

SOMO Identification

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

April 23, 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. 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. 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. Instead, they 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.

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 S (iop(3/33=1) keyword). 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 $\phi_i^\alpha \in \mathbb{R}^{1 \times n_{\text{basis}}}$ be the coefficient vector of the i -th occupied α orbital, and let $\Phi^\beta \in \mathbb{R}^{N \times n_{\text{basis}}}$ 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:

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

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:

$$\begin{aligned}\mathbf{v}_i^{\text{occ}} &= \phi_i^\alpha \cdot S \cdot (\Phi_{\text{occ}}^\beta)^T \\ \mathbf{v}_i^{\text{virt}} &= \phi_i^\alpha \cdot S \cdot (\Phi_{\text{virt}}^\beta)^T\end{aligned}$$

We then compute:

$$\begin{aligned}\|\mathbf{v}_i^{\text{occ}}\|^2 &= \text{projection of } \phi_i^\alpha \text{ onto occupied } \beta \text{ orbitals} \\ \|\mathbf{v}_i^{\text{virt}}\|^2 &= \text{projection of } \phi_i^\alpha \text{ onto virtual } \beta \text{ orbitals}\end{aligned}$$

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 contrib (%)**”, “**Top 2 contrib (%)**” and “**Top 3 contrib (%)**”.
- The **dominance ratio** 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 >20%” 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 index (1-based) 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:

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

2.2 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 Φ^α (for alpha-spin orbitals) and Φ^β (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 \mathbf{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:

$$\langle \Phi^\alpha | \Phi^\beta \rangle_S = \Phi^{\alpha T} \mathbf{S} \Phi^\beta \quad (1)$$

Thus, the cosine similarity between two molecular orbitals Φ^α and Φ^β accounting for the basis overlap matrix, \mathbf{S} , is given by:

$$\text{cosine similarity}(\Phi^\alpha, \Phi^\beta) = \frac{\Phi^{\alpha T} \mathbf{S} \Phi^\beta}{\sqrt{\Phi^{\alpha T} \mathbf{S} \Phi^\alpha} \sqrt{\Phi^{\beta T} \mathbf{S} \Phi^\beta}} \quad (2)$$

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). It is provided by the `scipy.optimize.linear_sum_assignment` function.

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.

2.3 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.

3 Examples

3.1 Formaldehyde (H_2CO)

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)	\mathbf{P}^2 β_{virt}	\mathbf{P}^2 β_{occ}	β MO*	Occ β	SOMO?	β MOs >20%
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. \mathbf{P}^2 β_{virt} and \mathbf{P}^2 β_{occ} are $\|\mathbf{v}_i^{\text{occ}}\|^2$ and $\|\mathbf{v}_i^{\text{virt}}\|^2$, 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 contributions above 20% 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)

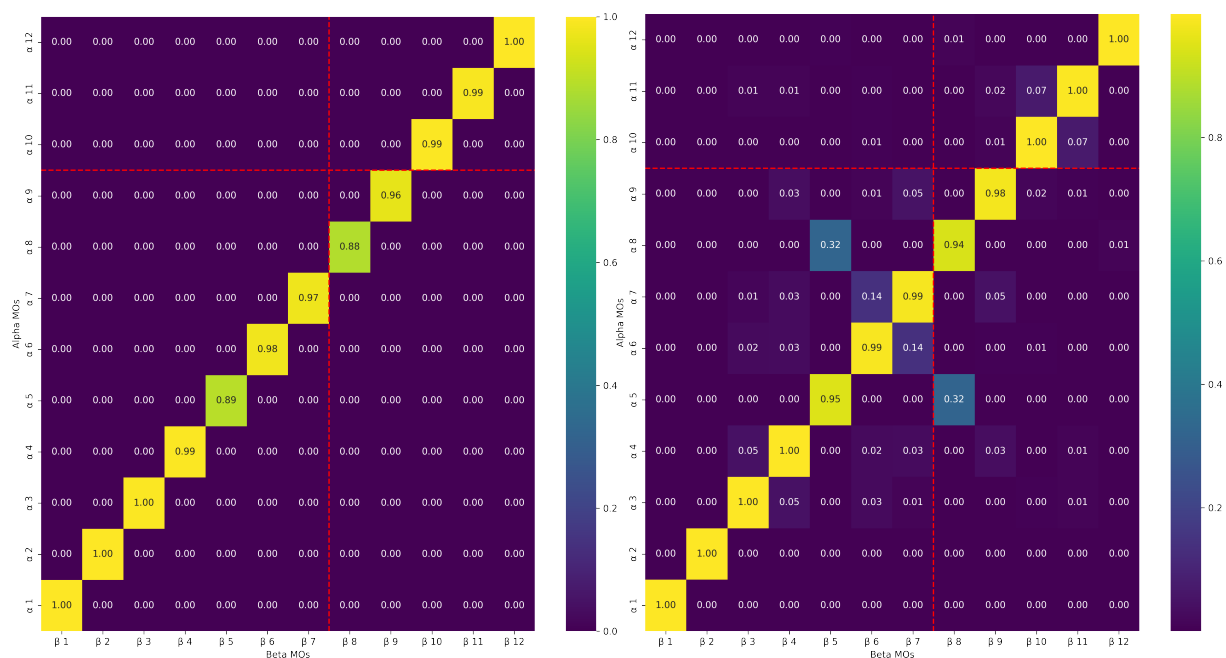


Figure 1: similarity of α and β MOs of the first triplet state of H_2CO around the HOMO-LUMO frontier. (a) Projection of α MOs onto the full space spanned by all β orbitals (only contributions above 20% were retained for clarity); (b) Cosine similarity. Dashed lines mark the HOMO/LUMO boundaries for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