My PhD thesis

Maximilien Alexandre Ambroise

July 7, 2021

This is a fancy front page: contains supervisors, university name, location, exam date ...

Licensing

(empty)

Contents

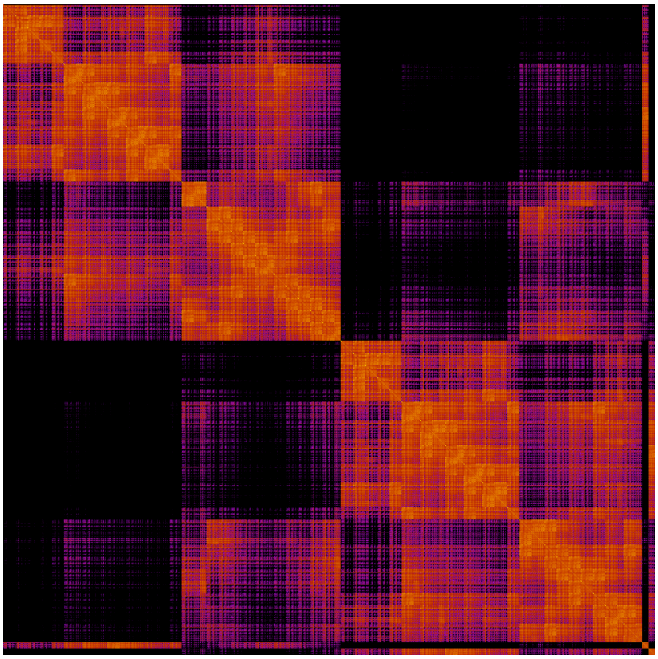


Figure 1: Adenine-guanine fock matrix Hartree FOck cc-pVTZ

0.1 Introduction

This is the introduction. Talk about sparsity. Show sparse matrix. How does sparsity arise, what can we do with it, where else does it emerge? State of arts: adc now, adc in the future

Chapter 1

Theory: The Basics

1.1 Ground Work

1.1.1 The Schrödinger Equation

1.1.2 Basis Sets

1.1.3 Electron Integrals

1.2 Hartree Fock

1.3 Post-Hartree Fock Ground State

1.3.1 Configuration Interaction

1.3.2 Perturbation Theory

1.3.3 Coupled Cluster

1.4 Post-Hartree Fock Excited State

1.4.1 Configuration Interaction

1.4.2 Coupled Cluster Linear Repsonse

1.4.3 Equation-of-Motion Coupled Cluster

1.4.4 Algebraic Diagrammatic Construction

Chapter 2

Local 0

While computational chemistry has manifested itself as a popular and widely used tool, its inherently steep scaling limits its applicibality to cheaper methods like DFT, orto small or middle sized molecules for post-Hartree Fock methods. Even Hartree Fock with O(N^3) expensive when comparing to the rest of the world of computer science (examples?). Very early on, with the development of CI methods, (Pulay) effort has been put into reducing the prefactor and computational complexity for QC methods. Example: One of the most time-consuming steps in Post-Hartree Fock is the formation of the molecular orbitals, e.g. formation of the OVOV block of the MO integral tensor as encountered in CC and MP2:

$$(ia \mid jb) = \sum_{b}^{vir} C_{\sigma b} \sum_{a}^{vir} C_{\nu a} \sum_{j}^{occ} C_{\lambda j} \sum_{i}^{occ} C_{\mu i} \, (\mu\nu \mid \lambda\sigma) \tag{2.1}$$

By efficiently refactoring the sums, the MO-AO integral transformation scales as O(N^4). There are two reasons for the large cost: first, and quite obviously, a rank-4 tensor scales very fast, hence O(N4), also becomes a bottle neck for tensor contractions as many indices. Secondly, for large basis sets with triple-zeta quality or higher, or basis sets with diffuse functions, the virtual orbital space is very large, and can be multiple times the size of the occupied space. Attempts to reduce scaling can be grouped into two groups: screening-based methods and domain-based methods. Screening-based methods recast existing equations into the AO basis and use the sparsity and fast decay between AOs to establish highly efficient screening algorithms to lower the scaling of integral transformation. Domain-based methods stay in the MO basis, but a localized one, and attempt to assign domains of virtual molecular orbitals to a single LMO or a pair of LMOs, to obtain a more compact representation. Other attempts at mitigating the cost of MO-AO transformation is to exploit the rank sparsity of the AO ERI tensor. Density fitting and Cholesky decompositions can refactor the ERI tensor into a product of two 3-dim tensors. Tensor Hypercontraction goes even further and decomposes into 4 2-dim tensor. Density does not inherently lower scaling of methods, but rather reduces the prefactor associated with integral transformation. In special cases, decomposition techniques allow a refactoring of the working equations into lower scaling. Examples include the coulomb part of the Fock-build (O3 to O2) and SOS-MP2, SOS-CC2 or SOS-ADC(2) (O5 to O4). Density fitting and local approximations can be combined, to yield the best of both world in what is known as local density fitting. All of the above methods have their fair share of problems, some more than others. We will first address principles of density fitting,

before looking at possible orbital representations, and how they can be used for reduced scaling. Also go to local density fitting, and finally how the methods are implemented for ground state (HF,MP2,CCSD) and excited state computations (CI, CCLR, ADC).

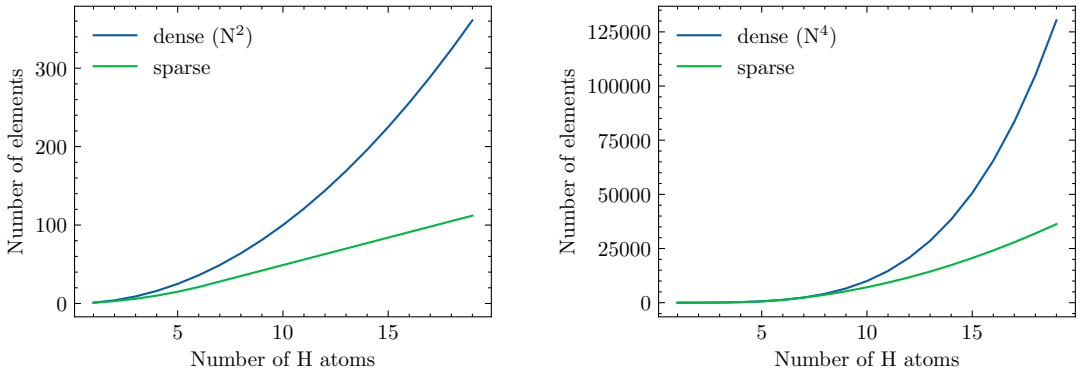
2.1 Sparsity in Electronic Structure Theory

Sparsity is a core concept in electronic structure theory. Many of the most commonly encountered matrices and tensors exhibit some form of sparsity.

2.1.1 Element-Wise Sparsity of Electron Integrals

Molecular electron integral evaluation can become prohibitively expensive for large systems, espcially the four-dimensional ERI tensor which formerly scales as $\mathcal{O}(N^4)$. It is therfore imperative to exploit the exponential decay of the GTO basis.

Consider a model system consisting of n hydrogen atoms arranged in a line, with a distance of 1 a_0 between one another, and a primitive 1s Gaussian function attached to each atom. Figure ... shows the scaling behaviour for the overlap and electron repulsion integrals of this system. A full line is used to show the number of total elements, while the dotted line represents the number of significant integrals with magitude $>$ 1e-10. From observing both graphs, it becomes apparent that for increasing number of atoms, many of the electron integrals can be ignored. Therefore, one only needs to store integrals above a certain threshold. This is also known as *element-wise sparsity*.

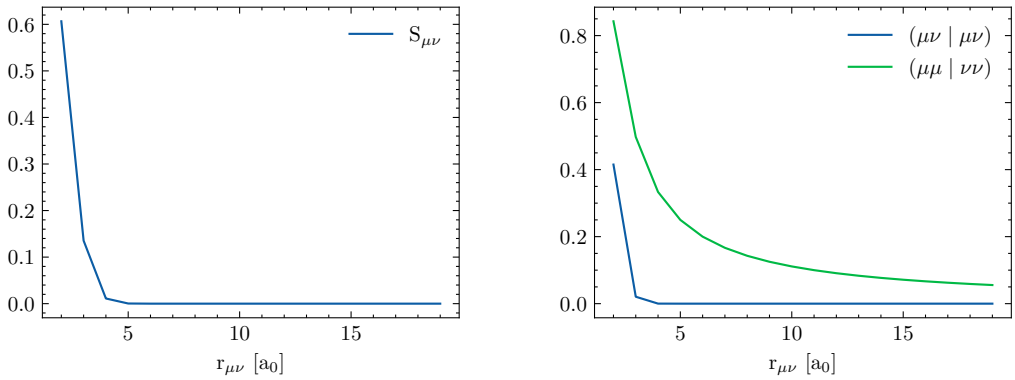


(a) (b)

Figure 2.1: (a) Number of significant entries (full line) in the overlap matrix for a hydrogen atom chain, with a threshold of 1e-10. The dotted line shows the total number of elements for the dense matrix, which scale as N^2. (b) Number of significant entries (full line) in the electron repulsion integral tensor for a hydrogen atom chain, with a threshold of 1e-10. The dotted line shows the total number of elements for the dense tensor, which scale as N^4.

Linear Scaling Overlap Integrals

While the overlap integrals formely scale with $\mathcal{O}(N^2)$, it can be shown that the number of significant elements scales *linearly*. First, consider the product of two 1s GTOSs χ_A



(a) (b)

Figure 2.2: (a) Magnitude of the overlap integral between two Gaussian 1s orbitals as a function of distance r (exponential decay). (b) Magnitude of the electron repulsion integral betwwen two Gaussian 1s orbitals as a function of r. The short range interaction $(\mu\nu \mid \mu\nu)$ decays at a muchfaster rate with e^{-r^2}, compared to the long range interaction with $1/R$.

and χ_B, centred at $\mathbf{A}$ and $\mathbf{B}$, with exponents α and β. The Gaussian product theorem (GPT) states that the result is itself also a (scaled) Gaussian function

$$\chi(A,\alpha)\chi(B,\beta) = e^{-\alpha|\mathbf{r}-\mathbf{A}|^2} e^{-\beta|\mathbf{r}-\mathbf{B}|^2} = \kappa\chi(P,\alpha+\beta) \tag{2.2}$$

with the scaling factor κ

$$\kappa = e^{-\frac{\alpha\beta}{\alpha+\beta}|\mathbf{A}-\mathbf{B}|^2} \tag{2.3}$$

and the centre-of-charge coordinate P

$$\mathbf{P} = \frac{\alpha\mathbf{A}+\beta\mathbf{B}}{\alpha+\beta} \tag{2.4}$$

Spatioal integration yields the expression for the overlap between χ_A and χ_B

$$S_{AB} = \int \kappa\chi_P dr = \kappa\left(\frac{\pi}{\alpha+\beta}\right)^{3/2} \tag{2.5}$$

The magnitude of the overlap integral is proportional to the scaling factor κ which decays exponentially with the distance between GTO centres. In the case of the model system given above, where $\alpha = \beta$, the distance at which the integral falls below a certain threshold ϵ is given by

$$d_s = \sqrt{\alpha^{-1} ln\left[\left(\frac{\pi}{2\alpha}\right)^3 \epsilon^{-1/2}\right]} \tag{2.6}$$

Which in our case is equal to 6.9 a_0. Each hydrogen atom therefore only has significant overlap with a finite number n_{max} of other centres. For atom chains with $n > n_{max}$, the number of non-zero elements in the overlap matrix will no longer scale as n^2, but *linearly* with nn_{max}. For more realistic, three-dimnensional molecular systems, the crossover is less clearly defined due to the non-uniform distribution of atoms and different GTO exponents. Nonetheless, if a system grows sufficiently large, the overlap integrals still scale linearly. Similar arguments can be brought forth for the kinetic-energy integrals as well.

Quadratic Scaling Electron Repulsion Integrals

Using the Guassian product theorem established above, we can express the two-electron repulsion integrals of four primtive 1s Gaussian functions $s(A,\alpha)$, $s(B,\beta)$, $s(C,\gamma)$ and $s(D,\delta)$ as

$$\begin{aligned} g_{ABCD} &= \int s(A,\alpha)s(B,\beta)\frac{1}{|\mathbf{r_1}-\mathbf{r_2}|}s(C,\gamma)s(D,\delta)dr \\ &= \int \kappa s(P,\alpha+\beta)\frac{1}{|\mathbf{r_1}-\mathbf{r_2}|}\lambda s(Q,\gamma+\delta) \end{aligned} \tag{2.7}$$

where $s(P,p)$ and $s(Q,q)$ are Gaussian distributions with

$$\mathbf{P} = \frac{\alpha\mathbf{A}+\beta\mathbf{B}}{\alpha+\beta}; \quad \mathbf{Q} = \frac{\gamma\mathbf{C}+\delta\mathbf{D}}{\gamma+\delta} \tag{2.8}$$

$$\kappa = e^{-p|\mathbf{A}-\mathbf{B}|^2}; \quad \lambda = e^{-q|\mathbf{C}-\mathbf{D}|^2} \tag{2.9}$$

$$p = \frac{\alpha\beta}{\alpha+\beta}; \quad q = \frac{\gamma\delta}{\gamma+\delta} \tag{2.10}$$

The coulomb integrals can then be evaluated as

$$g_{ABCD} = \sqrt{\frac{4\eta}{\pi}}S_{AB}S_{CD}F_0\left(\eta\left|\mathbf{P}-\mathbf{Q}\right|^2\right) \tag{2.11}$$

with the Boys function F_0 and the reduced exponent η given by

$$\eta = \frac{pq}{p+q} \tag{2.12}$$

The Boys function is an important function appearing in many expressions for molecular integral evaluation. There are two expressions that bound the Boys function

$$\begin{aligned} F_n(x) &\leq \frac{1}{2n+1} \quad \text{for small } x \\ F_n(x) &\leq \frac{(2n-1)!!}{2^{n+1}}\sqrt{\frac{\pi}{x^{2n+1}}} \quad \text{for large } x \end{aligned} \tag{2.13}$$

Using the Boys function's upper bounds, we can derive an upper bound for the electron repulsion integrals of our model system

$$g_{ABCD} \leq min\left\{\sqrt{\frac{4\eta}{\pi}}S_{AB}S_{CD}, \frac{S_{AB}S_{CD}}{|\mathbf{P}-\mathbf{Q}|}\right\} \tag{2.14}$$

The left-hand upperbound represents the short-range limit of the Boys function, and the right-hand one the long-range limit. In the short-range limit, i.e. for increasing distance R_{AB} or R_{CD}, the magnitude of g decreases *expoentially*. As shown in the previous section, the non-zero elements of the overlap integrals S_{AB} and S_{CD} scale linearly with system size, and therefore the number of significant electron repulsion integrals scales with N^2 in total. It should be noted, that in the long-range limit whith increasing distance R_{PQ} between

product densities, the number of elements in g will eventually scale linearly. However, the *algebraic* $1/R$ decay of the long-range interactions is so slow that it practically useless for the size of moleculaes that can be tackled with current technologies. In the case of the hydrogen atom chain, the integrals $(\mu\mu \mid \nu\nu)$ only fall below 1e-10 for R_{PQ} greater than 10^{10} a_0. While the long-range decay is impractical for use in the case of the electron repulsion integrals, there are instances such as in AO-MP2 where *bra* and *ket* decay as $1/R^4$. Knowing that the electron repulsion integrals are sparse is only the first step. One also has to develop a *screening* method to avoid computing small integrals, by finding a general upperbound. It has been shown (Roo1951) that g is positive-definite, fullfilling the relationship

$$\sum_{abcd} c_{ab} g_{abcd} c_{cd} > 0 \tag{2.15}$$

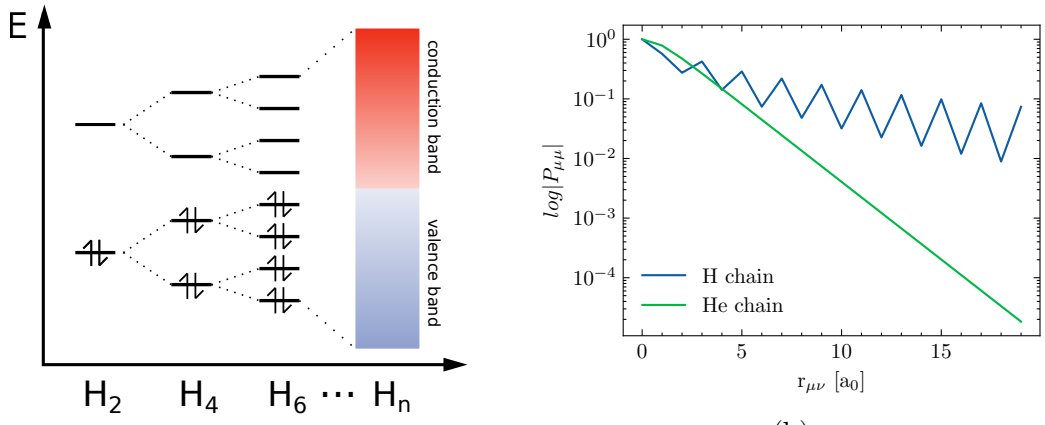
where c are one-electron orbital distrbutions. One can then apply the Schwarz inequality (Has1989) to obtain an upper bound expression for g

$$(\mu\nu \mid \sigma\lambda) \leq (\mu\nu \mid \mu\nu)\,(\sigma\lambda \mid \sigma\lambda) = Q_{\mu\nu} Q_{\sigma\lambda} \tag{2.16}$$

The matrix $\mathbb{Q}$ contains the square root of the short-range diagonal entries of g, and is also known as the Schwarz matrix. $\mathbf{Q}$ can be evaluated quickly with $\mathcal{O}(N^2)$ effort and inidividual integrals can be efficiently screened. It should be noted that Schwarz screening does not take into account the $1/R$ decay between product densities, which makes the method less useful in methods like AOMP2.

2.1.2 Element-Wise Sparsity of the Density Matrix

The decaying behaviour of the density matrix has been extensively studied in solids for atom-centred Bloch and Wannier functions. It was shown that for insulators, i.e. systems with large band-gaps, the contributions $P_{\mu\nu}$ decay exponentially with increasing distance $R_{\mu\nu}$, while for systems with small or no band gaps, such as metals, the elements decay algebraically. The same observations have been made for non-periodic systems using atomic orbitals as basis. For molecules with a large HOMO-LUMO gap, e.g. alkanes, the number of non-zero elements in the atomic orbital density matrix scales linearly with increasing system size. On the other hand, molecules with strong electron delocalization, such as conjugated polyenes, have have a small HOMO-LUMO gap, and the density matrix elements decay much slower. Consider again a chain of hydrogen atoms, equally spaced by a_0, each with one 1s Gaussian function, this time with N_{atom} atoms. Figure ... shows the MO diagram for increasing chain length. In the limit where $N_{atom} \to \infty$, the system takes on a band structure, similar to how they are encountered in a metal, with a smooth transition between occupied (valence) and virtual (conductance) band. In other words, the HOMO-LUMO gap becomes increasingly small. For hydrogen atom which each have one electron, the band is half filled, and the system is a conductor. If the Hydrogen atoms are replaced by Helium atoms, with two electrons per site, the band is fully filled and the system becomes an insulator. The magnitude of the density matrix elements $P_{\mu\nu}$ is plotted in Figure ... as a function of increasing distance between 1s functions. The elements decay much slower for the conducting hydrogen chain, while a rapid exponential decay can be observed in the case of the insulating helium chain.



(a) (b)

2.1.3 Diagrammatic Notation

Hollmann et al. (ref) have introduced a simple graphical representation to show contributing factors to the sparsity of a given matrix, tensor or tensor contraction. Each tensor index is represented as a vertex which contributes. Non-connected vertices each contribute $\mathcal{O}(N)$ elements to the overall expression. A sparsity relationship between two indices is represented as an *edge* connecting two vertices. In this case, the number of *pairs* scales as $\mathcal{O}(N)$. Consider the two electron integral tensor $(\mu\nu \mid \sigma\lambda)$. From the previous section, we know that the index pairs μ, ν and λ, σ are related by overlap. The diagrammtic representation takes the form:

$$\mu \overset{\mathrm{S}}{\longleftrightarrow} \nu \qquad \sigma \overset{\mathrm{S}}{\longleftrightarrow} \lambda$$

There are two pairs of connected vertices, which indicates that the integrals can be evaluated with $\mathcal{O}(N^2)$ effort, which is in agreement with the findings above. The S denotes the overlap relationship between vertices. For another example, consider the Hartree Fock expression for the exchange matrix

$$K_{\mu\nu} = (\mu\sigma \mid \nu\lambda)\, P_{\lambda\sigma} \tag{2.17}$$

Diagrammtically, the expression for $\mathbf{K}$ can be represented as

$$\mu \overset{\mathrm{S}}{\longleftrightarrow} \sigma \overset{\mathrm{P}}{\longleftrightarrow} \lambda \overset{\mathrm{S}}{\longleftrightarrow} \nu$$

The connection between σ and λ is also known as a "P-junction", which represents the sparsity relationship arising due to the exponential decay of density matrix elements. The sparsity graph is fully connect, which suggests that $\mathbf{K}$ can be evaluated with $\mathcal{O}(N)$ effort. This is indeed the case, as shown by the ONX or LinK method. For linear scaling to emerge, indices of an expression therefore need to be fully linked. This simple but important fact is also known the linked index rule (LIR). Diagrams also show which factors can influence the performance of the scaling, such as diffuseness of the atomic orbitals (slower S decay) or size of of the HOMO-LUMO gap (slower P decay). One can therefore conclude that the expression for $\mathbf{K}$ as given above, is less suitable for large basis sets and non-insulators.

2.1.4 Rank Sparsity

A positive semi-definite matrix **A** has the property that it can be decomposed as a product

$$\mathbf{A} = \mathbf{B}\mathbf{B}^T \tag{2.18}$$

where **A** has dimensions N by N, and **B** has dimensions N by $rank(A)$. The rank represents the number of linearly independent column vectors in matrix, and for $rank(A) < N$, the matrix is said to be rank-deficient. The decomposition matrix **B** therefore is more compact and needs less storage space than **A**. There are different ways to compute **B**, such as Cholesky decomposition or QR decomposition. The tensor $(\mu\nu \mid \sigma\lambda)$ can be represented as a N_{AO}^2 by $N_{AO}^2 0$ matrix with combined row indices $I = \mu + N_{AO} * \nu$ and column indices $J = \sigma + N_{AO} * \lambda$. Because the tensor has been shown to be postive semi-definite, there also exists a decomposition, such that

$$(\mu\nu \mid \sigma\lambda) = A_{(\mu\nu)(\sigma\lambda)} = B_{(\mu\nu)X} B_{(\sigma\lambda)X} \tag{2.19}$$

The rank of **A** is in general much smaller than the combined index range N_{AO}^2, and scales linearly rather than quadratically with the number of basis sets. The decomposition tensor **B** is therefore 3-dimensional, rather than 4-dimensional, which reduces the storage needed by an order of magnitude from $\mathcal{O}(N^4)$ to $\mathcal{O}(N^3)$, but only in the case where $(\mu\nu \mid \sigma\lambda)$ is dense. In the limit of large molecules, the NZEs of **B** also scale with $\mathcal{O}(N^2)$. Rather than for the molecular integrals in the AO basis, decomposition techniques are more useful for reducing the storage size of molecular integrals in the canonical MO basis

$$(ia \mid jb) = C_{\mu i} C_{\sigma a} B_{\mu\sigma X} B_{X\nu\lambda} C_{\nu j} C_{\lambda b} = B_{iaX} B_{Xjb} \tag{2.20}$$

The AO-MO transformation step is also drastically sped up, but remains a $\mathcal{O}(N^4)$ effort. Rank sparsity has therefore little impact on the overall scaling, but rather reduces the scaling *prefactor*. Over the years, different methods have been proposed to compute **C**, such as density fitting, Cholesky decompostion, pseudo-spectral methods, or tensor hypercontraction. Density matrices at different levels of theory (Hartree Fock, MP2, CC ...) also exhibit rank sparsity. Decomposition of such matrices play an important role in local molecular orbital schemes and low scaling electronic structure methods, as will be shown in later sections.

2.2 Density Fitting

The method of choice in this thesis for the decomposition of two-electron molecular integrals is *density fitting* (DF), also knwon as *resoulation of the identity* (RI). For a brief exploration of other popular methods, the reader is referred to (ANNEX).

2.2.1 Basics of Density Fitting

The two-electron integrals can be expressed in terms of the charge product densities $\rho_{\mu\nu} = \chi_\mu \chi_\nu$ as

$$(\mu\nu \mid \sigma\lambda) = \int\int \frac{\rho_{\mu\nu}(\mathbf{r_1})\rho_{\sigma\lambda}(\mathbf{r_2})}{\mathbf{r_1} - \mathbf{r_2}} d\mathbf{r_1} d\mathbf{r_2} \tag{2.21}$$

The charge densities ρ can be approximated by fitting them to a set of atom-centred auxiliary functions χ_P

$$\rho_{\mu\nu}(\mathbf{r}) = C_{P\mu\nu}\chi_P(\mathbf{r}) + \Delta\rho_{\mu\nu} \tag{2.22}$$

Or in the chemist's notation:

$$|\mu\nu) = C_{P\mu\nu}\,|P) + |\epsilon_{\mu\nu}) = |\tilde{\mu\nu}) + |\epsilon_{\mu\nu}) \tag{2.23}$$

where $C_{P\mu\nu}$ are the fitting coefficients, and $\Delta\rho_{\mu\nu}$ or $|\epsilon_{\mu\nu})$ is the error introduced by the fitting procedure. Eq. ... is known as the density fitting approximation (Whi73,Bae1973). The two-electron integrals then take the form

$$\begin{aligned}(\mu\nu \mid \sigma\lambda) &= (\widetilde{\mu\nu}|\widetilde{\sigma\lambda}) + \underbrace{(\widetilde{\mu\nu}|\epsilon_{\sigma\lambda}) + (\epsilon_{\mu\nu}|\widetilde{\sigma\lambda})}_{\text{first order}} + \underbrace{(\epsilon_{\mu\nu} \mid \epsilon_{\sigma\lambda})}_{\text{second order}} \\ &= (\widetilde{\mu\nu}|\widetilde{\sigma\lambda}) + \epsilon_J^{(1)} + \epsilon_J^{(2)}\end{aligned} \tag{2.24}$$

Here, $\epsilon_J^(1)$ and $\epsilon_J^(2)$ represent the first order (linear) and second order (quadratic) error. The fitting coefficients are then generally found by minimizing $\epsilon_J^{(2)}$. Substituting $(\epsilon_{\mu\nu}| = (\mu\nu - \widetilde{\mu\nu}|$ gives

$$\frac{\partial}{\partial C^P_{\mu\nu}}\left(\mu\nu - \widetilde{\mu\nu} \mid \sigma\lambda - \widetilde{\sigma\lambda}\right) = 0 \tag{2.25}$$

which then yields a set of linear equations

$$(\mu\nu \mid P) - \sum_Q C^Q_{\mu\nu}\,(Q \mid P) = 0 \tag{2.26}$$

Finding the fitting coefficients by minimizing $\epsilon_J^{(2)}$ has the important feature that $eps_J^{(1)} = 0$, which can be shown by substituting Eq. ... back into Eq. ... . The total electron integral error is therefore *quadratic* in the fitting error. Fitting procedures where the coefficients $C^P_{\mu\nu}$ satisfy Eq. ... are termed *robust* (Dun2000). Any restrictions posed on $C^P_{\mu\nu}$ makes ϵ_1 diffent from zero and the error scales linearly. Eq. ... needs the evaluation of the three-centre-two-electron (3c2e) and two-centre-two-electron (2c2e) integrals in the auxiliary basis set $\{P\}$

$$(\mu\nu \mid P) = \int\int \chi_\mu(\mathbf{r_1})\chi_\mu(\mathbf{r_1})\frac{1}{\mathbf{r_1} - \mathbf{r_2}}\chi_P(\mathbf{r_2})d\mathbf{r_1}d\mathbf{r_2} \tag{2.27}$$

$$(P \mid Q) = \int\int \chi_P(\mathbf{r_1})\frac{1}{\mathbf{r_1} - \mathbf{r_2}}\chi_Q(\mathbf{r_2})d\mathbf{r_1}d\mathbf{r_2} \tag{2.28}$$

The fitting coefficients are generally computed by inverting $(P \mid Q)$, which leads to the following approximation for the four-centre-two-electron integrals

$$(\mu\nu \mid \sigma\lambda) \approx (\mu\nu \mid P)\,(P \mid Q)^{-1}\,(Q \mid \sigma\lambda) \tag{2.29}$$

Matrix inversion is a $\mathcal{O}(N^3)$ computational effort. For more details on precision and best practices involving matrix inversion, see ... .

2.2.2 Scaling of the 3c2e Integrals

Using the diagrammatic representation introduced earlier, the 3c2e integral tensor reduces to

$$\mu \overset{\mathrm{S}}{\longleftrightarrow} \nu \qquad P$$

The number of non-zero elements elements therefore scales as $\mathcal{O}(N^2)$, just like for the 4c2e integrals. Similarly, the Schwarz inequality can be used to screen out small integrals

$$|(\mu\nu \mid P)| \leq |(\mu\nu \mid \mu\nu)|^{1/2} \, |(P \mid P)|^{1/2} \tag{2.30}$$

As mentioned above, Schwarz screening does not take into account increasing bra-ket distance. The long-range decay is too slow to be of any advantage in the case of the 4c2e integrals. However, it was found (Hol2015) that for an auxiliary density $\chi_P(\mathbf{r}$ with angular momentum l_P, the 3c2e integrals actually decay as $1/R^{-1-l_P}$ with increasing bra-ket distance, establishing a weak sparsity relationship between $(\mu\nu|$ and $|P)$

$$\mu \overset{\mathrm{S}}{\longleftrightarrow} \nu \qquad P$$
$$\underset{1/R^{-1-l_P}}{\vdots}$$

In principle, the 3c2e integrals can be evaluated with linear effort. Hollmann et al (Hol2015) have introduced a tight upperbound, known as the SVQl estimator, to exploit this faster decay. Due to the dependence on l_P, the screening is most effective with larger basis sets with high angular momentum functions.

The fitting coefficients evaluated as $C^P_{\mu\nu} = (\mu\nu \mid Q)\,(Q \mid P)^{-1}$ formerly scale with $\mathcal{O}(N^3)$

$$\mu \overset{\mathrm{S}}{\longleftrightarrow} \nu \qquad Q \qquad P$$

due to the inverse of $(P \mid Q)$ not being sparse.

2.2.3 Local Density Fitting: Principles

The long-range behaviour introduced by Eq. ... is often deemed "unphysical" (Tew2019). *Local density fitting* (LDF) methods circumvent this problem by forcing a more rapid decay of long-range contributions, either (a) by using a different metric in the fitting procedure Eq. ... or (b) by constructing domains $[\mu\nu]$ that exclude distant fitting functions P a priori. In both cases, Eq. ... is no longer fulfilled and the error in the electron integrals ... increases linearly with the fitting error, and the density fitting procedure is no longer robust. LDF methods therefore use a different expression for the electron integrals which includes the first order terms to remove the linear error

$$(\mu\nu \mid \sigma\lambda) \approx (\widetilde{\mu\nu}|\sigma\lambda) + (\mu\nu|\widetilde{\sigma\lambda}) - (\widetilde{\mu\nu}|\widetilde{\sigma\lambda}) \tag{2.31}$$

which is known as Dunlap's robust density fitting formula.

Type	Ref.	$g(r_{12})$
Overlap	[A]	1
Coulomb-Attenuated	[B]	$\frac{erfc(\omega r_{12})}{r_{12}}$
Yukawa	[C]	$\frac{e^{-\omega r_{12}}}{r_{12}}$
Gaussian-Damped	[D]	$\frac{e^{-\omega r_{12}^2}}{r_{12}}$

Table 2.1: Expressions for the operator g in different local metrics.

2.2.4 LDF (I): Short-Range Metrics

The first type of LDF methods replaces the fitting procedure in the Coulomb metric in Eq. ... by a more general expression

$$B^P_{\mu\nu} - C^Q_{\mu\nu} M_{QP} = 0 \tag{2.32}$$

where $B^P_{\mu\nu}$ and M_{PQ} are the 3-centre- and 2-centre-2-electron integrals given by

$$B^P_{\mu\nu} = \int\int \chi_\mu(\mathbf{r_1})\chi_\nu(\mathbf{r_1})g(\mathbf{r_1},\mathbf{r_2})\chi_P(\mathbf{r_2})d\mathbf{r_1}d\mathbf{r_2} \tag{2.33}$$

$$M_{PQ} = \int\int \chi_P(\mathbf{r_1})g(\mathbf{r_1},\mathbf{r_2})\chi_Q(\mathbf{r_2})d\mathbf{r_1}d\mathbf{r_2} \tag{2.34}$$

with g being the local metric. A list of known local metrics is given in Table ... . Earliest forms of density fitting actually first used an overlap metric to directly minimizing the norm of the residual $R_{\mu\nu} = (\mu\nu| - (\widetilde{\mu\nu}|$ using linear least squares (Baer), and the fitting coefficients are computed as

$$C^P_{\mu\nu} = S^{-1}_{PQ}(\mu\nu Q) \tag{2.35}$$

where S_{PQ} is the overlap in the auxiliary basis, and $(\mu\nu Q)$ are the 3-centre-**1-electron** overlap integrals. While the overlap metric has the most rapid decay and the quantities in Eq. can be evaluated in $\mathcal{O}(N)$ time, it has the worst accuracy of all metrics. One solution to this problem is to introduce a metric which is intermediate between overlap and coulomb fitting. Examples include the Yukawa, Coulomb- and Gaussian-attenuated metrics (Table ...). These intermediate metrics introduce a damping factor ω to control the sparsity and accuracy of the density fit. In the limit where $\omega \to 0$, and $\omega \to \infty$, one recovers the coulomb and overlap metric, respectively. Figure ... shows the decay behaviour of a local metric, using the Coulomb-attenuated metric as an example, for ω = 0.01, 0.1 and 1.0, compared to the overlap and the Coulomb metric. ...

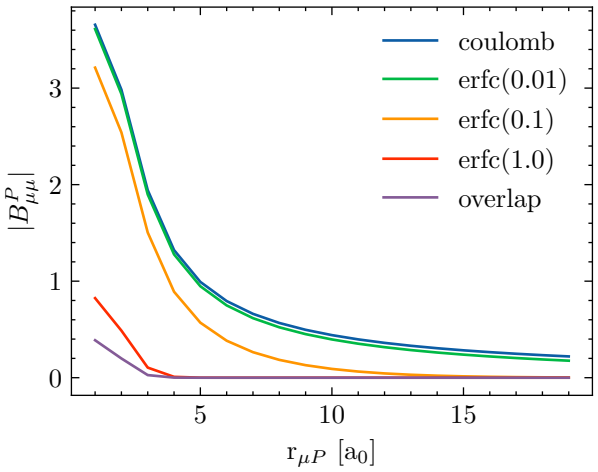


Figure 2.4: Some caption

2.2.5 LDF (II): Local Domains

The second method to force locality in density fitting consists in constructing local domains for each atom, pair of atoms or molecular orbital, and excluding any auxiliary functions that lie outside, which can drastically reduce the dimension of the fitting procedure.

Atomic Resolution of the Identity

The simplest example of a domain is one that includes a single atom. The atomic resolution of the identity (ARI) [] uses a fitting procedure where the sum over auxiliary function Q only includes those which are centred on the same atom A as the atomic orbital μ

$$|\widetilde{\mu\nu}) = \sum_{Q \cup A_\mu} (P \mid Q)^{-1}_{A_\mu} (\mu\nu \mid Q) \tag{2.36}$$

Each atom X has its own metric matrix inverse $(P \mid Q)^{-1}_X$ which takes the form

$$(P \mid Q)^{-1}_X = B_X \left((P \mid Q)_D + B_X (P \mid Q)_{OD} B_X \right)^{-1} B_X \tag{2.37}$$

where $(P \mid Q)_D$ and $(P \mid Q)_{OD}$ are the diagonal and off-diagonal part of $(P \mid Q)$ respectively. B_X is a so-called *bump matrix* which imposes a fast, but smooth decay between functions P and Q in order to avoid using all functions P for the fitting in Eq. ... . For further details, the reader is referred to the orginal publication. The bump matrix uses multiple distance criteria which make the ARI less of a black-box method.

Pair-Atomic Resolution of the Identity

A more popular and simple variant of atomic density fitting is the pair-atomic resolution of the identity (PARI) method (Mer2013). As the name implies, the domains include atom *pairs* rather than a single atom. Again expressing it in terms of the fitting procedure

$$(P \mid \mu\nu) = \sum_{Q \in A \cup B} (P \mid Q) C^Q_{\mu\nu} \quad \forall P \in A \cup B \tag{2.38}$$

The number of linear equations is equal to the number of non-zero pairs $\mu\nu$, which scales linearly. However, the PARI approach poses heavy constraints on the fitting coefficients, which leads to large integral errors. Merlot et al. proposed to increase the atomic pair domain with any atoms which lie between A and B. Alternatively, larger and more diffuse basis sets can be used. In both cases, performance is sacrificed for increased accuracy. The abesence of any distance dependent parameters or thresholds still make it an attractive method both for Hartree Fock and Post-Hartree Fock methods (refs).

LDF using Local Molecular Orbitals

Finally, domains can also be formed using local molecular orbitals instead of AOs. LMOs are larger than AOs, but are still generally centred on only a few atoms. The exact atomic sites can be determined for example by using a Mulliken population analysis. Consider the density fitting procedure as proposed by Polly et al. for their LDF-Hartree Fock method (pol2004)

$$(\mu i \mid P) = \sum_{Q \in [i]_{fit}} (P \mid Q)\, C^Q_{\mu i} \tag{2.39}$$

The fitting coefficients are determined individually for each AO-LMO pair $|\mu i)$, and include only those auxiliary functions centred on atoms in the fitting domain $[i]_{fit}$ for which the Mulliken charges are above a given threshold. Although the fitting coefficients need to be recomputed for each update of the MO coefficients, the number of $|\mu i)$ pairs scales linearly with system size. This type of local density fitting and variations thereof are predominantly used in pair-orbital specific local correlation methods, and will be explained in more detail further below.

2.2.6 LDF (III): Quasi-Robust Density Fitting

Local density fitting inmposes constraints on the fitting procedure, and the integral error consequently scales linearly with the fitting error. Using Dunlap's robust formula is deemed necessary in most cases to achieve acceptable accuracy, but reintroduces the slowly decaying 3c2e integrals. Furthermore, replacing the 4c2e integrals by Eq. ... greatly increases the complexity of expressions in electronic structure theory, which is still manageable for ground state methods, but quickly becomes cumbersome for excited states.

Quasi-robust density fitting (QRDF) aims to combine the exponential decay behaviour of LDF with accuracy comparable to standard density fitting, without the use of Dunlap's formula. Again, consider the fitting procedure

$$\sum_Q (P \mid Q)\, C^Q_{\mu\nu} = (P \mid \mu\nu) \tag{2.40}$$

The sets of auxiliary functions $\{P\}$ and $\{Q\}$ have different roles. The functions Q fit the charge density $|\mu\nu)$, while the P functions act as *test functions* where the electron integrals should be accurate, i.e. where $(X \mid \widetilde{\mu\nu}) \approx (X \mid \mu\nu)$. For two functions μ and ν not located on the same atom, their charge density $|\mu\nu)$ lies in the vacuum between them, and the atom-centred auxiliary functions may be ill-suited to fit $|\mu\nu)$. For this reason, the fitting procedure draws from all fitting functions $\{P\}$ spanning the whole molecule

to cancel out the linear error, which in consequence introduces long-range contributions in $C^P_{\mu\nu}$ in the coulomb metric, even if $|P)$ is not close to $|\mu\nu)$. The basic idea of QRDF is to only chose fitting functions $\{P\}$ close to $|\mu\nu)$ via overlap criteria, but still perform the fitting procedure in the coulomb metric.

The QRDF Fitting Procedure

For a set of given μ, ν, select a set of *fitting function* $\{P_{\mu\nu}\} \in \{P\}$ close to $|\mu\nu)$ according to the criteria

$$\left| \sum_R S^{-1}_{PR}(R\mu\nu) \right| > T \tag{2.41}$$

where S is the auxiliary overlap matrix, and $(R\mu\nu)$ are the 1-centre-3-electron overlap integrals. Next, choose a set of test functions $\{Q_{\mu\nu}\} \in \{P\}$ using

$$f(Q_{\mu\nu}, P_{\mu\nu}) < R \tag{2.42}$$

with

$$f(A, B) = \frac{\alpha\beta}{\alpha + \beta} |\mathbf{A} - \mathbf{B}|^2 \tag{2.43}$$

where for two auxiliary functions A and B, the values α, β are their smallest primitive exponents and $\mathbf{A}$, $\mathbf{B}$ are their respective positions. The fitting coefficients are then determined via

$$\sum_P (Q_{\mu\nu} \mid P_{\mu\nu})\, C^P_{\mu\nu} = (Q_{\mu\nu} \mid \mu\nu) \tag{2.44}$$

where the fitting coefficients are accurate within the set of test functions $\{Q_{\mu\nu}\}$. The linear equations in Eq. ... can be solved via QR decomposition of the rectangular matrix $(Q_{\mu\nu} \mid P_{\mu\nu})$. The QRDF scheme depends on two parameters, T and R. In the limit where $T \to 0$ and $R \to \infty$, the standard fitting procedure in the coulomb metric is recovered.

The fitting functions $\{P_{\mu\nu}\}$ are selected via overlap criteria and therefore scale linearly with the number of pairs $|\mu\nu)$, and consequently the same holds true for the number of test functions $\{Q_{\mu\nu}\}$ close to $\{P_{\mu\nu}\}$ chosen by Eq. ... . In the limit of large molecules, the size of the rectangular matrix in Eq. ... becomes constant and the fitting procedure can be evaluated with $\mathcal{O}(N)$ effort. However, a QR decomposition needs to be computed for each set of $|\mu\nu)$, leading to relatively high prefactor which makes the method unuseable for dense 3D structures like water clusters, as will be discussed in the results section. The QRDF method has been shown to deliver accuracies comparable to standard density fitting, without the use of Dunlap's formula, making it a very attractive alternative to other LDF schemes, especially if one wishes to reduce the complexity of expressions involving LDF.

2.2.7 Auxiliary Basis Sets

The density fitting approximation does not make any assumptions about the size or shape of the auxiliary basis set used. In principle, the fit is exact for the basis set containing all N^2_{AO} gaussian products $\chi_P = \chi_\mu \chi_\nu$. In practice, the product space is over-complete and can be represented by much smaller basis sets. Accurate results can be obtained for

auxiliary basis sets which are about 4 times larger than the principal basis set they are used with.

Auxiliary basis sets generally need more higher angular momentum functions than standard basis sets. Consider an isolated, unperturbed atom, with electrons occupying atomic orbitals with highest angular momentum l_{occ}. A minimal basis set for this atom contains functions of angular momenta 0 to l_{occ}. However, a minimal auxiliary basis set for fitting the product space $\chi_\mu^{0...l_{occ}}\chi_\nu^{0...l_{occ}}$ needs functions with maximum angular momentum $2l_{occ}$. For example, 2nd row elements ($l_{occ} = 1$) need an auxiliary basis set containing d-functions, and first row transition metals ($l_{occ} = 2$) even need g-functions. Similarly to standard basis sets, to describe atoms in molecules where the orbitals are subject to polarization effects, even higher angular momentum functions are needed to fit polarization functions. In practice, a principal basis set with maximum angular momentum l_{bas} is paired with an auxiliary basis set with highest angular momentum $l_{bas} + l_{occ}$.

Auxiliary basis sets have the drawback of being method-specific. There are two categories: auxiliary basis sets for density fitted Hartree Fock (DF-HF) and for density fitted correlated methods (e.g. DF-MP2, DF-CCSD, DF-ADC(2)). Auxiliary basis sets for DF-HF not only need to reproduce Hartree Fock energies, but also need to minimize theor impact on post-Hartree methods. An ill-suited auxiliary basis set leads to a deterioration of the virtual orbital space, and hence an increased error for correlated methods.

Optimization procedures often try to minimize the energy differences between the standard method and its density fitting approximation in a series of atomic calculations. For example, the jkfit family of basis sets (cc-pVXZ-JKFIT [Wei2002], def-XVP-JKFIT [Wei2007]) minimize the error

$$\Delta E_{HF} = E_{HF} - E_{DF-HF} \tag{2.45}$$

The RI basis set family (cc-pVXZ-RIFIT (Wei1998), def2-XVP-RIFIT (Ber1998)) minimize the same energy difference but for MP2 or Coupled Cluster.

Another disadvantage of auxiliary basis sets is that the accuracy of the fitting procedure cannot be easily controlled as a function of its composition (number of functions, angular momenta...), but rather extensive benchmarks are needed for each basis set that is introduced. An alternative approach was proposed by Aquilante et al. (Aqu2007) where the fitting basis sets are generated automatically by cholesky decomposition of the atomic 2-electron integrals

$$(\mu\nu \mid \sigma\lambda) = L_{\mu\nu}^X L_{\sigma\lambda}^X \tag{2.46}$$

The choleksy vectors $\mathbf{L}_{\mu\nu}$ indicate which product densities should be taken to construct the auxiliary basis. This type of atomic Cholesky decompositon (aCD) basis sets has the adavantage that the accuracy can be rigorously controlled by the decompostion threshold θ. To remove linear dependicies in the aCD basis set, another Cholesky decomposition can be performed to yield the atomic compact Cholesky decomposition (acCD) auxiliary basis set (Aqu2009).

2.3 Multipole Expansion of the Electron Integrals

The slow $1/R$ between the product densities $\Omega_{\mu\nu}$ and $\Omega_{\lambda\sigma}$ in the coulomb integrals is a major obstacle for achieving linear scaling in cases where no other sparsity relationship can be stablished between indices beclonging to separate charge densities, e.g. in the evaluation of the coulomb matrix **J** versus the exchange matrix **K**. Luckily, there are approximate methods for integral evaluation that can be computed with $\mathcal{O}(N)$ effort, known as *multipole methods*.

2.3.1 Classical and Non-Classical Electron Integrals

First, one needs to introduce the concept of classical and non-classical interactions. Two electron integrals are said to be *non-classical* if the two charge densities $\Omega_{\mu\nu}$ and $\Omega_{\sigma\lambda}$ overlap, and *classical* if the charge densties are well separated. In the latter case, the electron integrals represent classical interactions between disjoint point charges [ref], and can be approximated using multipole methods, whereas the non-classical contributions must be evaluated using the more expensive standard integral codes such as McMurchie-Davidson or Obara-Saika.

Two gaussian distributions Ω_P and Ω_Q are considered *well-separated* up to a target accuracy 10^{-k}, if their centre-to-centre distance R_{PQ} is larger than the sum of their extents ext_P and ext_Q:

$$R_{PQ} > ext_P + ext_Q \tag{2.47}$$

with the extent of a gaussian product P defined as

$$r_P = \frac{1}{\sqrt{p}} erfc^{-1}(10^{-k}) \tag{2.48}$$

where p is the reduced exponent as given in Equation 0. Another important thing to note is that the number of significant non-classical and classical integrals scale as $\mathcal{O}(N)$ and $\mathcal{O}(N^2)$ respectively (ref), which has important consequences as will be shown further below.

2.3.2 Multipole Expansion

For two well-separated charge distributions P and Q, the inverse interelectronic distance can be exapanded in terms of Legendre polynomials $\mathcal{P}$ as

$$\frac{1}{r_{12}} = \sum_{l=0}^{\infty} \frac{\Delta r_{12}^l}{R_{PQ}^l} \mathcal{P}_l cos\theta \tag{2.49}$$

with

$$cos\theta = \frac{\Delta \mathbf{r}_{12}^l \mathcal{R}_{QP}}{\Delta r_{12} R_{QP}} \tag{2.50}$$

$$\Delta \mathbf{r}_{12} = \mathbf{r}_{1P} \mathbf{r}_{2Q} \tag{2.51}$$

where $\mathbf{r}_{1P}$ and $\mathbf{r}_{2Q}$ are the distance between electron 1,2 and the centres P,Q. Equation ... is also known as the partial-wave expansion of the coulomb operator (Arfk1970).

Plugging Equation ... into Equation gives the bipolar multipole expansion of the two-electron integrals

$$g_{abcd} = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \sum_{j=0}^{\infty} \sum_{k=-j}^{j} M_{ab}^{lm}(P) T_{lm,jk} M_{cd}^{jk} \tag{2.52}$$

where $\mathbf{M}_{ab}^{lm}(P)$ is the multipole moment of the charge distribution P with total moment $l + k$, and $\mathbf{T}$ is the so-called interaction matrix. As such, the complicated 6-dimensional evaluation of g can be simply substituted by two 3-dimensional integrations of the multipole moments $\mathbf{M}$ at a much lower cost. For the lowest order expansion, where $l = m = 0$, and $j = k = 0$, the multipole moments and the interaction matrix become

$$M_{ab}^{00} = S_{ab} \tag{2.53}$$

$$M_{cd}^{00} = S_{cd} \tag{2.54}$$

$$T_{00,00} = 1/R_{PQ} \tag{2.55}$$

The zero order term of the multipole expansion therefore takes the form

$$g_{abcd} \approx \frac{S_{ab} S_{cd}}{R_{PQ}} \tag{2.56}$$

2.3.3 Fast Multipole Method

While the number of individual non-zero integrals still scales with $\mathcal{O}(N^2)$, the total contribution of all pair-wise interactions to the total energy (Hartree Fock, MP2 ...), can actually be evaluated in $\mathcal{O}(N)$.

For the sake of simplicity, consider system with point-charge particles with charge Z, in a 2-dimensional plane. The total interaction energy is given by

$$U = \sum_{i>j} \frac{Z_i Z_j}{r_{ij}} \tag{2.57}$$

Evaluating Equation ... as is takes a quadratic effort. In a first approximation, one can divide the plane into a grid of blocks of equal size, where each block contains a certain number of particles (Figure ...). Consider the interaction of a single particle i in its source block C with the other particles in the system. The interaction has two contributions: near-field (NF) contributions U_{NF} from the other particles in the source block, and the blocks immediately surrounding it, and far-field (FF) contributions U_{FF} from boxes that are well-separated from C. The NF interactions are evaluated directly by summing over all particles j in the near-field

$$U_i^{NF} = \sum_{j \in NF} \frac{Z_i Z_j}{R_{ij}} \tag{2.58}$$

while FF interactions are computed using multipole expansions $\mathbf{q}_{iC}$ and $\mathbf{q}_A$ of the FF boxes and the particle i

$$U_i^{FF} = \sum_{A \in FF} \mathbf{q}_{iC} \mathbf{T}_{CA} \mathbf{q}_A \tag{2.59}$$

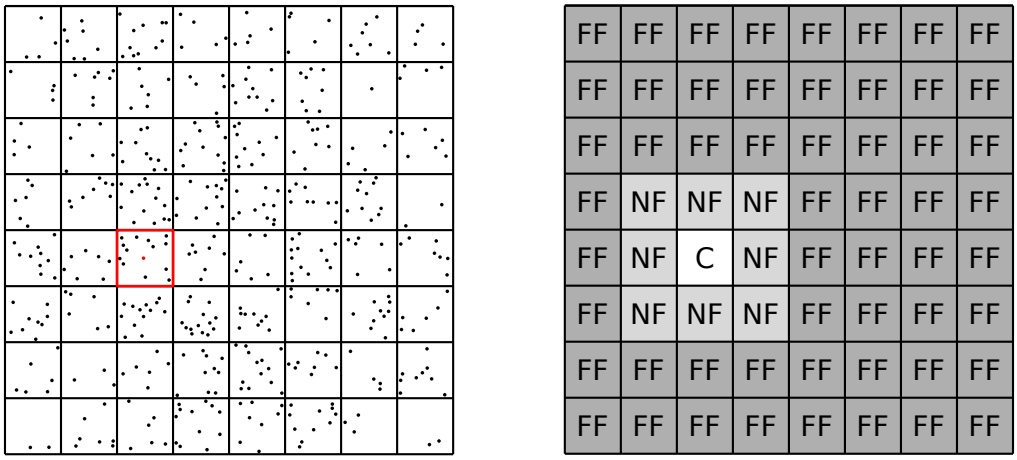


Figure 2.5: A nice caption

While evaluating the interaction energy at block-level rather than particle-level can considerably reduce the prefactor, the cost of this *single-level multipole level* is still quadratic, since for each particle i, there is a system-dependent number of FF boxes. The granularity of the blocks is the same, independent of how far away the blocks are. To achieve linear scaling, the crucial point to realize is that, the further one gets from the source block C, the smaller the single-particle interaction U_i becomes, and the less accurately it actually needs to be evaluated. This means that the farther one moves away from C, the larger the FF boxes can be. For this reason, *multi-level multipole methods* introduce a hierarchy of boxes (Figure ...), where at level 0, the whole system is in a single box, and for each subsequent level, the field is divided into fourths. FF boxes that are closest to C are evaluated at the highest level/granularity S. The region of FF boxes sourrounding the closest FF boxes are then treated at a lower level $S-1$, and so on, until all interactions have been computed. Because the multipole expansion is evaluated for increasing box size, it can be shown that the total number of boxes is constant for a single particle i. This is the basic idea on which the *Fast Multipole Method* (FMM) operates [refs], and it has quickly become one of the most important algorithms in scientific computing, as the problem of particle-particle interaction is not limited to the field of quantum chemistry. FMM can evaluate the total interaction energy U with linear computational complexity.

2.3.4 Continuous Fast Multipole Method

The fast multipole method does not work for continuous charge distrbutions like Gaussian functions, as their extents can be quite different from one another, making the separation into NF and FF contributions more difficult. Nonetheless, FMM has been generalized to the continous case, known as the continuous fast multipole method (CFMM) (ref). The principle is the same as in multi-level multipole methods, only special care needs to be taken to only include classical contributions into the FMM treatment. For further details, the reader is referred to the original publication.

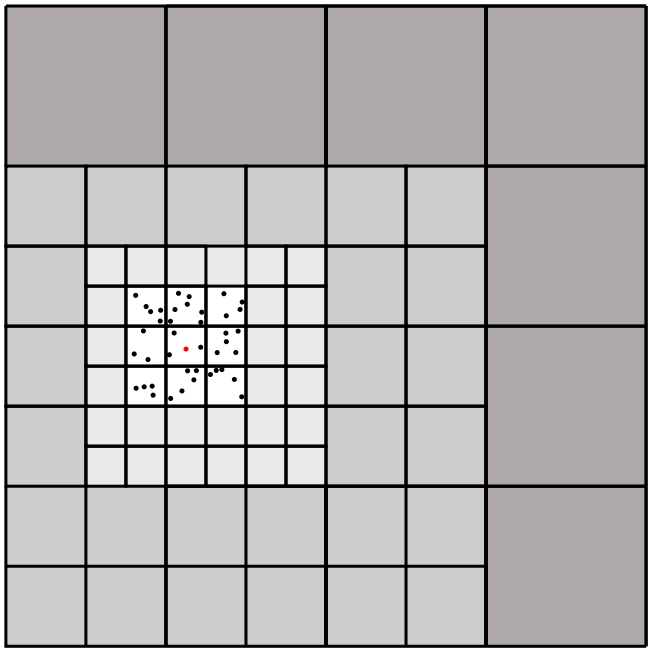


Figure 2.6: Another caption

2.4 The ABCs of LMOs: Orbital Representations

A small intro

Occupied and virtual molcular orbitals can generally be represented in two ways: canonical molecular orbitals (CMOs) and local molecular orbitals (LMOs). CMOs are the eigenvectors of the Fock matrix obtained by solving the eigenvalue problem

$$\mathbf{FC} = \mathbf{SC}\epsilon \tag{2.60}$$

where the eigevalues ϵ are known as the molecular orbital energies of the associated CMOs. However, CMOs are not unique in the sense that there are multiple molecular representations possible which yield the same electron density $\mathbf{P}$. Observables such as the electron density, or the total energy, are said to be invariant under unitary transformations (Fock V (1930) Z Phys 61:26–148). The CMOs $\mathbf{C}$ relate to other representations $\mathbf{L}$ as

$$L_{\mu\underline{i}} = U_{\underline{i}i}C_{\mu i} \tag{2.61}$$

where $\mathbf{U}$ is a unitary transformation matrix with $\mathbf{U}^\dagger\mathbf{U} = \mathbb{1}$. Typically, $\mathbf{U}$ is chosen to generate a set of molecular orbitals which are localized on as few atoms as possible, hence local molecular orbitals. While CMOs and LMOs agree on observables, they show differences for non-obervables, such as molecular orbital energy or orbital shape.

There are several reasons for choosing an LMO representation. First, as mentioned above, LMOs are used in local correlation methods, because CMOs are too delocalized, and electron correlation between LMO centres decay more rapidly. Secondly, they offer a more intuitive picture for chemists and help to interpret chemical phenomena (cite), e.g. invloving lone pairs or π bonds. Different representations can be used to interpret different phenomena, e.g. Boys LMOs vs NTOs.

Over the years, a myriad of different schemes has been proposed on how to find appropriate tranformation matrices **U**. We will now go over some examples.

2.4.1 Local Molecular Orbitals

Some words about it

2.4.2 LMOs by Reducing a Functional

One of the most popular methods for finding LMOs consists in maximizing a localization function $\eta(\phi)$ by successive rotation of the orbital space. The most prominent examples are Foster-Boys (FB)(1), Edmiston-Ruedenberg (ER) (2) and Pipek-Mezey (PM) (3). Their functionals can be written as

$$\zeta_{FB}(\chi) = \sum_i \langle \chi_i | \, \mathbf{r} \, | \chi i \rangle^2 \tag{2.62}$$

$$\zeta_{ER}(\chi) = \sum_i (\chi_i \chi_i \mid \chi_i \chi_i) \tag{2.63}$$

$$\zeta_{FB}(\chi) = \sum_i \sum_A \langle \chi_i | \, \mathbf{P}_A \, | \chi i \rangle^2 \tag{2.64}$$

The problem is generally solved using an iterative procedure consisting in consecutive pair-wise rotations, known as Jacobi sweeps (ALGO). These sweeps are rotated until convergence is reached, which may be slow. The methods differ within the procedure by how the rotational angle is computed, and scale diiferently with system size, with $\mathcal{O}(N^3)$ for FB, $\mathcal{O}(N^5)$ for ER and $\mathcal{O}(N^4)$ for PM. A faster alternative to Jacobi sweeps does also exist (4).

Over the years, PM has been the more popular choice of the three: like ER and unlike FB, it conserves σ-π separation (0), but it scales more favorably than ER.

Functional localization methods are most often used for rotating occupied MOs. Virtual MOs are often plagued by convergence issues and have a steep computational cost simply due to being much more numerous than occupied MOs (5). It is crucial that molecular localization should not take longer than the methods they are used for, and hence VMOs are often localized using separate methods.

EXAMPLES!! Ethylene

2.4.3 Projected Atomic Orbitals

A set of highly localized molecular orbitals can be obtained by projecting the CMOs onto the atomic orbital basis, known as projected atomic orbitals (PAO) (0). For a set of orthonormal occupied/virtual molecular orbitals $\{\Psi_i\}$ and $\{\Psi_a\}$, the projection operators are defined as (1)

$$\hat{P} = |\Psi_i\rangle \langle \Psi_i| = |\chi_\mu\rangle \, C_{\mu i} C_{\nu i} \, \langle \chi_\nu| \tag{2.65}$$

$$\hat{Q} = |\Psi_a\rangle \langle \Psi_a| = |\chi_\mu\rangle \, C_{\mu a} C_{\nu a} \, \langle \chi_\nu| \tag{2.66}$$

The projection operators are idempotent and mutually orthogonal with $\hat{P}\hat{Q} = \nvDash$. Applying the projection operators to the set of AOs

$$\hat{P}\,|\chi_{\mu'}\rangle \;= \sum_\mu |\chi_\mu\rangle\, P_{\mu\nu} S_{\nu\mu'} \;= L_{\mu I}\,|\chi_\mu\rangle \tag{2.67}$$

$$\hat{Q}\,|\chi_{\mu'}\rangle \;= \sum_\mu |\chi_\mu\rangle\, Q_{\mu\nu} S_{\nu\mu'} \;= L_{\mu A}\,|\chi_\mu\rangle \tag{2.68}$$

yields the set of occipied and virtual PAOs $\{\chi_I\}, \{\chi_A\}$. Both sets span a space of n_{AO} functions each, as opposed to n_{occ} and n_{vir}. As such, just like the AO basis, the PAO basis is redundant (non-orthogonal). CMOs are transformed to PAOs by the relationship

$$|\chi_I\rangle = (\mathbf{SC})_{Ii}\,|\Psi_i\rangle \tag{2.69}$$

$$|\chi_A\rangle = (\mathbf{SC})_{Aa}\,|\Psi_a\rangle \tag{2.70}$$

PAOs are centred on the atom on which their corresponding AO is localized. However, PAOs can still span multiple atoms. Methods which are entirely formulated in PAOs are rare but possible (1).

2.4.4 Subspace Projected Atomic Orbitals

Some applications need localized molecular orbitals that only span a certain region of a molecule, e.g. density matrix embedding theory (DMET) (refs) or local ADC (ref). The molecule is split into two subunits, and atoms are grouped into an active region A and an inactive region B according to specific selection criteria. Region A contains the molecular subunit of interest.

Most implementations use the Mulliken gross charges to find ... ? Not used for virtuals? Put it into AO-ADC Part?

2.4.5 Cholesky Molecular Orbitals

Sparsity of the atomic density matrix is crucial for achieving low-scaling electronic structure methods. Aquilante et al. proposed (0) to define a set of occupied molecular orbitals by Cholesky decomposition of the density matrix. Analysis of the resulting Cholesky molecular orbitals (CholMOs) showed their localized character inherited from the sparsity of the density matrix.

$$\mathbf{P} = \mathbf{L}\mathbf{L}^{\mathbf{T}} \tag{2.71}$$

Figure ... shows the sparsity of the occupied density matrix and the occupied cholesky molecular coefficient matrix of the linear alkane $H_{322}C_{160}$. The number of CholMOs is equal to the rank of the density matrix, which is equal to the number of occupied orbitals. The CholMOs are computed by an incomplete Cholesky decomposition with full row and column pivoting (ALGO). The unitary transformation matrix is given by

$$U_{i\underline{i}} = C_{\mu i} S_{\mu\nu} L_{\nu\underline{i}} \tag{2.72}$$

The decomposition algorithm scales with $\mathcal{O}(N^3)$ but can be made linearly scaling by using sparse matrix algebra. CholMOs have several advantages: the Cholesky decomposition is fast and non-iterative, and an initial guess for molecular orbitals is not needed.

The scheme can be extended to virtual orbitals as well, by CD of the virtual atomic density matrix $\mathbf{Q}$. The rank of $\mathbf{Q}$ is equal to the number of virtual orbitals $n_v ir$, therefore the prefactor of the incomplete CD increases with basis set size. Especially in the presence of diffuse functions, the rank reduction might not offer much of an advantage compared to simpler localization methods such as PAOs.

Moreover, orbitals obtained by CD are less localized than FB or ER LMOs, especially for small molecules. Low scaling is still possible using CholMOs in the context of LMO correlation methods, albeit with a larger prefactor.

CD is also used in the context of AO-MP2 to reduce the prefactor of inetgral transformation by using the rank sparsity of the pseudo-density matrices, as will be shown further below.

CholMOs can also used as an initial guess for iterative localization schemes to achive faster convergence.

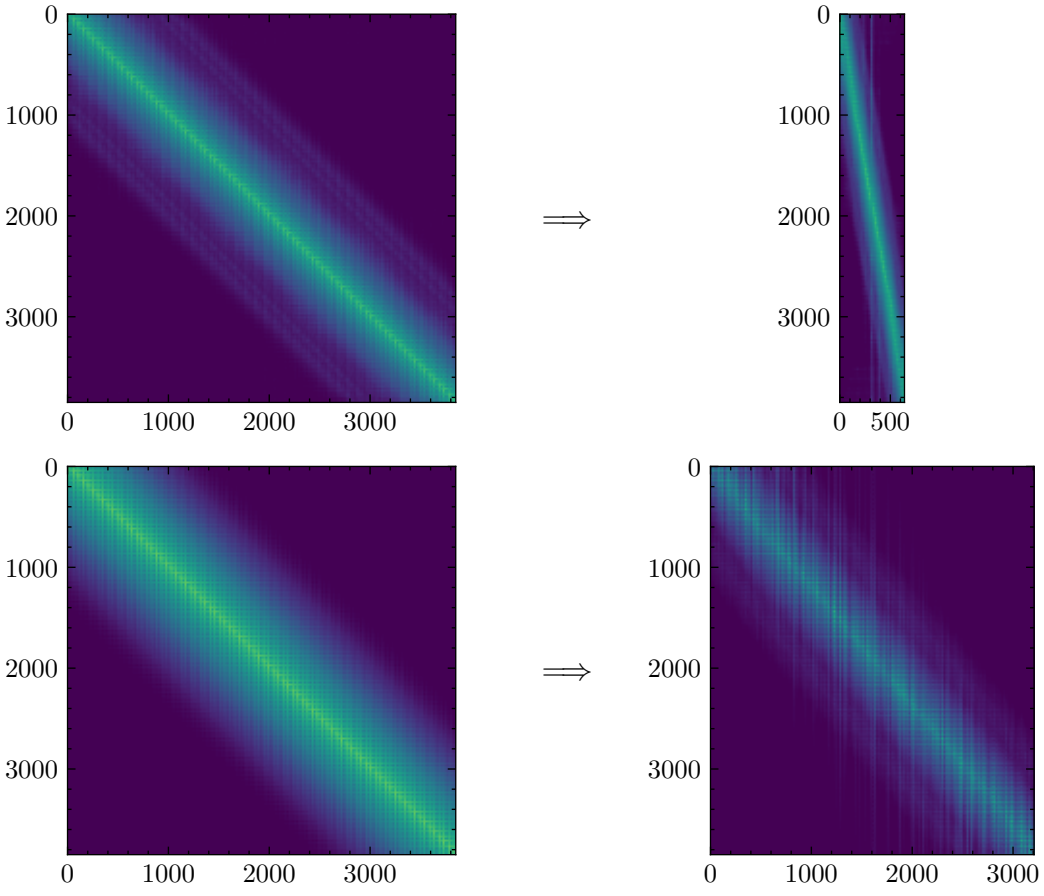


Figure 2.7: Something

2.4.6 Natural Orbitals

While the schemes described above try to generate a set of occupied and/or virtual molecular orbitals localized in space, natural orbital (NOs) methods try to generate a

set of "compact" orbitals, i.e. a minimal set of orbitals that can describe the problem at hand. The concept of natural orbitals was first introduced by Löwdin (A). The natural orbitals Θ_i of a wave function Ψ are defined as the eigenfunctions of the one-particle density operator $\hat{n}$

$$\hat{n} \left|\Theta_i\right\rangle = n_i \left|\Theta_i\right\rangle \tag{2.73}$$

where n_i are the occupation numbers of the associated orbital Θ_i. One can then choose a reduced orbital space $\{\tilde{\Psi}_i\}$ by only taking into account those orbitals with an occupation number above a certain threshold τ. The orbitals are "natural" in the sense that they are determined purely using Ψ, and are intrinsic to the system. NOs are computed by diagonalizing the one-particle density matrix at the desired level of theory (Hartree-Fock, MP, CIS, CC).

NOs are state-specific (Pok2019), meaning that NOs computed from the ground state densities may not be well suited to describe excited states, and NOs of different excited states might also greatly differ. As such, as will be later shown for local flavours of ADC, NOs need to be recomputed for each state.

Natural Orbitals in Hartree Fock Theory

In Hartree Fock theory, natural orbitals are mostly reserved for qualitative population and bond order analysis.

Natural atomic orbitals (NAOs) are computed by diagonalizing the blocks $P_{\mu_A \nu_A}$ of the atomic density matrix, where μ_A, ν_A are basis functions centred on atom A. NAOs are optimal for describing the electron density around individual atom centres (IUPAC). NAOs are also useful for obtaining a set of guess orbitals from densitiy matrices formed from the superposition of atomic densities (SAD) guess.

NHOs obtained from NAOs + off-diag NAOs

NBOs obtained from NHOs

Frozen Natural Orbitals

For large basis sets, CMOs are much more compact than the virtual orbital span, and the number of occupied NOs is not significantly lower than that of occupied CMOs. It is therefore sufficient to only compute the eigenfunctions of the virtual-virtual block of the one-particle density matrix, which are known as frozen natural orbitals (FNOs) (Bar1970). FNOs need information of the correlated wave function, and are therefore typically computed at a lower level of theory. For example, the easiest way to obtain a set of FNOs for CCSD or CCSD(T) computation is to diagonalize the virtual-virtual block of the MP2 density matrix (Sos1989, Tau2005, Tau2008)

$$D_{ab} = \frac{1}{2} sum_{cij} \frac{K_{ij}^{cb} K_{ij}^{ca}}{\epsilon_{ij}^{ab} \epsilon_{ij}^{ca}} \tag{2.74}$$

with

$$K_{ij}^{ab} = 2\,(ia \mid jb) - (ib \mid ja) \tag{2.75}$$

$$\epsilon_{ij}^{ab} = \epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b \tag{2.76}$$

The FNOs are then canonicalized (see ...). The combined set of occupied CMOs and virtual FNOs forms a very compact representation suitable for CC ground state and excited state calculations.

Natural Transition Orbitals

Consider the CIS eigenvalue problem for finding the excitation energies ω_n and their associated transition density matrices R_n

$$\mathbf{A_{CIS}} R_n = \omega_n R_n \tag{2.77}$$

The matrices $\mathbf{R}_n$ contain $n_{occ} n_{vir}$ expansion coefficients c_{ia} which show how much an orbital-virtual MO pair ia contributes to the excitation n. The numer of non-negligebale coefficients can be far from zero, making interpretations of the computed results difficult for some systems.

Natural transition orbitals (NTOs) were introduced to facilitate the qualititative description of an excited state and finding connections to experimental spectra (Luz1976, Mar2003, Mar2008). NTOs are typically obtained by computing the singular value decomposition (SVD) of the state densities $\mathbf{R}_n$

$$\mathbf{R} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\dagger \tag{2.78}$$

where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices with dimension $n_{occ} n_{NTO}$ and $n_{vir} n_{NTO}$, and Σ is a n_{NTO} by n_{NTO} matrix containg the singular values s on its diagonal. The CMOs $\{\Psi_i^{occ}, \Psi_a^{vir}\}$ are transformed to the NTO basis $\{\overline{\Psi}_k^{occ}, \overline{\Psi}_k^{vir}\}$ using

$$\left|\overline{\Psi}_k^{occ}\right\rangle = U_{ki} \left|\Psi_i^{occ}\right\rangle \tag{2.79}$$

$$\left|\overline{\Psi}_k^{vir}\right\rangle = V_{ka} \left|\Psi_a^{vir}\right\rangle \tag{2.80}$$

The singular value s_k show the contribution of an NTO pair k to the excited state. In most cases, the number of significant NTO pairs is significantly lower than $n_{occ} n_{vir}$ and at most equal to n_{occ}. NTOs are not limited to CIS, but can also be obtained by SVD decomposition of the singles-singles block of excited state densities from higher order methods such as ADC or CCLR.

Natural transition orbitals have also found use in local excited state correlation methods (Bau2017,Hof2017), where CIS NTOs are combined with MP2 NOs to obtain a compact orbital representation for ground and excited state coupled cluster calculations.

EXAMPLE!!! phenylalanine

2.4.7 Specific Virtual Orbitals

In most cases, using LMOs instead of CMOs does not offer any a priori advantage in terms of the computational complexity asscociated with correlated methods, and additional approximations are necessary. In local correlation methods, this is often done by truncating the VMO space. TRuncation of the VMOs has been an active field of research for "a long time", and several schemes have emerged over the years. A naive approach to truncate the virtual space would be to eliminate VMOs with orbital energies

above a certain threshold; however, this proved to be unusable in most contexts (ref). More successful methods for VMO truncation use the concept of what we will refer to as *specific virtual orbitals* (SVOs). SVOs are specific in the sense that each individual occupied LMO i or each pair of LMOs ij has their own set of SVOs a_i (orbital specific virtual orbitals) or a_{ij} (pair specific virtual orbitals) associated to it. The concept of SVOs naturally arises in the context of correlated methods such as the coupled electron pair approximation (CEPA) where the total energy is computed is computed as the sum of electron pair energies e

$$E_{CEPA} = \sum_{ij} e_{ij} \tag{2.81}$$

The electron pair energy decays rapidly as a function of the distance r between MO centres in an LMO basis. Distant virtual orbitals contribute less to the elctron pair energy as virtual orbitals close to ij. It has been shown early on that instead of using the whole virtual orbital span, one can correlate only a subset or reduced set of virtual orbitals with each electron pair (0,1,2,3) and still recover most of the correlation energy. In the limit of large molecules, the number of signiciant virtual orbitals for an electron pair becomes independent of system size (4). There are different ways to choose how to define the VMO subsets.

Domain Specific Virtual Orbitals

We call domain specific virtual orbitals (DSVOs) any type of virtual orbitals where the subsets are formed *a priori* by distance or partial charge criteria. Examples include the local MP2 and local CCSD implementations by Schütz et al.(AA,AB,AC)

First, occupied CMOs are localized by one of the methods described above. The method of choice in Ref [AA-AC] was the Foster-Boys scheme. Virtual CMOs are recast into the PAO basis. Each individual occuied LMO $|\Psi_i\rangle$ is then assigned a subset $[i]$ of PAOs, chosen by a Boughton-Pulay (BP) criterium (AD) or by population analysis (AE).

For a given electron pair ij, the pair domain is then formed by taking the union $[ij]$ $= [i] \cup [j]$. The set of all virtual pair domains $[ij]$ forms the DSVOs.

Alongside AOs, DSVOs were among the first orbital representations in which linear scaling correlated methods were formulated. Their dependency on distance criteria for selecting the pair domains makes them less rigorous than other methods.

Pair Natural Orbitals

First introduced under the guise of "pseudo-natural orbitals" (Edmiston), then rediscovered by Neese (CA,CB,CC), projected natural orbitals (PNOs) have risen in popularity in the recent years (refs). Similarly to DSVMOs, each electron pair has a set of PNOs associated to it. PNOs are formed by diagonalizing the MP2 pair density matrix for each LMO pair ij (hence "pair-natural")

$$\mathbf{D}^{ij} = \frac{1}{1+\delta_{ij}} \left(\tilde{\mathbf{t}}^{ij} \mathbf{t}^{ij} + tilde\mathbf{t}^{ij} \mathbf{t}^{ij\dagger} \right) \tag{2.82}$$

with

$$\tilde{\mathbf{t}}^{ij}_{ab} = 2\mathbf{t}^{ij}_{ab} - \mathbf{t}^{ji}_{ab} \tag{2.83}$$

The eigevalue decomposition of $\mathbf{D}$ then gives

$$\mathbf{D}^{ij}\mathbf{Q}^{ij} = n^{ij}\mathbf{Q}^{ij} \tag{2.84}$$

where $\mathbf{Q^{ij}}$ are the pair specific transformation matrices, and n^{ij} their occupation numbers. The Fock matrix in the LMO representation is not diagonal, and the MP2 amplitudes are approximated by

$$t_{ij}^{ab} = \frac{(ia \mid jb)}{\epsilon_a + \epsilon_b - f_{ii} - f_{jj}} \tag{2.85}$$

where f_{ii} are the diagonal entries of the Fock matrix in the LMO basis. The pair domains $[ij]$ are chosen by keeping the PNOs with an occupation number larger than a threshold τ_{PAO}. Therefore, accuracy is controlled by a single, distance-independent parameter, which is an advantage over other methods like DSVOs.

However, computing the PNOs requires a full MP2 calculation, and with density fitting scales with $\mathcal{O}(N^5)$. Moreover, even if the PNO basis is compact, the fact that each LMO pair has its own virtual orbital basis may lead to a prohibitively large number of PNOs for large molecules

Local PNOs:

Orbital Specific Virtuals

Closely related to PNOs are the orbital specific virtual orbitals (OSVs) (Yan2011). The OSVs for an LMO $|\Psi_i\rangle$ are obatined by taking the diagonal PNOs for the domain $[ii]$. The MP2 density matrix reduces to

$$\mathbf{D}^{ii} = 4\mathbf{t}^{ii}\mathbf{t_{ii}} \tag{2.86}$$

Instead of reducing the density matrix, one can just diagonalize $\mathbf{t}^{ij}$ instead.

$$\mathbf{t}^{ii}\mathbf{Q}^{ii} = t^{ii}\mathbf{Q}^{ii} \tag{2.87}$$

where t^{ii} are the eigenvalues, which are used to compute the compute the occupation numbers $n^{ii} = (t^{ii})^2$. OSVs for which $n^{ii} > \tau_{OSV}$ are included into the orbital specific domain $[i]$. Pair domains $[ij]$ are then formed as the union of $[i]$ and $[j]$ simalar to DSVOs.

OSVs have the advantage that the can be constructed with $\mathcal{O}(N^3)$ scaling provided that density fitting is used. However, OSVs are less compact than PNOs.

OSVs can be used to lower the computational complexity to construct PNOs. Several hybrid OSV-PNO schemes have been proposed whith a computaional complexity of $\mathcal{O}(N^4)$ (Kra2012, Hat2012), $\mathcal{O}(N^3)$ (Sch2013) and finally $\mathcal{O}(N)$ (Rip2013).

Local Pair Natural Orbitals

2.5 Electron Pairs

Nesbet theorem, number pairs etc...

Chapter 3

Local Correlation: Ground State

Some choice words here.....

3.1 Low-Scaling Self-Consistent Field Methods

Hartree Fock and density field theory, also grouped under the umbrella term of self-consistent field (SCF) methods, are the working horses in quantum chemsitry, and the underlying equations, the Roothan and Kohn-Sham equations, are well known and studied. Conventional formulations of HF and DFT, without inclusion of sparsity, scale with $\mathcal{O}(N^3)$ to $\mathcal{O}(N^4)$ which hampers extension to very large systems. There are three major bottle-necks: (1) computation of the coulomb matrix, (2) computation of the exchange matrix and (3) diagonalization of the Fock matrix. Over the last couple of decades, multiple different approaches have been proposed on how to lower the scaling of constructing the Fock matrix and circumvent matrix diagonalization. To this day, the field of low scaling SCF methods remains an active area of research in theoretical chemistry. The next sections will address the time-determining steps in detail.

3.1.1 The Coulomb Matrix

Consider again the expression for the coulomb matrix **J**

$$J_{\mu\nu} = (\mu\nu \mid \lambda\sigma)\, P_{\sigma\lambda} \tag{3.1}$$

Equation ... gives the following sparsity diagram:

$$\mu \overset{\text{S}}{\longleftrightarrow} \nu \qquad \sigma \overset{\text{P/S}}{\longleftrightarrow} \lambda$$

The counstruction of **J** has an inherent compuational complexity of $\mathcal{O}(N^2)$, eventhough the number of non-zero elements scales linearly due to the overlap relationship between μ and ν. Quadratic scaling algorithms are straight-forward to implement. The first method to construct **J** with $\mathcal{O}(N)$ effort was the continous fast multipole method (CFMM, cf Section .). For each element $\mu\nu$ the contributions $\sigma\lambda$ are split into near-field and far-field contributions. NF interactions are computed using standard integration techniques, while FF interactions are computed using multi-level multipole expansion. The linear

scaling consists even if the density matrix is not sparse. Other tree-like algorithms have also been proposed (Str1996,Cha1996). In all cases, computing the NF interactions are by far the most time-consuming step.

? J-engine ?

To speed up the evaluation of the non-classical contributions, one may introduce the density fitting approximation (Eq. ...). The coulomb matrix is then evaluated in several steps via an intermediate **d** as

$$J_{\mu\nu} = (\mu\nu \mid X)\, \tilde{d}_X \tag{3.2}$$

$$d_X = (Y \mid \mu\nu)\, P_{\nu\mu}; \quad \tilde{d}_X = (X \mid Y)^{-1}\, d_X \tag{3.3}$$

The computational effort remains unchanged, with $\mathcal{O}(N^2)$, but with a much lower prefactor, especially for larger, more diffuse basis sets (ref Weigend). The inversion of the metric matrix $(X \mid Y)$ scales cubically, and dominates the cost of the DF approximation for large molecules.

For long-range interactions, one could also consider using local density fitting, such as the atomic resolution of the idenstity or pair-atomic reslution of the identity. Unfortunately, LDF also only reduces the prefactor, rather than scaling. Furthermore, it is necessary to use the robust density fitting of the electron integrals (Eqaution ...) to recover quadratic scaling in the fitting error, due to the constraints imposed on the density fitting procedure. The coulomb matrix is then expressed as

$$J_{\mu\nu} = (\mu\nu \mid X)\, b_X + c^Y_{\mu\nu}\, [g_Y - \tilde{g}_Y] \tag{3.4}$$

$$g_X = C^X_{\mu\nu} P_{\mu\nu}; \quad \tilde{g}_X = (X \mid Y)\, b_Y; \quad b_X = C^X_{\mu\nu} P_{\mu\nu} \tag{3.5}$$

The robust LDF-J approximation is evaluated at an effort similar to standard DF-J. The only advantages to LDF in this case pertain to the fitting procedure itself. It is no longer necessary to invert the 2c2e integrals $(X \mid Y)$, and the fitting coefficients $C^X_{\mu\nu}$ can be evaluated in linear scaling fashion. However, it has been demonstrated that using Dunlap's robust formula in combination with local metrics leads to "attractive electron" states where the SCF energy may converge to very high positive values (Mer2013,Hol2014). The reason is that the the two-electron integral tensor in the robust LDF approximation is no longer postive semidefinite, but *indefinite*, which can lead to severe convergence problems. One way to circumvent this problem is to loosen the constraints on the density fitting procedure, or use a larger auxiliary basis set. In both cases, performance is compromised.

As shall be shown in the results section/chapter?, quasi-robust density offers a better alternative to robust LDF-J methods, with high accuracy, and without convergence problems.

MEMORY : ON-THE FLY

POISSON?

3.1.2 The Exchange Matrix

Exact Exchange

The expression for the exchange matrix is given by

$$K_{\mu\nu} = (\mu\sigma \mid \nu\lambda)\, P_{\lambda\sigma} \tag{3.6}$$

In the section where sparsity diagrams were introduced, it was demonstrotated that the indices of the exchange expression can be fully linked:

$$\mu \overset{\mathrm{S}}{\longleftrightarrow} \sigma \overset{\mathrm{P}}{\longleftrightarrow} \lambda \overset{\mathrm{S}}{\longleftrightarrow} \nu$$

The non-zero elements in $\mathbf{K}$ scale linearly and can also be evaluated with $\mathcal{O}(N)$. This property has been realized quite early on (Sch1996). However, a straight-forward implementation where the 4c2e integrals are directly contracted with the density matrix $\mathbf{P}$ using sparse matrix algebra does not give the desired results, when applying the standard $\mathcal{O}(N^2)$ Schwarz-screening to $(\mu\sigma \mid \nu\lambda)$. To lower the scaling of electron integral evaluation, it is important to design a screening algorithm which imposes the P junction between σ and λ which in turn leads to only a linear increase in the number of bra-ket pairs.

The ONX method by Schwegler (Sch1997) was the first $\mathcal{O}(N)$ scheme for constructing the exchange matrix, but did not exploit permutational symmetry, which lead to a four fold increase in the prefactor. More competitive methods were proposed later, such as Linear Exchange (LinK) by Ochsenfled et al (Och1997), or symmetrized ONX (SONX) (Sch2000) that could also be applied to small systems without major overhead.

For all approaches, an important step is the screening of the bra-ket pairs using a density-weighted integral estimate

$$|P_{\lambda\sigma}||\,(\mu\sigma \mid \mu\sigma)\,|^{1/2}|\,(\nu\lambda \mid \nu\lambda)\,|^{1/2} \leq \tau \tag{3.7}$$

which scales linearly for increasing size of systems whith large HOMO-LUMO gaps. Furthermore, shell ordering is very important to avoid the $\mathcal{O}(N^2)$ complexity of thresholding and enable early exit out of the shell loops during contruction of the exchange matrix. Similarly to the J kernel, the electron integrals need not to be held in memory, but can be recomputed on the fly.

Density Fitting

The downside of exact linear exchange algorithms is the steep $\mathcal{O}(N^4)$ scaling with increasing basis set size which delays the onset of the low-scaling regime. For this reason, considerable effort has been invested in recent years to also exploit rank sparsity by density fitting. The DF expression for the exchange matrix in the AO basis reads

$$C^X_{\mu\nu} = (X \mid Y)^{-1}\,(Y \mid \mu\nu) \tag{3.8}$$

$$K_{\mu\nu} = C^X_{\nu\lambda}\,(X \mid \mu\sigma)\,P_{\sigma\lambda} \tag{3.9}$$

In a straight-forward implementation using sparse matrix algebra, Equation ... and Equation .. are evaluated with $\mathcal{O}(N^3)$ and $\mathcal{O}(N^2)$ effort respectively. The non-zero elements of both the fitting coefficients C and the 3c2e integrals increase quadratically. Their storage can quickly become problematic for large basis sets if both are held in-core. In principle, both tensors can be recomputed batchwise on-the-fly to reduce memory-footprint, but in contrast to the DF-J kernel where only the 3c2e integrals need to be generated at each iteration, recomputing the fitting coefficients each time introduces a prefactor that is too large for an out-of-core DF-K kernel to be of any practical use.

For an efficient, direct evaluation of the exchange matrix using density fitting, an MO based approach is much more favourable (Wei2002). The MO-DF-K kernel is evaluated as

$$B_{\mu i}^{X} = C_{\nu i} \left(X \mid \mu\nu \right) \tag{3.10}$$

$$D_{\mu i}^{X} = \left(X \mid Y \right)^{-1/2} B_{\mu i}^{X} \tag{3.11}$$

$$K_{\mu\nu} = B_{\mu i}^{X} B_{\nu i}^{X} \tag{3.12}$$

with cubic computational complexity. The matrix elements of the exchange matrix are evaluated batch-wise over occupied blocks I. By contracting the 3c2e electron integrals with the coeffcient matrix $C_{\mu i}$ to form the half-transformed integrals $B_{\mu i}^{X}$, storage can be reduced from $N_{aux} N_{AO}^2$ to $N_{aux} N_{AO} N_{occ} / N_I$. The 3c2e integrals need to be recomputed for each block I, but in practice the number of blocks can be held quite small. The DF-MO-K method is especially well suited for small to medium sized molecules with large diffuse basis sets for post-HF calculations.

Local Density Fitting

For standard density fitting, one looses the linear scaling character of the exchange matrix, which can however be recovered by using LDF. Again, the Dunlap's robust density fitting needs to applied to get accurate results. The robust DF-K kernel can take the form

$$E_{\mu\nu}^{X} = C_{\mu\lambda}^{X} P_{\lambda\nu} \tag{3.13}$$

$$L_{\mu\nu} = \left(\mu\sigma \mid X \right) \left(X \mid Y \right) E_{\nu\sigma}^{Y} - \frac{1}{2} C_{\mu\sigma}^{X} E_{\nu\sigma}^{X} \tag{3.14}$$

$$K_{\mu\nu} = L_{\mu\nu} + L_{\nu\mu} \tag{3.15}$$

Alternatively, an LDF-K scheme based on LMOs is also possible (Pol2004,Mej2014). All steps can be evaluated with $\mathcal{O}(N)$. Over the years, many different LDF-K kernels have been proposed that approximate $C_{\mu\nu}^{X}$ based on LMO domains (Pol2004,Mej2014), atomic resolution of the identity (ARI) (Sod2007), pair-atomic resolution of the identity (PARI) (Mer2013) or concentric atomic density fitting (CADF) (Hol2017). Although the electron integrals are no longer positive semidefinite, LDF-K is not plagued by the same convergence problems as LDF-J, and it has been shown that LDF-K can be combined with standard DF-J to circumvent convergence problems (Man2015).

Semi-Numerical Exchange

3.1.3 Density Purification

3.2 Local Ground State Correlation Methods: MP2

Second-Order Møller Plesset is one of the simplest post-Hartree Fock methods available, but still scales as $\mathcal{O}(N^5)$. Since the seminal work of Saebo and Pulay (Pul1983,Sae1985), several different methods have been proposed which drastically reduce the computational complexity. Attempts can generally be grouped into two categories: AO-MP2 and LMO-MP2. While both schemes do have their differences, they share some of the problems asscated with computing the MP2 energy in a local basis.

First, the energy denominator in the MP2-amplitudes t make it difficult to reformulate the MP2 energy expressions in a different basis. AO-MP2 and LMO-MP2 take different approaches: AO-MP2 solves the problem using the Laplace quadrature, while LMO-MP2 methods usually use an orbital-invariant formulation of MP2 using the Hylleraas functional.

Second, steps involving the transformation of the AO 2-electron integrals to the Pseudo-AO or LMO basis still remain a major bottle-neck, even with sparsity involved. Both AO- and LMO-MP2 use screening criteria, additional domain restrictions, density fitting or similar methods to lower the cost of integral transformation. These additional procedures are crucial if one wishes to achieve a truly linear scaling MP2 method with a reduced overhead.

We will now address each point in detail in the next sections.

3.2.1 Atomic Orbital MP2

MP2 was first formulated in the AO basis in 1993 by Häser , and a linear scaling algorithm was presented by Scuseria and Ayala in 1999 (ref). AO-MP2 has since then been extended to DF-MP2 (ref) and SOS-MP2 (ref).

The Laplace Transform

In 1991, Almlöf showed (Alm1991) that the energy denominator in the MP2 amplitudes can be removed using an integral transform called the *Laplace Transform*

$$\frac{1}{\epsilon_a + \epsilon_b - \epsilon_i - \epsilon_j} = \int_0^\infty e^{-(\epsilon_a+\epsilon_b-\epsilon i-\epsilon_j)t} dt \tag{3.16}$$

The t-integration can be replaced (Has1993) by a finite summation using a functional approximation:

$$\frac{1}{\epsilon_a + \epsilon_b - \epsilon_i - \epsilon_j} \approx \sum_\alpha^n w^{(\alpha)} e^{-(\epsilon_a+\epsilon_b-\epsilon i-\epsilon_j)t^{(\alpha)}} \tag{3.17}$$

where $w^{(\alpha)}$ and $t^{(\alpha)}$ are the Laplace weights and exponents at the Laplace points α. Accuracy can be controlled by the number of Laplace points n. An efficient AO-MP2 implementation heavily relies on an accurate quadrature scheme to achive the desired accuracy using as few Laplace points as possible to reduce overhead caused by the repeated AO transformation at each step. In general, 5-8 Laplace points are needed to achieve milli-Hartree accuracy, and 10 to 15 points for μHartree accuracy. For more details, the reader is referred to section ... .

AO MP2 Equations

Using the Laplace transform, the energy expression for restricted canonical MP2 can be expressed as

$$\begin{aligned} E_{MP2} &= -\sum_{iajb} \frac{(ia \mid jb)\left[2\,(ia \mid ib) - (ib \mid ja)\right]}{\epsilon_a + \epsilon_b - \epsilon_i - \epsilon_j} \\ &\approx -\sum_{\alpha}^{n} \sum_{iajb} (ia \mid jb)\left[2\,(ia \mid ib) - (ib \mid ja)\right] w^{(\alpha)} e^{-(\epsilon_a+\epsilon_b-\epsilon i-\epsilon_j)t^{(\alpha)}} \end{aligned} \tag{3.18}$$

We can then proceed to factor out the coefficient matrices

$$\begin{aligned} &-\sum_{\alpha}^{n} \sum_{iajb} (ia \mid jb)\left[2\,(ia \mid ib) - (ib \mid ja)\right] w^{(\alpha)} e^{-(\epsilon_a+\epsilon_b-\epsilon i-\epsilon_j)t^{(\alpha)}} \\ &= -\sum_{\alpha}^{n} \sum_{iajb} \sum_{\substack{\mu\nu\lambda\sigma \\ \mu'\nu'\lambda'\sigma'}} w^{(\alpha)} e^{-(\epsilon_a+\epsilon_b-\epsilon i-\epsilon_j)t^{(\alpha)}} C_{\mu' i} C_{\sigma' a}\,(\mu'\sigma' \mid \nu'\lambda')\, C_{\nu' j} C_{\lambda' b} \\ &\qquad \times \left\{C_{\mu i} C_{\sigma a}\left[2\,(\mu\sigma \mid \nu\lambda) - (\mu\lambda \mid \nu\sigma)\right] C_{\nu j} C_{\lambda b}\right\} \\ &= -\sum_{\alpha}^{n} \sum_{\substack{\mu\nu\lambda\sigma \\ \mu'\nu'\lambda'\sigma'}} \underline{P}^{(\alpha)}_{\mu\mu'} \overline{P}^{(\alpha)}_{\sigma\sigma'}\,(\mu'\sigma' \mid \nu'\lambda')\, \underline{P}^{(\alpha)}_{\nu\nu'} \overline{P}^{(\alpha)}_{\lambda\lambda'} \left[2\,(\mu\sigma \mid \nu\lambda) - (\mu\lambda \mid \nu\sigma)\right] \end{aligned} \tag{3.19}$$

with the occupied and virtual *pseudo* or *Laplace* density matrices

$$\begin{aligned} \underline{P}^{(\alpha)}_{\mu\mu'} &= \sum_{i} C_{\mu i} e^{0.25 ln(w^{(\alpha)}) + \epsilon_i t^{(\alpha)}} C_{\mu' i} \\ \overline{P}^{(\alpha)}_{\mu\mu'} &= \sum_{i} C_{\sigma a} e^{0.25 ln(w^{(\alpha)}) - \epsilon_a t^{(\alpha)}} C_{\sigma' i} \end{aligned} \tag{3.20}$$

Introducing the *pseudo-AO* transformed electron integrals

$$\left(\underline{\mu}\overline{\sigma} \mid \underline{\nu}\overline{\lambda}\right)^{(\alpha)} = \underline{P}^{(\alpha)}_{\mu\mu'} \overline{P}^{(\alpha)}_{\sigma\sigma'}\,(\mu'\sigma' \mid \nu'\lambda')\, \underline{P}^{(\alpha)}_{\nu\nu'} \overline{P}^{(\alpha)}_{\lambda\lambda'} \tag{3.21}$$

the energy expression for AO-MP2 reads

$$E_{AO-MP2} = -\sum_{\alpha}^{n} \sum_{\mu\nu\lambda\sigma} \left(\underline{\mu}\overline{\sigma} \mid \underline{\nu}\overline{\lambda}\right)^{(\alpha)} \left[2\,(\mu\sigma \mid \nu\lambda) - (\mu\lambda \mid \nu\sigma)\right] \tag{3.22}$$

For t = 0, $\underline{\mathbf{P}}^{(\alpha)}$ and $\overline{\mathbf{P}}^{(\alpha)}$ are equal to the Hartree Fock density matrices. The Laplace matrices also fulfill similar relationships

$$\underline{\mathbf{P}}^{(\alpha)} \mathbf{S} \overline{\mathbf{P}}^{(\alpha)} = \mathbf{0} \tag{3.23}$$

$$\underline{\mathbf{P}}^{(\alpha)} \mathbf{S} + \overline{\mathbf{P}}^{(\alpha)} \mathbf{S} = \mathbf{I}_{exp} \tag{3.24}$$

where $\mathbf{I}_{exp}$ is a diagonal matrix with trace

$$Tr[\mathbf{I}_{exp}] = \sum_{i} e^{0.25 ln(w^{(\alpha)}) + \epsilon_i t^{(\alpha)}} + \sum_{a} e^{0.25 ln(w^{(\alpha)}) - \epsilon_a t^{(\alpha)}} \tag{3.25}$$

The entries of the pseudo-density also decay exponentially as function of the distance between pseudo-AO centres.

Quadratic Scaling AO-MP2

Using the linked index rule, we can easily find the computational complexity of the AO-MP2 method. From the energy expression in Eq. we find that we compute the dot product between two different tensors, the AO-ERIs $(\mu\sigma \mid \nu\lambda)$ and the pseudo-AO-ERIs $\left(\underline{\mu}\overline{\sigma} \mid \underline{\nu}\overline{\lambda}\right)^{(\alpha)}$. The scaling is thus determined by the sparsity of those two tensors. We know from a previous discussion that the ERIs can be computed with $\mathcal{O}(N^2)$ effort. The pseudo-AO ERIs are computed by transforming the ERIs with the pseudo-density matrices, whose indices μ, ν are connected by a P junction. The diagrammatic expression for the pseudo-AO ERIs in Eq. ... is given by

$$\begin{aligned} \mu &\overset{P}{\longleftrightarrow} \mu' \overset{S}{\longleftrightarrow} \sigma' \overset{P}{\longleftrightarrow} \sigma \\ \nu &\overset{P}{\longleftrightarrow} \nu' \overset{S}{\longleftrightarrow} \lambda' \overset{P}{\longleftrightarrow} \lambda \end{aligned} \tag{3.26}$$

Two vertices indicate an $\mathcal{O}(N^2)$ effort for evaluating Eq. ... . Therefore, the inherent asymptotic scaling of AO-MP2, without any other further approximations, is $\mathcal{O}(N^2)$ as well. Similarly to the AO ERIs, a quadratic scaling evaluation of the pseudo-AO ERIs can be achived using a Schwarz-like screening, as first advocated by Almlöf. Defining the screening matrices

$$\begin{aligned} Q_{\mu\nu} &= \mid (\mu\nu \mid \mu\nu) \mid^{1/2} \\ X_{\mu\nu} &= \mid \left(\underline{\mu}\nu \mid \underline{\mu}\nu\right) \mid^{1/2} \\ Y_{\mu\nu} &= \mid (\mu\overline{\nu} \mid \mu\overline{\nu}) \mid^{1/2} \\ Z_{\mu\nu} &= min\left(\sum_{\sigma} A_{\mu\sigma} \mid \overline{P}_{\sigma\nu} \mid; \sum_{\sigma} B_{\mu\sigma} \mid \underline{P}_{\sigma\nu} \mid\right) \end{aligned} \tag{3.27}$$

gives an upper-bound for each transformation step in Eq. ..., for example

$$(\mu'\sigma' \mid \nu'\lambda') \leq Q_{\mu'\sigma'} Q_{\nu'\lambda'} \tag{3.28}$$

$$\left(\underline{\mu}\sigma' \mid \nu'\lambda'\right) \leq X_{\mu'\sigma'} Q_{\nu'\lambda'} \tag{3.29}$$

$$\left(\underline{\mu}\overline{\sigma} \mid \underline{\nu}\overline{\lambda}\right) \leq Z_{\mu\sigma} Z_{\nu\lambda} \tag{3.30}$$

and an eficient screening protocol can be devised (Has1993) to get quadratic scaling AO-MP2.

Linear Scaling AO-MP2

For the two-electron repulsion integrals, the $1/R$ decay between the charge densities $(\mu\sigma|$ and $|\nu\lambda)$ is too slow to be of any use even for large systems. However, it has been shown (Aya1999) that *bra* and *ket* in the Laplace integral tensor $e^{(\alpha)}$ decays much faster with $1/R^3$. Here, we follow the discussion in Ref Lam2005a.

For two non-overlapping charge densities $(\mu\sigma|$ and $|nu\lambda)$ the following inequality holds

$$(\mu\sigma \mid \nu\lambda) = (\mu\sigma| \frac{1}{\mathbf{r}_{12}} |\nu\lambda) \leq \frac{1}{R} \left| \sum_{n=0}^{\infty} \frac{(\mu\sigma| \left(\mathbf{r}_1 - \mathbf{r}_2\right)^n |\nu\lambda)}{R^n} \right| \tag{3.31}$$

We then introduce the following abbreviation for the nth order 1-centre multipole integrals

$$M_{\mu\sigma}^{(n)} = \int \chi_\mu(r_1)\mathbf{r}_1^n\chi_\sigma(r_1)dr \tag{3.32}$$

where M^0 are the overlap integrals, M^1 are the dipole integrals etc. We can then rewrite equation (...) as a multipole expansion

$$\begin{aligned}(\mu\sigma \mid \nu\lambda) \leq\ & R^{-1}\left|M_{\mu\sigma}^{(0)}M_{\nu\lambda}^{(0)}\right| + R^{-2}\left|M_{\mu\sigma}^{(1)}M_{\nu\lambda}^{(0)} - M_{\mu\sigma}^{(0)}M_{\nu\lambda}^{(1)}\right| \\ & + R^{-3}\left|M_{\mu\sigma}^{(2)}M_{\nu\lambda}^{(0)} - 2M_{\mu\sigma}^{(1)}M_{\nu\lambda}^{(1)} + M_{\mu\sigma}^{(0)}M_{\nu\lambda}^{(2)}\right| \\ & + R^{-4}\left|M_{\mu\sigma}^{(3)}M_{\nu\lambda}^{(0)} - 3M_{\mu\sigma}^{(2)}M_{\nu\lambda}^{(1)} + 3M_{\mu\sigma}^{(1)}M_{\nu\lambda}^{(2)} - M_{\mu\sigma}^{(0)}M_{\nu\lambda}^{(3)}\right| \\ & + \mathcal{O}(R^{-5})\end{aligned} \tag{3.33}$$

From equation ..., we know that $M_{\underline{\mu}\overline{\sigma}}^{(0)} = S_{\underline{\mu}\overline{\sigma}} = 0$. The multipole expansion for the pseudo-AO ERIs $e^{(\alpha)}$ is therefore reduced to

$$\begin{aligned}\left(\underline{\mu}\overline{\sigma} \mid \underline{\nu}\overline{\lambda}\right) \leq\ & R^{-3}\left|-2M_{\underline{\mu}\overline{\sigma}}^{(1)}M_{\underline{\nu}\overline{\lambda}}^{(1)}\right| \\ & + R^{-4}\left|-3M_{\underline{\mu}\overline{\sigma}}^{(2)}M_{\underline{\nu}\overline{\lambda}}^{(1)} + 3M_{\underline{\mu}\overline{\sigma}}^{(1)}M_{\underline{\nu}\overline{\lambda}}^{(2)}\right| + \mathcal{O} \\ & + \mathcal{O}(R^{-5})\end{aligned} \tag{3.34}$$

which shows the $1/R^3$ dependence of the tensor $\left(\underline{\mu}\overline{\sigma} \mid \underline{\nu}\overline{\lambda}\right)$. Combined with the $1/R$ decay of the AO ERIs, this leads to an overall $1/R^4$ behaviour for the AO-MP2 energy. This long-range decay can be exploited to introduce a sparsity relationship between the bra and ket quantities, and reduce the scaling of AO-MP2 from $\mathcal{O}(N^2)$ to $\mathcal{O}(N^1)$. In the original paper by Ayala and Scuseria, this decay was accounted for by introducing an interaction domain centred on each atomic orbital μ in the form of a sphere. For the integrals $\left(\underline{\mu}\overline{\mu} \mid \underline{\nu}\overline{\nu}\right)$, the domain $\mathcal{D}(\mu)$, comprises all charge distributions $\sigma\lambda$ for which

$$\left(\underline{P}_{\mu\sigma}S_{\sigma\lambda}\overline{P}_{\lambda\mu}\right) \geq \epsilon \tag{3.35}$$

The radius R_μ of the interaction sphere is defined by the maximum distance between μ and the charge density $\sigma\lambda$ in its domain. One can then screen long-range behaviour for the interaction sphere μ and ν by the distance criterium

$$r_{\mu\nu} - R_\mu - R_\nu \geq r_0 \tag{3.36}$$

The biggest drawback of the scheme above is that the thresholding parameters r_0 and ϵ are system-dependent. A more rigorous screening method has been proposed by Lambrecht et al. known as multipole based integral estimates (MBIE) (Lam2005,Lam2005a). MBIEs offer a tight upper bound for the AO and pseudo-AO electron integrals by using the multipole expansion in Eq. ... and replacing the higher order terms $\mathcal{O}(R^{-5})$ by lower-order ones. ALternative: QQR screening

Cholesky Decomposition of Pseudo-Densities

As with any method formulated entirely in an AO basis, AO-MP2 suffers from $\mathcal{O}(N^4)$ scaling with increasing basis set N. The cost associated with larger basis sets can be mitigated by Cholesky decomposition of the pseudo-density matrices (CDD) (Zie2009). Similar to the orbital localization technique described in Section ..., where the (incomplete) CD of the occupied and virtual Hartree Fock density matrices yields a set of occupied and virtual Cholesky molecular orbitals, the CD of the pseudo-density matrices $\underline{P}^{(\alpha)}$ and $\overline{P}^{(\alpha)}$ yields a set of Cholesky pseudo-molecular orbitals:

$$\underline{\mathbf{P}}^{(\alpha)} = \underline{\mathbf{L}}^{(\alpha)}\underline{\mathbf{L}}^{(\alpha)T} \tag{3.37}$$

$$\overline{\mathbf{P}}^{(\alpha)} = \overline{\mathbf{L}}^{(\alpha)}\overline{\mathbf{L}}^{(\alpha)T} \tag{3.38}$$

The pseudo-molecular orbitals show a local behaviour inherited from the sparsity of the pseudo-density matrices. It has been observed however (Lue2017), that the pseudo-MOs L are not always very well localized. A more localized set of MOs can be obtained by using the orthogonalized pseudo-density matrices, for example in the case of $\underline{\mathbf{P}}^{(\alpha)}$:

$$\underline{\mathbf{P}}^{(\alpha)}_{orth} = \mathbf{S}^{1/2}\underline{\mathbf{P}}^{(\alpha)}\mathbf{S}^{1/2} \tag{3.39}$$

The pseudo-MO coefficients are then obtained as

$$\underline{L}^{(\alpha)} = \mathbf{S}^{-1/2}\underline{\mathbf{L}}^{(\alpha)}_{orth} \tag{3.40}$$

The square root and inverse square root of the overlap matrix $\mathbf{S}$ are most reliably found by cholesky decomposition. The number of occupied and virtual pseudo-MOs is given by the rank of the occupied/virtual pseudo-density matrices, which in turn is equal or less than the number of occupied/virtual CMOs.

One can then formulate the CDD-AO-MP2 energy expression

$$E_{CDD-AO-MP2} = -\sum_{\alpha}^{n}\sum_{\underline{i}\overline{a}\underline{j}\overline{b}} \left(\underline{i}\overline{a} \mid \underline{j}\overline{b}\right)^{(\alpha)} \left[2\left(\underline{i}\overline{a} \mid \underline{j}\overline{b}\right)^{(\alpha)} - \left(\underline{i}\overline{b} \mid \underline{j}\overline{a}\right)^{(\alpha)}\right] \tag{3.41}$$

whith the pseudo-MO integrals

$$\left(\underline{i}\overline{a} \mid \underline{j}\overline{b}\right)^{(\alpha)} = \underline{L}^{(\alpha)}_{\mu\underline{i}}\overline{L}^{(\alpha)}_{\sigma\overline{a}} \left(\mu\sigma \mid \nu\lambda\right) \underline{L}^{(\alpha)}_{\nu\underline{j}}\overline{L}^{(\alpha)}_{\lambda\overline{b}} \tag{3.42}$$

CDD-AO-MP2 therefore reduces the sizes of the tensors from N^4_{AO} to $N^2_{occ}N^2_{vir}$, while still being sparse. Similar to AO-MP2, Schwarz screening and interaction domain can be introduced to obtain quadratic and linear scaling CDD-AO-MP2.

Density Fitting in AO-MP2

To reduce the prefactor associated with integral transformation, either from AOs to pseudo-AOs, or from AOs to pseudo-MOs, on can furthermore introduce density fitting (Zie2009,Mau2014).The transformed 3c2e integrals are given at each Laplace point α by

$$\left(X \mid \underline{\mu}\overline{\nu}\right)^{(\alpha)} = \left(X \mid \mu'\nu'\right) \underline{P}^{(\alpha)}_{\mu\mu'}\overline{P}^{(\alpha)}_{\nu\nu'} \tag{3.43}$$

which are evaluated with $\mathcal{O}(N^2)$ cost. Using local density fitting approxaimations, this step can be reduced to linear.

SOS-AO-DF-MP2

From section ..., we know that SOS-MP2 is a cost-efficient variant of MP2 with excellent accuracy. Starting from equation ..., we omit the same-spin contributions and also apply the density fitting approximation to arrive at the energy expression for the AO-DF-SOS-MP2 (Mau2014,Gla2020)

$$\begin{aligned} E_{AO-DF-SOS-MP2} = -c_{os} \sum_{\alpha=1}^{nlap} \sum_{\mu\nu\sigma\lambda} & \left(\underline{\nu\overline{\sigma}} \mid X\right)^{(\alpha)} \left(X \mid Y\right)^{-1} \left(Y \mid \underline{\nu\overline{\lambda}}\right)^{(\alpha)} \\ & \left(\mu\sigma \mid X'\right) \left(X' \mid Y'\right) \left(Y' \mid \nu\lambda\right) \end{aligned} \tag{3.44}$$

Introducing the intermediates

$$Z_{XY}^{(\alpha)} = \left(X \mid \underline{\mu\overline{\sigma}}\right)^{(\alpha)} \left(\mu\sigma \mid Y\right) \tag{3.45}$$

$$\tilde{Z}_{XY}^{(\alpha)} = \left(X \mid R\right)^{-1} Z_{RX}^{(\alpha)} \tag{3.46}$$

we arrive at a very compact expression

$$E_{AO-DF-SOS-MP2} = -c_{os} \sum_{\alpha=1}^{nlap} \sum_{XY} \tilde{Z}_{XY}^{(\alpha)} \tilde{Z}_{YX}^{(\alpha)} \tag{3.47}$$

Without local density fitting, the time determining step is the computation of $\mathbf{Z}^{(\alpha)}$ The sparse map of the intermediate is given by

$$\begin{array}{ccccc} X & & \mu' & \overset{\mathrm{S}}{\leftrightarrow} & \nu' & \\ & \mathrm{P}\updownarrow & & & \updownarrow\mathrm{P} & \\ & & \mu & \underset{\mathrm{S}}{\longleftrightarrow} & \nu & Y \end{array}$$

which suggests that the AO-DF-SOS-MP2 has an overall asymptotic scaling of $\mathcal{O}(N^3)$. With local density fitting, the graph can however become fully connected

$$\begin{array}{ccccc} & & \mathrm{LDF} & & \\ & \swarrow & \downarrow & \searrow & \\ X & & \mu' & \overset{\mathrm{S}}{\leftrightarrow} & \nu' \\ & \mathrm{P}\updownarrow & & & \updownarrow\mathrm{P} \\ & & \mu & \underset{\mathrm{S}}{\longleftrightarrow} & \nu \quad Y \\ & & \uparrow & \nwarrow & \uparrow \\ & & & & \mathrm{LDF} \end{array}$$

where "LDF" is the sparsity relationship introduced between the auxiliary density X and the product density $(\mu\nu|$, which is metric-specific. In the case of quasi-robust density fitting, LDF = S, and the intermediates $\mathbf{Z}^{(\alpha)}$ can be constructed with linear effort. For weaker decay behaviour, such as the error function coulomb-attenuated metric, the scaling is intermediate between linear and quadratic (Gla2020).

3.2.2 Local Molecular Orbital MP2

Problems with PAO based methods:

While linear scaling MP2 was first achived using an atomic orbital formulation, the first low-scaling MP2 implementations were actually formulated in a local molecular orbital basis with domain-specific virtual orbitals (Pul1983-Sae1987). SEPA, electron pairs etc, NESBETS theorem ....

Laplace LMP2

In the local molecular orbital basis, the Fock matrix is no longer diagonal, and the amplitudes t_{ia}^{jb} can no longer be easily expressed in a local basis, due to the energy denominator. AO-MP2 tackle this problem by virtue of the Laplace transform. Similarly, one can obtain an energy expression in the LMO basis. The Laplace decomposed MP2 energy is given by

$$E_{MP2} = \sum_{\alpha}^{n} \sum_{iajb} |w^{(\alpha)}| e^{(\epsilon_i+\epsilon_j-\epsilon_a-\epsilon_j)t^{(\alpha)}} \left[2\,(ia \mid jb) - (ib \mid ja)\right](ia \mid jb) \tag{3.48}$$

Introducing the unitary occupied and virtual LMO-MO transformation matrix **U**

$$|i\rangle = U_{i\underline{i}}\,|\underline{i}\rangle \tag{3.49}$$

$$|a\rangle = U_{i\overline{a}}\,|\overline{a}\rangle \tag{3.50}$$

which is factorized out, Equation ... becomes

$$\begin{aligned} E_{MP2} &= \sum_{\alpha}^{n} \sum_{iajb} \sum_{\underline{i}\overline{a}\overline{b}} \sum_{\underline{k}\overline{c}\underline{l}\overline{d}} |w^{(\alpha)}| e^{(\epsilon_i+\epsilon_j-\epsilon_a-\epsilon_j)t^{(\alpha)}} U_{i\underline{i}} U_{a\overline{a}} \left[2\left(\underline{i}\overline{a} \mid \underline{j}\overline{b}\right) - \left(\underline{j}\overline{b} \mid \underline{j}\overline{b}\right)\right] U_{j\underline{j}} U_{b\overline{b}} \\ &\qquad U_{i\underline{k}} U_{a\overline{c}} \left(\underline{k}\overline{c} \mid \underline{l}\overline{d}\right) U_{j\underline{l}} U_{b\overline{d}} \\ &= \sum_{\alpha}^{n} \sum_{\underline{i}\overline{a}\overline{b}} \left[2\left(\underline{i}\overline{a} \mid \underline{j}\overline{b}\right) - \left(\underline{j}\overline{b} \mid \underline{j}\overline{b}\right)\right] \sum_{\underline{k}\overline{c}\underline{l}\overline{d}} X_{\underline{i}\underline{k}}^{(\alpha)} Y_{\overline{a}\overline{c}}^{(\alpha)} \left(\underline{k}\overline{c} \mid \underline{l}\overline{d}\right) X_{\underline{j}\underline{l}}^{(\alpha)} Y_{\overline{b}\overline{d}}^{(\alpha)} \\ &= \sum_{\underline{i}\overline{a}\overline{b}} \left[2\left(\underline{i}\overline{a} \mid \underline{j}\overline{b}\right) - \left(\underline{j}\overline{b} \mid \underline{j}\overline{b}\right)\right] \mathcal{T}_{\underline{i}\overline{a}\underline{j}\overline{b}} \end{aligned} \tag{3.51}$$

with the Laplace amplitudes $\mathcal{T}$ and the Laplace matrices

$$X_{\underline{i}\underline{k}}^{(\alpha)} = \sum_{i} U_{i\underline{i}} |w^{(\alpha)}|^{1/4} e^{\epsilon_i t^{(\alpha)}} U_{k\underline{k}} \tag{3.52}$$

$$Y_{\overline{a}\overline{c}}^{(\alpha)} = \sum_{a} U_{a\overline{a}} |w^{(\alpha)}|^{1/4} e^{-\epsilon_a t^{(\alpha)}} U_{c\overline{c}} \tag{3.53}$$

Equation ... is the general expression for the MP2 energy in a local molecular orbital basis, where both the occupied and virtual orbitals are *orthogonal*. The situation changes slightly when using non-orthogonal PAOs, as PAOs and CMOs are no longer related by a unitary transformation:

$$|I\rangle = P_{Ii}\,|i\rangle \tag{3.54}$$

$$|i\rangle = P_{Ii} S^{-1}_{IJ} |J\rangle \tag{3.55}$$

with **S** being the overlap matrix in the PAO basis. Due to the non-orthogonality of the PAOs, entries in the overlap matrix can become very small which might lead to numerical instability when computing its inverse. For this reason, the inverse $\mathbf{S}^{-1}$ is substituted by a more general *pseudo-inverse* **V** with the property (see ...)

$$\mathbf{SVS} = V \tag{3.56}$$

The Laplace matrices will then take the following form instead (Kat2008)

$$\mathbf{X}^{(\alpha)} = \mathbf{V}\mathbf{A}^{(\alpha)}\mathbf{V}^{(\dagger)} \tag{3.57}$$

$$\mathbf{Y}^{(\alpha)} = \mathbf{V}\mathbf{B}^{(\alpha)}\mathbf{V}^{(\dagger)} \tag{3.58}$$

$$A^{(\alpha)}_{IK} = \sum_i P_{Ii} |w^{(\alpha)}|^{1/4} e^{\epsilon_i t^{(\alpha)}} P_{kK} \tag{3.59}$$

$$B^{(\alpha)}_{AB} = \sum_a Q_{Aa} |w^{(\alpha)}|^{1/4} e^{-\epsilon_a t^{(\alpha)}} Q_{Aa} \tag{3.60}$$

In practice, only the virtual orbital space is transformed to PAOs, while the occupied space is kept in an orthogonal local orbital representation.

NESBETS THEOREM

Orbital Invariant MP2

Alternatively, the local MP2 amplitudes can be determined iteratively via an *orbital-invariant* formulation of the MP2 energy expression. Orbital-invariant MP2 predates LT-MP2 by a decade and is based on the *Hylleraas functional* (Hyl,Pul1986). The Hylleraas functional form of the energy is given by minimizing

$$E^{(2)} = min \left[2 \left\langle \Psi^{(1)} \right| \mathbf{H} - E_0 \left| Psi^{(0)} \right\rangle - \left\langle \Psi^{(1)} \right| \mathbf{H}_0 - E_0 \left| \Psi^{(1)} \right\rangle \right] \tag{3.61}$$

In the case of MP2, the qunatities in Equation ... take the form

$$\left\langle \Psi^{(1)} \right| \mathbf{H} - E_0 \left| \Psi^{(0)} \right\rangle = \frac{1}{4} \sum_{ijab} t_{ijab} \langle ij|ab\rangle \tag{3.62}$$

$$\left\langle \Psi^{(1)} \right| \mathbf{H}_0 - E_0 \left| \Psi^{(1)} \right\rangle = \frac{1}{8} \sum_{ijabc} t_{iajb} f_{cb} t_{iajc} - \frac{1}{8} \sum_{ijkab} t_{iajb} f_{jk} t_{iakb} \tag{3.63}$$

Minimization of the MP2 Hylleraas functional, with respect to the amplitudes **t** yields a set of linear equations given by

$$R_{iajb} = \langle ij| |ab\rangle + \sum_c \left(t_{ijab} f_{cb} + f_{ac} t_{iacb} \right) - \sum_k \left(t_{iakb} f_{kj} + f_{ik} t_{kajb} \right) = 0 \tag{3.64}$$

where **R** is the residual. The amplitudes **t** are then no longer computed directly by a closed expresion, but iteratively by solving the system of equations, in a similar vein to coupled cluster. This method is said to be orbital invariant, because any molecular orbital representation can be used. For a set of orthogonal MOs $\underline{i}$ and $\overline{a}$, the quantities in

Equation ... are simply replaced by their local equivalent. As was the case in LT-LMP2, if PAOs are to be used for the virtual orbital space, the non-orthogonality needs to be taken into consideration. For a mixed LMO-PAO basis, Equation ... reads

$$\begin{aligned} R_{\underline{i}A\underline{j}B} = \langle \underline{ij} | \, |AB\rangle + \sum_C \Big(f_{AC} t_{\underline{i}C\underline{j}L} S_{LB} + S_{AC} t_{\underline{i}C\underline{j}D} f_{DB} \Big) \\ - \sum_{\underline{k}} \Big(f_{\underline{ik}} S_{AC} t_{\underline{k}C\underline{j}D} S_{DB} + f_{\underline{kj}} S_{AC} t_{\underline{i}C\underline{j}D} S_{DB} \Big) = 0 \end{aligned} \tag{3.65}$$

For specific virtual orbitals, the equations are solved individually for each electron pair ij to obtain their amplitude $\mathbf{t}_{ij}$ and then compute the pair correlation energy.

why not AO?

Quadratic Scaling LMP2

Similar to CEPA, the MP2 energy can be computed as a sum of electron pair energies

$$E_{MP2} = \sum_i j e_{ij} \tag{3.66}$$

$$e_{ij} = \left(2t_{ij}^{ab} - t_{ij}^{ba}\right) (ia \mid jb) \tag{3.67}$$

where the amplitudes t_{ijab} are computed according to Eq. ... . The occupied molecular orbitals ij are localized using e.g. Foster-Boys, Pipiek-Mezey or a Cholesky decomposition of the density matrix. Virtual orbitals are generally localized by projection onto the atomic orbital space (PAOs) and subsequently assigning them to pair domains $[ij]$ (domain specific virtuals), or by diagonalizing the MP2 density matrix for each electron pair (pair natural orbitals). In all cases, the number of SVOs scales as $\mathbf{O}(1)$ for each electron pair ij, in the limit of large molecules (see ...).

For well localized orbitals, the electron pair correlation e_{ij} decays with $1/r_{ij}^6$ whith the distance between orbital centres. Electron pairs are generally divided into four groups: strong pairs ($r_{ij} < 1a_0$), weak pairs ($1 < r_{ij} \leq 8a_0$), distant pairs ($8 < r_{ij} \leq 15a_0$), and very distant pairs ($15 < r_{ij}$) (Sch1999). Other than by distance criteria, electron pairs can also be grouped by their pair energy (Nee2009). Figure ... shoes the number of significant electron pairs in each category for glycine chains. The number of strong, weak and distant pairs scale as $\mathcal{O}(N)$, while the number of very distant pairs scales quadratically.

The major bottle-neck in LMP2 is, as usual, the transformation of the 2 electron integrals from the AO basis into the local basis

$$\left(\underline{i}\overline{a} \mid \underline{j}\overline{b}\right) = L_{\mu\underline{i}} L_{\sigma\overline{a}} \, (\mu\sigma \mid \nu\lambda) \, L_{\nu\underline{j}} L_{\lambda\overline{b}} \tag{3.68}$$

The expression above translates into the sparsity diagram

$$\begin{array}{ccc} \mu & \overset{\mathrm{S}}{\longleftrightarrow} & \sigma \\ \updownarrow & & \updownarrow \\ \underline{i} & \longleftrightarrow & \overline{a} \end{array} \qquad \begin{array}{ccc} \nu & \overset{\mathrm{S}}{\longleftrightarrow} & \lambda \\ \updownarrow & & \updownarrow \\ \underline{j} & \longleftrightarrow & \overline{b} \end{array}$$

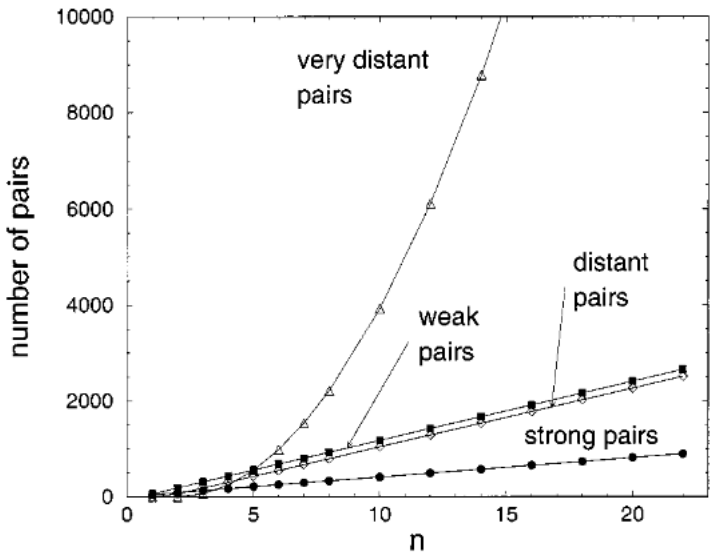


Figure 3.1: Taken from Sch1999

which indicates that the MO integrals can be evaluated with $\mathcal{O}(N^2)$ effort without further approximations. One thing to note is that the quadratic scaling is also obatined, even if the sparsity relationships $i \leftrightarrow a$ and $j \leftrightarrow b$ did not exist, i.e. where virtual orbitals are localized, but not grouped into (apir) domains. The major disadavantage of such non-pair specific methods is that the virtual orbital space is less compact, which leads to a high overhead for integral transformation involving virtual orbitals, which could be the reason that there are no examples in literature using such a scheme. Establishing an a priori sparsity relationship between occupied and virtual space allows to more easily reach the low-scaling regime.

Linear Scaling LMP2

It has been found [Sae1987] early on that the quadratic scaling very distant electron pairs can safely ignored without major impact on the total correlation energy. (Distance or Connectivity ciriteria) Distant pairs can also be approximated either by a multipole expansion (Het1998) or empirically (Rau1995), which further lowers the prefactor of the method. As a consequence, this establishes a sparsity relationship between i and j, and the sparsity diagram for the MO integrals becomes fully connected

$$\begin{array}{ccc} \mu & \overset{\mathrm{S}}{\longleftrightarrow} & \sigma \\ \updownarrow & & \updownarrow \\ \underline{i} & \longleftrightarrow & \overline{a} \end{array} \qquad \begin{array}{ccc} \nu & \overset{\mathrm{S}}{\longleftrightarrow} & \lambda \\ \updownarrow & & \updownarrow \\ \underline{j} & \longleftrightarrow & \overline{b} \end{array}$$

$$\underbrace{\hphantom{\underline{i} \longleftrightarrow \overline{a} \quad \underline{j}}}_{1/R^6}$$

and linear scaling LMP2 therefore becomes possible (Sch1999).

Instead of using distance criteria, can use screening:

Density Fitting for LMP2

While specific virtual orbitals form a very compact representation of the virtual space, the fact that each electron pair has their own orthogonal vitual orbital basis means that the total number of virtuals can become exceedingly large, and consequently increases the cost associated with the AO-MO transformation step. The most expensive step then becomes

$$(\underline{i}\overline{a} \mid P) = L_{\mu\underline{i}} \, (\mu\nu \mid P) \, L_{\nu\overline{a}} \tag{3.69}$$

Transformation of the 3c2e integrals scales with $\mathcal{O}(N^2)$. Linear scaling can be achieved by instroducing an orbital-specific fitting domain $[i]_fit$, e.g. by assigning all auxiliary functions P on atoms with a Mulliken charge above a given threshold for the local orbital i (Pin2015), or by using a Boughton-Pulay like scheme (Wer2015). This yields the sparsity diagramm

$$\begin{array}{ccccc} P & & \mu & \overset{\mathrm{S}}{\longleftrightarrow} & \sigma \\ \uparrow & & \updownarrow & & \updownarrow \\ \text{LDF} & \longrightarrow & \underline{i} & \longleftrightarrow & \overline{a} \end{array}$$

As opposed to SOS-AO-MP2, where density fitting can give a more favourable factorization of the energy expression, the MO integrals need to fully assembled for LMP2 in order to solve the linear equations (...). The assembly is done in two steps

$$B^X_{\underline{i}\overline{a}} = \sum_{Y \in [i]_{fit} \cup [j]_{fit}} (X \mid Y)^{-1/2} \, (Y \mid \underline{i}\overline{a}) \tag{3.70}$$

$$\left(\underline{i}\overline{a} \mid \underline{j}\overline{b}\right) = \sum_{X \in [i]_{fit} \cup [j]_{fit}} B^X_{\underline{i}\overline{a}} B^X_{\underline{j}\overline{b}} \tag{3.71}$$

The two steps are repeated for each electron pair ij, and the sum runs over all auxiliary functions P in the unified fitting domain $[i]_{fit} \cup [j]_{fit}$, which enforces linear scaling for these steps as well.

3.2.3 Atomic Orbital Coupled Cluster

3.2.4 Local Coupled Cluster

3.2.5 FNO Coupled Cluster ??

3.3 Critical Stance on AO vs LMO

Distance criteria, or Mulliken/Löwdin pouplation AO only for closed expressions (not for CCSD++) LOcal: ij cirteria, ij-¿ab criteria

Chapter 4

Local Correlation: Excited State

The previous chapter demonstrated how it is possible to achieve low to linear computational complexity for Hartree Fock as well as Møller-Plesset perturbation theory and Coupled Cluster. Highly optimized code has been developped for all of them, exploiting distributed memory parallelism and/or accelerators such as GPUs. The local treatment of electron correlation can naturally be extended to excited states as well. The development of low-scaling excited states has been steadily progressing since the early 2000s, and many competing flavors have emerged over the years both for the Algebraic Diagrammtic Construction method, Coupled Cluster linear response, and Equation-of-Motion Coupled Cluster. Local excited state methods need to take into considerations both local electron correlation, as well as the local character of the excited state itself.

4.1 Low Scaling ADC, CCLR and EOM-CC using Molcular Orbitals

Virtually all existing low-scaling implementations of ADC, CCLR and EOM-CCSD use some form of local or compact molecular orbital representation, to varying degrees of success. The major problem that these methods face is the non-locality of certain excited states such as charge transfer states (Figure ...), which can involve occupied and virtually orbitals which are localized on entirely different parts in the system. Clearly, truncating virtual orbitals spatially is no longer a valid option, and makes a straight-forward extension of LMO-methods difficult, becuase they cut out far-away contributions. Similarly, the excited state is often not properly described by the electronic ground state (pair-)densities and their associated (pair) natural representation. Over the years, various strategies have been proposed to adapt exisiting LMO, NO and PNO schemes to excited states as well.

4.1.1 Orbital Invariant ADC/CCLR/EOM-CCSD

Similarly to ground state MP and CC methods, the working equations can be reformulated to work with any orbital representation. Consider for example the general ADC(2)

working excitations for computing the singles and doubles vectors

$$\begin{aligned} r_{ia} &= A_{ia,jb}u_{jb} + A_{ia,jbkc}u_{jbkc} \\ r_{iajb} &= A_{iajb,kc}u_{kc} + A_{iajb,kcdl}u_{kcdl} \end{aligned} \tag{4.1}$$

Transforming to the local basis is straight-forward: the trial vectors u_{ia}, as well as the molecular electron integrals $(ia \mid jb)$ and MP2 amplitudes t_{iajb} in the Jacobian $\mathbf{A}$ (see Section ...) are simply replaced by their local counterparts $u_{\underline{i}\overline{a}}$, $\left(\underline{i}\overline{a} \mid \underline{j}\overline{b}\right)$ and $t_{\underline{i}\overline{a}\underline{j}\overline{b}}$. The local MP2 amplitudes are computed using either the Hylleraas functional (Equation ...) or the Laplace transform (Equation ...). Similar things apply for an orbital-invariant formulation of CCLR and EOM-CCSD.

ADC(2) and CC2 variants also allow an on-the-fly computation of the doubles part (see ...). Here, the orbital invariant formaulation becomes less straight-forward becuase the doubles-doubles block of the ADC and CC2 Jacobian matrix is no longer diagonal, similar to the Fock matrix. Fortunately, the Laplace transform can be applied to circumvent this problem, in this case shown for ADC(2)

$$\begin{aligned} r_{ia}(\omega) &= A_{ia,jb}u_{jb} + A_{ia,jbkc}\frac{A_{iajb,kc}u_{kc}}{\omega - \epsilon_a - \epsilon_b + \epsilon_i + \epsilon_j} \\ &= A_{ia,jb}u_{jb} - \sum_{\alpha}^{n} |w^{(\alpha)}| e^{(\omega - \epsilon_a - \epsilon_b + \epsilon_i + \epsilon_j)tpa} A_{ia,jbkc}A_{,kc}u_{kc} \end{aligned} \tag{4.2}$$

Using a similar apporach to Equation ..., the local ADC(2) equations are then given by

$$r_{\underline{i}\overline{a}}(\omega) = A_{\underline{i}\overline{a},\underline{j}\overline{b}}u_{\underline{j}\overline{b}} - A_{\underline{i}\overline{a},\underline{j}\overline{b}\underline{k}\overline{c}} \sum_{\alpha}^{n} e^{\omega t^{(\alpha)}} X^{(\alpha)}_{\underline{j}\underline{j}'} Y^{(\alpha)}_{\overline{b}\overline{b}'} A_{\underline{j}'\overline{b}'\underline{k}'\overline{c}',\underline{l}\overline{d}} u_{\underline{l}\overline{d}} X^{(\alpha)}_{\underline{k}\underline{k}'} Y^{(\alpha)}_{\overline{c}\overline{c}'} \tag{4.3}$$

where the Laplace matrices $\mathbf{X}$ and $\mathbf{Y}$ read

$$X^{(\alpha)}_{\underline{i}\underline{k}} = \sum_{i} U_{i\underline{i}} |w'^{(\alpha)}|^{1/4} e^{\epsilon_i t'^{(\alpha)}} U_{k\underline{k}} \tag{4.4}$$

$$Y^{(\alpha)}_{\overline{a}\overline{c}} = \sum_{a} U_{a\overline{a}} |w'^{(\alpha)}|^{1/4} e^{-\epsilon_a t'^{(\alpha)}} U_{c\overline{c}} \tag{4.5}$$

Note that the Laplace parameters $w'^{(\alpha)}$ and $t'^{(\alpha)}$ are different from the ones used in the MP2 amplitudes (—), due to the presence of the eigenvalue *omega* in the denominator. Each time the eigenvalue changes, the Laplace parameters need to be recomputed to obtain an accurate approximation.

An orbital invariant reformulation is not needed by every method. NOs, PNOs and NTOs can be *canonicalized* by diagonalizing the occupied-occupied and virtual-virtual block of the Fock matrix in the truncated NO/PNO/NTO basis to get a smaller set of canonical molecular orbitals and orbital energies. Because these types of representations generally do not depend on distance criteria, they are unaffected by the delocalized nature of the CMOS, as they seek compactness rather than locality.

4.1.2 State Specificity for Local Molecular Orbitals

The most challenging part in extending domain-specific virtual orbital methods to excited states lies in determing a suitable excitation domain in which to expand the virtual space. The first implementations of local excited state EOM-CCSD (Kor2003,Cra2002) and CC2-LR (Kat2003) constructed the domains using a Mulliken-charge like analysis of the CIS coeffcients $r_i a$. The CIS coefficients are first transformed to the LMO-PAO basis

$$r_{\underline{i}A} = U_{\underline{i}i} r_{ia} Q_{Aa} \tag{4.6}$$

To determine the importance w of each LMO/PAO, the squares of the norms of the coefficients are summed up row- and column-wise

$$\begin{aligned} w_{\underline{i}} &= \sum_{A} |r_{\underline{i}A}|^2 \\ w_A &= \sum_{\underline{i}} |r_{\underline{i}A}|^2 \end{aligned} \tag{4.7}$$

The LMOs/PAOs are then ordered by decreasing weight. Their weights are then summed up until a certain threshold T_{LMO}/T_{PAO} is reached (typically around 0.995 to 0.9999). The excited state orbital domains $[\underline{i}]_{ES}$ containg the relevant virtual orbitals are then contructed by applying the BP algorithm to a set of "excited natural orbitals" (Kor2003)

$$\phi^*_{\underline{i}} = \sum_{\overline{a}} r_{\underline{i}A} \phi_A \tag{4.8}$$

The full orbital domain of i is then given as the union of its ground-state and excited state domain $[\underline{i}] = [i]_{GS} \cup [i]_{ES}$. The virtual orbital weights w_A can be used to impose further restrictions on the virtual orbital space. Finally, the pair domains ij are formed as the union $[\underline{i}] \cup [\underline{j}]$. In general, only the computation of the doubles part, which is time-determining, is subject to domain-restrictions, while the singles part is computed without domain lists.

The method however has the major flaw that the orbital domains are highly senitive to the CIS transition density, which does not describe the excited state very accurately. Some orbitals can be dropped in the domain construction which might become important for doubles contributions. LMO methods face an interesting chicken-or-egg problem where they need information from the excited state wave function, to accurately compute properties of said function. There are several ways to mitigate this problem. In their local CC2-LR implementation, Kats and Schütz (Kat2009) use the Laplace transform (Equation ...) to recompute the doubles amplitudes on the fly, which allows to adapt the excited state domains dynamically during the optimization procedure. Starting from the CIS transition density, the domains are recomputed at each step by analyzing the state vector $r_{\underline{i}A}(\omega)$ as described above. This greatly increased accuracy compared to canonical calculations with energy differences well below 0.1 eV.

Mester et al. proposed a more pragmatic approach, where they first analyse the CIS state vector to extract the important LMOs and PAOs. They then augment the domains $[i]_{EX}$ by adding all remaining molecular orbitals that has a significant Mulliken charge on an atom that is also significant for i. This is based on the assumption that, although CIS might not be a good approximations, the important orbitals should still be close by.

Nonetheless, the LMO method is again plagued by spurious distance dependent thresholds and Mulliken charge thresholds. Nowadays, local excited state methods are mostly dominated by PNOs, NOs, or NTOs.

4.1.3 State Specificity for Natural Orbitals

NO methods achieve performance by dropping virtual natural orbitals with low occupation numbers. The first implementations of EOM-CC and CCLR in the NO representation used natural orbitals obtained from the diagonalization of the ground state MP2 density matrix (Lan2010,Kum2017). The excited state character was not taken into account, but still a reasonable speed-up could be observed. However, it was shown (Kum2017) that properties like the polarizability are much more sensitive to the truncation of the virtual orbitals than the ground state correlation energy, with the error increasing linearly as a function of the number of dropped virtual natural orbitals. While VNOs with low occupation numbers, i.e. diffuse character, can be safely ignored for the ground state correlation energy, diffuse VNOs play a much more important role for response properties, and hence fewer VNOs can be omitted. Better results could be obtained by simply truncating the virtual CMOs instead, which invalidates the use of VNOs.

In their NO-CC2 and NO-ADC(2) implementations, Mester et al. (Mes2017, Mes2018, Mes2019) proposed to compute a set of occupied and virtual NOs by diagonalizing occupied and virtual state-averaged densities

$$\mathbf{D}_{ij} = \frac{1}{2}\left(\mathbf{D}_{ij}^{MP2} + \mathbf{D}_{ij}^{CIS(D)}\right) \tag{4.9}$$

$$\mathbf{D}_{ab} = \frac{1}{2}\left(\mathbf{D}_{ab}^{MP2} + \mathbf{D}_{ab}^{CIS(D)}\right) \tag{4.10}$$

where $\mathbf{D}^{MP2}$ is the MP2 ground state density and $\mathbf{D}^{CIS}$ is the state-specific CIS(D) excited state density. Their restricted expressions read

$$D_{ij}^{MP2} = \sum_{kab}\left(2t_{ik}^{ab}t_{jk}^{ab} - t_{ik}^{ab}t_{jk}^{ab}\right) \tag{4.11}$$

$$D_{ab}^{MP2} = \sum_{ijc}\left(2t_{ij}^{ca}t_{ij}^{cb} - t_{ij}^{ca}t_{ij}^{bc}\right) \tag{4.12}$$

$$D_{ij}^{CIS(D)} = \sum_{a} c_i^a c_j^a + \sum_{kab}\left(2t_{ik}^{ab}t_{jk}^{ab} - t_{ik}^{ab}t_{jk}^{ab}\right) \tag{4.13}$$

$$D_{ab}^{CIS(D)} = \sum_{i} c_i^a c_i^b + \sum_{ijc}\left(2c_{ij}^{ca}c_{ij}^{cb} - c_{ij}^{ca}c_{ij}^{bc}\right) \tag{4.14}$$

where c_i^a are the CIS coefficients and the c_{ij}^{ab} are the CIS(D) doubles coefficients

$$c_{ij}^{ab} = \frac{\sum_c \left[(ac \mid bj)\, c_i^c + (ac \mid bi)\, c_j^c\right] - \sum_k \left[(kj \mid ai)\, c_k^b + (kj \mid bj)\, c_k^a\right]}{D_{ij}^{ab} + \omega_{CIS}} \tag{4.15}$$

The state-averaged density needs to be recomputed and diagonalized for each state because $\mathbf{D}^{CIS(D)}$ depends on the excitation energy ω. While the CIS(D) density is much

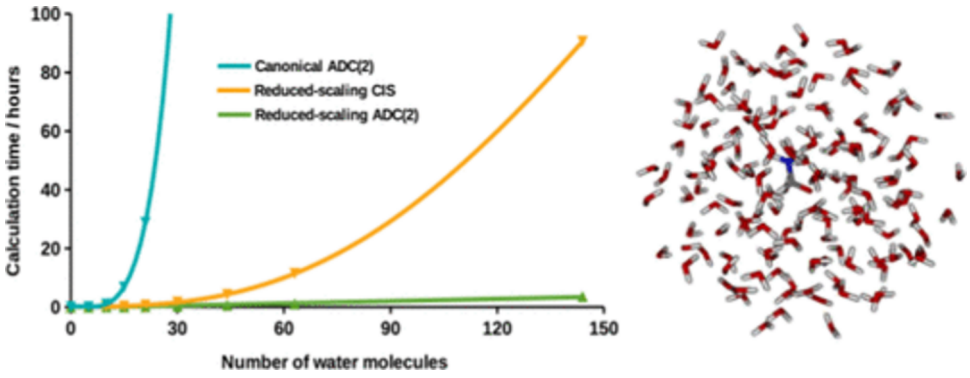


Figure 4.1: This is actually really cool

easier to compute than the ADC(2) or CC2-LR state density, it still scales with $\mathcal{O}(N^5)$. To mitigate the computational complexity, the density is constructed in a truncated orbital space: first, a set of occupied and virtual LMOs are chosen according to the CIS weighting criteria w described in the previous section. The basis is augmented by spatially close orbitals, and then canonicalized to yield a highly compact orbital molecular space which lowers the cost of constructing the CIS(D) densities.

In combination with natural auxiliary functions, this hybrid NO-LMO scheme can reduce the timings for CC2 and ADC(2) to such a drastic extent that the CIS pre-iterations become the time-determining step (Figure ...), with an additional error of only 2-4 meV. The reduced scaling however comes at a high prefactor for computing multiple different excitation energies.

4.1.4 State Specificity for Pair Natural Orbitals

Pair natural orbital methods face the same problems as NOs, where PNOs with low occupation numbers are considerably more important for response properties than for ground state properties. In a similar vein, excited state PNOs can be generated by considering lower level excited state electron pair densities (Hel2011). PNO methods have been succuessfully extended to ADC(2), CC2-LR (Hel2013), ADC(2)-x (Hel2014) and CCSD-LR (ref) by using CIS(D) or CIS(D)-like densities

$$D_{ij}^{ab} = \sum_{c} \left(2b_{ij}^{ab} - b_{ij}^{ba}\right) b_{ij}^{ab} + \left(2b_{ij}^{ab} - b_{ij}^{ba}\right) b_{ij}^{ba} \tag{4.16}$$

where $\mathbf{b}_{ij}$ are state-specific modified pair amplitudes which are not uniquely defined. Again, these methods come at the cost of a higher prefactor due to the relatively high cost of constructing PNOs. Nonetheless, it was shown that the computational complexity can be lowered to $\mathcal{O}(N^3)$ for PNO-CCSD-LR.

Efforts have also been made to develop PNO response methods which are more economical for computing larger excitation manifolds by removing the state-specificity. Instead of taking individual excited state densities, Peng et al. (Pen2018) proposed to generate a set of *state-averaged* PNOs obtained by diagonalization of the average excited

state density over an N-state manifold

$$\mathbf{D}_{ij} = \frac{1}{N} \sum_{k}^{N} \mathbf{D}_{ij}^{(k)} \tag{4.17}$$

A production-qaulity implementation has not yet been shown which uses this approach.

In their perturbed pair-natural orbital (PNO++) approach for CCLR, Cunha and Crawford (Cun2021) incorporate the external perturbation into the electron pair density

$$D_{ij}^{ab} = \sum_{c} \left(2x_{ij}^{ab} - x_{ij}^{ba}\right) x_{ij}^{ab} + \left(2x_{ij}^{ab} - x_{ij}^{ba}\right) x_{ij}^{ba} \tag{4.18}$$

where $\mathbf{x}$ are perturbed amplitudes given by

$$x_{ij}^{ab} = \frac{\overline{B}}{\overline{H}_{aa} + \overline{H}_{bb} - \overline{H}_{ii} - \overline{H}_{jj} + \omega} \tag{4.19}$$

with an external perturbation $\overline{B}$ and the similarity transformed Hamiltonian $\overline{H}$.

Finally, there are also the *back-transformed* PNOs, or bt-PNOs, where the ground state PNO-quantities like the amplitudes are transformed back to the canonical basis and used in the canonical working equations (Dut2016).

In the end, most local excited state methods using natural orbitals differ by how they redefine the amplitudes $\mathbf{t}$ for the individual excited states or the whole perturbed molecular system. It is still very much an active field of research.

4.1.5 State Specificity for Natural Transition Orbitals

The last method to obtain a compact representation of excited states is via natural transition orbitals. NTOs are the equivalent of NOs for excited states, and represent a compact representations of their dominant contribution (Figure ...). Baudin and Kristensen have developped two different CC2LR schemes based on NTOs called LoFEX (local framework for calculating excitation energies) (Bau2016) and CornFLEX (correlated natural transition orbital framework for calculating excitation energies) (Bau2017).

Again, one needs information avout the excited state to effciently compute its properties. The LoFEX method starts with a time-dependent Hartree Fock caculation and generates a set of NTOs by decomposition of the TDHF transition vectors $\mathbf{r}$ by diagonalization

$$\mathbf{r}\mathbf{r}^{\dagger}\mathbf{U} = \lambda_o \mathbf{U} \tag{4.20}$$

$$\mathbf{r}^{\dagger}\mathbf{r}\mathbf{V} = \lambda_v \mathbf{V} \tag{4.21}$$

Equations ... are alternative ways to compute the occupied and virtual NTO transformation matrices $\mathbf{U}$ and $\mathbf{V}$, rather than by singular value decomposition. A set of dominant NTO pairs is then chosen for which their occupation numbers are above a given treshold τ_{LoFEX}. The non-dominant NTOs are not discarded, but rather localized. The idea is to construct a sourrounding excitation orbital space (XOS) containing LMOs that are important for correlation effects of the NTOs. A first guess to the XOS is chosen based on distance criteria and Löwdin charges. The CC2LR are then solved in that basis, and new NTOs

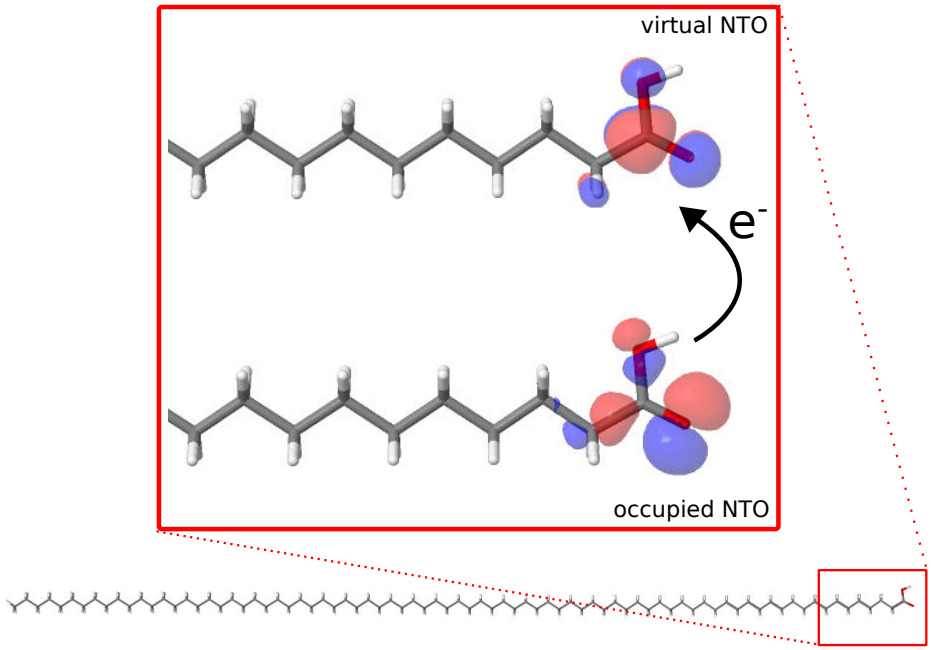


Figure 4.2: Dominant natural transition orbital pair for the lowest excitation of the carboxylic acid $C_{79}H_{159}COOH$ ($\pi \rightarrow \pi^*$ transition). The span of the NTOs is very small compared to the rest of the molecule, and the compactness can be used to drastically speed up excited state caculations.

are computed from the CC2 transition vector and added to the XOS. This procedure is repeated until the excitation energy ω for that state has converged. While the guess XOS is first formed using distance cirtieria, the subsequent optimization procudure makes the method much more robust and black-box. Even for relatively small molecules, LoFEX can obtain considerable speed-ups. The main disadvantage is that LoFEX does not give any leverage for very delocalized excitations.

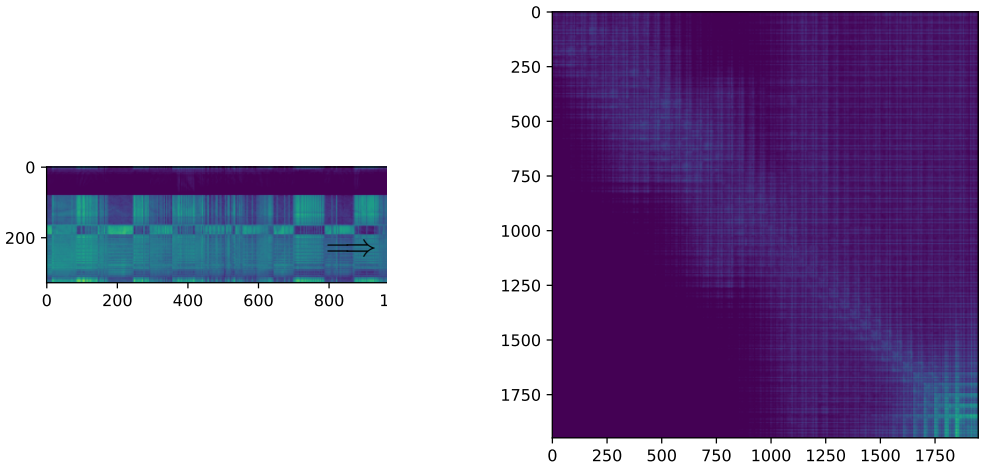
The CornFLEX method constructs a set of CIS(D)-like NTOs (CIS(D')-NTOs) which is obtained from diagonalizing a CIS(D)-like density matrix in the CIS-NTO basis. As opposed to CIS-NTOs, the CIS(D') NTOs also include correlation effects and are a more robust representation that the simple ad-hoc extension of CIS-NTOs using LMOs. Speed-ups can be observed in CornFLEX even for delocalized excitations.

4.2 Atomic Orbital Configuration Interaction Singles

-¿ AO TDSCF (Kussmann)

The methods presented in the previous section all work similarly. They first start by approximating the targeted excited states with a lower level of theory using CIS or CIS(D). They then solve higher order equations in the basis obtained from that approximation and may also dynamically augment the correlation domain while optimizing the excitation energies. The methods work on the principle of orbital *compactness* rather than sparsity

At the moment of writing, CIS is the only excited state method which is traditionally evaluated in the AO basis. While CIS does not give qualitatively good results, it is still



(a) (b)

Figure 4.3: Logarithm of the absolute values of the matrix elements in the transition densities in the MO (left) and AO basis (right) for the lowest excited state for the carboxylic acid $C_{79}H_{159}COOH$. The excitation domain is entirely localized on the carboxylic group. Using sparse matrix algebra, significant speed-ups can be obtained for CIS in the AO basis

a very important stepping stone for higher order methods, as was demonstrated in the previous section. Omitting the zero-order contributions, the CIS working equations are given by

$$u_{ia} = \left[2\,(ia \mid jb) - (ib \mid ja)\right] u_{jb} \tag{4.22}$$

Factoring out the MO coefficient matrices:

$$\begin{aligned} u_{ia} &= C_{\mu i} C_{\sigma a} \left[2\,(\mu\sigma \mid \nu\lambda) - (\mu\lambda \mid \nu\sigma)\right] C_{\nu j} C_{\lambda b} u_{jb} \\ &= C_{\mu i} C_{\sigma a} \left[2\,(\mu\sigma \mid \nu\lambda) - (\mu\lambda \mid \nu\sigma)\right] P_{\nu\lambda} \end{aligned} \tag{4.23}$$

where $\mathbf{P}$ is the non-symmetric transition density in the AO basis. The CIS working equations can be reduced to the construction of a "pseudo"-Fock matrix which has a coulomb and an exchange part. The Fock matrix is then transformed to the MO basis:

$$F_{\mu\nu} = J_{\mu\nu} + K_{\mu\nu} \tag{4.24}$$

$$u_{ia} = C_{\mu i} F_{\mu\nu} C_{\mu a} \tag{4.25}$$

For localized excitations, the AO transition density is sparse (Figure ...), and similar approximation can be used as in Hartree Fock, e.g. LinK, CFMM, or LDF. CIS can therefore be evaluated with $\mathcal{O}(N)$ computational effort.

OTHER: https://www.kth.se/blogs/pdc/2018/11/scalability-strong-and-weak-scaling/

Part I

Annex

- ERI deomposition: cholesky, THC, pseudo-spectral - The evil matrix inversion: considerations - mulliken, boughton pulay