My PhD thesis

Maximilien Alexandre Ambroise

June 22, 2021

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

Licensing

(empty)

Contents

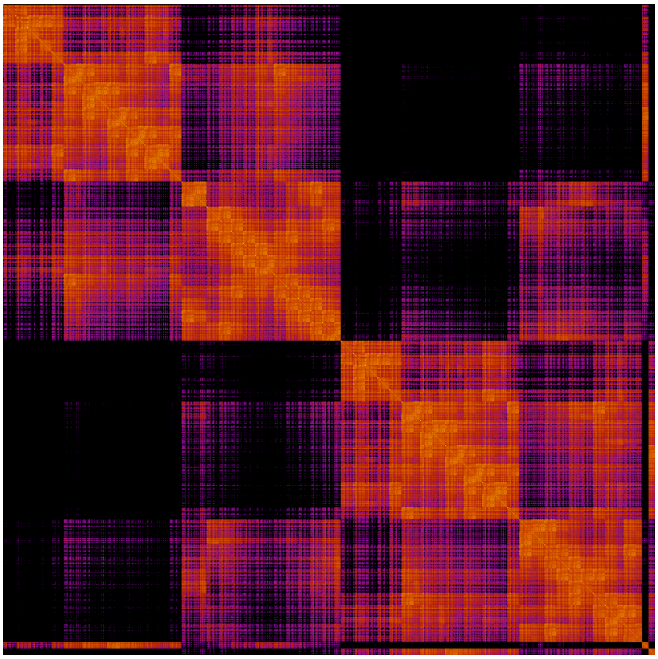


Figure 1: Adenine-guanine fock matrix Hartree FOck cc-pVTZ

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

Part I
Theory: The Basics

Part II
Theory: Reduced-Cost QC

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:

$$(ia \mid jb) = \sum_{b}^{vir} C_{\sigma b} \sum_{a}^{vir} C_{\nu a} \sum_{j}^{occ} C_{\lambda j} \sum_{i}^{occ} C_{\mu i} \, (\mu\nu \mid \lambda\sigma) \tag{1}$$

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).

7 Density Fitting

About density fitting. Principles

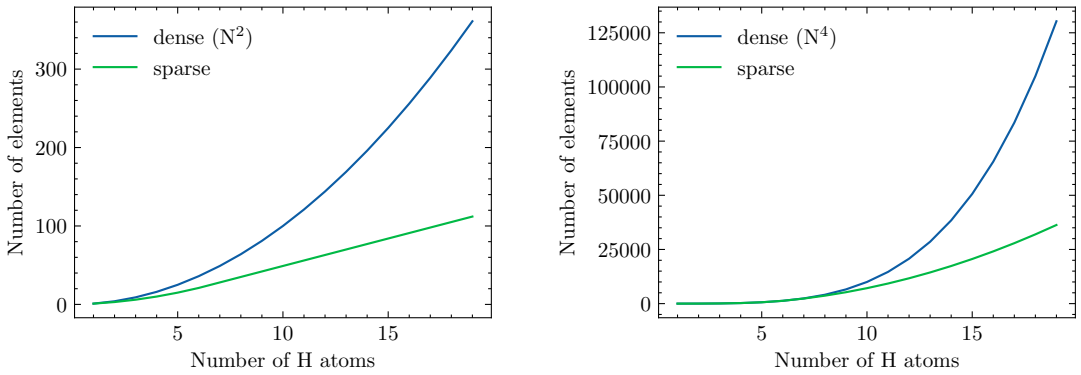
8 Sparsity in Electronic Structure Theory

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

8.1 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*.



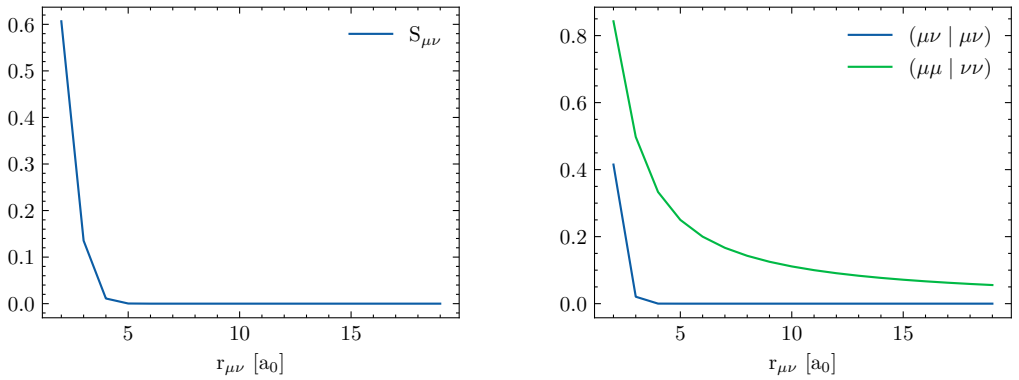
(a) (b)

Figure 2: (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.

8.1.1 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 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

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



(a) (b)

Figure 3: (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 $1/R$.

with the scaling factor κ

$$\kappa = e^{-\frac{\alpha\beta}{\alpha+\beta}|\mathbf{A}-\mathbf{B}|^2} \tag{3}$$

and the centre-of-charge coordinate P

$$\mathbf{P} = \frac{\alpha\mathbf{A} + \beta\mathbf{B}}{\alpha + \beta} \tag{4}$$

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

$$S_{AB} = \int \kappa\chi_P dr = \kappa \left(\frac{\pi}{\alpha+\beta}\right)^{3/2} \tag{5}$$

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

$$d_s = \sqrt{\alpha^{-1} ln \left[\left(\frac{\pi}{2\alpha}\right)^3 \epsilon^{-1/2}\right]} \tag{6}$$

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.

8.1.2 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

$$\begin{aligned} g_{ABCD} &= \int s(A,\alpha)s(B,\beta)\frac{1}{|\mathbf{r_1}-\mathbf{r_2}|}s(C,\gamma)s(D,\delta)dr \\ &= \int \kappa s(P,\alpha+\beta)\frac{1}{|\mathbf{r_1}-\mathbf{r_2}|}\lambda s(Q,\gamma+\delta) \end{aligned} \tag{7}$$

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

$$\mathbf{P} = \frac{\alpha\mathbf{A}+\beta\mathbf{B}}{\alpha+\beta}; \quad \mathbf{Q} = \frac{\gamma\mathbf{C}+\delta\mathbf{D}}{\gamma+\delta} \tag{8}$$

$$\kappa = e^{-p|\mathbf{A}-\mathbf{B}|^2}; \quad \lambda = e^{-q|\mathbf{C}-\mathbf{D}|^2} \tag{9}$$

$$p = \frac{\alpha\beta}{\alpha+\beta}; \quad q = \frac{\gamma\delta}{\gamma+\delta} \tag{10}$$

The coulomb integrals can then be evaluated as

$$g_{ABCD} = \sqrt{\frac{4\eta}{\pi}}S_{AB}S_{CD}F_0\left(\eta\left|\mathbf{P}-\mathbf{Q}\right|^2\right) \tag{11}$$

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

$$\eta = \frac{pq}{p+q} \tag{12}$$

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

$$\begin{aligned} F_n(x) &\leq \frac{1}{2n+1} \quad \text{for small } x \\ F_n(x) &\leq \frac{(2n-1)!!}{2^{n+1}}\sqrt{\frac{\pi}{x^{2n+1}}} \quad \text{for large } x \end{aligned} \tag{13}$$

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

$$g_{ABCD} \leq min\left\{\sqrt{\frac{4\eta}{\pi}}S_{AB}S_{CD}, \frac{S_{AB}S_{CD}}{|\mathbf{P}-\mathbf{Q}|}\right\} \tag{14}$$

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

$$\sum_{abcd} c_{ab} g_{abcd} c_{cd} > 0 \tag{15}$$

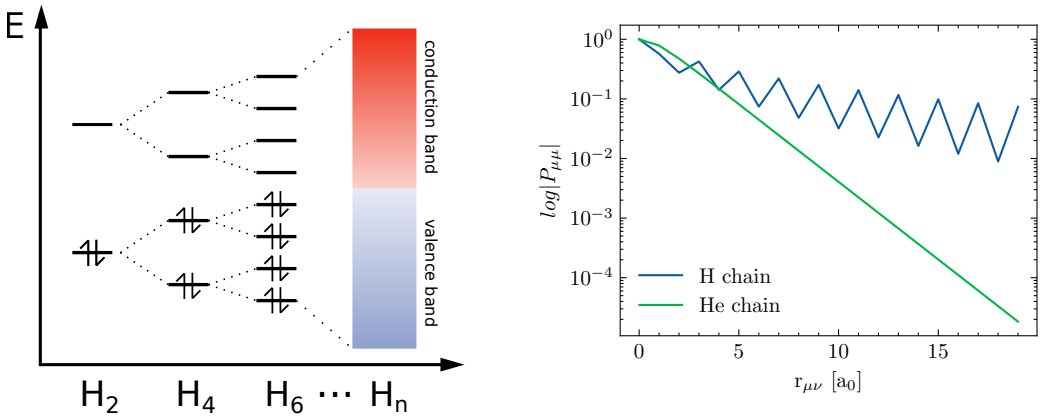
where c are one-electron orbital distrbutions. One can then apply the Schwarz inequality (Has1989) to obtain an upper bound expression for g

$$(\mu\nu \mid \sigma\lambda) \leq (\mu\nu \mid \mu\nu)\,(\sigma\lambda \mid \sigma\lambda) = Q_{\mu\nu} Q_{\sigma\lambda} \tag{16}$$

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.

8.2 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)

8.3 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

$$K_{\mu\nu} = (\mu\sigma \mid \nu\lambda)\, P_{\lambda\sigma} \tag{17}$$

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.

8.4 Rank Sparsity

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

$$\mathbf{A} = \mathbf{B}\mathbf{B}^T \tag{18}$$

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

$$(\mu\nu \mid \sigma\lambda) = A_{(\mu\nu)(\sigma\lambda)} = B_{(\mu\nu)X} B_{(\sigma\lambda)X} \tag{19}$$

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

$$(ia \mid jb) = C_{\mu i} C_{\sigma a} B_{\mu\sigma X} B_{X\nu\lambda} C_{\nu j} C_{\lambda b} = B_{iaX} B_{Xjb} \tag{20}$$

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.

9 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).

9.1 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

$$(\mu\nu \mid \sigma\lambda) = \int\int \frac{\rho_{\mu\nu}(\mathbf{r_1})\rho_{\sigma\lambda}(\mathbf{r_2})}{\mathbf{r_1} - \mathbf{r_2}} d\mathbf{r_1} d\mathbf{r_2} \tag{21}$$

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

$$\rho_{\mu\nu}(\mathbf{r}) = C_{P\mu\nu}\chi_P(\mathbf{r}) + \Delta\rho_{\mu\nu} \tag{22}$$

Or in the chemist's notation:

$$|\mu\nu) = C_{P\mu\nu}\,|P) + |\epsilon_{\mu\nu}) = |\tilde{\mu\nu}) + |\epsilon_{\mu\nu}) \tag{23}$$

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

$$\begin{aligned}(\mu\nu \mid \sigma\lambda) &= (\widetilde{\mu\nu}|\widetilde{\sigma\lambda}) + \underbrace{(\widetilde{\mu\nu}|\epsilon_{\sigma\lambda}) + (\epsilon_{\mu\nu}|\widetilde{\sigma\lambda})}_{\text{first order}} + \underbrace{(\epsilon_{\mu\nu} \mid \epsilon_{\sigma\lambda})}_{\text{second order}} \\ &= (\widetilde{\mu\nu}|\widetilde{\sigma\lambda}) + \epsilon_J^{(1)} + \epsilon_J^{(2)}\end{aligned} \tag{24}$$

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

$$\frac{\partial}{\partial C^P_{\mu\nu}}\left(\mu\nu - \widetilde{\mu\nu} \mid \sigma\lambda - \widetilde{\sigma\lambda}\right) = 0 \tag{25}$$

which then yields a set of linear equations

$$(\mu\nu \mid P) - \sum_Q C^Q_{\mu\nu}\,(Q \mid P) = 0 \tag{26}$$

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 $\{P\}$

$$(\mu\nu \mid P) = \int\int \chi_\mu(\mathbf{r_1})\chi_\mu(\mathbf{r_1})\frac{1}{\mathbf{r_1} - \mathbf{r_2}}\chi_P(\mathbf{r_2})d\mathbf{r_1}d\mathbf{r_2} \tag{27}$$

$$(P \mid Q) = \int\int \chi_P(\mathbf{r_1})\frac{1}{\mathbf{r_1} - \mathbf{r_2}}\chi_Q(\mathbf{r_2})d\mathbf{r_1}d\mathbf{r_2} \tag{28}$$

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

$$(\mu\nu \mid \sigma\lambda) \approx (\mu\nu \mid P)\,(P \mid Q)^{-1}\,(Q \mid \sigma\lambda) \tag{29}$$

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

9.2 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

$$|(\mu\nu \mid P)| \leq |(\mu\nu \mid \mu\nu)|^{1/2} \, |(P \mid P)|^{1/2} \tag{30}$$

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 $|P)$

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

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.

9.3 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

$$(\mu\nu \mid \sigma\lambda) \approx (\widetilde{\mu\nu}|\sigma\lambda) + (\mu\nu|\widetilde{\sigma\lambda}) - (\widetilde{\mu\nu}|\widetilde{\sigma\lambda}) \tag{31}$$

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 1: Expressions for the operator g in different local metrics.

9.4 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

$$B^P_{\mu\nu} - C^Q_{\mu\nu} M_{QP} = 0 \tag{32}$$

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

$$B^P_{\mu\nu} = \int\int \chi_\mu(\mathbf{r_1})\chi_\nu(\mathbf{r_1})g(\mathbf{r_1},\mathbf{r_2})\chi_P(\mathbf{r_2})d\mathbf{r_1}d\mathbf{r_2} \tag{33}$$

$$M_{PQ} = \int\int \chi_P(\mathbf{r_1})g(\mathbf{r_1},\mathbf{r_2})\chi_Q(\mathbf{r_2})d\mathbf{r_1}d\mathbf{r_2} \tag{34}$$

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

$$C^P_{\mu\nu} = S^{-1}_{PQ}(\mu\nu Q) \tag{35}$$

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. ...

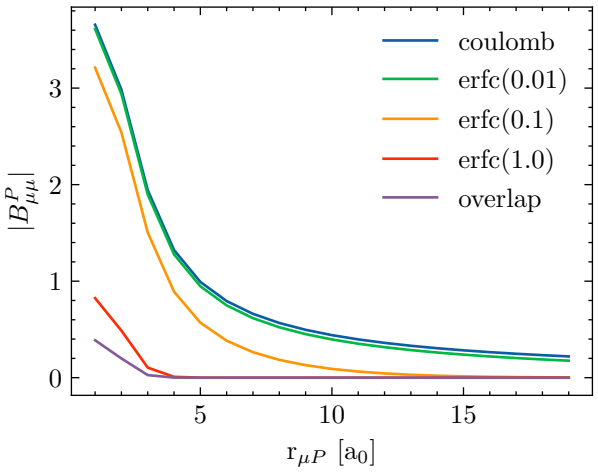


Figure 5: Some caption

9.5 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.

9.5.1 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

$$|\widetilde{\mu\nu}) = \sum_{Q \cup A_\mu} (P \mid Q)^{-1}_{A_\mu} (\mu\nu \mid Q) \tag{36}$$

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

$$(P \mid Q)^{-1}_X = B_X \left((P \mid Q)_D + B_X (P \mid Q)_{OD} B_X \right)^{-1} B_X \tag{37}$$

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.

9.5.2 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

$$(P \mid \mu\nu) = \sum_{Q \in A \cup B} (P \mid Q) \, C^Q_{\mu\nu} \quad \forall P \in A \cup B \tag{38}$$

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).

9.5.3 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)

$$(\mu i \mid P) = \sum_{Q \in [i]_{fit}} (P \mid Q)\, C^Q_{\mu i} \tag{39}$$

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 (refs).

9.6 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

$$\sum_Q (P \mid Q)\, C^Q_{\mu\nu} = (P \mid \mu\nu) \tag{40}$$

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.

9.6.1 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

$$\left| \sum_R S^{-1}_{PR}(R\mu\nu) \right| > T \tag{41}$$

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

$$f(Q_{\mu\nu}, P_{\mu\nu}) < R \tag{42}$$

with

$$f(A, B) = \frac{\alpha\beta}{\alpha + \beta} |\mathbf{A} - \mathbf{B}|^2 \tag{43}$$

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

$$\sum_P (Q_{\mu\nu} \mid P_{\mu\nu})\, C^P_{\mu\nu} = (Q_{\mu\nu} \mid \mu\nu) \tag{44}$$

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.

9.7 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

$$\Delta E_{HF} = E_{HF} - E_{DF-HF} \tag{45}$$

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

$$(\mu\nu \mid \sigma\lambda) = L_{\mu\nu}^X L_{\sigma\lambda}^X \tag{46}$$

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).

10 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

$$\mathbf{FC} = \mathbf{SC}\epsilon \tag{47}$$

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

$$L_{\mu\underline{i}} = U_{\underline{i}i}C_{\mu i} \tag{48}$$

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 $\mathbf{U}$. We will now go over some examples.

10.1 Local Molecular Orbitals

Some words about it

10.2 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

$$\zeta_{FB}(\chi) = \sum_i \langle\chi_i|\,\mathbf{r}\,|\chi i\rangle^2 \tag{49}$$

$$\zeta_{ER}(\chi) = \sum_i (\chi_i\chi_i \mid \chi_i\chi_i) \tag{50}$$

$$\zeta_{FB}(\chi) = \sum_i \sum_A \langle\chi_i|\,\mathbf{P}_A\,|\chi i\rangle^2 \tag{51}$$

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

10.3 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)

$$\hat{P} = |\Psi_i\rangle \langle\Psi_i| = |\chi_\mu\rangle C_{\mu i} C_{\nu i} \langle\chi_\nu| \tag{52}$$

$$\hat{Q} = |\Psi_a\rangle \langle\Psi_a| = |\chi_\mu\rangle C_{\mu a} C_{\nu a} \langle\chi_\nu| \tag{53}$$

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

$$\hat{P}\,|\chi_{\mu'}\rangle = \textstyle\sum_\mu |\chi_\mu\rangle\, P_{\mu\nu} S_{\nu\mu'} = L_{\mu I}\,|\chi_\mu\rangle \tag{54}$$

$$\hat{Q}\,|\chi_{\mu'}\rangle = \textstyle\sum_\mu |\chi_\mu\rangle\, Q_{\mu\nu} S_{\nu\mu'} = L_{\mu A}\,|\chi_\mu\rangle \tag{55}$$

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. CMOs are transformed to PAOs by the relationship

$$|\chi_I\rangle = (\mathbf{SC})_{Ii}\,|\Psi_i\rangle \tag{56}$$

$$|\chi_A\rangle = (\mathbf{SC})_{Aa}\,|\Psi_a\rangle \tag{57}$$

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 (1).

PAOs also arise in the context of AO-MP2, when rearranging the canoncial equations into an AO basis, as will be shown later.

10.4 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?

10.5 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.

$$\mathbf{P} = \mathbf{L}\mathbf{L}^{\mathbf{T}} \tag{58}$$

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

$$U_{i\underline{i}} = C_{\mu i} S_{\mu\nu} L_{\nu\underline{i}} \tag{59}$$

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.

10.6 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

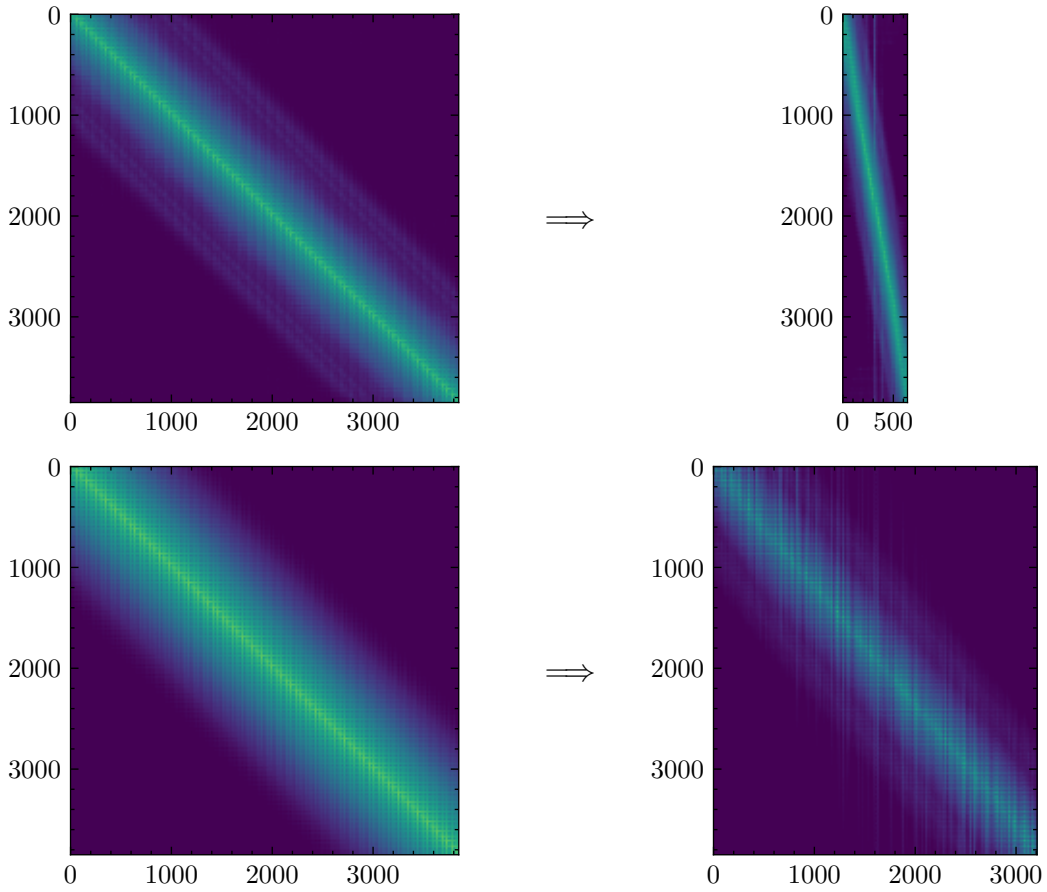


Figure 6: Something

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 $\hat{n}$

$$\hat{n} \left|\Theta_i\right\rangle = n_i \left|\Theta_i\right\rangle \tag{60}$$

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)

$$D_{ab} = \frac{1}{2} sum_{cij} \frac{K_{ij}^{cb} K_{ij}^{ca}}{\epsilon_{ij}^{ab} \epsilon_{ij}^{ca}} \tag{61}$$

with

$$K_{ij}^{ab} = 2\,(ia \mid jb) - (ib \mid ja) \tag{62}$$

$$\epsilon_{ij}^{ab} = \epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b \tag{63}$$

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

$$\mathbf{A_{CIS}} R_n = \omega_n R_n \tag{64}$$

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 $\mathbf{R}_n$

$$\mathbf{R} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\dagger} \tag{65}$$

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

$$\left|\overline{\Psi}_k^{occ}\right\rangle = U_{ki} \left|\Psi_i^{occ}\right\rangle \tag{66}$$

$$\left|\overline{\Psi}_k^{vir}\right\rangle = V_{ka} \left|\Psi_a^{vir}\right\rangle \tag{67}$$

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

10.7 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 apprximations are necessary. In LMO 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 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

$$E_{CEPA} = \sum_{ij} e_{ij} \tag{68}$$

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.

Include this: Pair apprximations Electron pairs are first grouped by distance r between occuied LMOs: strong pairs ($1 < r \leq 8a_0$), weak pairs ($8 < r \leq 15a_0$) and very distant pairs ($15 < r$). Very distant: lower level of theory, or neglected Talk about how they are transformed? (4)

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 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 $[ij]$ $= [i] \bigcup [j]$. 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

$$\mathbf{D}^{ij} = \frac{1}{1+\delta_{ij}} \left(\tilde{\mathbf{t}}^{ij} \mathbf{t}^{ij} + tilde \mathbf{t}^{ij} \mathbf{t}^{ij\dagger} \right) \tag{69}$$

with

$$\tilde{\mathbf{t}}^{ij}_{ab} = 2\mathbf{t}^{ij}_{ab} - \mathbf{t}^{ji}_{ab} \tag{70}$$

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

$$\mathbf{D}^{ij} \mathbf{Q}^{ij} = n^{ij} \mathbf{Q}^{ij} \tag{71}$$

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

$$t^{ab}_{ij} = \frac{(ia \mid jb)}{\epsilon_a + \epsilon_b - f_{ii} - f_{jj}} \tag{72}$$

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

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

$$\mathbf{D}^{ii} = 4\mathbf{t}^{ii}\mathbf{t}_{\mathbf{ii}} \tag{73}$$

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

$$\mathbf{t}^{ii}\mathbf{Q}^{ii} = t^{ii}\mathbf{Q}^{ii} \tag{74}$$

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).

11 Local Density Fitting

12 Low-Scaling Hartree Fock Methods

Talk about bottle necks DF, CFMM, LinK, CADF, PARI, ...

12.1 Density Purification

13 Local Correlation Methods I: MP2

13.1 Local Formulation of MP2

13.2 Orbital Invariant MP2

13.3 AO-MP2

13.4 Scaling: Contraction

13.5 Scaling: MO Transformation

We will now present how the concepts above are used. Merge them?

Orbital invariant MP2: Hylleraas functional General Local MP2: Laplace transform AO-MP2

scaling cons

14 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 assciated 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.

14.1 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).

14.1.1 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*

$$\frac{1}{\epsilon_a + \epsilon_b - \epsilon_i - \epsilon_j} = \int_0^\infty e^{-(\epsilon_a+\epsilon_b-\epsilon i-\epsilon_j)t} dt \tag{75}$$

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

$$\frac{1}{\epsilon_a + \epsilon_b - \epsilon_i - \epsilon_j} \approx \sum_\alpha^n w^{(\alpha)} e^{-(\epsilon_a+\epsilon_b-\epsilon i-\epsilon_j)t^{(\alpha)}} \tag{76}$$

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 ... .

14.1.2 AO MP2 Equations

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

$$\begin{aligned} E_{MP2} &= -\sum_{iajb} \frac{(ia \mid jb)\left[2\,(ia \mid ib) - (ib \mid ja)\right]}{\epsilon_a + \epsilon_b - \epsilon_i - \epsilon_j} \\ &\approx -\sum_{\alpha}^{n} \sum_{iajb} (ia \mid jb)\left[2\,(ia \mid ib) - (ib \mid ja)\right] w^{(\alpha)} e^{-(\epsilon_a+\epsilon_b-\epsilon i-\epsilon_j)t^{(\alpha)}} \end{aligned} \tag{77}$$

We can then proceed to factor out the coefficient matrices

$$\begin{aligned} &-\sum_{\alpha}^{n} \sum_{iajb} (ia \mid jb)\left[2\,(ia \mid ib) - (ib \mid ja)\right] w^{(\alpha)} e^{-(\epsilon_a+\epsilon_b-\epsilon i-\epsilon_j)t^{(\alpha)}} \\ &= -\sum_{\alpha}^{n} \sum_{iajb} \sum_{\substack{\mu\nu\lambda\sigma \\ \mu'\nu'\lambda'\sigma'}} w^{(\alpha)} e^{-(\epsilon_a+\epsilon_b-\epsilon i-\epsilon_j)t^{(\alpha)}} C_{\mu' i} C_{\sigma' a}\,(\mu'\sigma' \mid \nu'\lambda')\, C_{\nu' j} C_{\lambda' b} \\ &\qquad \times \left\{C_{\mu i} C_{\sigma a}\left[2\,(\mu\sigma \mid \nu\lambda) - (\mu\lambda \mid \nu\sigma)\right] C_{\nu j} C_{\lambda b}\right\} \\ &= -\sum_{\alpha}^{n} \sum_{\substack{\mu\nu\lambda\sigma \\ \mu'\nu'\lambda'\sigma'}} \underline{P}^{(\alpha)}_{\mu\mu'} \overline{P}^{(\alpha)}_{\sigma\sigma'}\,(\mu'\sigma' \mid \nu'\lambda')\, \underline{P}^{(\alpha)}_{\nu\nu'} \overline{P}^{(\alpha)}_{\lambda\lambda'}\left[2\,(\mu\sigma \mid \nu\lambda) - (\mu\lambda \mid \nu\sigma)\right] \end{aligned} \tag{78}$$

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

$$\begin{aligned} \underline{P}^{(\alpha)}_{\mu\mu'} &= \sum_{i} C_{\mu i} e^{0.25 ln(w^{(\alpha)}) + \epsilon_i t^{(\alpha)}} C_{\mu' i} \\ \overline{P}^{(\alpha)}_{\mu\mu'} &= \sum_{i} C_{\sigma a} e^{0.25 ln(w^{(\alpha)}) - \epsilon_a t^{(\alpha)}} C_{\sigma' i} \end{aligned} \tag{79}$$

Introducing the *pseudo-AO* transformed electron integrals

$$\left(\underline{\mu}\overline{\sigma} \mid \underline{\nu}\overline{\lambda}\right)^{(\alpha)} = \underline{P}^{(\alpha)}_{\mu\mu'} \overline{P}^{(\alpha)}_{\sigma\sigma'}\,(\mu'\sigma' \mid \nu'\lambda')\, \underline{P}^{(\alpha)}_{\nu\nu'} \overline{P}^{(\alpha)}_{\lambda\lambda'} \tag{80}$$

the energy expression for AO-MP2 reads

$$E_{AO-MP2} = -\sum_{\alpha}^{n} \sum_{\mu\nu\lambda\sigma} \left(\underline{\mu}\overline{\sigma} \mid \underline{\nu}\overline{\lambda}\right)^{(\alpha)} \left[2\,(\mu\sigma \mid \nu\lambda) - (\mu\lambda \mid \nu\sigma)\right] \tag{81}$$

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

$$\underline{\mathbf{P}}^{(\alpha)} \mathbf{S} \overline{\mathbf{P}}^{(\alpha)} = \mathbf{0} \tag{82}$$

$$\underline{\mathbf{P}}^{(\alpha)} \mathbf{S} + \overline{\mathbf{P}}^{(\alpha)} \mathbf{S} = \mathbf{I}_{exp} \tag{83}$$

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

$$Tr[\mathbf{I}_{exp}] = \sum_{i} e^{0.25 ln(w^{(\alpha)}) + \epsilon_i t^{(\alpha)}} + \sum_{a} e^{0.25 ln(w^{(\alpha)}) - \epsilon_a t^{(\alpha)}} \tag{84}$$

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

14.1.3 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

$$\begin{aligned} \mu &\overset{P}{\longleftrightarrow} \mu' \overset{S}{\longleftrightarrow} \sigma' \overset{P}{\longleftrightarrow} \sigma \\ \nu &\overset{P}{\longleftrightarrow} \nu' \overset{S}{\longleftrightarrow} \lambda' \overset{P}{\longleftrightarrow} \lambda \end{aligned} \tag{85}$$

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

$$\begin{aligned} Q_{\mu\nu} &= \mid (\mu\nu \mid \mu\nu) \mid^{1/2} \\ X_{\mu\nu} &= \mid \left(\underline{\mu}\nu \mid \underline{\mu}\nu\right) \mid^{1/2} \\ Y_{\mu\nu} &= \mid (\mu\overline{\nu} \mid \mu\overline{\nu}) \mid^{1/2} \\ Z_{\mu\nu} &= min\left(\sum_{\sigma} A_{\mu\sigma} \mid \overline{P}_{\sigma\nu} \mid; \sum_{\sigma} B_{\mu\sigma} \mid \underline{P}_{\sigma\nu} \mid\right) \end{aligned} \tag{86}$$

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

$$(\mu'\sigma' \mid \nu'\lambda') \leq Q_{\mu'\sigma'} Q_{\nu'\lambda'} \tag{87}$$

$$\left(\underline{\mu}\sigma' \mid \nu'\lambda'\right) \leq X_{\mu'\sigma'} Q_{\nu'\lambda'} \tag{88}$$

$$\left(\underline{\mu}\overline{\sigma} \mid \underline{\nu}\overline{\lambda}\right) \leq Z_{\mu\sigma} Z_{\nu\lambda} \tag{89}$$

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

14.1.4 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

$$(\mu\sigma \mid \nu\lambda) = (\mu\sigma| \frac{1}{\mathbf{r}_{12}} |\nu\lambda) \leq \frac{1}{R} \left| \sum_{n=0}^{\infty} \frac{(\mu\sigma| \left(\mathbf{r}_1 - \mathbf{r}_2\right)^n |\nu\lambda)}{R^n} \right| \tag{90}$$

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

$$M_{\mu\sigma}^{(n)} = \int \chi_\mu(r_1)\mathbf{r}_1^n\chi_\sigma(r_1)dr \tag{91}$$

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

$$\begin{aligned}(\mu\sigma \mid \nu\lambda) \leq\ & R^{-1}\left|M_{\mu\sigma}^{(0)}M_{\nu\lambda}^{(0)}\right| + R^{-2}\left|M_{\mu\sigma}^{(1)}M_{\nu\lambda}^{(0)} - M_{\mu\sigma}^{(0)}M_{\nu\lambda}^{(1)}\right| \\ & + R^{-3}\left|M_{\mu\sigma}^{(2)}M_{\nu\lambda}^{(0)} - 2M_{\mu\sigma}^{(1)}M_{\nu\lambda}^{(1)} + M_{\mu\sigma}^{(0)}M_{\nu\lambda}^{(2)}\right| \\ & + R^{-4}\left|M_{\mu\sigma}^{(3)}M_{\nu\lambda}^{(0)} - 3M_{\mu\sigma}^{(2)}M_{\nu\lambda}^{(1)} + 3M_{\mu\sigma}^{(1)}M_{\nu\lambda}^{(2)} - M_{\mu\sigma}^{(0)}M_{\nu\lambda}^{(3)}\right| \\ & + \mathcal{O}(R^{-5})\end{aligned} \tag{92}$$

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

$$\begin{aligned}\left(\underline{\mu}\overline{\sigma} \mid \underline{\nu}\overline{\lambda}\right) \leq\ & R^{-3}\left|-2M_{\underline{\mu}\overline{\sigma}}^{(1)}M_{\underline{\nu}\overline{\lambda}}^{(1)}\right| \\ & + R^{-4}\left|-3M_{\underline{\mu}\overline{\sigma}}^{(2)}M_{\underline{\nu}\overline{\lambda}}^{(1)} + 3M_{\underline{\mu}\overline{\sigma}}^{(1)}M_{\underline{\nu}\overline{\lambda}}^{(2)}\right| + \mathcal{O} \\ & + \mathcal{O}(R^{-5})\end{aligned} \tag{93}$$

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

$$\left(\underline{P}_{\mu\sigma}S_{\sigma\lambda}\overline{P}_{\lambda\mu}\right) \geq \epsilon \tag{94}$$

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

$$r_{\mu\nu} - R_\mu - R_\nu \geq r_0 \tag{95}$$

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

14.1.5 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:

$$\underline{\mathbf{P}}^{(\alpha)} = \underline{\mathbf{L}}^{(\alpha)}\underline{\mathbf{L}}^{(\alpha)T} \tag{96}$$

$$\overline{\mathbf{P}}^{(\alpha)} = \overline{\mathbf{L}}^{(\alpha)}\overline{\mathbf{L}}^{(\alpha)T} \tag{97}$$

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 $\underline{\mathbf{P}}^{(\alpha)}$:

$$\underline{\mathbf{P}}^{(\alpha)}_{orth} = \mathbf{S}^{1/2}\underline{\mathbf{P}}^{(\alpha)}\mathbf{S}^{1/2} \tag{98}$$

The pseudo-MO coefficients are then obtained as

$$\underline{L}^{(\alpha)} = \mathbf{S}^{-1/2}\underline{\mathbf{L}}^{(\alpha)}_{orth} \tag{99}$$

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

$$E_{CDD-AO-MP2} = -\sum_{\alpha}^{n}\sum_{\underline{i}\overline{a}\underline{j}\overline{b}} \left(\underline{i}\overline{a} \mid \underline{j}\overline{b}\right)^{(\alpha)} \left[2\left(\underline{i}\overline{a} \mid \underline{j}\overline{b}\right)^{(\alpha)} - \left(\underline{i}\overline{b} \mid \underline{j}\overline{a}\right)^{(\alpha)}\right] \tag{100}$$

whith the pseudo-MO integrals

$$\left(\underline{i}\overline{a} \mid \underline{j}\overline{b}\right)^{(\alpha)} = \underline{L}^{(\alpha)}_{\mu\underline{i}}\overline{L}^{(\alpha)}_{\sigma\overline{a}} \left(\mu\sigma \mid \nu\lambda\right) \underline{L}^{(\alpha)}_{\nu\underline{j}}\overline{L}^{(\alpha)}_{\lambda\overline{b}} \tag{101}$$

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.

14.1.6 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

$$\left(X \mid \underline{\mu}\overline{\nu}\right)^{(\alpha)} = \left(X \mid \mu'\nu'\right) \underline{P}^{(\alpha)}_{\mu\mu'}\overline{P}^{(\alpha)}_{\nu\nu'} \tag{102}$$

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

14.1.7 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)

$$E_{AO-DF-SOS-MP2} = -c_{os} \sum_{\alpha=1}^{nlap} \sum_{\mu\nu\sigma\lambda} \left(\underline{\nu\overline{\sigma}} \mid X\right)^{(\alpha)} \left(X \mid Y\right)^{-1} \left(Y \mid \underline{\nu\overline{\lambda}}\right)^{(\alpha)} \left(\mu\sigma \mid X'\right) \left(X' \mid Y'\right) \left(Y' \mid \nu\lambda\right) \tag{103}$$

Introducing the intermediates

$$Z_{XY}^{(\alpha)} = \left(X \mid \underline{\mu\overline{\sigma}}\right)^{(\alpha)} \left(\mu\sigma \mid Y\right) \tag{104}$$

$$\tilde{Z}_{XY}^{(\alpha)} = \left(X \mid R\right)^{-1} Z_{RX}^{(\alpha)} \tag{105}$$

we arrive at a very compact expression

$$E_{AO-DF-SOS-MP2} = -c_{os} \sum_{\alpha=1}^{nlap} \sum_{XY} \tilde{Z}_{XY}^{(\alpha)} \tilde{Z}_{YX}^{(\alpha)} \tag{106}$$

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).

14.2 SVO-MP2 flavours

Among the eraliest

Use a different approach than Laplace: Hylleraas Functional

Sparsity relation ship [ij] ¡-¿ [ab]

14.3 NTO-MP2 ?

15 Low-Scaling Correlated Excited State Methods

Local CC2, PNO-ADC, Mester-ADC, Mester-CC2, NTO-CC2, CornFlex

15.1 CC2

15.2 ADC

Part III
Benchmarking: Timings and

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

Part IV
Annex

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