Laboratoire de Physique & Chimie des Nano-Objets

SOMOs Identification

Romuald Poteau, romuald.poteau@univ-tlse3.fr, LPCNO

April 28, 2025

1 Introduction

To identify singly occupied molecular orbitals (**SOMOs**) in open-shell systems, we employed two complementary schemes: (i) orbital projection analysis, and (ii) cosine similarity mapping. Both approaches compare the sets of α and β molecular orbitals obtained from unrestricted calculations. SOMOs are expected to show strong projections onto the β virtual space and minimal overlap with the β occupied space — a characteristic signature of magnetic (unpaired) molecular orbitals in open-shell systems. The projection scheme evaluates the squared overlap of each α orbital with the β orbital space using the atomic orbital overlap matrix, allowing for a quantitative decomposition of each α orbital across the β manifold. It is completed by the standard projection of MOs on a reduced space, namely the projection of α occupied MOs on two sub-spaces, defined by β-occupied MOs and β-virtual MOs, respectively. In contrast, the cosine similarity approach measures the angular similarity between α and β orbitals based on their MO coefficients, identifying pairs of orbitals with nearly identical spatial character. Together, these methods help identify SOMOs as occupied α orbitals that lack a clear counterpart among the occupied β orbitals.

2 Similarity schemes

2.1 Projection of occupied α MOs onto the β orbital space

Given a Gaussian log file from an unrestricted DFT calculation, we extract the molecular orbital (MO) coefficients for both α and β orbitals (`pop=full` keyword), along with the AO overlap matrix $\boldsymbol{S}$ (`iop(3/33=1)` keyword) and the printing of the basis set (`gfprint gfinput` keywords). The analysis focuses on evaluating how each occupied α orbital projects onto the full space spanned by all β orbitals, which includes both occupied and virtual ones.

Let $\mathbf{\Phi}_i^\alpha \in \mathbb{R}^{1\times n_{\text{basis}}}$ be the coefficient vector of the i-th occupied α orbital, and let $\mathbf{\Phi}^\beta \in \mathbb{R}^{N\times n_{\text{basis}}}$ be the matrix of all β orbitals stored row-wise, where $N = n_\beta$ is the total number of β orbitals. The projection vector is computed as:

$$\mathbf{v}_i =< \phi_i^\alpha | \phi^\beta >= \mathbf{\Phi}_i^\alpha \cdot \boldsymbol{S} \cdot (\mathbf{\Phi}^\beta)^T \in \mathbb{R}^{1\times N}$$

The squared norm $\|\mathbf{v}_i\|^2$ gives the total overlap of the α orbital with the β space.

To differentiate between the contributions from occupied and virtual β orbitals, we split the projection:

$$\mathbf{v}_i^{\text{occ}} =< \phi_i^\alpha | \phi_{\text{occ}}^\beta >= \mathbf{\Phi}_i^\alpha \cdot \boldsymbol{S} \cdot (\mathbf{\Phi}_{\text{occ}}^\beta)^T$$
$$\mathbf{v}_i^{\text{virt}} =< \phi_i^\alpha | \phi_{\text{virt}}^\beta >= \mathbf{\Phi}_i^\alpha \cdot \boldsymbol{S} \cdot (\mathbf{\Phi}_{\text{virt}}^\beta)^T$$

We then compute:

- $\|\mathbf{v}_i^{\text{occ}}\|^2$ = projection of $|\phi_i^\alpha >$ onto occupied β orbitals
- $\|\mathbf{v}_i^{\text{virt}}\|^2$ = projection of $|\phi_i^\alpha >$ onto virtual β orbitals

The total projection norm is decomposed to analyze how concentrated or spread the projection is across β orbitals:

- The three largest values among the squared projections v_{ij}^2 are summed to compute **"Top 1 (%)"**, **"Top 2 (%)"** and **"Top 3 (%)"**. **Top 1 (%)** can also bee seen as a **dominance ratio**, *i.e.* the quantity is defined as the largest single squared projection divided by the total projection norm: $\max_j v_{ij}^2 / \|\mathbf{v}_i\|^2$
- The "β **MOs >15%**" column lists all β orbitals contributing more than the specified percentage to the squared projection norm, along with their contribution in the format $[j, p_j]$, where j is the β-MO index and p_j the percentage contribution. For the most important contribution, it is nothing else than the **dominance ratio**. It provides a direct quantitative decomposition of each α orbital onto the β orbital basis. Each entry explicitly identifies the β orbital(s) that significantly compose the corresponding α orbital, along with their respective percentage contributions

An orbital is flagged as a **SOMO candidate** if its projection onto the virtual β space exceeds 0.5 and its projection onto the occupied β space is below 0.5:

$$\|\mathbf{v}_i^{\text{virt}}\|^2 > 0.5 \quad \text{and} \quad \|\mathbf{v}_i^{\text{occ}}\|^2 < 0.5$$

This criterion is named **"SOMO P2v?"** in the output and in a saved spreadsheet file. In some cases, where the mixing of the projection onto occupied and virtual β MOs makes the identification not straightforward, a secondary, less robust, criterion has been defined. A SOMO candidate, named "**SOMO dom. β MO?**", is identified when the dominance ratio is associated to a virtual β MO.

2.2 Diagonalization of the projection of α occupied orbitals onto β sub-spaces

In order to further analyze the nature of singly occupied molecular orbitals (SOMOs) and their relation to the β spin manifold, a complementary diagonalization procedure was implemented.

Starting from the set of occupied α orbitals ϕ_i^α, two separate projections are constructed, namely the projection onto the occupied β orbitals ϕ_{occ}^β and the projection onto the virtual β orbitals ϕ_{virt}^β.

Given the atomic orbital overlap matrix $\boldsymbol{S}$, the rectangular projection matrices are defined as:

$$\boldsymbol{A}_{\text{occ}} = \boldsymbol{\Phi}_i^\alpha \cdot \boldsymbol{S} \cdot (\boldsymbol{\Phi}_{\text{occ}}^\beta)^T, \quad \boldsymbol{A}_{\text{virt}} = \boldsymbol{\Phi}_i^\alpha \cdot \boldsymbol{S} \cdot (\boldsymbol{\Phi}_{\text{virt}}^\beta)^T.$$

From these, the symmetric projection matrices are formed:

$$\boldsymbol{P}_{\text{occ}} = \boldsymbol{A}_{\text{occ}} \boldsymbol{A}_{\text{occ}}^T, \quad \boldsymbol{P}_{\text{virt}} = \boldsymbol{A}_{\text{virt}} \boldsymbol{A}_{\text{virt}}^T.$$

The matrices $\boldsymbol{P}_{\text{occ}}$ and $\boldsymbol{P}_{\text{virt}}$ are diagonalized to obtain their eigenvalues and eigenvectors.

The eigenvalues of $\boldsymbol{P}_{\text{occ}}$ quantify how strongly a linear combination of occupied α orbitals projects onto the occupied β space. Similarly, the eigenvalues of $\boldsymbol{P}_{\text{virt}}$ measure the projection onto the virtual β space. Eigenvectors with low eigenvalues for $\boldsymbol{P}_{\text{occ}}$ but significant projection onto β virtual orbitals are strong candidates for SOMOs.

2.3 Cosine similarity of MOs

The identification of singly occupied molecular orbitals (SOMOs) can also be achieved through the computation of the cosine similarity between pairs of molecular orbitals (MOs) derived from unrestricted spin density functional theory (DFT) calculations. Specifically, we computed similarities between α and β spin orbitals, taking into account the non-orthogonality of the basis set used in quantum chemical calculations.

Let us denote two molecular orbital coefficient vectors as $\boldsymbol{\Phi}_i^\alpha$ (for alpha-spin orbitals) and $\boldsymbol{\Phi}_j^\beta$ (for beta-spin orbitals). Each vector has dimensions corresponding to the number of basis functions used in the calculation, denoted by n_{basis}. Given the overlap matrix $\boldsymbol{S}$ (dimension $n_{\text{basis}} \times n_{\text{basis}}$), obtained from the quantum chemistry calculation, the scalar product between two coefficient vectors accounting for basis overlap is defined as:

$$\langle \phi_i^\alpha | \phi_j^\beta \rangle = \boldsymbol{\Phi}_i^{\alpha T} \boldsymbol{S} \boldsymbol{\Phi}_j^\beta \tag{1}$$

Thus, the cosine similarity between two molecular orbitals $\mathbf{\Phi}_i^\alpha$ and $\mathbf{\Phi}_j^\beta$ accounting for the basis overlap matrix, $\boldsymbol{S}$, is given by:

$$\text{cosine similarity}(\phi_i^\alpha, \phi_j^\beta) = \frac{\mathbf{\Phi}_i^{\alpha T} \boldsymbol{S} \mathbf{\Phi}_j^\beta}{\sqrt{\mathbf{\Phi}_i^{\alpha T} \boldsymbol{S} \mathbf{\Phi}_i^\alpha} \sqrt{\mathbf{\Phi}_j^{\beta T} \boldsymbol{S} \mathbf{\Phi}_j^\beta}} \tag{2}$$

The similarity matrix constructed from these cosine similarities was then used to optimally match α and β orbitals employing the Hungarian algorithm, ensuring maximal global similarity. Also known as the Kuhn–Munkres algorithm, it is a classic method used to solve the assignment problem: given a cost matrix, it finds the optimal one-to-one assignment (or matching) that minimizes (or maximizes) the total cost (or similarity).

Orbital pairs with high cosine similarity, particularly those involving occupied alpha-spin orbitals matched to virtual beta-spin orbitals (or vice versa), can also be identified as potential candidates for SOMOs. This method provides a robust and quantitatively precise approach to identifying SOMOs in unrestricted DFT calculations, facilitating detailed analyses of electronic structures in open-shell systems.

2.4 Brief discussion

The projection technique quantifies how much each α orbital overlaps with the entire β orbital space by computing the squared norm of the projection vector using the AO overlap matrix. This provides an absolute, physically meaningful measure of orbital mixing, especially relevant when analyzing partial spin contamination or magnetic character. In contrast, cosine similarity evaluates the angle between two orbital vectors, yielding a dimensionless similarity score between -1 and 1. It's more suited for comparing the shape of orbitals than their actual physical contribution to each other. While cosine similarity is useful for clustering and pattern recognition, the projection approach is generally more precise when it comes to quantifying actual contributions and mixing between spin orbitals, especially in systems with open-shell or near-degenerate character. As regards the diagonalization-based projection strategy, it allows the detection of SOMO candidates with negligible coupling to occupied β orbitals, provides a detailed inspection of how α occupied orbitals distribute onto the β manifold, and offers deeper insights of orbital reorganization effects in open-shell systems.

3 Examples

3.1 Formaldehyde (H_2CO) in its lowest triplet state

In an all-electron basis set, there are 9 occupied α MOs, ϕ_{occ}^α, and 7 occupied β MOs, ϕ_{occ}^β. As summarized in the previous section, `gSOMOs` provides several tools to find two two SOMOs among the nine ϕ_{occ}^α. Table 1 presents simplified projection data of occupied α orbitals onto β orbitals for the lowest triplet state, T_1, of formaldehyde. It is adapted from the dataframe created by the `project_occupied_alpha_onto_beta()` function. Orbitals identified as SOMO indicate significant projection onto virtual β orbitals and negligible projection onto occupied β orbitals.

α MO	Occ α	Energy (Ha)	P² $\beta_{\mathbf{virt}}$	P² $\beta_{\mathbf{occ}}$	β MO*	Occ β	SOMO?	β MOs >15%
9	O	-0.204	0.996	0.004	9	V	Y	9: 96.2%
8	O	-0.367	0.895	0.105	8	V	Y	8: 88.4%
7	O	-0.438	0.005	0.995	7	O	N	7: 97.5%
6	O	-0.487	0.000	1.000	6	O	N	6: 97.9%
5	O	-0.529	0.106	0.894	5	O	N	5: 89.4%
4	O	-0.672	0.002	0.998	4	O	N	4: 99.4%
3	O	-1.096	0.001	0.999	3	O	N	3: 99.6%
2	O	-10.238	0.000	1.000	2	O	N	2: 100.0%
1	O	-19.240	0.000	1.000	1	O	N	1: 100.0%

Table 1: Projection of α molecular orbitals onto β space for formaldehyde (H_2CO), highlighting SOMOs. **P²** $\boldsymbol{\beta_{\mathbf{virt}}}$ and **P²** $\boldsymbol{\beta_{\mathbf{occ}}}$ are $\|\mathbf{v}_i^{\mathrm{occ}}\|^2$ and $\|\mathbf{v}_i^{\mathrm{virt}}\|^2$, respectively (see text).

A heatmap can be generated to visualize the main projection contributions of α molecular orbitals onto β orbitals. The color intensity reflects the percentage contribution of each β orbital to the total projection norm of a given α orbital. Only above 15% were retained for clarity. Red dashed lines indicate the HOMO–LUMO frontier for both spin channels (Figure 1a). This analysis is very close to the heatmap generated after the cosine similarity between α and β MOs (Figure 1b)

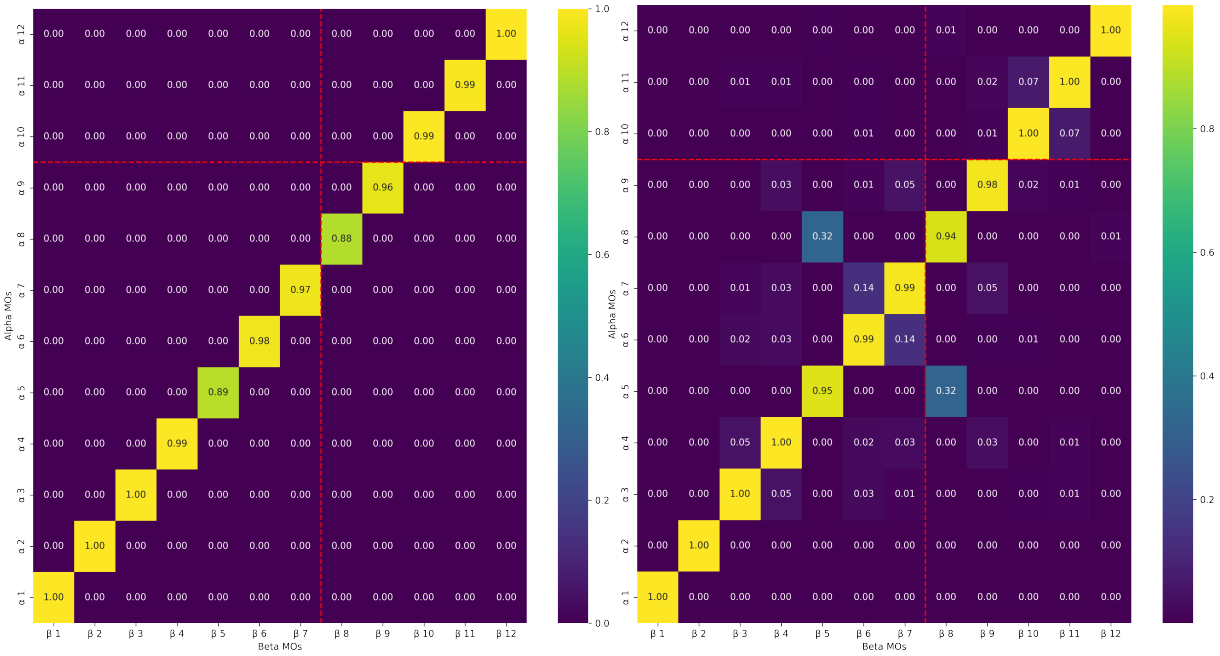

Figure 1: similarity of α and β MOs of the first triplet state of H_2CO around the HOMO-LUMO frontier. (a) Projection of α MOs onto the full space spanned by all β orbitals (only contributions above 20% were retained for clarity); (b) Cosine similarity. Dashed lines mark the HOMO/LUMO boundaries for α (horizontal) and β (vertical) spin orbitals.

And finally, and still in a projection scheme, the analysis of the eigenvalues of $\boldsymbol{P}_{\text{occ}}$ and $\boldsymbol{P}_{\text{virt}}$ agree with the previous similarity analysis schemes. Figure 2 reveals that two occupied α MOs do not project on the $\phi^{\beta}_{\text{occ}}$ subspace, whereas two MOs - hopefully the same - project on the $\phi^{\beta}_{\text{virt}}$ subspace. The MO analysis performed on the two eigenvectors of $\boldsymbol{P}_{\text{occ}}$ and the two eigenvectors of $\boldsymbol{P}_{\text{virt}}$ and reported in Table 2 show without doubt that the counterpart of the 8th and 9th α MOs are the 8th and 9th unoccupied β MOs.

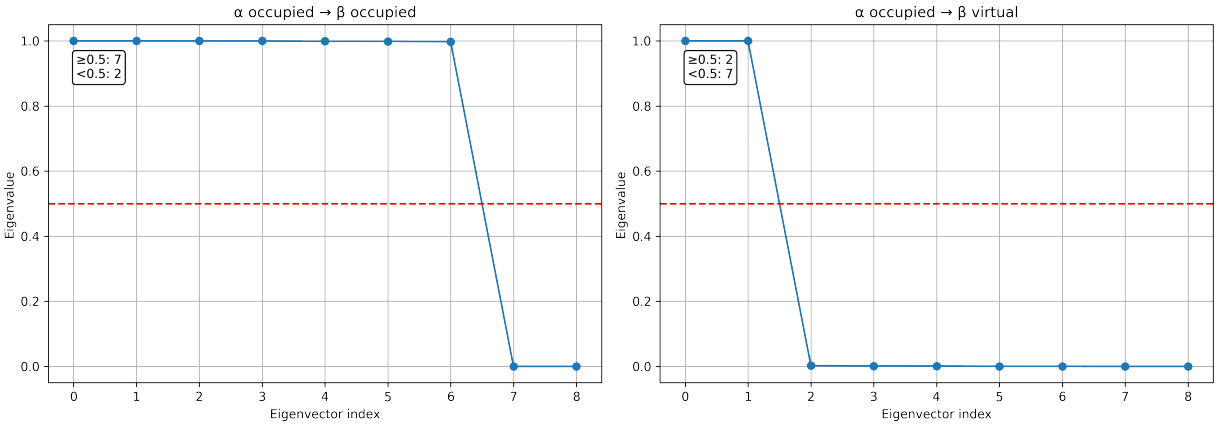

Figure 2: Eigenvalues of $\boldsymbol{P}_{\text{occ}}$ and $\boldsymbol{P}_{\text{virt}}$ for H_2CO.

SOMO Candidate #1	SOMO Candidate #2
α occupied contributions:	α occupied contributions:
• α 9 (99.5%)	• α 8 (89.4%)
β virtual projections:	β virtual projections:
• β 9 (96.4%)	• β 8 (98.4%)

Table 2: SOMO candidates for H_2CO, as given by the analyzis of the eigenvectors of $\boldsymbol{P}_{\text{occ}}$ and $\boldsymbol{P}_{\text{virt}}$.

3.2 Iron complex

In some cases, the identification of SOMOs by visual inspection can become very challenging, making a tool like gSOMOs particularly valuable. The dispersion of SOMO contributions over multiple α and β orbitals can arise from near-degeneracy effects in the frontier orbital region, possibly amplified by minor symmetry breaking or delocalization effects inherent to DFT. This behavior is typical for open-shell transition metal complexes with dense manifolds of occupied and virtual states This is the case for the quintet state of the iron complex shown in Figure 3. The cosine similarity method correctly identifies three SOMOs but fails for the fourth, which only imperfectly projects onto the virtual β space - whereas it correctly identifies MOs 169, 92 and 194, the fourth SOMOs is identified as MO 164. In contrast, the projection scheme performs well, especially through the diagonalization of $\boldsymbol{P}_{\text{occ}}$ and $\boldsymbol{P}_{\text{virt}}$ and the associated analysis. Figure 4 shows that four null eigenvalues are found after the diagonalization of $\boldsymbol{P}_{\text{occ}}$, associated to four eigenvalues close to 1 after the diagonalization of $\boldsymbol{P}_{\text{virt}}$. The decomposition of the SOMO candidates is reported in Table 3. The four SOMOs are reported in Figure 5. The projection scheme isolates clear contributions for each SOMO candidate, mainly involving a few α-occupied and β-virtual orbitals. $SOMO_1$ shows a mixed α-character (orbitals 187 and 164) projecting mainly onto β-orbital 194. $SOMO_2$ has a more distributed α-character but projects onto β-orbitals 192 and 193. $SOMO_3$ is dominated by orbital 186, projecting onto β-orbital 198, while $SOMO_4$ is mainly from orbital 168, with projections onto β-orbitals 193 and 192. $SOMO_1$ clearly shows why the cosine similarity misidentified the dominant α_{occ}-β_{virt} pair: although orbital 187 contributes most (44.2%), a significant mixing with orbital 164 (27.3%) leads to an overemphasis on the 164 $\rightarrow$ 194 projection. This explains why the cosine similarity method failed here, while the projection-based approach provides a more reliable identification. Despite some mixing, the dominant contributions are clearly identified, confirming the robustness of the projection analysis.

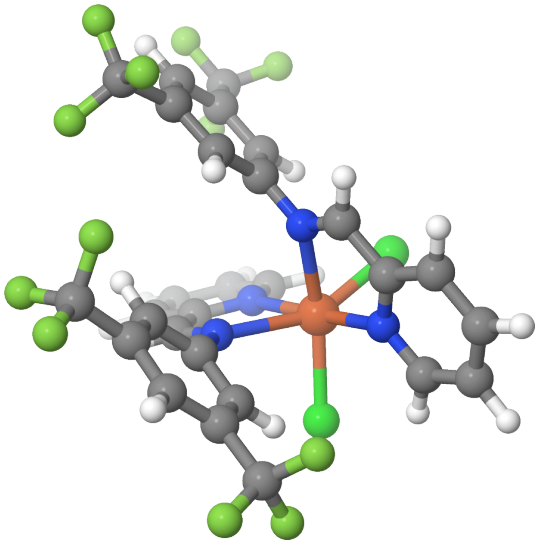

Figure 3: Iron complex in its fully optimized quintet state geometry.

4 Conclusion

In summary, the combined use of orbital projection analysis and cosine similarity mapping available in gSOMOs provides a robust framework for the identification and characterization of SOMOs in open-shell systems. The projection approach quantifies the physical overlap between α and β spin channels, while the similarity mapping offers a complementary perspective. Together, these methods capture subtle near-degeneracies and possible weak spin contamination effects, offering a detailed understanding of the electronic structure. This workflow can be readily extended to the analysis of open-shell transition metal complexes, radicals, and any other systems exhibiting open-shell character.

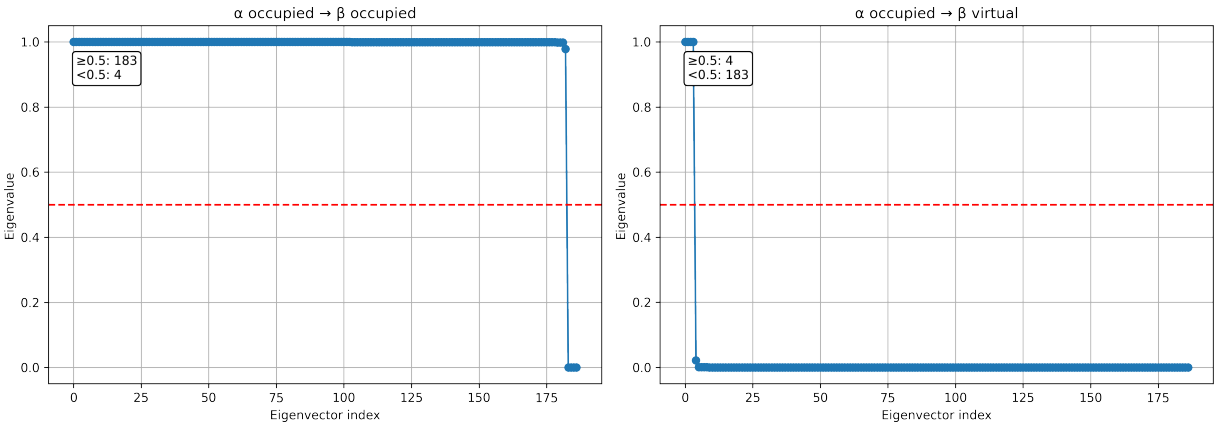

Figure 4: Eigenvalues of $\boldsymbol{P}_{\text{occ}}$ and $\boldsymbol{P}_{\text{virt}}$ for the iron complex.

SOMO Candidate #1	SOMO Candidate #2
α occupied contributions: • α 187 (44.2%) • α 164 (27.3%) β virtual projections: • β 194 (73.3%) • β 196 (16.1%)	α occupied contributions: • α 169 (41.1%) • α 186 (21.6%) • α 165 (15.7%) β virtual projections: • β 192 (53.1%) • β 193 (26.9%)
SOMO Candidate #3	**SOMO Candidate #4**
α occupied contributions: • α 186 (30.0%) β virtual projections: • β 198 (73.0%)	α occupied contributions: • α 168 (51.8%) • α 183 (16.3%) β virtual projections: • β 193 (41.6%) • β 192 (26.7%)

Table 3: SOMO candidates for the iron complex, as given by the analyzis of the eigenvectors of $\boldsymbol{P}_{\text{occ}}$ and $\boldsymbol{P}_{\text{virt}}$.

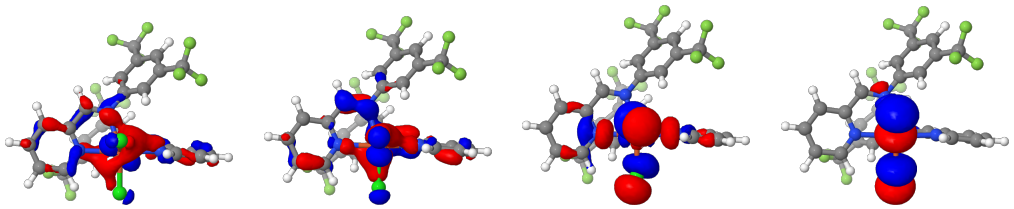

Figure 5: SOMOs of the iron complex