ERI permutational symmetry.

Although the advantage of exploiting ERI permutational symmetry has been well established for small molecular systems, the situation is unclear for calculations on medium to large systems. For these cases, ERI permutational symmetry will be useful only if the magnitudes of each contribution an ERI makes to **K** decay at a similar rate, which depends on the structure of **D** as seen in Eq. (2.6). Specifically, this is due to the different ways small contributions to **K** are avoided in methods that do not use permutational symmetry as compared to those that do use symmetry. For example, in ONX small contributions to **K** are avoided with the criterion

$$|D_{cd}||(\phi_a\phi_c|\phi_a\phi_c)|^{1/2}|(\phi_b\phi_d|\phi_b\phi_d)|^{1/2} \leq \texttt{TwoENeglect}\,. \tag{2.7}$$

However, in methods that exploit permutational symmetry all four of the contributions in Eqn. (2.6) are considered together, so the criterion

$$\begin{aligned} &\mathrm{Max}\left[|D_{bd}|\,,|D_{bc}|\,,|D_{ad}|\,,|D_{ac}|\right] \times \\ &\qquad |(\phi_a\phi_b|\phi_a\phi_b)|^{1/2}|(\phi_c\phi_d|\phi_c\phi_d)|^{1/2} \leq \ \texttt{TwoENeglect} \end{aligned} \tag{2.8}$$

must be satisfied instead.

In Ref. [75], it is argued that in order for a given ERI, $(\phi_a\phi_b|\phi_c\phi_d)$, to be considered significant it is necessary that ϕ_a and ϕ_b (and likewise ϕ_c and ϕ_d) have significant overlap. This implies that the basis function pair ϕ_a and ϕ_b are spatially close. Since the ERI is contracted with D_{bd}, D_{bc}, D_{ad}, and D_{ac}, this also implies that the index pairs of these density matrix elements all share similar (but not necessarily small) spatial separations. A central premise of Ref. [75] is that since all four elements of **D** involve similar separations, it is not unreasonable to assume that they also share similar magnitudes. However, this assumption is contradicted by Figures 1.1 and 1.2, which demonstrate that elements of **D** sharing identical spatial separations vary in magnitude by up to five orders.

```
do a = 1,N
   N_a^dis = 0
   do b = 1,N
      if (overlap of φ_a with φ_b > TwoENeglect) then
         N_a^dis =  N_a^dis + 1
         add φ_b to distribution list a
      endif
   enddo b
   sort distribution list a by decreasing overlap
enddo a
```

Figure 2.1: Ordering the list of significant distributions.

Although the evaluation of ERIs dominates the formation of **K**, there are additional steps that have a significant computational expense. For instance, once the ERIs have been calculated it is necessary to contract them with the appropriate elements of **D** and accumulate the contributions to form **K**. In some cases these steps, referred to here as integral digestion, can represent up to 20 percent of the total time needed to form **K**. This is important because if the observed permutational symmetry factor is less than four, then in methods that use permutational symmetry there will be extra integral digestions that do not occur in ONX. For example, if a given calculation reveals that using permutational symmetry only reduces the number of calculated ERIs by a half rather than the full factor of four, then this means that there will also be a two-fold *increase* in the number of integral digestions. We also note that when sparse matrix storage routines are used to represent **D** and **K**, which is a necessity for large-scale calculations, the relative cost of extra integral digestions increases because accessing elements of a sparse matrix is less efficient than for a dense matrix.

2.2 Methods

In order to investigate the importance of exploiting ERI permutational symmetry in formation of the HF exchange matrix, we have incorporated ERI permutational symmetry as described in Ref.[75] into ONX, which we will refer to here as symmetrized ONX (SONX) to indicate that we have used the same ERI routines as in ONX. In the following, we briefly outline the main components of the method as it is implemented in SONX.

The loop skip-out statements introduced in Chapter 1 can be combined with ERI permutational symmetry through ordered distribution lists in conjunction with ordered lists of density matrix elements. The formation of these ordered distribution and density lists is outlined in Figs. 2.1 and 2.2 respectively.

As shown in Fig. 2.3, the construction of **K** begins with an outer loop over the sorted distribution list $\{\phi_a \phi_b\}_a$, where the subscript indicates the index that is held fixed. The density ordered list corresponding to the index a is then used to select the index c, for which

```
do a = 1,N
   N_a^den = 0
   do c = 1,N
      if( |D_ac| * |(φ_a φ_a|φ_a φ_a)|^(1/2) * |(φ_c φ_c|φ_c φ_c)|^(1/2) > TwoENeglect) then
         N_a^den =  N_a^den + 1
         add φ_c to density list a
      endif
   enddo c
   sort density list a by decreasing |D_ac| * |(φ_c φ_c|φ_c φ_c)|^(1/2)
enddo a
```

Figure 2.2: Ordering the list of significant elements of the density matrix.

the ordered distribution list $\{\phi_c\phi_d\}_c$ is iterated over. Each $|\phi_c\phi_d)$ in the distribution list $\{\phi_c\phi_d\}_c$ that satisfies the criterion

$$|D_{ac}||(\phi_a\phi_b|\phi_a\phi_b)|^{1/2}|(\phi_c\phi_d|\phi_c\phi_d)|^{1/2} \leq \texttt{TwoENeglect} \tag{2.9}$$

is stored in a temporary ERI list specific to a. When Eq. (2.9) is not satisfied, the innermost loop can be aborted safely because the distribution list has been ordered. Due to the sorted density lists, the loop over index c may also be aborted safely when the inner loops do not alter the temporary ERI list. This process is repeated in a second set of loops except now the density ordered list corresponding to index b is used to select index d, and the resulting distributions $|\phi_c\phi_d)$ are stored in a temporary ERI list specific to b.

Once the temporary ERI lists have been formed, they are sorted and merged together so that all redundancies are removed. The combined ERI list is then iterated over and the ERIs $(\phi_a\phi_b|\phi_c\phi_d)$ are computed and contracted with $\mathbf{D}$ to form $\mathbf{K}$.

```
  do $a$ = 1,$N$
    do $b$ = 1,$N_a^{\mathrm{dis}}$
      do $c$ = 1,$N_a^{\mathrm{den}}$
        do $d$ = 1,$N_c^{\mathrm{dis}}$
          if ($|D_{ac}| * |(\phi_a\phi_b|\phi_a\phi_b)|^{1/2} * |(\phi_c\phi_d|\phi_c\phi_d)|^{1/2}$ < TwoENeglect) then
            goto 100
          else
            add $|\phi_c\phi_d)$ to ERI list $a$
          endif
        enddo $d$
100     continue
        if( ERI list $a$ has not changed ) goto 200
      enddo $c$
200   continue
      do $d$ = 1,$N_b^{\mathrm{den}}$
        do $c$ = 1,$N_d^{\mathrm{dis}}$
          if ($|D_{bd}| * |(\phi_a\phi_b|\phi_a\phi_b)|^{1/2} * |(\phi_c\phi_d|\phi_c\phi_d)|^{1/2}$ < TwoENeglect) then
            goto 300
          else
            add $|\phi_c\phi_d)$ to ERI list $b$
          endif
        enddo $c$
300     continue
        if( ERI list $b$ has not changed ) goto 400
      enddo $d$
400   continue
      combine ERI list $a$ and $b$ into single ERI list
      sort and remove duplicates in ERI list
      do $c,d$ = 1, Length of ERI list
        compute $(\phi_a\phi_b|\phi_c\phi_d)$
        contract $(\phi_a\phi_b|\phi_c\phi_d)$ with $D_{bd}$, $D_{bc}$, $D_{ad}$, and $D_{ac}$
        add contributions to $K_{ac}$, $K_{ad}$, $K_{bc}$, and $K_{bd}$
      enddo $c,d$
    enddo $b$
  enddo $a$
```

Figure 2.3: The LinK algorithm of Ref. [75] as implemented in SONX.

Table 2.1: The total number of ERIs calculated by ONX and SONX in a RHF/STO-2G calculation of a cluster of 30 water molecules and with different levels of `TwoENeglect`.

	Number of ERIs		
`TwoENeglect`	ONX	SONX	Ratio
1.0e-5	1.7e6	1.0e6	1.7
1.0e-7	6.3e6	3.3e6	1.9
1.0e-9	1.6e7	7.5e6	2.1
1.0e-11	3.3e7	1.4e7	2.3
1.0e-13	5.7e7	2.2e7	2.6
1.0e-15	8.7e7	3.1e7	2.8

Table 2.2: The total number of ERIs calculated by ONX and SONX in a RHF/STO-2G calculation of circumcoronene, $C_{24}H_{12}$, and with different levels of `TwoENeglect`.

	Number of ERIs		
`TwoENeglect`	ONX	SONX	Ratio
1.0e-5	1.6e7	5.8e6	2.8
1.0e-7	3.3e7	1.1e7	3.0
1.0e-9	5.0e7	1.5e7	3.3
1.0e-11	6.7e7	1.9e7	3.5
1.0e-13	8.3e7	2.3e7	3.6
1.0e-15	9.6e7	2.6e7	3.7

2.3 Results and discussion

In the following sections we present several ONX and SONX benchmark calculations, which illustrate the speedups achieved by exploiting ERI permutational symmetry in computation of the HF exchange matrix. All calculations have been carried out within the MondoSCF[76] suite of programs on a single 332 MHz 604e PowerPC processor.

2.3.1 Reduction in ERIs

It is first of interest to determine to what extent utilizing permutational symmetry actually reduces the number of ERIs that need to be calculated. This can be determined by taking the ratio of the number of ERIs that are calculated by ONX versus the number calculated by SONX for identical systems and levels of accuracy.

To examine this ratio in detail we have performed a series of RHF/STO-2G calculations on a three dimensional cluster of 30 water molecules and on circumcoronene ($C_{25}H_{12}$) at different levels of integral prescreening, denoted by `TwoENeglect`. As can be seen from Table 2.1, the actual reduction in the number of ERIs for the water cluster with reasonable values of `TwoENeglect` (10^{-5} to 10^{-10}) is only one half of the full permutational factor. For the circumcoronene calculation, shown in Table 2.2, the reduction in ERIs is increased and approaches the full permutational factor of four with similar values of `TwoENeglect`.

It appears that the reductions in ERIs for the cluster of 30 water molecules and circumcoronene differ substantially because the long-range behavior of $\mathbf{D}$ is controlled by the system's HOMO-LUMO gap, as discussed in Chapter 1. For much larger systems, similar differences in the asymptotic decay of $\mathbf{D}$ are illustrated in Figs. 1.1 and 1.2. Note that the choice of basis set can also have an effect on the HOMO-LUMO gap, which in turn alters the long-range exchange interactions[9]. However, similar calculations show that the observed reduction in ERIs is less dependent on the choice of basis sets than on the physical properties of the system.

2.3.2 Reduction in CPU time

Although ONX and SONX involve somewhat different approaches in terms of how the ERIs are looped over, both are based on the same implementation of the HGP method for computing ERIs[19]. As a result, a meaningful comparison can be made between the CPU times that each method requires.

Table 2.3: The total CPU time required to form one RHF exchange matrix for a cluster of 50 water molecules by ONX and SONX with different basis sets and with `TwoENeglect` $= 10^{-7}$. Speedup is the ratio of the ONX and the SONX CPU times and $\tilde{K}$ is the average basis set contraction length for the ERIs calculated in SONX[77].

	CPU Time (sec)			
Basis Set	ONX	SONX	Speedup	$\tilde{K}$
STO-2G	20	13	1.5	1.9
STO-3G	90	50	1.8	2.6
STO-6G	1163	613	1.9	4.8
3-21G	91	86	1.1	1.4
6-31G	290	209	1.4	1.8
6-31G*	599	491	1.2	1.5
6-31G**	979	854	1.1	1.4
Dunning DZ	621	474	1.3	1.8
Dunning-Hay SV	697	462	1.5	1.9

References for basis sets used:
STO-2G, STO-3G, and STO-6G [78], 3-21G [79], 6-31G [80], 6-31G* and 6-31G** [81] , Dunning DZ and Dunning-Hay SV [82]

In Table 2.3 the total CPU time incurred by ONX and SONX to compute the RHF exchange matrix of a cluster of 50 water molecules is shown for the same series of basis sets as in Table 2.1. The corresponding set of calculations for the circumcoronene system is given in Table 2.4. A comparison of the observed speedups obtained for these systems reveals that the reduction in CPU time is fairly dependent on the choice of basis sets. Given that the ratio of ERIs does not change significantly between different basis sets, this observation is somewhat surprising. However, it may be explained by noting that when contracted basis functions are used, as in Eq. (2.1), the computation of ERIs involves a four-fold summation

Table 2.4: The total CPU time required to form one RHF exchange matrix for circumcoronene by ONX and SONX with different basis sets and with `TwoENeglect` $= 10^{-7}$. Speedup is the ratio of the ONX and the SONX CPU times and $\tilde{K}$ is the average basis set contraction length for the ERIs calculated in SONX.

	CPU Time (sec)			
Basis Set	ONX	SONX	Speedup	$\tilde{K}$
STO-2G	24	9	2.7	1.9
STO-3G	94	30	3.1	2.6
STO-6G	1076	315	3.4	4.7
3-21G	144	72	2.0	1.4
6-31G	350	148	2.4	1.9
6-31G*	1838	706	2.6	1.4
6-31G**	1971	783	2.5	1.4
Dunning DZ	1316	588	2.2	1.9
Dunning-Hay SV	1415	550	2.6	2.0

over K^4 primitive ERIs

$$(\phi_a\phi_b|\phi_c\phi_d) = \sum_i^{K_a}\sum_j^{K_b}\sum_k^{K_c}\sum_l^{K_d} d_{ai}d_{bj}d_{ck}d_{dl}\,(\varphi_{ai}\varphi_{bj}|\varphi_{ck}\varphi_{dl})\,. \tag{2.10}$$

For basis sets that involve large values of K, the cost of evaluating individual ERIs rapidly increases as $\mathcal{O}(K^4)$. This is in contrast to other steps in formation of the exchange matrix, such as integral digestion, that do not depend on K since they operate solely on contracted quantities. The high cost of ERI contraction is relevant to the comparison of ONX to SONX because when K is large, any reduction in the number of ERIs via permutational symmetry will have a much larger effect on the total CPU time than when K is small. In Tables 2.3 and 2.4 the average contraction length, $\tilde{K}$, for each basis set is shown. $\tilde{K}$ is less than K because products of primitive basis functions are discarded based on the criterion

$$\int d\mathbf{r}\; \varphi_{ai}(\mathbf{r})\varphi_{bi}(\mathbf{r}) \leq \texttt{DistNeglect}, \tag{2.11}$$

where `DistNeglect`=`TwoENeglect`$\times 10^{-2}$. Inspection of Tables 2.3 and 2.4 reveals that the benefit of using ERI permutational symmetry correlates well with $\tilde{K}$.

2.4 Conclusions

In this chapter, the effect of ERI permutational symmetry on the efficient formation of the HF exchange matrix has been examined. By incorporating the use of permutational symmetry in ONX, this effect has been measured in terms of the number of computed ERIs as well as the total CPU time required. An unexpected finding was that, depending on the basis set contraction length, a reduction in ERIs does not necessarily translate into a corresponding reduction in CPU time. A similar conclusion was not reached in Ref. [75]

because their comparison was based solely on the number of computed ERIs rather than on the total CPU time.

In systems such as three-dimensional clusters of water molecules, permutational symmetry is shown to reduce the number of ERIs computed by a factor of two. For basis sets that have long contraction lengths, a factor of two reduction in the CPU time is also observed. However, when basis sets involving low levels of contraction are used the reduction in the number of computed ERIs may be masked by the increased overhead needed to incorporate the permutational symmetry.

In systems that involve long range exchange interactions, such as graphitic sheets, the reduction in ERIs readily approaches a full factor of four. Although the observed CPU times remain basis set dependent, the use of permutational symmetry appears to be advantageous in all cases.

Chapter 3

Multipole acceptability criterion

3.1 Introduction

New computational methods have greatly extended the applicability of electronic structure theory. In particular, hierarchical methods such as treecodes [83, 84, 85, 86, 87] and the fast multipole method [88, 89, 90, 91, 92, 93], originally developed to solve the classical N-body problem, have recently been generalized to accelerate computation of the Coulomb matrix [22, 94, 95, 24, 96]. In addition, a number of non-multipole based methods have appeared for linear scaling computation of the Hartree-Fock exchange matrix [37, 38, 97, 75].

The efficiency of hierarchical multipole based methods depends to a large extent on how aggressively the multipole approximation is applied. When the approximation is used to represent a sufficiently large number of direct interactions a savings can be achieved. However, these savings do not come for free. As expansions are truncated at a finite order, errors accumulate, and must be controlled to ensure a reliable and accurate method.

Truncation errors are controlled primarily with a multipole acceptability criterion (MAC) in treecodes, or with a well separated index in the fast multipole method. Both MACs and well separated indices serve a similar purpose, which is to estimate the truncation error in a given multipole expansion. Since treecodes and fast multipole methods have their origin in classical particle simulations, error estimates have traditionally been developed for systems of interacting point (monopole) charges. However, in electronic structure calculations translations of higher order moments are encountered. In these situations it is common practice to use empirical error estimates based on expansions of monopole moments [22, 95, 24]. In continuous fast multipole methods such as CFMM[22] or GvFMM[95] an empirical error estimate is acceptable as interactions are tested on the basis of worst-case cell-cell interactions, where "cell" refers to a collection of particles represented by a multipole expansion. However in the quantum chemical tree code (QCTC)[24], MACs are applied on a finer level to multipole-cell interactions. In this case, substantial benefits may be gained by using non-empirical, numerically tight MACs.

The importance of tight MACs may be even greater for the direct evaluation of electron repulsion integrals (ERIs). For example, large savings can be achieved when individual ERIs are computed in a multipole representation [98, 99, 100, 101]. In electronic structure methods, ERIs are usually computed between products of Gaussian-type basis functions, also referred to as Gaussian distributions. The Gaussian product theorem states that the

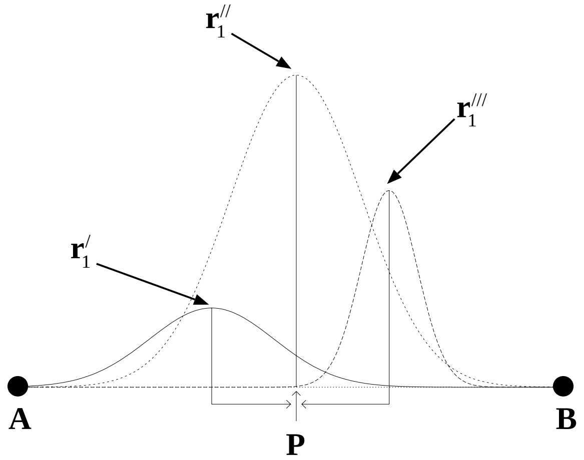

Figure 3.1: A set of primitive distributions located at $\mathbf{r}_1'$, $\mathbf{r}_1''$, and $\mathbf{r}_1'''$, which are translated with the multipole approximation to a common center $\mathbf{P}$ and contracted.

product of two Gaussians can be expressed exactly as a finite sum of one-center Gaussians. For example, the product of two s-type Gaussians is

$$e^{\zeta_a(\mathbf{r}-\mathbf{A})^2} e^{\zeta_b(\mathbf{r}-\mathbf{B})^2} = e^{\xi(\mathbf{A}-\mathbf{B})^2} e^{\zeta_{r_1}(\mathbf{r}-\mathbf{r}_1)^2}, \tag{3.1}$$

where $\xi = \zeta_a\zeta_b/\zeta_{r_1}$, $\mathbf{r}_1 = (\zeta_a\mathbf{A} + \zeta_b\mathbf{B})/\zeta_{r_1}$, and $\zeta_{r_1} = \zeta_a + \zeta_b$. In addition, basis functions may consist of a contraction of K primitive functions where each primitive has a different exponent, ζ_a. Rather than a single center, $\mathbf{r}_1$, contracted distributions form finite expansions at K^2 centers between the originating basis functions. Therefore, evaluation of a contracted ERI involves computation of K^4 primitive ERIs, which is a computationally demanding requirement.

One approach for avoiding the contraction problem involves using the multipole approximation to translate sets of primitive distributions about common centers [102, 67, 103]. This is illustrated in Figure 3.1, which shows a typical set of primitive distributions centered at $\mathbf{r}_1'$, $\mathbf{r}_1''$, and $\mathbf{r}_1'''$ between basis functions at $\mathbf{A}$ and $\mathbf{B}$. For those ERIs that involve well separated distributions, contractions can be precomputed at the level of basis function products instead of on primitive ERIs, which leads to a reduction in computational work from $\mathcal{O}(K^4)$ to $\mathcal{O}(1)$. As contraction lengths of high quality basis sets can be on the order of ten, this method offers the possibility of large improvements in the computation of ERIs.

In addition to avoiding contraction steps in ERI evaluation, the multipole representation also allows a straightforward reduction in the complexity of ERI computation with respect to their total angular symmetry. For a given total angular symmetry $\mathcal{L}$, spherical multipole interaction tensors can be used that involve only $\mathcal{O}(\mathcal{L}^2)$ work as compared to the $\mathcal{O}(\mathcal{L}^4)$ work required in corresponding Cartesian representations.

Since electronic structure methods involve continuous distributions, errors may also arise from penetration effects when a multipole representation is used [94]. This situation is illustrated in Figure 3.2. However, preliminary results indicate that penetration errors are secondary to truncation errors[104].

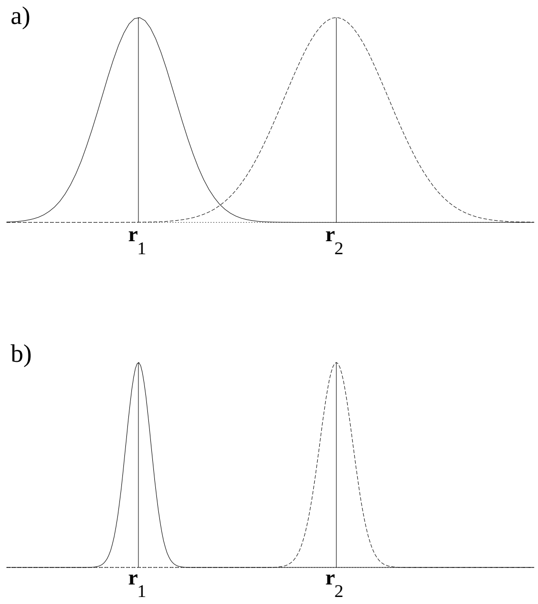

Figure 3.2: A schematic of penetration effects that can occur between continuous distributions. In a) there is significant penetration between the two distributions that is not accounted for in a multipole representation. In b) the distributions do not involve significant penetration and a multipole representation is acceptable.

Despite the important advantages to both hierarchical and direct methods for ERI evaluation, no attention has been given to the development of a tight non-empirical MAC for the high order moments encountered in quantum chemistry. In this chapter, truncated multipole expansions are examined in detail. Existing error estimates for the expansion of $|\mathbf{r}_1 - \mathbf{r}_2|^{-1}$ are reviewed and extended to high angular symmetry moments. In this development emphasis is placed on balancing numerical accuracy with computational simplicity so that the resulting MAC can be used effectively on the fly. The chapter is organized as follows: In the next Section, the multipole limit of Gaussian based ERIs is discussed. In Section 3.3.1, existing methods for the expansion and corresponding error estimate of $|\mathbf{r}_1 - \mathbf{r}_2|^{-1}$ are reviewed. Then, in Section 3.3.2, the error estimates are extended to high order moments and simplified to a form suitable for use as a MAC. In Section 3.4, several model systems similar to those found in electronic structure theory are used to demonstrate the performance of the new MAC. These results are discussed in Section 3.5, and in Section 3.6, we present our conclusions.

3.2 The multipole limit of ERIs

In modern quantum chemistry, the basis functions of choice are the Cartesian Gaussian type functions (CGTFs)[74],

$$\varphi_a(\mathbf{r}) = (r_x - A_x)^{l_a} (r_y - A_y)^{m_a} (r_z - A_z)^{n_a} \times \exp\left[-\zeta_a (\mathbf{r} - \mathbf{A})^2\right], \tag{3.2}$$

which enable the analytic evaluation of electron repulsion integrals (ERIs) [74, 15, 16, 17, 18, 19, 105, 20],

$$(\varphi_a \varphi_b | \varphi_c \varphi_d) = \int\int d\mathbf{r} d\mathbf{r}' \varphi_a(\mathbf{r}) \varphi_b(\mathbf{r}) \frac{1}{|\mathbf{r} - \mathbf{r}'|} \varphi_c(\mathbf{r}') \varphi_d(\mathbf{r}'). \tag{3.3}$$

Throughout this chapter, basis functions with $l_a + m_a + n_a = 0, 1, 2$ are referred to as s, p, and d-type respectively.

It is often convenient to represent products of CGTFs, which we will refer to as distributions, with exact one center expansions in Hermite Gaussian type functions (HGTFs):

$$\varphi_a(\mathbf{r}) \varphi_b(\mathbf{r}) = \sum_{L=0}^{l_a+l_b} \sum_{M=0}^{m_a+m_b} \sum_{N=0}^{n_a+n_b} e^{ab}_{LMN} \Lambda^{\mathbf{r}_1}_{LMN}(\mathbf{r}), \tag{3.4}$$

where e^{ab}_{LMN} are the expansion coefficients obtained with the McMurchie-Davidson two term recurrence relations[17] and multiplied by the appropriate radial overlap. The HGTFs centered at $\mathbf{r}_1$ are

$$\Lambda^{\mathbf{r}_1}_{LMN}(\mathbf{r}) = \frac{\partial^L}{\partial r^L_{1x}} \frac{\partial^M}{\partial r^M_{1y}} \frac{\partial^N}{\partial r^N_{1z}} \exp\left[-\zeta (\mathbf{r} - \mathbf{r}_1)^2\right]. \tag{3.5}$$

In the multipole limit, HGTFs are related to derivatives of the Dirac delta function by

$$\begin{aligned} \delta^{\mathbf{r}_1}_{LMN}(\mathbf{r}) &= \frac{\partial^L}{\partial r^L_{1x}} \frac{\partial^M}{\partial r^M_{1y}} \frac{\partial^N}{\partial r^N_{1z}} \delta(\mathbf{r} - \mathbf{r}_1) \\ &= \lim_{\zeta \to \infty} \left(\frac{\zeta}{\pi}\right)^{3/2} \Lambda^{\mathbf{r}_1}_{LMN}(\mathbf{r}), \end{aligned} \tag{3.6}$$

which enables ERIs over HGTFs to be evaluated with the Cartesian multipole interaction tensor[94, 106] as

$$\begin{aligned} (\Lambda^{\mathbf{r}_1}_{LMN} | \Lambda^{\mathbf{r}_2}_{L'M'N'}) &= (-1)^{L+M+N} \left(\frac{\pi}{\sqrt{\zeta_{ab}\zeta_{cd}}}\right)^3 \\ &\times \frac{\partial^{L+L'}}{\partial r^{L+L'}_{12x}} \frac{\partial^{M+M'}}{\partial r^{M+M'}_{12y}} \frac{\partial^{N+N'}}{\partial r^{N+N'}_{12z}} \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|}. \end{aligned} \tag{3.7}$$

3.3 Error estimates

Since matrix elements in electronic structure theory involve summation of interactions spanning many orders of magnitude, we focus on estimates of absolute error. Relative error estimates are non-optimal because computing small interactions to a high relative accuracy leads to unnecessary computational work that has little or no effect on the total accumulated error of each matrix element.

Figure 3.3: The interaction of two particles located at $\mathbf{r}_1$ and $\mathbf{r}_2$, which is represented with a bipolar expansion about the centers $\mathbf{P}$ and $\mathbf{Q}$. The translations $|\mathbf{r}_1 - \mathbf{P}|$ and $|\mathbf{r}_2 - \mathbf{Q}|$ are assumed to be smaller than the interaction distance $|\mathbf{P} - \mathbf{Q}|$.

3.3.1 Estimates for $|\mathbf{r}_1 - \mathbf{r}_2|^{-1}$

The truncated bipolar expansion of $|\mathbf{r}_1 - \mathbf{r}_2|^{-1}$ about centers $\mathbf{P}$ and $\mathbf{Q}$, as in Figure 3.3, can be carried out in several different representations [106, 107, 108, 109, 110]. For example, in terms of the associated Legendre polynomials, $P_l^m(x)$, the truncated expansion is[25]

$$\begin{aligned} \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} &\approx \sum_{l_p=0}^{p} (-1)^{l_p} \sum_{m_p=-l_p}^{l_p} O_{l_p}^{m_p}[\mathbf{r}_1 - \mathbf{P}] \\ &\times \sum_{l_q=0}^{q} \sum_{m_q=-l_q}^{l_q} M_{l_p+l_q}^{m_p+m_q}[\mathbf{P} - \mathbf{Q}]\, O_{l_q}^{m_q}[\mathbf{r}_2 - \mathbf{Q}] \end{aligned} \tag{3.8}$$

where

$$O_l^m[\mathbf{r}] = \frac{r^l}{(l+m)!} P_l^m(\cos\theta_{\mathbf{r}})\, e^{-im\phi_{\mathbf{r}}} \tag{3.9}$$

and

$$M_l^m[\mathbf{R}] = \frac{(l-m)!}{R^{l+1}} P_l^m(\cos\theta_{\mathbf{R}})\, e^{im\phi_{\mathbf{R}}}. \tag{3.10}$$

The orders of expansion about the new centers $\mathbf{P}$ and $\mathbf{Q}$ are p and q respectively. Expansions in both Cartesian and spherical representations that are truncated at the same level yield equivalent errors.

Recently, the error of expanding $|\mathbf{r}_1 - \mathbf{r}_2|^{-1}$ in the context of the fast multipole method has been thoroughly investigated[111, 112, 113]. For expansions truncated at order p and q, the absolute error

$$\epsilon = \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} - \sum_{l_p=0}^{p} (-1)^{l_p} \sum_{m_p=-l_p}^{l_p} O_{l_p}^{m_p} [\mathbf{r}_1 - \mathbf{P}]$$

$$\times \sum_{l_q=0}^{q} \sum_{m_q=-l_q}^{l_q} M_{l_p+l_q}^{m_p+m_q} [\mathbf{P} - \mathbf{Q}] O_{l_q}^{m_q} [\mathbf{r}_2 - \mathbf{Q}] \quad (3.11)$$

may be partitioned as

$$\epsilon = \epsilon_1 + \epsilon_2 + \epsilon_{12} \quad (3.12)$$

where

$$\epsilon_1 = \sum_{l=p+1}^{\infty} \frac{|\mathbf{r}_1 - \mathbf{P}|^l}{|\mathbf{P} - \mathbf{r}_2|^{l+1}} P_l(\cos\theta_1), \quad (3.13)$$

$$\epsilon_2 = \sum_{l=q+1}^{\infty} \frac{|\mathbf{r}_2 - \mathbf{Q}|^l}{|\mathbf{Q} - \mathbf{r}_1|^{l+1}} P_l(\cos\theta_2), \quad (3.14)$$

and $\cos\theta_1 = \left(\widehat{\mathbf{P} - \mathbf{r}_1} \cdot \widehat{\mathbf{P} - \mathbf{r}_2}\right)$. Equations (3.13) and (3.14) correspond to one-center expansion errors where the translations $(\mathbf{r}_1 - \mathbf{P})$ and $(\mathbf{r}_2 - \mathbf{Q})$ in Figure 3.3 are considered independently. The quantity ϵ_{12} represents a cross term in the bipolar expansion error that becomes significant only with $p, q \geq 10$[113]. Equations (3.13) and (3.14) provide a numerically tight estimate and are useful for estimating the total error in the fast multipole method for a given expansion order.

To balance accuracy and computational simplicity, we replace the Legendre polynomials in Eqs. (3.13) and (3.14) with the bound $P_l(\cos\theta) \leq 1$ [25, 114]. This is equivalent to assuming a one-dimensional worst-case orientation where translations are in opposite directions pointing away from each other. The summations in Eqs. (3.13) and (3.14) can then be evaluated exactly to yield the compact forms

$$\epsilon_1 \leq \left(\frac{1}{|\mathbf{r}_2 - \mathbf{P}| - |\mathbf{r}_1 - \mathbf{P}|}\right) \left(\frac{|\mathbf{r}_1 - \mathbf{P}|}{|\mathbf{r}_2 - \mathbf{P}|}\right)^{p+1}, \quad (3.15)$$

and

$$\epsilon_2 \leq \left(\frac{1}{|\mathbf{r}_1 - \mathbf{Q}| - |\mathbf{r}_2 - \mathbf{Q}|}\right) \left(\frac{|\mathbf{r}_2 - \mathbf{Q}|}{|\mathbf{r}_1 - \mathbf{Q}|}\right)^{q+1}. \quad (3.16)$$

Further simplifications of ϵ_1 and ϵ_2 are possible by assuming that translations are much smaller than interaction distances (see Figure 3.3);

$$|\mathbf{r}_1 - \mathbf{P}| \ll |\mathbf{P} - \mathbf{Q}| \text{ and } |\mathbf{r}_2 - \mathbf{Q}| \ll |\mathbf{P} - \mathbf{Q}|. \quad (3.17)$$

This assumption leads to the following approximate expression for the total error

$$\epsilon \approx \frac{|\mathbf{r}_1 - \mathbf{P}|^{p+1}}{|\mathbf{r}_2 - \mathbf{P}|^{p+2}} + \frac{|\mathbf{r}_2 - \mathbf{Q}|^{q+1}}{|\mathbf{r}_1 - \mathbf{Q}|^{q+2}} \quad . \quad (3.18)$$

Equation (3.18) is a tight and relatively cheap estimate of the absolute error introduced by truncating the bipolar expansion of $|\mathbf{r}_1 - \mathbf{r}_2|^{-1}$.

3.3.2 Extension to arbitrary moments

In the evaluation of HGTF-based ERIs, Eq. (3.7), $|\mathbf{r}_1 - \mathbf{r}_2|^{-1}$ can be replaced with a bipolar expansion about two new centers as in Eq. (3.8). The corresponding error that the truncated expansion involves can be estimated by replacing $|\mathbf{r}_1 - \mathbf{r}_2|^{-1}$ in Eq. (3.7) with Eq. (3.18) to yield

$$\begin{aligned} \epsilon \approx\ & (-1)^{L+M+N+L'+M'+N'} \left(\frac{\pi}{\sqrt{\zeta_{ab}\zeta_{cd}}}\right)^3 \\ & \times \frac{\partial^L}{\partial r_{1x}^L} \frac{\partial^M}{\partial r_{1y}^M} \frac{\partial^N}{\partial r_{1z}^N} \frac{\partial^{L'}}{\partial r_{2x}^{L'}} \frac{\partial^{M'}}{\partial r_{2y}^{M'}} \frac{\partial^{N'}}{\partial r_{2z}^{N'}} \\ & \times \left(\frac{|\mathbf{r}_1 - \mathbf{P}|^{p+1}}{|\mathbf{r}_2 - \mathbf{P}|^{p+2}} + \frac{|\mathbf{r}_2 - \mathbf{Q}|^{q+1}}{|\mathbf{r}_1 - \mathbf{Q}|^{q+2}} \right) . \end{aligned} \tag{3.19}$$

Equation (3.19) is prohibitively expensive for evaluation on the fly. To achieve an inexpensive MAC we once again ignore angular effects and assume a one-dimensional orientation to arrive at

$$\begin{aligned} \epsilon \approx\ & \left(\frac{\pi}{\sqrt{\zeta_{ab}\zeta_{cd}}}\right)^3 \\ & \times \left[(-1)^{\mathcal{L}'} \mathcal{M}_1^p [\mathcal{L}, \mathcal{L}'] + (-1)^{\mathcal{L}} \mathcal{M}_2^q [\mathcal{L}', \mathcal{L}] \right], \end{aligned} \tag{3.20}$$

where

$$\begin{aligned} \mathcal{M}_1^p [0,0] &= \frac{|\mathbf{r}_1 - \mathbf{P}|^{p+1}}{|\mathbf{r}_2 - \mathbf{P}|^{p+2}}, \\ \mathcal{M}_1^p [i,0] &= \mathcal{M}_1^p [i-1,0] \frac{(p-i+2)}{|\mathbf{r}_1 - \mathbf{P}|}, \quad i > 0 \\ \mathcal{M}_1^p [i,j] &= \mathcal{M}_1^p [i,j-1] \frac{(p+j+1)}{|\mathbf{r}_2 - \mathbf{P}|}, \quad j > 0 . \end{aligned} \tag{3.21}$$

By contraction with the MD two-term expansion coefficients, as in Eq. (3.4), and summing the magnitude of each component, a MAC applicable to CGTF based integrals involving distributions of angular symmetry $l_{ab} = (l_a + l_b + m_a + m_b + n_a + n_b)$ is

$$\begin{aligned} & \mathrm{MAC}[l_{ab}, l_{cd}] = \\ & \quad \sum_{i=0}^{l_{ab}} \sum_{j=0}^{l_{cd}} c_i^{ab} c_j^{cd} \left(|\mathcal{M}_1^p [i,j]| + |\mathcal{M}_2^q [j,i]| \right) \end{aligned} \tag{3.22}$$

where

$$c_i^{ab} = \max_{L+M+N=i} \left[\left| e_{LMN}^{ab} \right| \right] \left(\frac{\pi}{\zeta_{ab}} \right)^{3/2} . \tag{3.23}$$

It should be noted that Eq. (3.23) can be precomputed and contracted with the appropriate translation distances ($|\mathbf{r}_1 - \mathbf{P}|$ and $|\mathbf{r}_2 - \mathbf{Q}|$) at the level of distributions, which enables Eq. (3.22) to be efficiently applied at the level of interactions.

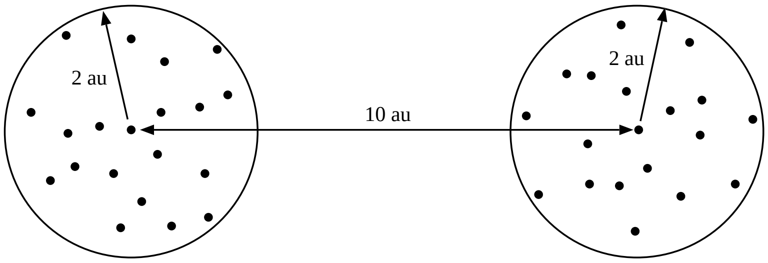

Figure 3.4: A schematic of the distribution spheres used to investigate the performance of the MACs. The distribution centers are represented by the points within each sphere.

3.4 Results

Previous empirical estimates of the truncated multipole expansion error have been based on approximating translation errors of arbitrary angular symmetry moments with expressions derived for the lowest order case [22, 95, 24]. To compensate for neglecting the effect of higher order terms, an empirically determined safety or fudge factor is usually included in the estimate. In terms of the MAC developed here, this approach is equivalent to using the $i = j = 0$ component only in Eq. (3.22) to describe higher angular symmetry moments. In the following sections we examine the overall performance as well as the specific cases where $\mathrm{MAC}[l_{ab}, l_{cd}]$ offers a significant advantage over $\mathrm{MAC}[0, 0]$.

3.4.1 Statistical performance

To investigate the statistical performance of the MACs we have constructed several model systems. Each system consists of 100 distributions of particular angular symmetry. The distributions were formed from two-center products of normalized CGTFs that were chosen randomly within a sphere of a given radius. The exponents of the CGTFs were selected as $\zeta = 10^n$ with n distributed uniformly in the range $[-1, 3]$. The radial overlap of the CGTF products used were required to be greater than 10^{-7}. Random expansion centers for each distribution were chosen to be within a given distance from the original distribution centers. Hereafter, the distance between expansion and distribution centers, $|\mathbf{r}_1 - \mathbf{P}|$ in Figure 3.1, will be referred to as the translation distance.

The relative performance of $\mathrm{MAC}[l_{ab}, l_{cd}]$ and $\mathrm{MAC}[0, 0]$ were compared by observing the order of magnitude deviation of each MAC from the exact multipole expansion error,

$$\mathcal{D} = \mathrm{Log}_{10}\left(\frac{\mathrm{error}}{\mathrm{MAC}}\right). \tag{3.24}$$

Positive values of $\mathcal{D}$ correspond to an underestimation and negative values to an overestimation of the expansion error.

Tables 3.1 and 3.2 summarize the statistical behavior of $\mathrm{MAC}[l_{ab}, l_{cd}]$ and $\mathrm{MAC}[0, 0]$ for a variety of integral types. In each case the systems consist of the two interacting spheres shown in Figure 3.4, which have radii of 2 atomic units (au) and centers that are separated by 10 au. Random translation distances for each distribution were chosen to be within the

Table 3.1: Statistics of MAC[l_{ab}, l_{cd}] for various integral types with $p = l_{ab}$ and $q = l_{cd}$. Distributions are in spheres of radius 2 au, which have centers separated by 10 au. Translations are restricted to [0,2] and exponents to [0.1,1000]. Max, Avg, and δ are the maximum, average, and standard deviation of $\mathcal{D}$.

Type	p	q	Max($\mathcal{D}$)	Avg($\mathcal{D}$)	δ
$(ss\|ss)$	0	0	0.14	-0.5	0.4
$(ps\|ss)$	1	0	0.17	-0.5	0.5
$(pp\|ss)$	2	0	0.20	-0.5	0.5
$(ps\|ps)$	1	1	0.21	-0.7	0.5
$(pp\|ps)$	2	1	0.23	-0.8	0.5
$(pp\|pp)$	2	2	0.24	-1.0	0.6
$(dd\|dd)$	4	4	0.19	-1.2	0.6

range $[0, 2]$ au, such that the resulting expansion centers remain within the distribution sphere.

For the systems examined in Tables 3.1 and 3.2, the multipole expansions were truncated at the same level ($p = l_{ab}$ and $q = l_{cd}$) as the largest non-zero untranslated distribution moment. In Tables 3.3 and 3.4 the effects of increasing the expansion level on the performance of MAC[l_{ab}, l_{cd}] and MAC[$0, 0$] are shown for the integral types $(pp|pp)$ and $(dd|dd)$.

3.4.2 Effect of translation distance

For Tables 3.1 to 3.4, the model systems employed involve a mixture of many different exponent sizes, translation distances, as well as orientations. It is therefore difficult to determine the exact conditions for which the MACs perform poorly. In order to examine these issues in more detail we have constructed a simple one-dimensional system. The system consists of a single integral of type $(pp|pp)$ located along the z-axis. The centers of the $(pp|$ and the $|pp)$ distributions are separated by 10 au. Figure 3.5 illustrates the behavior of $\mathcal{D}$ for MAC[l_{ab}, l_{cd}] and MAC[$0, 0$] as the translation distance, $|\mathbf{r}_1 - \mathbf{P}|$, is increased. This dependence of the MACs on the translation distance holds for other integral types and

Table 3.2: Statistics of MAC[$0, 0$] for various integral types. Distributions are in spheres of radius 2 au, which have centers separated by 10 au. Translations are restricted to [0,2] and exponents to [0.1,1000].

Type	p	q	Max($\mathcal{D}$)	Avg($\mathcal{D}$)	δ
$(ss\|ss)$	0	0	0.14	-0.5	0.4
$(ps\|ss)$	1	0	0.68	-0.5	0.5
$(pp\|ss)$	2	0	0.53	-0.6	0.6
$(ps\|ps)$	1	1	1.67	-0.5	0.5
$(pp\|ps)$	2	1	2.14	-0.5	0.6
$(pp\|pp)$	2	2	2.42	-0.5	0.7
$(dd\|dd)$	4	4	4.63	-0.2	0.8

Table 3.3: Statistics of MAC$[l_{ab}, l_{cd}]$ for the integral types $(pp|pp)$ and $(dd|dd)$ with increasing q. Distributions are in spheres of radius 2 au, which are separated by 10 au. Translations of the $(pp|$ and $(dd|$ distributions are limited to the range $[0, 1]$ and the $|pp)$ and $|dd)$ distributions to the range $[0, 4]$.

Type	p	q	Max($\mathcal{D}$)	Avg($\mathcal{D}$)	δ
$(pp\|pp)$	2	2	0.13	-0.9	0.5
	2	4	0.24	-1.0	0.5
	2	6	0.29	-0.9	0.5
	2	8	0.31	-0.9	0.5
$(dd\|dd)$	4	4	-0.15	-1.3	0.5
	4	6	-0.13	-1.2	0.5
	4	8	-0.28	-1.4	0.5

Table 3.4: Statistics of MAC$[0, 0]$ for the integral $(pp|pp)$ and $(dd|dd)$ with increasing q. Distributions are in spheres of radius 2 au, which are separated by 10 au. Translations of the $(pp|$ and $(dd|$ distributions are limited to the range $[0, 1]$ and the $|pp)$ and $|dd)$ distributions to the range $[0, 4]$.

Type	p	q	Max($\mathcal{D}$)	Avg($\mathcal{D}$)	δ
$(pp\|pp)$	2	2	1.35	-1.0	0.6
	2	4	2.31	-0.8	0.7
	2	6	2.33	-0.5	0.8
	2	8	2.34	-0.4	0.8
$(dd\|dd)$	4	4	4.32	-0.9	0.7
	4	6	5.20	-0.7	0.9
	4	8	5.33	0.2	1.0

orientations as well and becomes more pronounced for MAC$[0, 0]$ as the angular symmetry is increased. In Table 3.5 the behavior of MAC$[0, 0]$ is demonstrated for three-dimensional systems of $(pp|pp)$ integrals, for which the random translations have been restricted to the range $[t_{\min}, t_{\max}]$.

3.5 Discussion

Inspection of Table 3.1 reveals that MAC$[l_{ab}, l_{cd}]$ performs quite well for the integral types studied, and is on average less than an order of magnitude from the exact expansion error. In particular, $\mathcal{D}$ is essentially always less than zero, which indicates that the error is not underestimated. The average value of $\mathcal{D}$ is shifted below zero because angular factors have been neglected.

As shown in Table 3.2, MAC$[0, 0]$ is not able to effectively represent the expansion error for higher angular symmetry moments. In the highest angular symmetry case considered here, $(dd|dd)$, MAC$[0, 0]$ underestimates the expansion error by nearly five orders of magnitude. Although the average deviation of MAC$[0, 0]$ remains small, the large positive values

Table 3.5: Statistics of MAC$[0,0]$ for the integral type $(pp|pp)$. Distributions are in spheres of radius 3 au, which have centers separated by 10 au. Translations are restricted to the interval $[t_{\min}, t_{\max}]$.

$t_{\min}$	$t_{\max}$	Max($\mathcal{D}$)	Avg($\mathcal{D}$)	δ
0.00	0.25	4.4	0.5	0.8
0.25	0.50	1.4	-0.1	0.6
0.50	0.75	0.9	-0.3	0.5
0.75	1.00	0.8	-0.5	0.5
1.00	2.00	0.8	-0.7	0.6
2.00	3.00	0.7	-0.8	0.6

of $\mathcal{D}$ make it useless as an estimate for higher order moments.

Figure 3.5 dramatically shows how the translation distance can affect the performance of the MAC. As expected, MAC$[l_{ab}, l_{cd}]$ remains accurate for the entire range of translation distances. However, for MAC$[0,0]$ it becomes clear that large positive values of $\mathcal{D}$ occur primarily for small translations. As the translation distance is increased the performance of MAC$[0,0]$ also increases. This behavior of MAC$[0,0]$ is demonstrated again for a more general three-dimensional distribution of $(pp|pp)$ type integrals in Table 3.5. By restricting minimum as well as maximum allowed translation distances, the ability of MAC$[0,0]$ to estimate expansion errors of high order moments can be improved to an acceptable level.

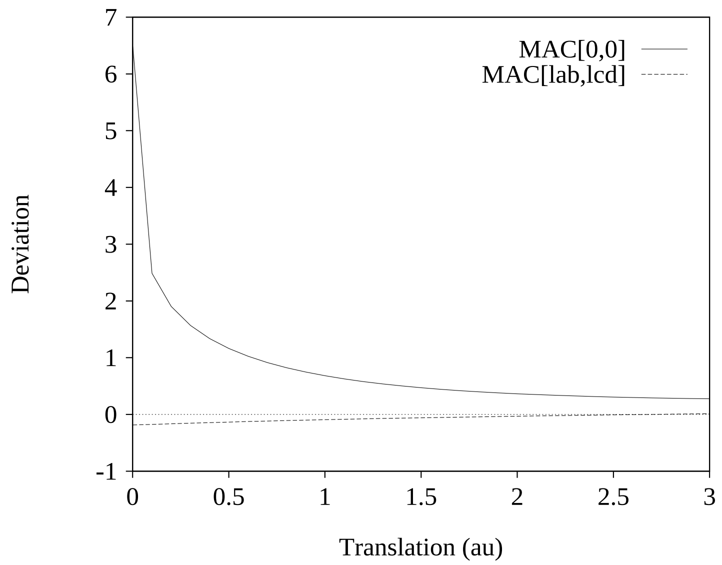

Figure 3.5: Deviation, $\mathcal{D}$, of MAC$[l_{ab}, l_{cd}]$ and MAC$[0,0]$ for the integral type $(pp|pp)$ at various translations of $(pp|$ in one-dimension. The distribution exponents are $\zeta = 1.0$ and expansions are truncated at $p = q = 2$.

3.6 Conclusions

Fast methods for computing long range interaction potentials are rapidly changing the field of electronic structure theory. Importantly, these methods are able to achieve a reduced computational complexity without sacrificing numerical accuracy. However, their efficiency depends critically on accurate estimates of the errors incurred. This situation is quite similar to direct SCF methods [70], which were greatly enhanced with the advent of rigorous integral prescreening techniques[14, 115].

The new MAC developed here represents a significant improvement for estimating the expansion error of high order moments. By enabling the aggressive application of the multipole approximation, the MAC is expected to have a large impact on hierarchical methods such as QCTC[24] and direct methods such as ONX[97] and LinK[75]. Although the emphasis has been on moments originating from continuous charge distributions, the basic results apply as well to multipole interactions that arise in semiempirical[116] and classical [93] methods.

The increases in accuracy afforded by the new MAC need to be weighed against the increased cost of its evaluation. The largest improvements in accuracy are shown to occur with systems that involve small translation distances, which are characteristic of the finest levels of spatial decomposition in hierarchical multipole methods, and of multipole based methods for direct ERI evaluation in general.

Chapter 4

Multipole accelerated exchange

4.1 Introduction

As mentioned in Chapter 2, the use of contracted CGTFs

$$\phi_a\left(\mathbf{r}\right)=\sum_i^{K_a} d_{ai}\varphi_{ai}\left(\mathbf{r}\right) \tag{4.1}$$

leads to expensive four-fold contraction loops in the evaluation of ERIs

$$(\phi_a\phi_b|\phi_c\phi_d)=\sum_i^{K_a}\sum_j^{K_b}\sum_k^{K_c}\sum_l^{K_l} d_{ai}d_{bj}d_{ck}d_{dl}(\varphi_{ai}\varphi_{bj}|\varphi_{ck}\varphi_{dl}). \tag{4.2}$$

Over the past 20 years, many different methods have been developed to efficiently compute ERIs [15, 16, 17, 18, 19, 105, 20]. For example, by employing specialized recursion relations some fraction of the ERI evaluation can be moved outside of the contraction summations[19, 20]. In addition, it is possible to reduce the average contraction length, $\tilde{K}$, of basis function products,

$$\phi_a(\mathbf{r})\phi_b(\mathbf{r})=\sum_i^{K_a}\sum_j^{K_b}\varphi_{ai}(\mathbf{r})\varphi_{bj}(\mathbf{r}), \tag{4.3}$$

by prescreening with the criterion

$$\int d\mathbf{r}\ \varphi_{ai}(\mathbf{r})\varphi_{bj}(\mathbf{r})\leq \texttt{DistNeglect}. \tag{4.4}$$

These methods, though, still result in a significant amount of computational work that grows as $\mathcal{O}(\tilde{K}^4)$.

In Chapter 3, an attractive approach for avoiding the contraction problem was mentioned that is based on the multipole approximation[102, 67, 103]. By decoupling well separated distributions $(\phi_a\phi_b|$ and $|\phi_c\phi_d)$, the multipole approximation enables contractions to be performed independently of ERI evaluation.

In the following sections, we describe how the multipole approximation used in conjunction with the previously described MAC can be included in ONX and SONX, which we will

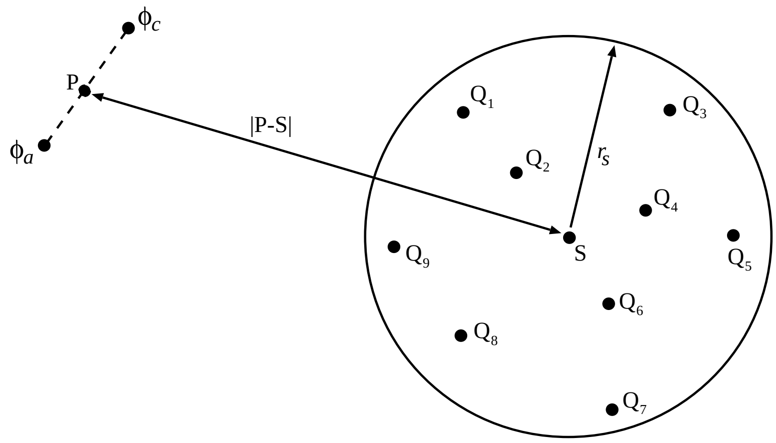

Figure 4.1: The clustering scheme used for the MAC. $\mathbf{P}$ and $\mathbf{Q}_n$ are expansion centers that the primitive distributions originating from ρ_{ac} and ρ_{bd} respectively are translated to. The sphere with the center $\mathbf{S}$ and radius r_s is used in the MAC to represent the entire set $\{\rho_{bd}\}$.

refer to as multipole accelerated ONX (MAONX) and multipole accelerated SONX (MASONX). By eliminating much of the work associated with basis set contraction, these methods offer an extremely competitive approach for computation of the HF exchange matrix for both large systems and highly-contracted basis sets.

4.2 Methods

4.2.1 Multipole expansions

Rather than use the multipole expansion of $|\mathbf{r}_1 - \mathbf{r}_2|^{-1}$ shown in Eq. (3.8), which involves complex arithmetic, we instead use a real arithmetic implementation, which has been simplified by removing redundancies with the symbolic manipulation features of *Mathematica* [117] along with the `FortranAssign` utility in `Format.m` [118] The resulting expansions are expressed as explicit FORTRAN statements, which operate at peak efficiencies on modern computational platforms [24].

4.2.2 Translation of primitive distributions

The use of multipole expansions alone will not reduce the computational complexity of ERI evaluation with respect to basis function contraction lengths. In order to reduce the cost of contraction, the primitive distributions need to be translated to a common center, as in Fig. 3.1. This center is chosen so that the average translation distance is minimized. Once the primitive distribution moments share a center, their moments can then be combined independent of ERI evaluation, and the high cost of contraction avoided.

4.2.3 Multipole acceptability criteria

In order to ensure that a given contracted ERI can be safely represented with the multipole approximation, the MAC described in Chapter 3 is used. However, rather than applying the MAC to each ERI, as was done for the model systems in Chapter 3, we have adopted the following scheme to cluster ERIs into groups. This is necessary because the MAC requires computation of the distance between interacting distributions, which is too computationally expensive to apply to every ERI. For sake of clarity, the clustering method is described in terms of ONX, but can be easily extended to SONX as well.

In ONX elements of the exchange matrix,

$$K_{ab} = \sum_{cd} D_{cd}(\phi_a\phi_c|\phi_b\phi_d), \tag{4.5}$$

are formed through the ordered distribution lists

$$\rho_{ac}(\mathbf{r}) = \phi_a(\mathbf{r})\phi_c(\mathbf{r}) \tag{4.6}$$

and

$$\rho_{bd}(\mathbf{r}') = \phi_a(\mathbf{r}')\phi_c(\mathbf{r}'), \tag{4.7}$$

where the the indices c and d are fixed and a and b run over the possible distributions in each set. In the calculation of large systems, when the centers of the basis functions ϕ_c and ϕ_d are far apart, the sets $\{\rho_{ac}\}$ and $\{\rho_{bd}\}$ form two spatially distinct groups. Instead of applying a different MAC to each pair ρ_{ac} and ρ_{bd}, a single MAC can be computed between each ρ_{ac} and the entire set $\{\rho_{bd}\}$. This is accomplished by determining the smallest sphere that encompasses all of the centers of $\{\rho_{bd}\}$. In the evaluation of Eq. (3.22), the following substitutions are then made

$$|\mathbf{r}_2 - \mathbf{P}|\,, |\mathbf{r}_1 - \mathbf{Q}| \Rightarrow |\mathbf{P} - \mathbf{S}| - r_s\ , \tag{4.8}$$

where $\mathbf{S}$ is the center of the sphere and r_s is its radius. In addition, the translation distance $|\mathbf{r}_2 - \mathbf{Q}|$ and the moment c_i^{bd} are replaced by their maximum values over all $\{\rho_{bd}\}$. This clustering, which is illustrated in Fig. 4.1, enables calculation of a single MAC for each ρ_{ac} that represents the worst-case error in the interactions with $\{\rho_{bd}\}$.

4.2.4 Penetration acceptability criteria

In addition to MAC, penetration acceptability criteria (PAC) must be satisfied in order for an ERI to be accurately calculated in a multipole representation [94]. Unlike point charges common in classical physics, CGTF charge distributions have extents, which can lead to contributions that the multipole expansion can not reproduce. These errors are significant when the tails of interacting distributions penetrate each other, as in Figure 3.2a. This type of error is not accounted for by the MAC and is specific to systems of continuous charge distributions.

We have found empirically that it is sufficient to estimate penetration errors by considering the lowest order contributions only. Using the notation introduced in Chapter 3, the penetration errors are dominated by

$$\epsilon = e^{ac}_{000} e^{bd}_{000} \left[u^{ac}_{bd} F_0\left(\tau\right) - \frac{\pi^3}{\left(\zeta_{ac}\zeta_{bd}\right)^{3/2}} \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} \right], \tag{4.9}$$

```
do a = a_f, a_l
   P = expansion center of ρ_ac
   S = center of sphere around {ρ_bd}
   r_s = radius of sphere
   if( |P − S| − r_s < 0 ) goto 352
   compute MAC
   if(MAC >TwoENeglect) goto 352
   τ = ζ_ac ζ_bd/(ζ_ac + ζ_bd)(|P − S| − r_s)^2
   if(τ >20.0) goto 351
   compute PAC
   if(PAC >TwoENeglect) goto 352
351   continue
   mark ρ_ac for multipole ERI evaluation
   goto 353
352   continue
   mark ρ_ac for direct ERI evaluation
353   continue
enddo a
```

Figure 4.2: Applying the MAC and PAC in MAONX. The above code is evaluated immediately following the "301 continue" line in Figure 1.4. The values of **P**, **S**, r_s, and the majority of the MAC can be precomputed at the level of distributions.

where

$$u_{bd}^{ac} = \frac{2\pi^{5/2}}{\zeta_{ac}\zeta_{bd}\sqrt{\zeta_{ac}+\zeta_{bd}}}, \tag{4.10}$$

and

$$\tau = \frac{\zeta_{ac}\zeta_{bd}}{\zeta_{ac}+\zeta_{bd}} \left|\mathbf{r}_1 - \mathbf{r}_2\right|^2 . \tag{4.11}$$

In Eq. (4.9),

$$F_0(\tau) = \frac{1}{2}\sqrt{\frac{\pi}{\tau}}\,\mathrm{erf}\left(\sqrt{\tau}\right) = \int_0^1 e^{-\tau u^2} du \tag{4.12}$$

is the 0^{th} order reduced incomplete gamma function[119, 120]. By introducing the complementary error function and the moments defined in Eq. (3.23) the penetration acceptability criterion is

$$\mathrm{PAC} = c_0^{ac} c_0^{bd} \left(\frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|}\right) \mathrm{erfc}\left(\sqrt{\tau}\right). \tag{4.13}$$

If more accurate PAC are needed, contributions from higher order terms can be included by differentiating Eq. (4.13) in a manner equivalent to those used in development of the MAC.

For the same reasons mentioned for the MAC, the PAC is too expensive to apply to every ERI. We again use the clustering scheme described above by making the substitution

$$|\mathbf{r}_1 - \mathbf{r}_2| \Rightarrow |\mathbf{r}_1 - \mathbf{S}| - r_s \tag{4.14}$$

in Eqs. (4.11) and (4.13). Fig. 4.2 illustrates how the MAC and PAC are applied together in MAONX. In order to further reduce the computational cost, simple estimates are placed before the actual MAC and PAC to screen out ERIs that are obviously either in the multipole or direct regime. For example, in Fig. 4.2, the conditional

$$\texttt{if}\ \ (|\mathbf{P}-\mathbf{S}|-r_s<0)\ \ \texttt{goto 352} \tag{4.15}$$

is used to skip the MAC and PAC when ρ_{ac} is inside the $\{\rho_{bd}\}$ sphere and

$$\texttt{if}\ \ (\tau>20.0)\ \ \texttt{goto 351} \tag{4.16}$$

is used to skip the PAC when separations are too large for penetration to be an issue. The above approach is similar in MASONX except that the code corresponding to Fig. 4.2 is applied twice, once after the line "`200 continue`", and once after the line "`400 continue`" in Fig. 2.3.

4.3 Results and discussion

MAONX and MASONX calculations within MONDOSCF[76] were performed with a variety of basis sets on a series of three-dimensional water clusters and graphitic sheets. The geometries of the water clusters and the graphitic sheets are identical to those used in Chapter 1. In each case the multipole expansions have been truncated at the same order as the total angular symmetry of the distributions involved (l_p and l_q), and the thresholding values used in the MAC and PAC were set equal to `TwoENeglect`. For primitive distribution prescreening, `DistNeglect`=`TwoENeglect`$\times 10^{-2}$ was used in Eq. (4.4). All timings were obtained on a single 322 MHz 604e PowerPC processor.

There are two key factors that determine to what extent formation of the exchange matrix can be accelerated by use of the multipole approximation. These factors are the percentage of all interactions that can be accurately computed in a multipole representation, and the relative speed of calculating an interaction with multipole versus direct ERI evaluation.

The percentage of interactions computed with the multipole approximation in MAONX is shown in Figs. 4.3 and 4.4 for the series of water clusters and graphitic sheets. This percentage reaches 75 to 80 percent of the total interactions as system sizes are increased. Similar results are obtained with MASONX. Based on these observed trends, if the time needed to compute a multipole interaction can be made small relative to direct ERI evaluation then, at best, a four to five fold decrease in the total CPU time can be expected.

The observed speed of multipole versus direct ERI evaluation depends on a number of issues, including the various implementational details of each method. However, the most significant factor is the number of primitive ERIs that are avoided in a multipole representation. This, of course, is related to the average basis set contraction length $\tilde{K}$. In Fig. 4.5, the ratio of direct to multipole time for ERI evaluation is shown as a function of $\tilde{K}$. Since the computational cost of direct ERI evaluation grows as $\mathcal{O}(\tilde{K}^4)$, the benefits of multipole evaluation, which is $\mathcal{O}(1)$ with respect to $\tilde{K}$, can be quite large.

MAONX and MASONX timings for the cluster of 50 water molecules calculated with different basis sets are summarized in Table 4.1. Along with the total time required to form

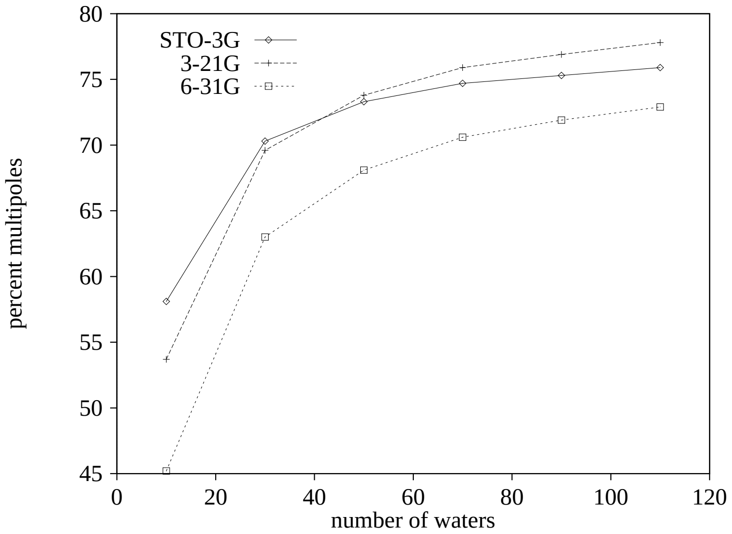

Figure 4.3: The percentage of interactions computed with multipole expansions in the RHF calculation of the exchange matrix for a series of three-dimensional water cluster

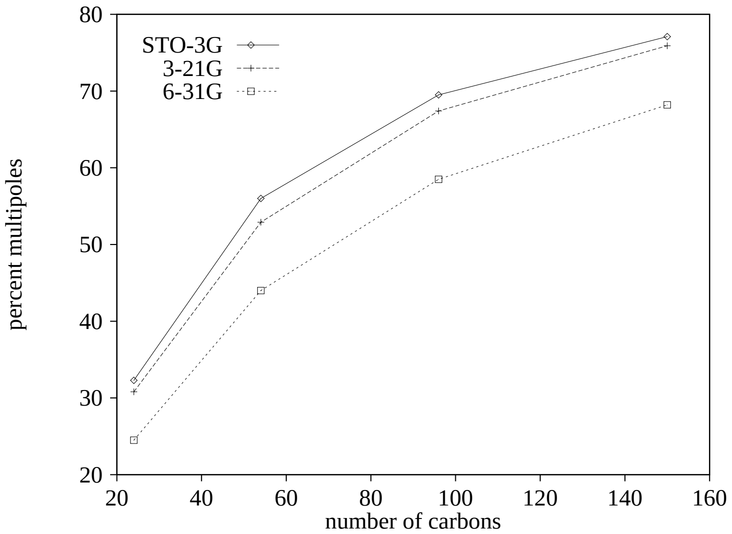

Figure 4.4: The percentage of interactions computed with multipole expansions in the RHF calculation of the exchange matrix for a series of graphitic sheets.

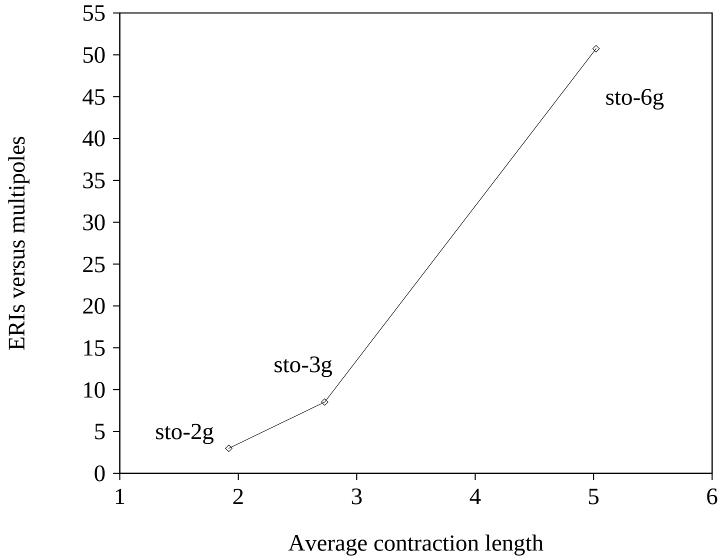

Figure 4.5: The ratio of the time needed to compute an ERI by direct evaluation versus with a multipole expansion[121] for increasing basis set contraction length.

Table 4.1: The total CPU time required to form one RHF exchange matrix for a cluster of 50 water molecules by MAONX and MASONX with different basis sets and with `TwoENeglect` $= 10^{-7}$. $\tilde{K}$ is the average basis function contraction length[77]. The ONX and SONX times included for reference and the multipole speedups are in parenthesis.

	CPU Time (sec)				
Basis Set	ONX	SONX	MAONX	MASONX	$\tilde{K}$
STO-2G	20	13	13(1.5)	10(1.2)	1.9
STO-3G	90	50	35(2.6)	24(2.1)	2.6
STO-6G	1163	613	255(4.6)	147(4.1)	4.7
3-21G	91	86	72(1.3)	88(1.0)	1.4
6-31G	290	209	149(1.9)	158(1.3)	1.9
6-31G*	599	491	375(1.6)	414(1.2)	1.4
6-31G**	979	854	646(1.5)	759(1.1)	1.4
Dunning DZ	621	474	297(2.1)	332(1.4)	1.9
Dunning-Hay SV	697	462	299(2.3)	292(1.6)	2.0

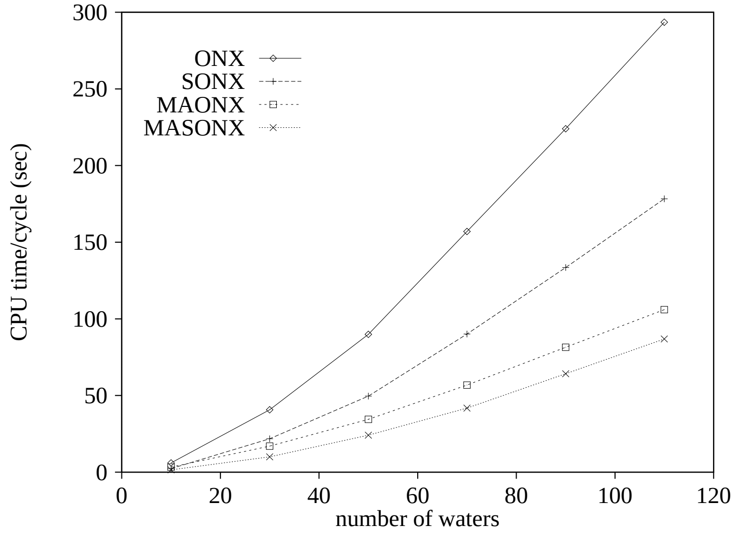

Figure 4.6: RHF/STO-3G water cluster timings with `TwoENeglect` $= 10^{-7}$.

a single exchange matrix, the speedup of MAONX over ONX and MASONX over SONX are shown in parenthesis. For the largest contracted basis set used, STO-6G, both MAONX and MASONX approach the expected four to five fold decrease. As shown in Fig. 4.5, multipole acceleration increases as a power of $\tilde{K}$. In Figs. 4.6 to 4.11, similar accelerations are observed for the series of water clusters and graphitic sheets.

It is interesting to note from Table 4.1 that the observed MAONX speedup is always larger than in MASONX and for many cases MAONX is the fastest method available. In particular this is true for basis sets with small contraction lengths. As was discussed in Chapter 2, this is due to the increased overhead associated with exploiting ERI permutational symmetry.

4.4 Conclusions

In this chapter, we have described a new approach for the rapid evaluation of contracted ERIs with the multipole approximation. Implementations in the linear scaling methods MAONX and MASONX indicate that large computational savings are possible for calculations that involve highly contracted basis sets. For example, MAONX is 4.6 times faster than ONX for computing the exchange matrix of a cluster of 50 water molecules with the STO-6G basis set. With basis sets that have lower levels of contraction, a factor of two speedup is typically observed.

Since the multipole approximation has no effect on the overhead needed to incorporate ERI permutational symmetry, methods that do not use permutational symmetry enjoy larger benefits than those that do. A similar trade-off of reduction in ERIs versus increases in overhead was observed in the comparison of ONX and SONX in Chapter 2. It is interesting to note, though, that the benefits of permutational symmetry are further diminished because, with the multipole approximation, even highly contracted ERIs are evaluated inexpensively, whereas the overhead remains unchanged.

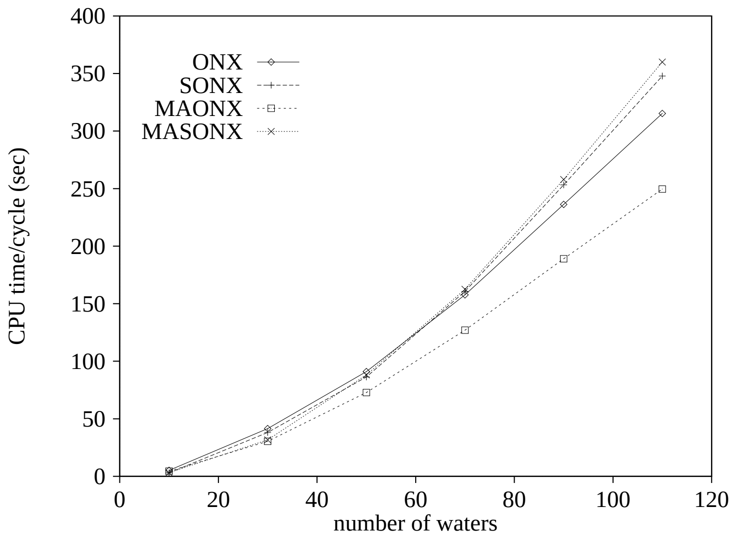

Figure 4.7: RHF/3-21G water cluster timings with `TwoENeglect` $= 10^{-7}$.

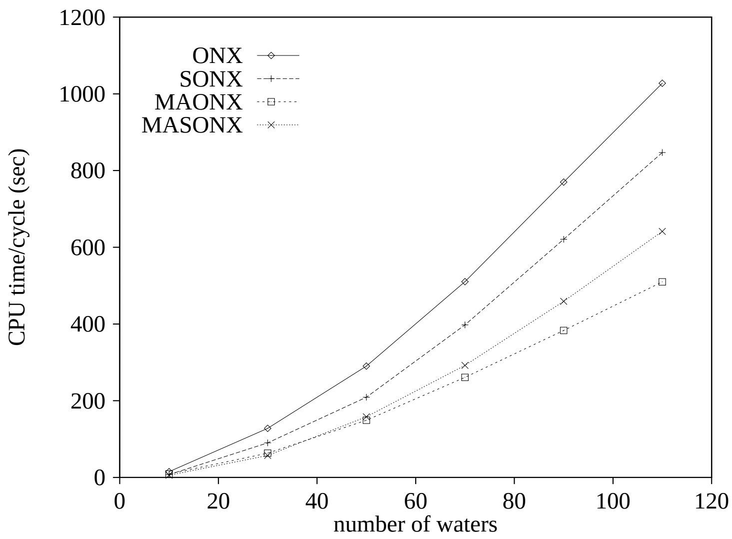

Figure 4.8: RHF/6-31G water cluster timings with `TwoENeglect` $= 10^{-7}$.

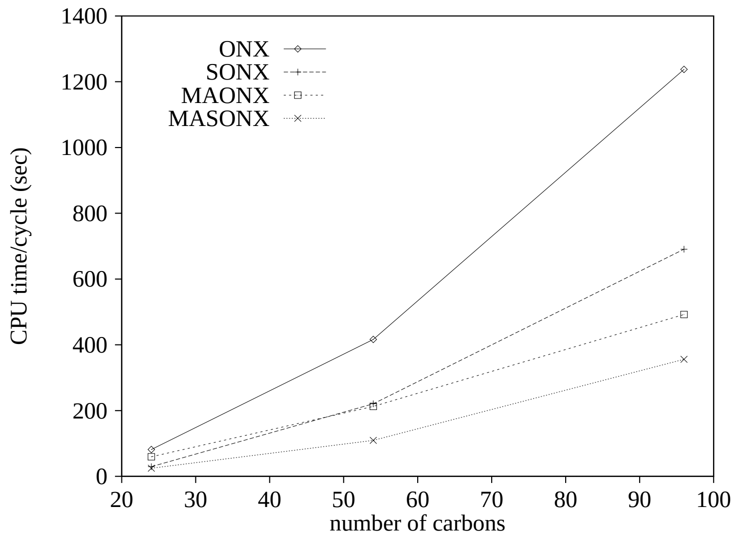

Figure 4.9: RHF/STO-3G graphitic sheet timings with `TwoENeglect` $= 10^{-7}$.

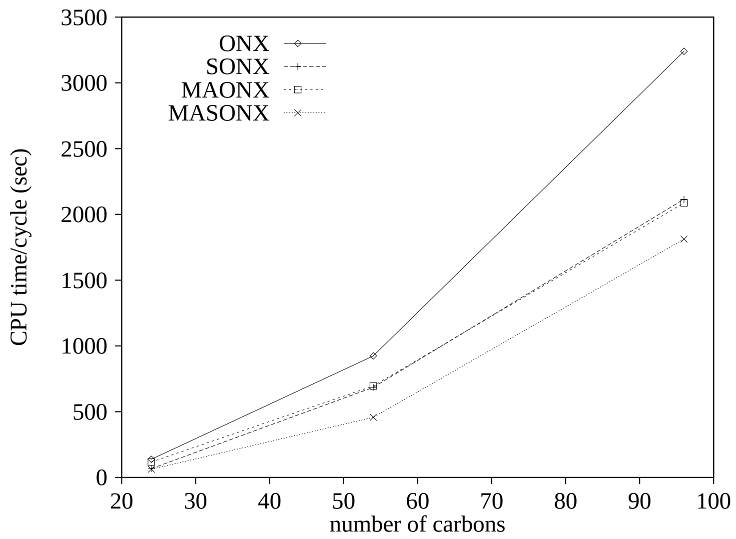

Figure 4.10: RHF/3-21G graphitic sheet timings with `TwoENeglect` $= 10^{-7}$.

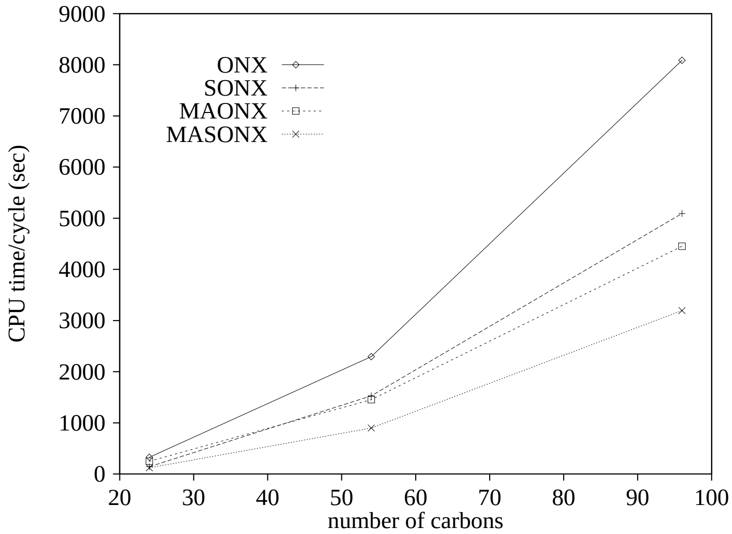

Figure 4.11: RHF/6-31G graphitic sheet timings with `TwoENeglect` $= 10^{-7}$.

Chapter 5

Analytical Gradients

5.1 Introduction

Probably the greatest use of Hartree-Fock (HF) theory is the accurate prediction of molecular structure. Even with relatively modest size basis sets it is not uncommon to achieve accuracies within 0.01 to 0.05Å in bond lengths and 1 to 2 degrees in bond angles[3]. Although analytical derivatives are not required to compute the various structural properties, their use greatly facilitates the process.

In the following sections, we review the calculation of the analytical first derivative of the HF energy with respect to a nuclear displacement and demonstrate how the previously described linear scaling method MASONX for computing the HF exchange matrix can be adapted to compute the exchange component of the derivative in $\mathcal{O}(N)$ CPU time.

5.2 HF energy gradient

The electronic energy of a closed shell system in the HF approximation is

$$E^{HF} = \sum_{ab} D_{ab} h_{ab} + \sum_{ab} \left[D_{ab} J_{ab} - \frac{1}{2} D_{ab} K_{ab} \right] \tag{5.1}$$

where $\mathbf{h}$ is the core Hamiltonian and

$$J_{ab} = \sum_{cd} D_{cd} \left(\phi_a \phi_b | \phi_c \phi_d \right) , \tag{5.2}$$

$$K_{ab} = \sum_{cd} D_{cd} \left(\phi_a \phi_c | \phi_b \phi_d \right) , \tag{5.3}$$

are the Coulomb and the exchange matrices, respectively. The density matrix, $\mathbf{D}$, is formed from the molecular orbital coefficients

$$D_{ab} = 2 \sum_{\mu} C_{a\mu} C^*_{b\mu} \ . \tag{5.4}$$

The total derivative of the HF energy with respect to the nuclear coordinate A can be expressed as

$$\frac{dE^{HF}}{dA} = \frac{\partial E^{HF}}{\partial A} + \sum_{\alpha} \frac{\partial E^{HF}}{\partial \alpha} \frac{\partial \alpha}{\partial A}. \tag{5.5}$$

where α represents all the parameters that have an implicit dependence on A. The term $\partial E^{HF}/\partial A$ in Eq. (5.5) refers to the derivative of all components in E^{HF} that depend explicitly on A. Possible candidates for α include: the molecular orbital coefficients ($\mathbf{C}$), the basis function exponents (ζ), and the basis function centers. Since a necessary condition for finding a solution to the HF equations is

$$\frac{\partial E^{HF}}{\partial C_{ab}} = 0, \tag{5.6}$$

$\alpha = \mathbf{C}$ does not need to be considered. In addition, it is standard practice to hold the basis function exponents fixed during the calculation,

$$\frac{\partial \zeta}{\partial A} = 0, \tag{5.7}$$

so $\alpha = \zeta$ need not be considered either. The basis function centers are neither held constant nor optimized because they are usually required to follow the nuclear centers. The total derivative of E^{HF} with respect to the nuclear coordinate A may, therefore, be expressed as

$$\begin{aligned} \frac{dE^{HF}}{dA} = & \sum_{ab} D_{ab} \frac{\partial h_{ab}}{\partial A} + \sum_{abcd} D_{ab} D_{cd} \frac{\partial \left(\phi_a \phi_b | \phi_c \phi_d\right)}{\partial A} - \frac{1}{2} \sum_{abcd} D_{ab} D_{cd} \frac{\partial \left(\phi_a \phi_c | \phi_b \phi_d\right)}{\partial A} \\ & + \sum_{ab} \frac{\partial D_{ab}}{\partial A} h_{ab} + 2 \sum_{abcd} \frac{\partial D_{ab}}{\partial A} D_{cd} \left(\phi_a \phi_b | \phi_c \phi_d\right) - \sum_{abcd} \frac{\partial D_{ab}}{\partial A} D_{cd} \left(\phi_a \phi_c | \phi_b \phi_d\right). \end{aligned} \tag{5.8}$$

By expanding the derivatives of $\mathbf{D}$ in the last three terms of Eq. (5.8) with the corresponding derivatives of $\mathbf{C}$ and collecting terms,

$$\begin{aligned} \frac{dE^{HF}}{dA} = & \sum_{ab} D_{ab} \frac{\partial h_{ab}}{\partial A} + \sum_{abcd} D_{ab} D_{cd} \frac{\partial \left(\phi_a \phi_b | \phi_c \phi_d\right)}{\partial A} - \frac{1}{2} \sum_{abcd} D_{ab} D_{cd} \frac{\partial \left(\phi_a \phi_c | \phi_b \phi_d\right)}{\partial A} \\ & + 4 \sum_{ab} \sum_{\mu} \frac{\partial C_{a\mu}}{\partial A} C_{b\mu} h_{ab} + 4 \sum_{abcd} \sum_{\mu} \frac{\partial C_{a\mu}}{\partial A} C_{b\mu} D_{cd} \left(\phi_a \phi_b | \phi_c \phi_d\right) \\ & - 2 \sum_{abcd} \sum_{\mu} \frac{\partial C_{a\mu}}{\partial A} C_{b\mu} D_{cd} \left(\phi_a \phi_c | \phi_b \phi_d\right). \end{aligned} \tag{5.9}$$

After rearrangement,

$$\begin{aligned} \frac{dE^{HF}}{dA} = & \sum_{ab} D_{ab} \frac{\partial h_{ab}}{\partial A} + \sum_{abcd} D_{ab} D_{cd} \frac{\partial \left(\phi_a \phi_b | \phi_c \phi_d\right)}{\partial A} - \frac{1}{2} \sum_{abcd} D_{ab} D_{cd} \frac{\partial \left(\phi_a \phi_c | \phi_b \phi_d\right)}{\partial A} \\ & + 4 \sum_{ab} \sum_{\mu} \frac{\partial C_{a\mu}}{\partial A} \left[h_{ab} + \sum_{cd} D_{cd} \left(\phi_a \phi_b | \phi_c \phi_d\right) - \frac{1}{2} \sum_{cd} D_{cd} \left(\phi_a \phi_c | \phi_b \phi_d\right) \right] C_{b\mu}. \end{aligned} \tag{5.10}$$

Inspection of Eq. (5.10) reveals that the terms enclosed in [...] are simply the Fock matrix, for which

$$\left[\mathbf{h} + \mathbf{J} - \frac{1}{2}\mathbf{K}\right] \mathbf{C} = \mathbf{FC} = \epsilon \mathbf{SC}. \tag{5.11}$$

Equation 5.11 can be used in place of the last term in Eq. (5.10), which leads to

$$\begin{aligned} \frac{dE^{HF}}{dA} &= \sum_{ab} D_{ab} \frac{\partial h_{ab}}{\partial A} + \sum_{abcd} D_{ab} D_{cd} \frac{\partial \left(\phi_a \phi_b | \phi_c \phi_d\right)}{\partial A} \\ &\quad - \frac{1}{2} \sum_{abcd} D_{ab} D_{cd} \frac{\partial \left(\phi_a \phi_c | \phi_b \phi_d\right)}{\partial A} + 4 \sum_{ab} \sum_{\mu} \epsilon_\mu \frac{\partial C_{a\mu}}{\partial A} S_{ab} C_{b\mu}, \end{aligned} \tag{5.12}$$

where ϵ_μ are the molecular orbital energies. At first it may appear that the gradient of the HF energy requires the derivatives of $\mathbf{C}$, which would be quite expensive to calculate[122, 123]. Fortunately, in the particular case of a HF wavefunction, the following relations

$$\sum_{ab} C_{a\mu} S_{ab} C_{b\nu} = \delta_{\mu\nu} \tag{5.13}$$

and

$$\sum_{ab} \frac{\partial C_{a\mu}}{\partial A} S_{ab} C_{b\mu} + \sum_{ab} C_{a\mu} \frac{\partial S_{ab}}{\partial A} C_{b\mu} + \sum_{ab} C_{a\mu} S_{ab} \frac{\partial C_{b\mu}}{\partial A} = 0 \tag{5.14}$$

can be used to give

$$\begin{aligned} \frac{dE^{HF}}{dA} &= \sum_{ab} D_{ab} \frac{\partial h_{ab}}{\partial A} + \sum_{abcd} D_{ab} D_{cd} \frac{\partial \left(\phi_a \phi_b | \phi_c \phi_d\right)}{\partial A} \\ &\quad - \frac{1}{2} \sum_{abcd} D_{ab} D_{cd} \frac{\partial \left(\phi_a \phi_c | \phi_b \phi_d\right)}{\partial A} - \sum_{ab} \tilde{D}_{ab} \frac{\partial S_{ab}}{\partial A} \end{aligned} \tag{5.15}$$

where

$$\tilde{D}_{ab} = 2 \sum_{\mu} \epsilon_\mu C_{a\mu} C_{b\mu} \tag{5.16}$$

is the energy weighted density matrix. As can be seen in Eq. (5.15), the gradient of the HF energy requires the calculation of one and two-electron integrals and their partial derivatives.

When basis functions consisting of CGTFs

$$\begin{aligned} \phi_a^{lmn}(\mathbf{r}) &= (x - A_x)^l (y - A_y)^m (z - A_z)^n \\ &\quad \times \exp\left[-\zeta_a (\mathbf{r} - \mathbf{A})\right] \end{aligned} \tag{5.17}$$

are used, their derivatives can be readily evaluated as

$$\frac{\partial \phi_a^{lmn}}{\partial A_x} = 2\zeta_a \phi_a^{l+1mn} - l\ \phi_a^{l-1mn}, \tag{5.18}$$

which is a scaled sum of two CGTFs with angular symmetries that have been raised and lowered by one. Therefore, the derivative of an ERI over CGTFs involves the calculation of two additional ERIs over higher and lower angular symmetry CGTFs. For example,

$$\frac{\partial \left(\phi_a^{100} \phi_b^{000} | \phi_c^{000} \phi_d^{000}\right)}{\partial A_x} = 2\zeta_a \left(\phi_a^{200} \phi_b^{000} | \phi_c^{000} \phi_d^{000}\right) - \left(\phi_a^{000} \phi_b^{000} | \phi_c^{000} \phi_d^{000}\right). \tag{5.19}$$

5.3 Order N exchange energy gradient

The third term in Eq. (5.15) is quite similar to the calculation of the HF exchange matrix, the only differences being the calculation of ERI derivatives and the additional contraction with **D**. As might be expected, the previously described methods for linear scaling computation of **K** can be modified to compute

$$\frac{\partial E_K}{\partial A} = -\frac{1}{2}\sum_{abcd} D_{ab} D_{cd} \frac{\partial(\phi_a\phi_c|\phi_b\phi_d)}{\partial A} \ , \tag{5.20}$$

which we will refer to as the exchange component of the gradient.

The general way Eq. (5.20) is written is usually referred to as a matrix driven approach, emphasizing that the summations over elements of **D** determine the order of ERI derivative evaluation. It is often helpful, though, to consider the equivalent expressions in an integral driven approach:

$$\frac{\partial E_K}{\partial A} = \frac{\partial E_K}{\partial A} - \frac{1}{2} D_{ab} D_{cd} \frac{\partial(\phi_a\phi_c|\phi_b\phi_d)}{\partial A}. \tag{5.21}$$

The ERI permutational symmetry, given in Eq. (2.5), can then be readily incorporated into Eq. (5.21) to yield

$$\frac{\partial E_K}{\partial A} = \frac{\partial E_K}{\partial A} - \left[D_{ac}D_{bd} + D_{ad}D_{cb}\right] \frac{\partial(\phi_a\phi_b|\phi_c\phi_d)}{\partial A}. \tag{5.22}$$

Insignificant contributions to the derivative can be avoided with the estimate

$$\frac{1}{2}\,|D_{ab}D_{cd}|\;|(\phi_a\phi_c|\phi_a\phi_c)'|^{1/2}\;|(\phi_b\phi_d|\phi_b\phi_d)'|^{1/2} \leq \texttt{GradNeglect} \tag{5.23}$$

when Eq. (5.21) is used, and

$$\begin{aligned} &\mathrm{Max}\left[|D_{ac}D_{bd}|, |D_{ad}D_{cb}|\right] \\ &\quad \times |(\phi_a\phi_b|\phi_a\phi_b)'|^{1/2}\;|(\phi_c\phi_d|\phi_c\phi_d)'|^{1/2} \leq \texttt{GradNeglect} \end{aligned} \tag{5.24}$$

when Eq. (5.22) is used. In most ERI evaluation routines the derivatives of an ERI with respect to all four centers are computed together[124], so the primes in Eqs. (5.23) and (5.24) are used to indicate that contributions from differentiating each center are included in the estimate.

Equations (5.23) and (5.24) may look somewhat reminiscent of direct SCF integral prescreening criteria that control errors in the HF energy[125] rather than in elements of the Fock matrix[70]. In the direct SCF case, the use of such criteria is known to lead to unstable convergence in the SCF process and is not recommended. However, in the calculation of the gradient, the SCF equations have already been solved and similar problems need not concern us.

5.4 Results and discussion

In order to demonstrate linear scaling computation of the exchange component of the gradient, we have implemented Eq. (5.22) in a modified version of MASONX, which we will refer

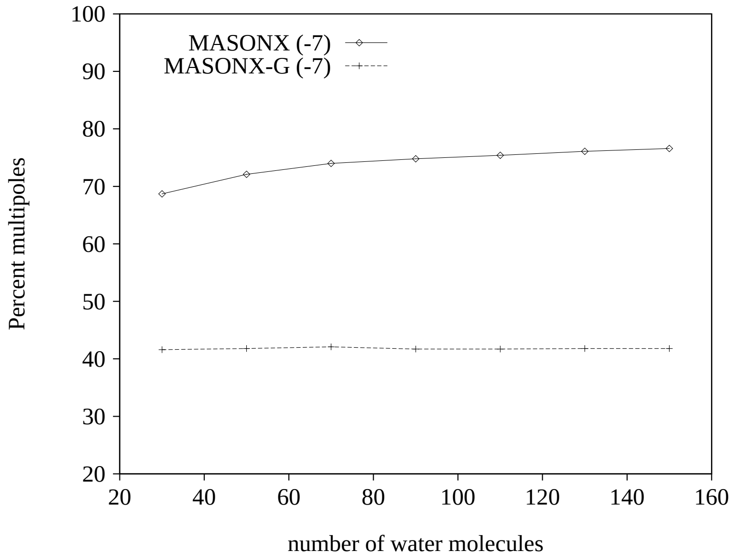

Figure 5.1: The percentage of ERIs treated with multipoles in the RHF/STO-3G exchange matrix and in the exchange component of the gradient. The notation (-7) is used to indicate that `GradNeglect`$= 10^{-7}$.

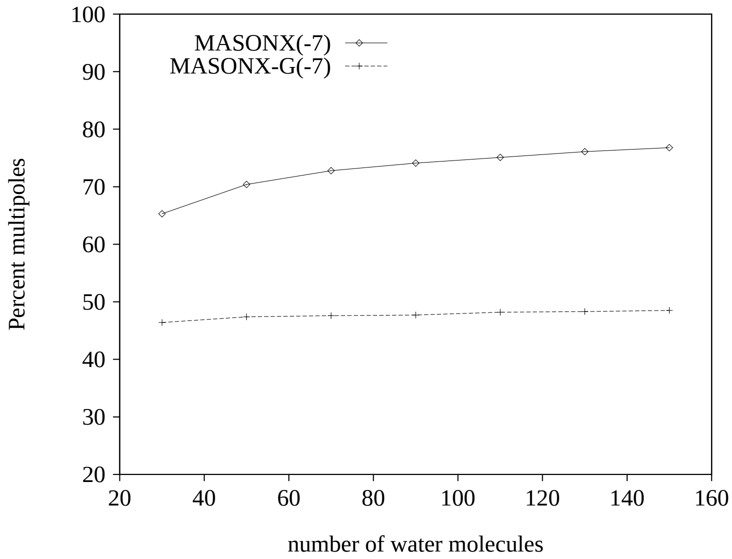

Figure 5.2: The percentage of ERIs treated with multipoles in the RHF/3-21G exchange matrix and in the exchange component of the gradient.

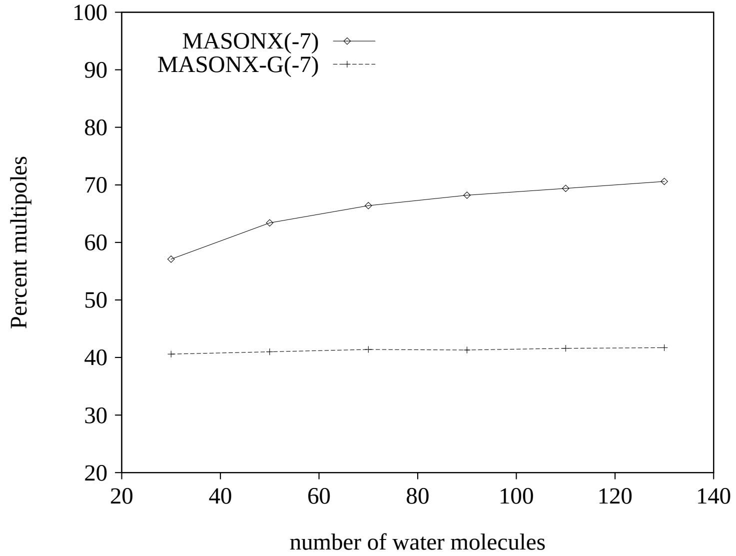

Figure 5.3: The percentage of ERIs treated with multipoles in the RHF/6-31G exchange matrix and in the exchange component of the gradient.

to as MASONX-G. In the following, MASONX-G benchmark calculations are presented on the same series of three-dimensional water clusters used in the previous chapters. All timings have been obtained on a single 322 MHz 604e PowerPC processor.

Since the multipole based methods introduced in Chapters 3 and 4 for accelerating the computation of $\mathbf{K}$ are applicable to any method involving contracted ERIs, the same techniques have been used in MASONX-G to rapidly compute ERI derivatives. In Figs. 5.1 to 5.3, the percentage of ERI derivatives that are evaluated with the multipole approximation for the series of water clusters is shown. For comparison, the percentage of multipole based ERIs evaluated by MASONX is also given. The MASONX-G percentages are roughly one half of those obtained from MASONX because of the differences in the thresholding criteria. The additional element of $\mathbf{D}$ that occurs in Eq. (5.24) enables the long-range interactions to be neglected more rapidly in the gradient evaluation than in the formation of $\mathbf{K}$. Since the majority of non-multipole based ERI derivatives involve short-range interactions, the percentage of multipole based ERI derivatives is reduced overall.

MASONX-G and MASONX timings for the series of water clusters are shown in Figs. 5.4 to 5.6. In MASONX-G, two levels of `GradNeglect` (10^{-7} and 10^{-9}) have been used. Although the gradient involves the calculation of ERIs over higher angular symmetry CGTFs, the timings for MASONX-G and MASONX are roughly equivalent for the same levels of `GradNeglect`. This is, once again, due to the different thresholding criteria that are used in both methods. In Figs. 5.7 to 5.9 the RMS and the maximum absolute error in exchange component of the gradient are given for the series of water clusters computed with `GradNeglect`$= 10^{-7}$. All errors have been determined by comparison with the results of the `GradNeglect`$= 10^{-9}$ calculations. Since the long-range interactions fall off rapidly in the exchange component of the gradient, its evaluation with respect to a given nuclear coordinate only depends on the local environment. As a result, the errors do not increase with system size.

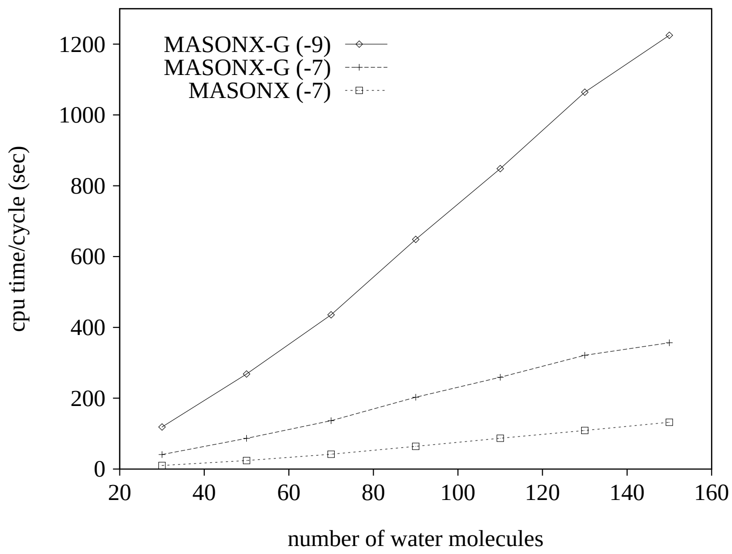

Figure 5.4: RHF/STO-3G water cluster times for formation of the exchange matrix and the exchange component of the gradient.

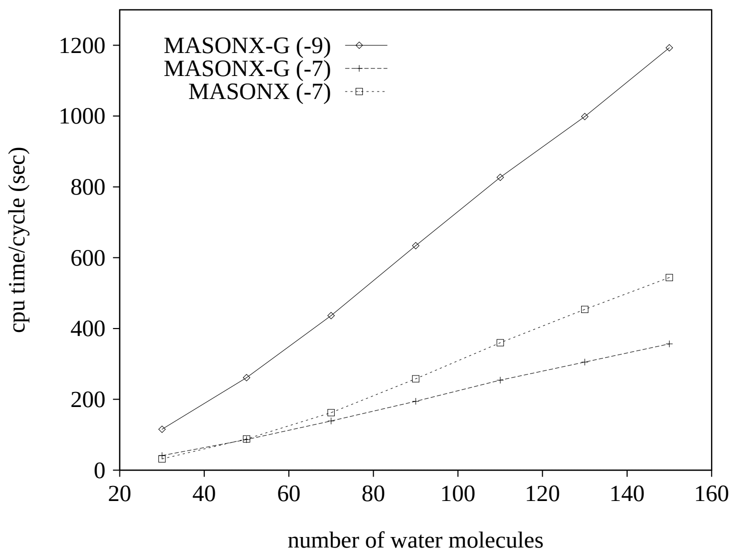

Figure 5.5: RHF/3-21G water cluster times for formation of the exchange matrix and the exchange component of the gradient.

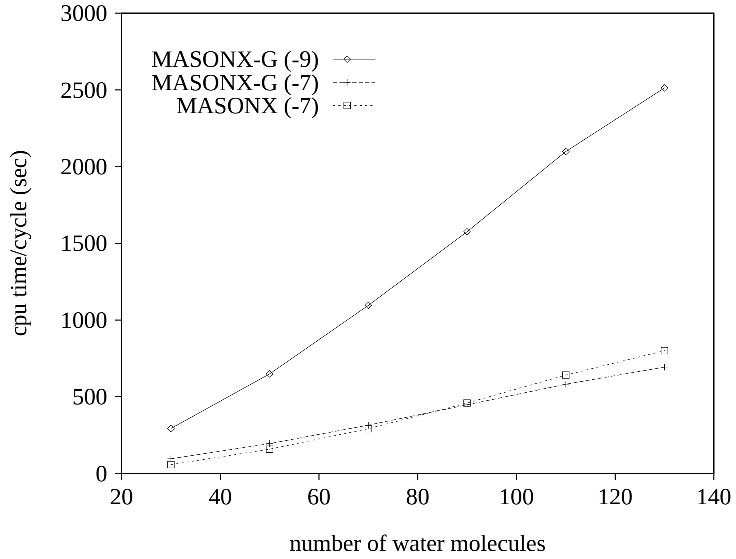

Figure 5.6: RHF/6-31G water cluster times for formation of the exchange matrix and the exchange component of the gradient.

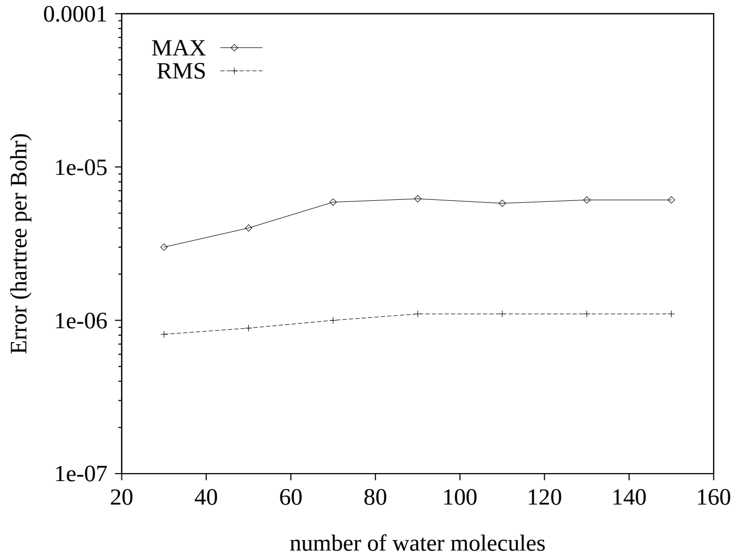

Figure 5.7: The absolute and RMS error in the RHF/STO-3G exchange energy gradient calculated with `GradNeglect` $= 10^{-7}$ on a series of water clusters.

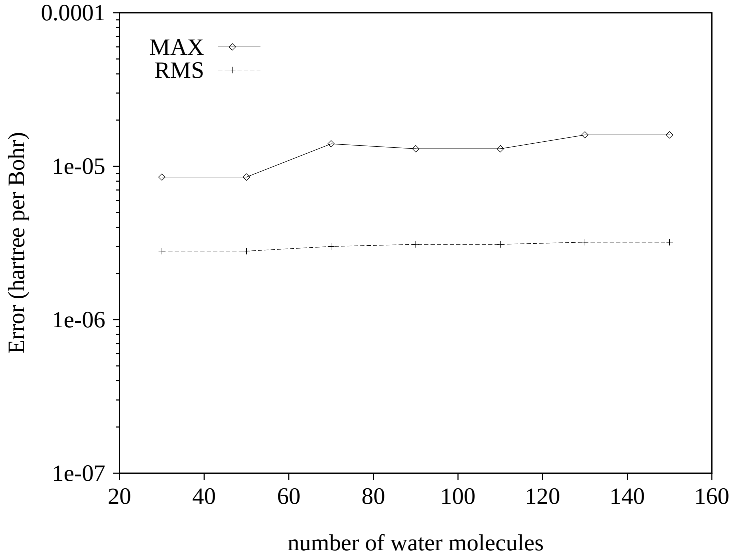

Figure 5.8: The absolute and RMS error in the RHF/3-21G exchange energy gradient calculated with `GradNeglect` $= 10^{-7}$ on a series of water clusters.

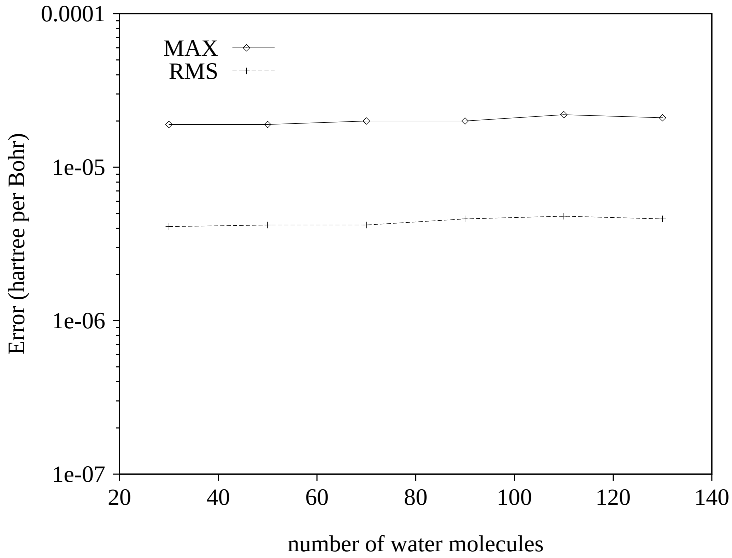

Figure 5.9: The absolute and RMS error in the RHF/6-31G exchange energy gradient calculated with `GradNeglect` $= 10^{-7}$ on a series of water clusters.

5.5 Conclusions

In this chapter, the analytical gradient of the Hartree-Fock energy with respect to a nuclear coordinate has been reviewed and linear scaling computation of the exchange component of the gradient is demonstrated. MASONX-G calculations on three-dimensional clusters of water molecules indicate that, since long-range gradient contributions fall off rapidly, formation of the exchange gradient requires CPU time proportional to that needed to form the exchange matrix. In addition, errors in the gradient that arise from integral prescreening are shown to be local and do not increase with system size.

Bibliography

[1] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry.* Mc Graw–Hill Inc., New York, first, revised edition, 1989.

[2] A. D. Becke. Density-functional thermochemistry. IV. A new dynamical correlation functional and implications for exact-exchange mixing. *J. Chem. Phys.*, 104(3):1040–1046, 1996.

[3] W. J. Hehre, L. Radom, P. von Schleyer, and J. A. Pople. *Ab Initio Molecular Orbital Theory.* J. Wiley, New York, first edition, 1986.

[4] C. C. J. Roothaan. New developments in molecular orbital theory. *Rev. Mod. Phys.*, 23:69, 1951.

[5] J. A. Pople and R. K. Nesbet. Self-consistent orbitals for radicals. *J. Chem. Phys.*, 22:571, 1954.

[6] P. O. Löwdin. Quantum theory of electronic structure of molecules. *Ann. Rev. Phys. Chem.*, 11:107, 1960.

[7] R. McWeeny. Some recent advances in density matrix theory. *Rev. Mod. Phys.*, 32(2):335–369, 1960.

[8] X. P. Li, R. W. Nunes, and D. Vanderbilt. Density-matrix electronic structure method with linear system-size scaling. *Phys. Rev. B*, 47(16):10891–10894, 1993.

[9] M. Challacombe. A simplified density matrix minimization for linear scaling SCF theory. *J. Chem. Phys.*, 1998. accepted.

[10] V. Dyczmons. No N^4-dependence in the calculation of large molecules. *Theoret. Chim. Acta*, 28(3):307–310, 1973.

[11] R. Ahlrichs. Methods for efficient evaluation of integrals for Gaussian type basis sets. *Theoret. Chim. Acta*, 33(2):157–167, 1974.

[12] J. Almlöf, K. Faegri, and K. Korsell. Principles for a direct SCF approach to LCAO-MO *ab initio* calculations. *J. Comp. Chem.*, 3(3):385–399, 1982.

[13] D. Cremer and J. Gauss. An unconventional SCF method for calculations on large molecules. *J. Comput. Chem.*, 7(3):274–282, 1986.

[14] M. Häser and R. Ahlrichs. Improvements on the direct SCF method. *J. Comp. Chem.*, 10(1):104–111, 1989.

[15] M. Dupuis, J. Rys, and H. F. King. Evaluation of molecular integrals over Gaussian basis functions. *J. Chem. Phys.*, 65(1):111–116, 1976.

[16] J. .A. Pople and W. J. Hehre. Computation of electron repulsion integrals involving contracted gaussian basis functions. *J. Comp. Phys.*, 27(2):161–168, 1978.

[17] L. McMurchie and E. R. Davidson. One- and two-electron integrals over Cartesian Gaussian functions. *J. Comp. Phys.*, 26:218–231, 1978.

[18] S. Obara and A. Saika. Efficient recursive computation of molecular integrals over Cartesian Gaussian functions. *J. Chem. Phys.*, 84(7):3963–3974, 1985.

[19] M. Head-Gordon and J. A. Pople. A method for two-electron Gaussian integral and integral derivative evaluation using recurrence relations. *J. Chem. Phys.*, 89(9):5777–5786, 1988.

[20] P. M. W. Gill. Molecular integrals over Gaussian basis functions. *Adv. Quantum Chem.*, 25:141–205, 1994.

[21] D. L. Strout and G. E. Scuseria. A quantitative study of the scaling properties of the Hartree-Fock method. *J. Chem. Phys.*, 102(21):8448–8452, 1995.

[22] C. A. White, B.G. Johnson, P.M.W. Gill, and M. Head-Gordon. The continuous fast multipole method. *Chem. Phys. Letters*, 230(1,2):8–16, 1994.

[23] M. C. Strain, G. E. Scuseria, and M. J. Frisch. Achieving linear scaling for the electronic quantum Coulomb problem. *Science*, 271(5245):51–53, 1996.

[24] M. Challacombe and E. Schwegler. Linear scaling computation of the Fock matrix. *J. Chem. Phys.*, 106(13):5526–5536, 1997.

[25] C. A. White and M. Head-Gordon. Derivation and efficient implementation of the fast multipole method. *J. Chem. Phys.*, 101(8):6593–6605, 1994.

[26] C. A. White and M. Head-Gordon. A J matrix engine for density functional theory calculations. *J. Chem. Phys.*, 104(7):2620–9, 1996.

[27] C. A. White and M. Head-Gordon. Fractional tiers in fast multipole method calculations. *Chem. Phys. Letters*, 257:647–650, 1996.

[28] C. A. White and M. Head-Gordon. Rotating around the quartic angular momentum barrier in fast multipole method calculations. *J. Chem. Phys.*, 105(12):5061–67, 1996.

[29] Jeremy P. Dombroski, Stephen W. Taylor, and Peter M. W. Gill. KWIK: Coulomb energies in O(N) work. *J. Phys. Chem.*, 100(15):6272–76, 1996.

[30] S. Goedecker and L. Colombo. Efficient linear scaling algorithm for tight-binding molecular dynamics. *Phys. Rev. Lett.*, 73(1):122–5, 1994.

[31] E. Hernández, M. J. Gillan, and C. M. Goringe. Linear-scaling density-functional-theory technique: the density-matrix approach. *Phys. Rev. B*, 53(11):7147–57, 1996.

[32] W. Yang and T. S. Lee. A density-matrix divide-and-conquer approach for electronic structure calculations of large molecules. *J. Chem. Phys.*, 103(13):5674–5678, 1995.

[33] A. F. Voter, J. D. Kress, and R. N. Silver. Linear-scaling tight binding from a truncated-moment approach. *Phys. Rev. B*, 53(19):12733–41, 1996.

[34] W. Kohn. Density functional and density matrix method scaling linearly with the number of atoms. *Phys. Rev. Lett.*, 76(17):3168–3171, 1996.

[35] A. P. Horsfield and A. M. Bratkovsky. O(N) tight-binding methods with finite electronic temperature. *Phys. Rev. B*, 53(23):15381–84, 1996.

[36] S. L. Dixon and K. M. Merz. Semiempirical molecular orbital calculations with linear system size scaling. *J. Chem. Phys.*, 104(17):6643–49, 1996.

[37] E. Schwegler and M. Challacombe. Linear scaling computation of the Hartree-Fock exchange matrix. *J. Chem. Phys.*, 105(19):2726, 1996.

[38] J. C. Burant, G. E. Scuseria, and M. J. Frisch. A linear scaling method for Hartree-Fock exchange calculations of large molecules. *J. Chem. Phys.*, 105(19):8969–8972, 1996.

[39] P. Ordejón, D. A. Drabold, R. M. Martin, and M. P. Grumbach. Linear system-size scaling methods for electronic-structure calculations. *Phys. Rev. B*, 51(3):1456, 1995.

[40] F. Mauri, G. Galli, and R. Car. Orbital formulation for electronic-structure calculations with linear system-size scaling. *Phys. Rev. B*, 47:9973–9976, 1993.

[41] P. Ordejón and D. A. Drabold. Unconstrained minimization approach for electronic computations that scales linearly with system size. *Phys. Rev. B*, 48:14646–14649, 1993.

[42] E. B. Stechel, A. R. Williams, and P. J. Feibleman. N-scaling algorithm for density-functional calculations of metals and insulators. *Phys. Rev. B*, 49(15):10088–10101, 1994.

[43] X. Chen, J. M. Langlois, and W. A. Goddard III. Dual-spance approach for denisty-functional calculations of two- and three-dimensional crystals using Gaussian basis functions. *Phys. Rev. B*, 54(4):2348–2361, 1995.

[44] J. Kim, F. Mauri, and G. Galli. Total-energy global optimizations using nonorthogonal localized orbitals. *Phys. Rev. B*, 52(3):1640–1648, 1995.

[45] P. Ordejón, E. Artacho, and J. M. Soler. Self-consistent order-n density-functional calculations for very large systems. *Phys. Rev. B*, 53(16):10441–10444, 1996.

[46] J. P. Stewart. Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations. *Int. J. Quant. Chem.*, 58(2):133–146, 1996.

[47] A. Canning, G. Galli, F. Mauri, A. DeVita, and R. Car. O(N) tight-binding molecular dynamics on massively parallel computers: an orbital decomposition approach. *Comp. Phys. Comm.*, 94(2,3):89–102, 1996.

[48] A. D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.*, 98(7):5648–52, 1993.

[49] B.G. Johnson, P.M.W. Gill, and J.A. Pople. The performance of a family of density functional methods. *J. Chem. Phys.*, 98(7):5612–5626, 1993.

[50] C. W. Bauschlicher. A comparison of the accuracy of different functionals. *Chem. Phys. Letters*, 246:40–4, 1995.

[51] G. S. Tschumper, J. T. Fermann, and H. F. Schaefer III. Structures, thermochemistry, and electron affinities of the PF_n and PF_n^- series, $n = 1-6$. *J. Chem. Phys.*, 104:3676, 1996.

[52] V. Barone, C. Adamo, and F. Mele. Comparisons of conventional and hybrid density functional approaches. cationic hydrides of first-row transition metals as a case study. *Chem. Phys. Letters*, 249:290–296, 1996.

[53] P. O. Löwdin. Quantum theory of many-particle systems. II. Study of the ordinary Hartree-Fock approximation. *Phys. Rev.*, 97(6):1490–1508, 1955.

[54] P. R. Surján, M. Kertész, A. Karpfen, and J. Koller. *Ab initio* numerical studies on density-matrix asymptotics in extended systems. *Phys. Rev. B*, 27(12):7583–7588, 1983.

[55] P. Maslen, C. Ochsenfeld, C. White, M. Lee, and M. Head-Gordon. Locality and sparsity of ab initio one-particle density matrices and localized orbitals. *J. Phys. Chem. A*, 102(12):2215–2222, 1998.

[56] R. Dovesi, C. Pisani, and C. Roetti. Exact-exchange Hartree-Fock calculations for periodic systems. II. Results for graphite and hexagonal boron nitride. *Int. J. Quant. Chem.*, 17:517–529, 1980.

[57] H. J. Monkhorst and M. Kertész. Exact-exchange asymptotics in polymer Hartree-Fock calculations. *Phys. Rev. B*, 24(6):3015–3024, 1981.

[58] M. Causà, R. Dovesi, C. Pisani, and V. R. Saunders. Treatment of the exchange interactions in Hartree-Fock linear combination of atomic orbital calculations of periodic systems. *J. Phys. Chem.*, 92:909–913, 1988.

[59] C. Pisani, E. Aprà, and M. Causà. Density matrix of crystalline systems. I. Long-range behavior and related computational problems. *Int. J. Quant. Chem.*, 38:395–417, 1990.

[60] C. Pisani, M. Causà, and R. Orlando. Density matrix of crystalline systems. II. The influence of structural and computational parameters. *Int. J. Quant. Chem.*, 38:419–433, 1990.

[61] W. Kohn. *Phys. Rev.*, 115:809, 1959.

[62] J. des Cloizeaux. Energy bands and projection operators in a crystal: analytic and asymptotic properties. *Phys. Rev. A*, 135:685, 1964.

[63] W. Kohn. Density functional theory for systems of very many atoms. *Int. J. Quantum Chem.*, 56:229–232, 1995.

[64] J. M. Ziman. *Models of Disorder: The Theoretical Physics of Homogeneously Disordered Systems.* Cambridge University Press, Cambridge, England, 1979.

[65] H. Böhm and R. Ahlrichs. A study of short-range repulsions. *J. Chem. Phys.*, 77(4):2028, 1982.

[66] F. Vigne Maeder and P. Claverie. The exact multicenter multipolar part of a molecular charge distribution and its simplified representations. *J. Chem. Phys.*, 88(8):4934–4948, 1988.

[67] I. Panas, J. Almlöf, and M. W. Feyereisen. *Ab Initio* methods for large systems. *Int. J. Quant. Chem.*, 40(6):797–807, 1991.

[68] V. Termath and N. C. Handy. A Kohn-Sham method involving the direct determination of the Coulomb potential on a numerical grid. *Chem. Phys. Letters*, 230(1,2):17–24, 1994.

[69] G. R. Ahmadi and J. Almlöf. The Coulomb operator in a Gaussian product basis. *Chem. Phys. Letters*, 246(4-5):364–370, 1995.

[70] J. Almlöf, K. Faegri, and K. Korsell. Principles for a direct SCF approach to LCAO-MO *ab-initio* calculations. *J. Comp. Chem.*, 3(3):385–399, 1982.

[71] E. Schwegler and M. Challacombe. Early studies. 1996.

[72] M. Dupuis, A. Marquez, and E. R. Davidson. *HONDO 95.3 from CHEM-Station.* IBM Corporation, Neighborhood Road, Kingston, NY. 12401, 1995.

[73] Coordinate files available upon request.

[74] S. F. Boys. Electronic wave functions.I. A general method of calculation for the stationary states of any molecular system. *Proc. Roy. Soc.*, A200:542, 1950.

[75] C. Ochsenfeld, C. A. White, and M. Head-Gordon. Linear and sub-linear scaling formation of Hartree-Fock type exchange matrices. *J. Chem. Phys.*, 109(5):1663–1669, 1998.

[76] M. Challacombe and E. Schwegler. MONDOSCF, a suite of programs for linear scaling SCF theory, unpublished.

[77] The observed average contraction length of the basis functions is less than the formal contraction length because insignificant primitive distributions have been discarded.

[78] W.J. Hehre, R.F. Stewart, and J.A. Pople. Self-consistent molecular-orbital methods. I. Use of Gaussian expansions of Slater-type atomic orbital. *J. Chem. Phys.*, 51(6):2657–2664, 1969.

[79] J.S. Binkley, J.A. Pople, and W.J. Hehre. Self-consistent molecular orbital methods. XXI. Small split-valence basis sets for first-row elements. *J. Am. Chem. Soc.*, 102(3):939–947, 1980.

[80] W.J. Hehre, R. Ditchfield, and J.A. Pople. Self-consistent molecular orbital methods. XII. Further extensions of Gaussian-type basis sets for use in molecular orbital studies of organic molecules. *J. Chem. Phys.*, 56(5):2257–2261, 1972.

[81] P.C. Hariharan and J.A. Pople. The influence of polarization functions on molecular orbital hydrogenation energies. *Theoret. Chim. Acta*, 28(3):213–222, 1973.

[82] T.H. Dunning and P.J. Hay. In H.F. Schaefer III, editor, *Methods of Electronic Structure Theory*, volume 3. Plenum Press, 1977.

[83] A. W. Appel. An efficient program for many-body simulation. *SIAM J. Sci. Stat. Comput.*, 6(1):85–103, January 1985.

[84] J. Barnes and P. Hut. A hierarchical $\mathcal{O}(NlogN)$ force-calculation algorithm. *Nature*, 324(6096):446–449, 1986.

[85] L. Hernquist. Hierarchical N-body methods. *Comp. Phys. Comm.*, 48(1):107–115, 1988.

[86] J. Makino. Vectorization of a treecode. *J. Comput. Phys.*, 87(1):148–160, 1990.

[87] M. S. Warren and J. K. Salmon. In *Proceedings SUPERCOMPUTING '93*, pages 12–21, Los Alamitos, CA, 1993. IEEE Computer Society Press.

[88] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Comp. Phys.*, 73(2):325–348, 1987.

[89] K. E. Schmidt and M. A. Lee. Implementing the fast multipole method in three dimensions. *J. Stat. Phys.*, 63(5–6):1223–1235, 1991.

[90] J. A. Board, J. W. Causey, James F. Leathrum, A. Windemuth, and Klaus Schulten. Accelerated molecular dynamics simulation with the parallel fast multipole algorithm. *Chem. Phys. Lett*, 198(1–2):89–94, 1992.

[91] H. G. Petersen, D. Soelvason, J. W. Perram, and E. R. Smith. The very fast multipole method. *J. Chem. Phys.*, 101(10):8870–8876, 1994.

[92] R. Zhou and B.J. Berne. A new molecular dynamics method combining the reference system propagator algorithm with a fast multipole method for simulating proteins and other complex system. *J. Chem. Phys.*, 103(i21):9444–9459, 1995.

[93] R. Kutteh and J. B. Nicholas. Efficient dipole iteration in polarizable charged systems using the cell multipole method and application to polarizable water. *Comp. Phys. Comm.*, 86(3):227–235, 1995.

[94] M. Challacombe, E. Schwegler, and J. Almlöf. Fast assembly of the Coulomb matrix: A quantum chemical tree code. *J. Chem. Phys.*, 104(11):4685–98, 1996.

[95] M. C. Strain, G. E. Scuseria, and M. J. Frisch. Achieving linear scaling for the electronic quantum Coulomb problem. *Science*, 271(5245):51–53, 1996.

[96] J. M. Perez and W. Yang. Fast evaluation of the Coulomb energy for electron densities. *J. Chem. Phys.*, 107(4):1218–1226, 1997.

[97] E. Schwegler, M. Challacombe, and M. Head-Gordon. Linear scaling computation of the Fock matrix. II. Rigorous bounds on exchange integrals and incremental Fock builds. *J. Chem. Phys.*, 106(23):9703–9717, 1997.

[98] R. G. Parr. Three remarks on molecular orbital theory of complex molecules. *J. Chem. Phys.*, 33(4):1184–1199, 1960.

[99] J. Cizek. A new method of approximative calculation of polycentric integrals used in quantum mechanical study of molecular structure. *Mol. Phys.*, 6(1):19, 1963.

[100] B. Voigt. On bridging the gap between the INDO and the NDDO schemes. *Theoret.Chim. Acta*, 31:289–295, 1973.

[101] M. J. S. Dewar and W. Thiel. A semiemperical model for the two-center repulsion integrals in the NDDO approximation. *Theoret.Chim. Acta*, 46:89–104, 1977.

[102] J. Cioslowski. Principles of ab initio SCF calculations with minimal storage requirements. *Comp. Phys. Comm.*, 53:117–122, 1989.

[103] T. R. Adams, R. D. Adamson, and P. M. W. Gill. A tensor approach to two-electron matrix elements. *J. Chem. Phys.*, 107(1):124–131, 1997.

[104] E. Schwegler and M. Challacombe. *work in progress*, 1998.

[105] P. Gill, B. Johnson, and J. Pople. Two-electron repulsion integrals over Gaussian s functions. *Int. J. Quantum Chem.*, 40(6):745–752, 1991.

[106] M. Challacombe, E. Schwegler, and J. Almlöf. Recurrence relations for calculation of the Cartesian multipole tensor. *Chem. Phys. Letters*, 241:67, 1995.

[107] B. C. Carlson and G. S. Rushbrooke. *Proc. Cambridge Philos. Soc.*, 46(12):626–633, 1950.

[108] J. Cipriani and B. Silvi. Cartesian expressions for electric multipole moment operators. *Mol. Phys.*, 45(2):259–272, 1982.

[109] I. Panas. Practical expressions for the two-center multipole expansion of r_{12}^{-1}. *Int. J. Quant. Chem.*, 53:255–263, 1995.

[110] W. W. Schulz. Note on the computation of the multipole tensor in spherical harmonics. *Chem. Phys. Lett.*, 254:337–340, 1996.

[111] H. G. Petersen, D. Soelvason, J. W. Perram, and E. R. Smith. Error estimates for the fast multipole methods. I. The two-dimensional case. *Proc. R. Soc. Lond. A*, 448:389–400, 1995.

[112] H.G. Petersen, E. R. Smith, and D. Soelvason. Error estimates for the fast multipole method. II. The three-dimensional case. *Proc. Roy. Soc. A*, 448:401–418, 1995.

[113] D. Sølvason and H. G. Petersen. Error estimates for the fast multipole method. *J. of Stat. Phys.*, 86(1,2):391–420, 1997.

[114] W. D. Elliot and J. A. Board Jr. Technical Report 94-001, Duke University Dept. of Electrical Engineering, 1994. Available at http://www.ee.duke.edu/Research/SciComp/papers.html.

[115] P. M. W. Gill, B. G. Johnson, and J. A. Pople. A simple yet powerful upper bound for Coulomb integrals. *Chem. Phys. Letters*, 217(1,2):65–68, 1994.

[116] T. S. Lee, D. M. York, and W. Yang. Linear-scaling semiempirical quantum calculations for macromolecules. *J. Chem. Phys.*, 105(7):2744–2750, 1996.

[117] S. Wolfram. *Mathematica, A System for Doing Mathematics by Computer.* Addison-Wesley, 2nd edition, 1991.

[118] M. Sofroniou. *The Mathematica Journal*, 3(3):74–80, 1993.

[119] I. Shavitt. The Gaussian function in calculations of statistical mechanics and quantum mechanics. *Meth. Comp. Phys.*, 2:1–45, 1963.

[120] P. J. Davis. Gamma function and related functions. In M. Abramowitz and I. A. Stegun, editors, *Handbook of Mathematical Functions*, pages 260–262. Dover Publications, Inc., New York, ninth edition, 1987.

[121] The CPU time needed to apply the MAC and PAC has been included in the multipole expansion time.

[122] J.A. Pople, R. Krishnan, H.B. Schlegel, and J.S. Binkley. Derivative studies in Hartree-Fock and Moller-Plesset theories. *Int. J. Quant. Chem. Symp.*, 13:225–241, 1979.

[123] M. Frisch, M. Head-Gordon, and J. Pople. Direct anaytic SCF second derivatives and electronic field properties. *Chem. Phys.*, 141(2,3):189–196, 1990.

[124] Translation invariance can be used to reduce the number of centers differentiated from four to three. For example, see Ref. [19].

[125] G. Karlström. On the possibility of improving the integral evaluation section in ab initio Hartree-Fock calculations on large molecules. *J. Comput. Chem.*, 2(1):83–86, 1981.