# Supporting Information for:

## Localized Molecular Orbital-Based Embedding Scheme for Correlated Methods

Giovanni Macetti[(1)*], Erna K. Wieduwilt[(1)], Xavier Assfeld[(2)], Alessandro Genoni[(1)*]

(1) Université de Lorraine & CNRS, Laboratoire de Physique et Chimie Théoriques (LPCT), UMR CNRS 7019, 1 Boulevard Arago, F-57078 Metz, France.

(2) Université de Lorraine & CNRS, Laboratoire de Physique et Chimie Théoriques (LPCT), UMR CNRS 7019, Boulevard des Aiguilletes, BP 70239, F-54506 Vandoeuvre-lès-Nancy, France.

* Correspondence to:

- Alessandro Genoni, Université de Lorraine & CNRS, Laboratoire de Physique et Chimie Théoriques (LPCT), UMR CNRS 7019, 1 Boulevard Arago, F-57078 Metz, France. E-mail: alessandro.genoni@univ-lorraine.fr; Phone: +33 (0)3 72 74 91 70; Fax: +33 (0)3 72 74 91 87.
- Giovanni Macetti, Université de Lorraine & CNRS, Laboratoire de Physique et Chimie Théoriques (LPCT), UMR CNRS 7019, 1 Boulevard Arago, F-57078 Metz, France. E-mail: giovanni.macetti@univ-lorraine.fr; Phone: +33 (0)3 72 74 91 70; Fax: +33 (0)3 72 74 91 87.

## S1. Details about theory, transfer and libraries of Extremely Localized Molecular Orbitals

### S1.1 Basic theory of Extremely Localized Molecular Orbitals

ELMOs (Extremely Localized Molecular Orbitals) are molecular orbitals strictly localized on small molecular fragments, such as atoms, bonds and functional groups. They can be obtained by defining *a priori* a localization scheme that partitions the molecules under exam into subunits that may overlap.[1,2] Due to this partition, a local basis-set $\beta_i = \{|\chi_{i\nu}\rangle\}_{\nu=1}^{M_i}$ is automatically defined for each fragment *i*. These basis-sets comprise only basis functions centered on the atoms that belong to the subunits and they are used to expand the ELMOs of the different fragments. For example, the generic $\gamma$-th ELMO for the *i*-th fragment can be expressed as follows:

$$|\,\varphi_{i\gamma}\rangle = \sum_{\nu=1}^{\mathrm{M}_i} C_{i\nu,i\gamma}\ |\chi_{i\nu}\rangle \qquad \text{(S1)}$$

According to Stoll,[1] the system under examination may be described through a single Slater determinant built up with the extremely localized molecular orbitals defined by equation (S1), leading to the so-called ELMO wave function:

$$|\Psi_{ELMO}\rangle = \frac{1}{\sqrt{(2N)!\,det[\mathbf{S}]}}\ \hat{A}\left[\prod_{i=1}^{f}\prod_{\gamma=1}^{n_i}\varphi_{i\gamma}\,\bar{\varphi}_{i\gamma}\right] \qquad \text{(S2)}$$

where $\hat{A}$ is the usual antisymmetrizer, $n_i$ is the number of occupied ELMOs for the *i*-th fragment, $\varphi_{i\gamma}$ is a spinorbital with spatial part $\varphi_{i\gamma}$ and spin part $\alpha$ and $\bar{\varphi}_{i\gamma}$ is a spinorbital with spatial part $\varphi_{i\gamma}$ and spin part $\beta$. Furthermore, $det[\mathbf{S}]$ is the determinant of the overlap-matrix between the occupied ELMOs, which appears because of the ELMOs non-orthogonality.

ELMOs are obtained by minimizing the energy associated with $|\Psi_{ELMO}\rangle$ with respect to the expansion coefficients $\{C_{i\nu,i\gamma}\}$ in (S1). This is equivalent to solving a set of modified Hartree-Fock equations (namely, the Stoll equations) for each fragment in a self-consistent way:[1]

$$\hat{F}_i|\varphi_{i\gamma}\rangle = \varepsilon_{i\gamma}\,|\varphi_{i\gamma}\rangle \qquad \text{(S3)}$$

with $\hat{F}_i$ as the modified Fock operator corresponding to the *i*-th subunit:

$$\hat{F}_i = \left(1 - \hat{\rho} + \hat{\rho}_i^{\dagger}\right)\hat{F}\left(1 - \hat{\rho} + \hat{\rho}_i\right) \qquad \text{(S4)}$$

In the previous equation $\hat{F}$ is the traditional Fock operator, $\hat{\rho}$ is the global density operator depending on all the occupied ELMOs of the molecule, and $\hat{\rho}_i$ is the density operator for the *i*-th subunit depending only on the occupied ELMOs of the fragment.

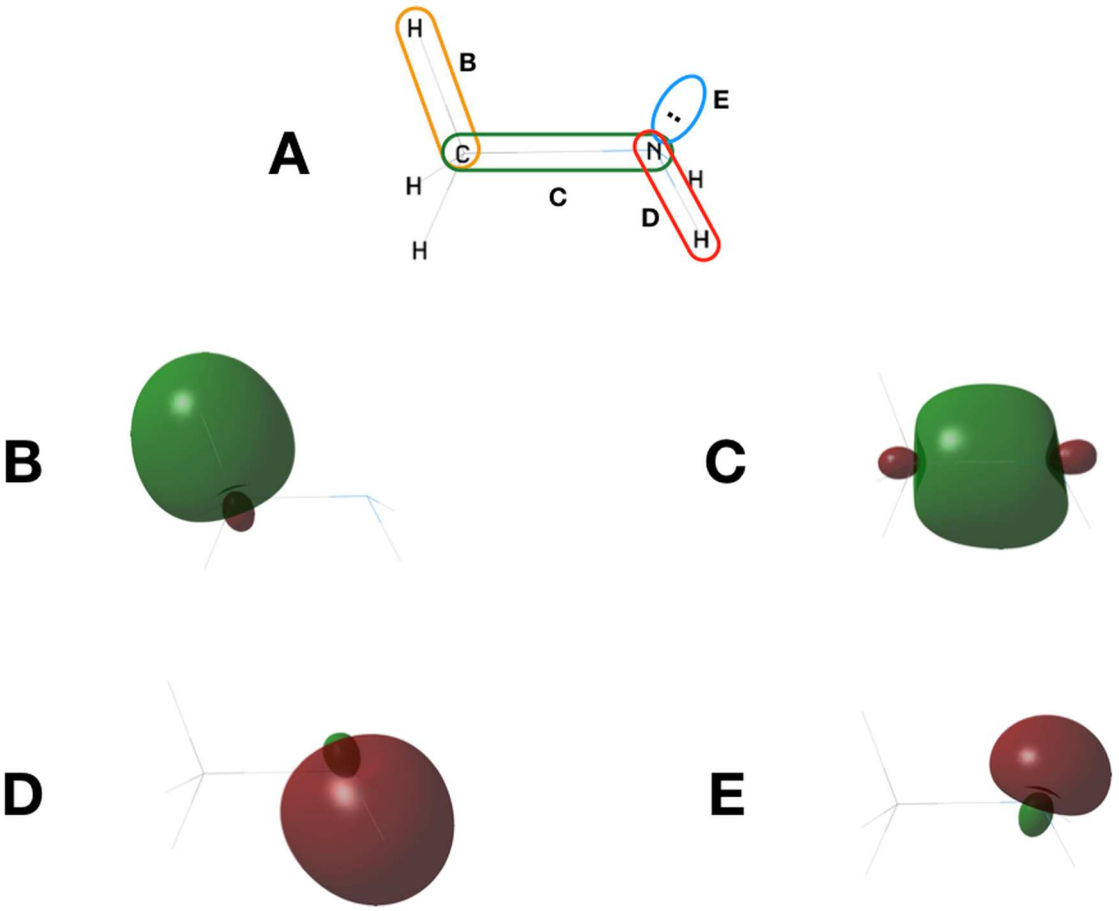


**Figure S1.** Example of Extremely Localized Molecular Orbitals computed for the methylamine molecule using a localization scheme corresponding to the Lewis structure of the system: (A) Lewis structure of methylamine, also explicitly indicating the ELMOs depicted in Figures S1B-S1E, (B) ELMO describing one of the C-H bonds of the methyl group, (C) ELMO associated with the C-N bond, (D) ELMO corresponding to one of the N-H bonds of the amino group and (E) ELMO describing the lone-pair of the nitrogen atom. For the sake of simplicity, ELMOs associated with the core electrons of carbon and nitrogen atoms are omitted because they simply correspond to spherical distributions around the nuclei. All the orbitals were computed using the cc-pVDZ basis-set considering a 0.1 a.u. (C-H, C-N and N-H ELMOs) or a 0.2 a.u. (lone-pair ELMO) isosurface

As an example, in Figure S1 we depicted the ELMOs that one can compute for methylamine by adopting a localization scheme strictly corresponding to the Lewis structure of the molecule (Figure S1A). Namely, one can obtain i) ELMOs describing the C-H bonds of the methyl group (Figure S1B); ii) an ELMO associated with the central C-N bond (Figure S1C); iii) ELMOs describing the N-H bonds (Figure S1D); iv) an ELMO describing the lone-pair of the nitrogen atom (Figure S1E).

**S1.2 Transferability and rotation of ELMOs.**

Because of their extreme localization, ELMOs are orbitals easily transferable from molecule to molecule,[3,4] and particularly from a model system (which is usually the small molecule on which the orbital is originally determined) to the target system under examination. To accomplish this task, it is necessary to follow the method proposed by Philipp and Friesner.[5] This method allows the definition of a proper rotation matrix **P** that transforms the ELMO coefficients (see equation (S1)) obtained in the geometry of the model molecule to the ELMO coefficients in the geometry of the target system. This rotation matrix is obtained by defining *i)* a reference frame (**a**, **c**, **d**) in the model system and *ii)* a reference frame (**a**', **c**', **d**') in the target molecule (see Figure S2).

Both the reference frames originate from the choice of two atomic triads (one for the model system and one for the target molecule) that guarantee the uniqueness of the rotation. Indicating the triads for the model and target molecules as ($A_1$, $A_2$, $A_3$) and ($A_1$', $A_2$', $A_3$'), respectively, the vectors that define the two reference frames are: **a** (**a**'), which is the position vector of $A_2$ ($A_2$') relative to $A_1$ ($A_1$') (see Figure S2), while **c** (**c**') and **d** (**d**') are given by the following vector products:

$$\begin{cases} \mathbf{c} = \mathbf{a} \times \mathbf{b} & (\mathbf{c}' = \mathbf{a}' \times \mathbf{b}') \\ \mathbf{d} = \mathbf{c} \times \mathbf{a} & (\mathbf{d}' = \mathbf{c}' \times \mathbf{a}') \end{cases} \qquad \text{(S5)}$$

where **b** (**b'**) is the position vector of $A_3$ ($A_3$') with respect to $A_1$ ($A_1$') (see Figure S2).

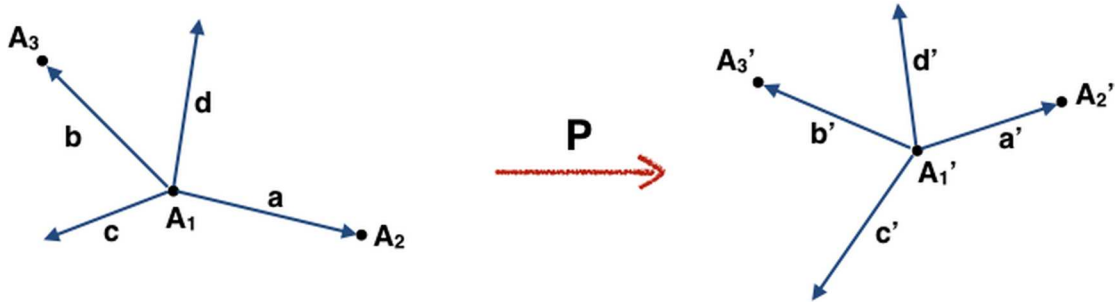


**Figure S2.** Schematic representation of the reference frames and of the atomic triads that are required to define matrix **P** associated with the rotation of the ELMOs from the geometry of the model molecule (left) to the geometry of the target system (right).

For ELMOs strictly localized only on one atom (i.e., ELMOs corresponding to core or lone-pair electrons) the atomic triads simply correspond to the atom on which the ELMO is localized and usually by two other bonded atoms. For ELMOs associated with two-centre bonds, the triads correspond to the atoms that form the bond plus an atom describing the local dissymmetry of the bond under examination.[6] For ELMOs localized on three-centre (e.g., ELMOs used to describe

situations in which it is important to take into account the delocalization of the electronic structure, such as in peptide bonds or aromatic rings), the triads of atoms are automatically selected. For ELMOs localized on more than three atoms, the definition of triads (and, consequently, of reference frames) that simultaneously take into account the orientation of all the atoms in the subunit is impossible and this is the reason why all the ELMOs available in the current databanks are localized at the largest on three atoms (see the next subsection).

Rotation matrix **P** that transforms reference frame (**a**, **c**, **d**) to reference frame (**a**', **c**', **d**') is the crucial matrix that enables the definition and construction of all the rotation matrices for all kinds of basis functions and associated ELMO coefficients. In fact, if we exclude the *s*-type atomic orbitals, which are invariant because of their spherical symmetry, it is easy to show[3] that the *p*-type basis functions (and the related coefficients) can be exactly rotated by matrix **P**, while basis functions (and corresponding coefficients) with angular momentum greater than 1 can be transformed employing matrices defined as a function of **P**.

### S1.3 Libraries of Extremely Localized Molecular Orbitals

As a consequence of the ELMOs transferability, libraries of Extremely Localized Molecular Orbitals have been recently constructed.[7] The ELMO libraries allow the description of all the possible fragments of water molecule and of the twenty natural amino acids in all their possible protonation states and forms (namely, N-terminal, C-terminal and non-terminal forms). All the ELMOs in the databanks were calculated on proper model systems taking into account the chemical environment of the considered fragments. The ELMO libraries are currently available for five basis-sets (i.e., 6-31G, 6-311G, 6-31G(d,p), 6-311G(d,p) and cc-pVDZ) and comprise molecular orbitals strictly localized on one-atom subunits (corresponding to core or lone pair electrons), two-atom fragments (corresponding to bonding electrons), but also molecular orbitals strictly localized on three-atom subunits, which allow to describe the delocalized nature of the electronic structure in some particular situations (e.g., in carboxylic groups, peptide bonds or aromatic rings).

The in-house program *ELMOdb*[7] handles the transfer of ELMOs from the databanks to the examined target structures: it analyses PDB files of polypeptides or proteins and, for each of their residues, executes the transfer procedure for one subunit at a time. The *ELMOdb* software also enables to read *ad hoc* ELMOs corresponding to subunits of molecules that are not included in the current version of the libraries. These ELMOs must be computed on proper model molecules and afterwards stored in specific folders from which the *ELMOdb* program can read them.

The ELMO libraries and the associated *ELMOdb* program are available upon motivated request to one of the corresponding authors (A.G.) of the present paper.

## S2. Supplementary figures and tables

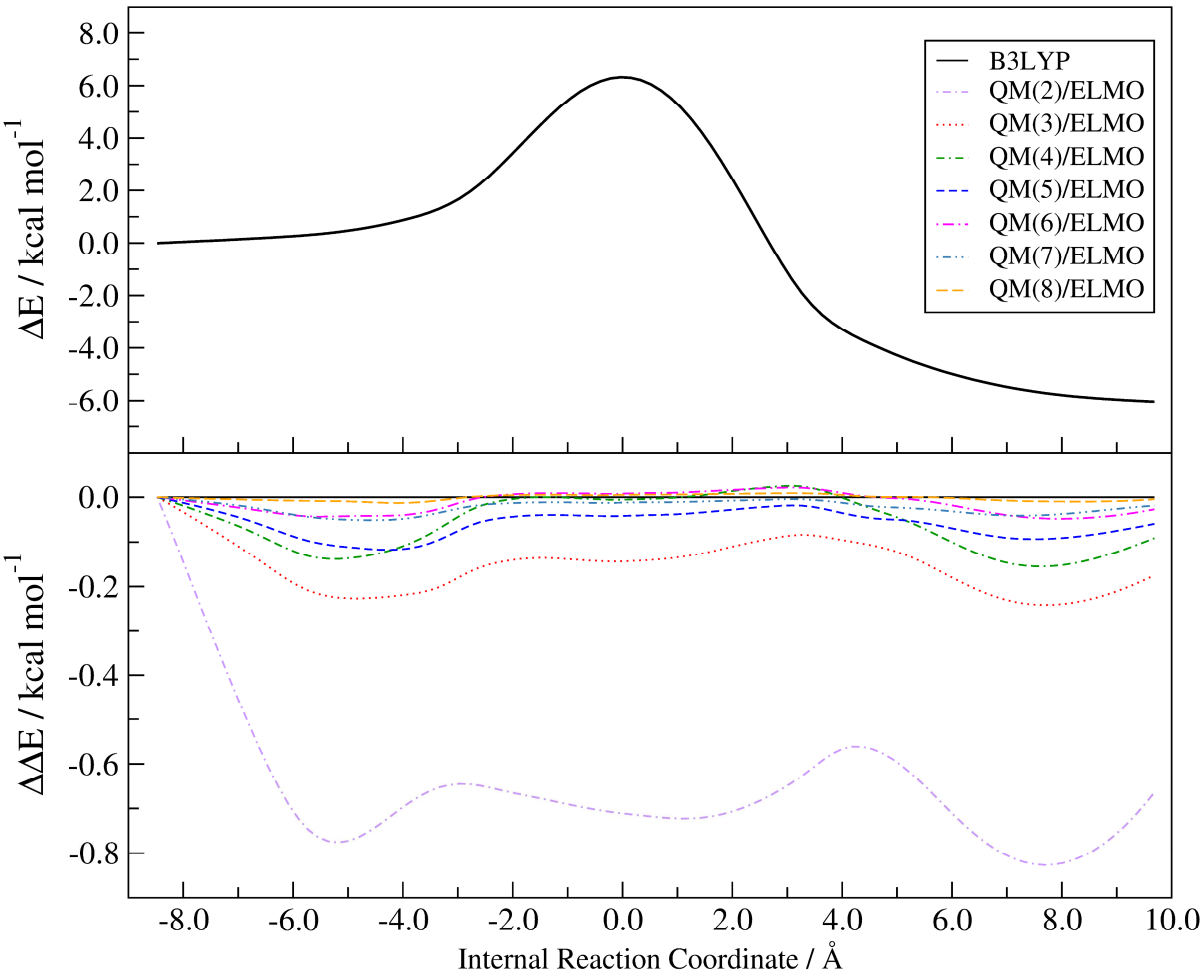


**Figure S3.** $S_N2$ reaction energy profile at B3LYP level (top panel) and deviations from it when B3LYP/ELMO calculations are performed with QM regions of different size (bottom panel); all the curves refer to the cc-pVDZ basis-set.

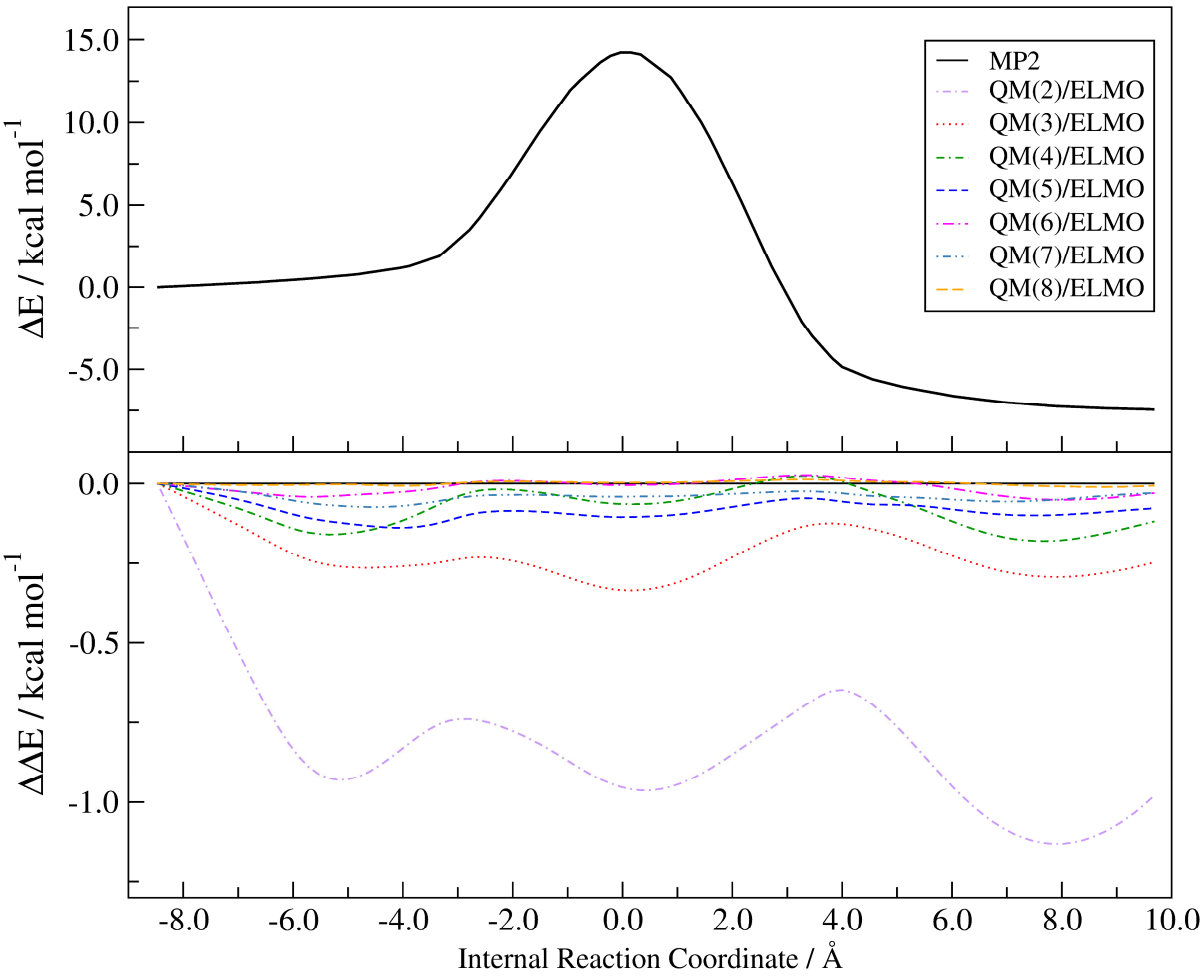


**Figure S4.** $S_N2$ reaction energy profile at MP2 level (top panel) and deviations from it when MP2/ELMO calculations are performed with QM regions of different size (bottom panel); all the curves refer to the cc-pVDZ basis-set.

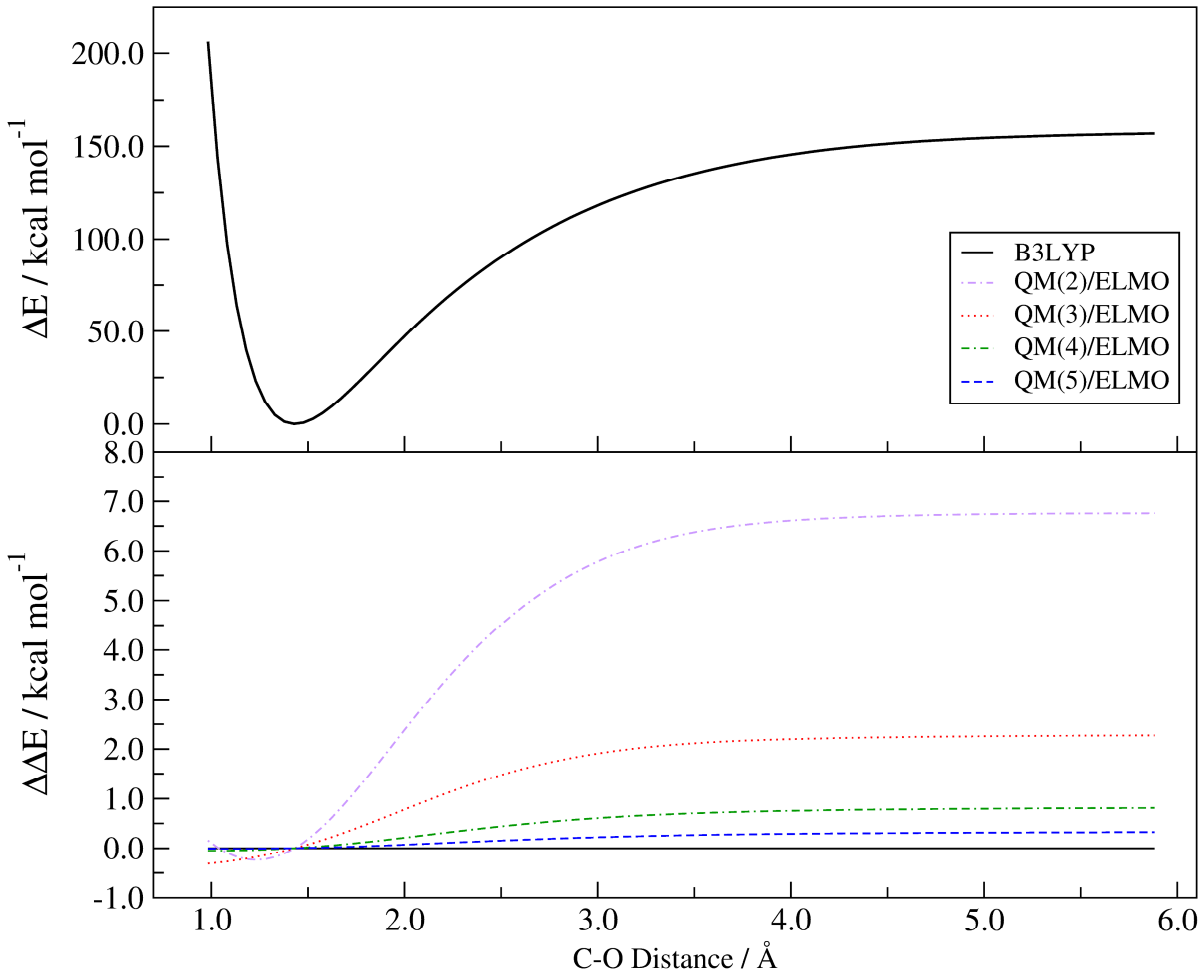


**Figure S5.** Dissociation energy profile of the terminal C-OH bond in sorbitol at B3LYP level (top panel) and deviations from it when B3LYP/ELMO calculations are performed with QM regions of different size (bottom panel); all the curves refer to the cc-pVDZ basis-set.

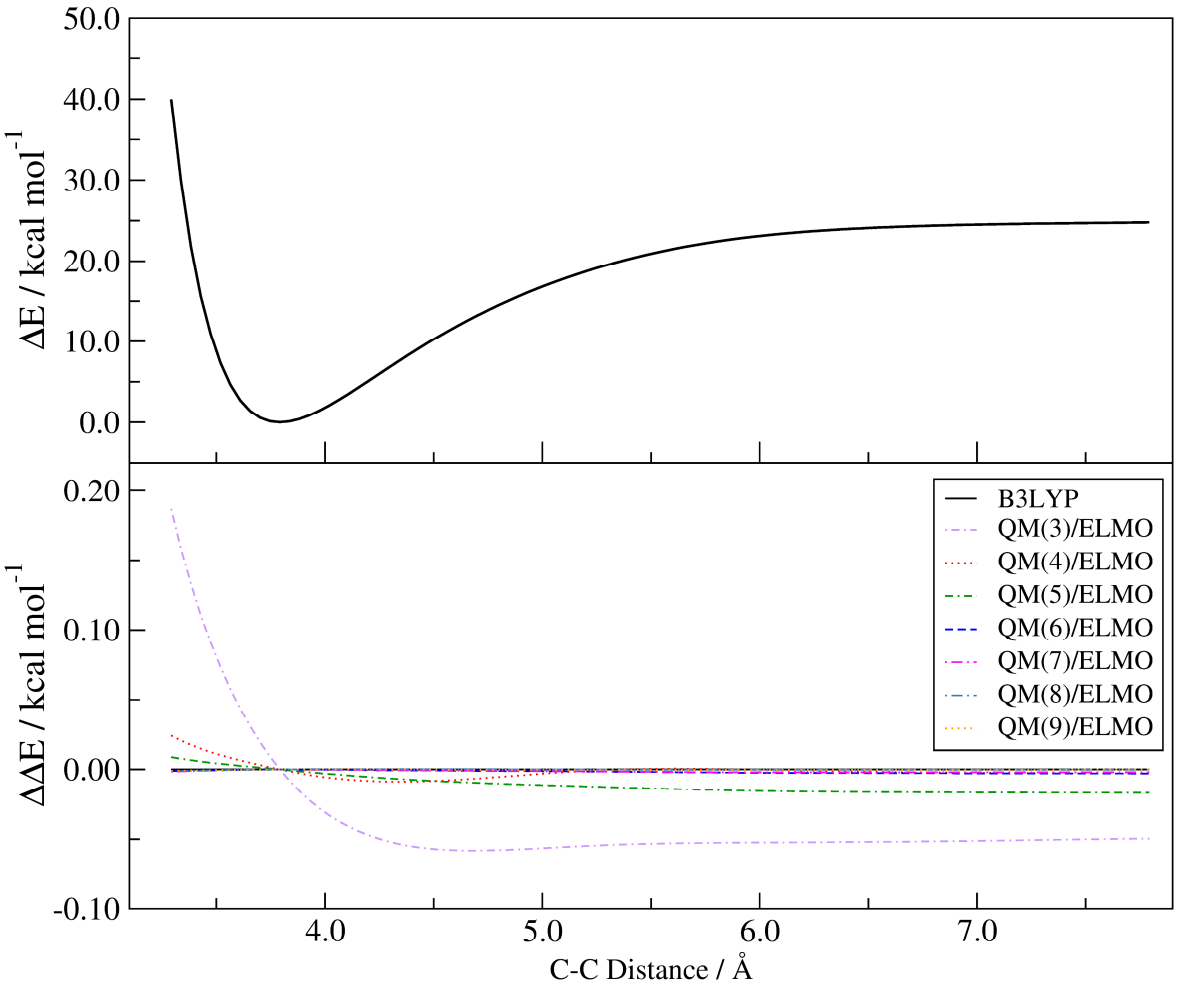


**Figure S6.** Interaction energy profile for the formic acid - decanoic acid dimer at B3LYP level (top panel) and deviations from it when B3LYP/ELMO calculations are performed with QM regions of different size (bottom panel); all the curves refer to the cc-pVDZ basis-set.

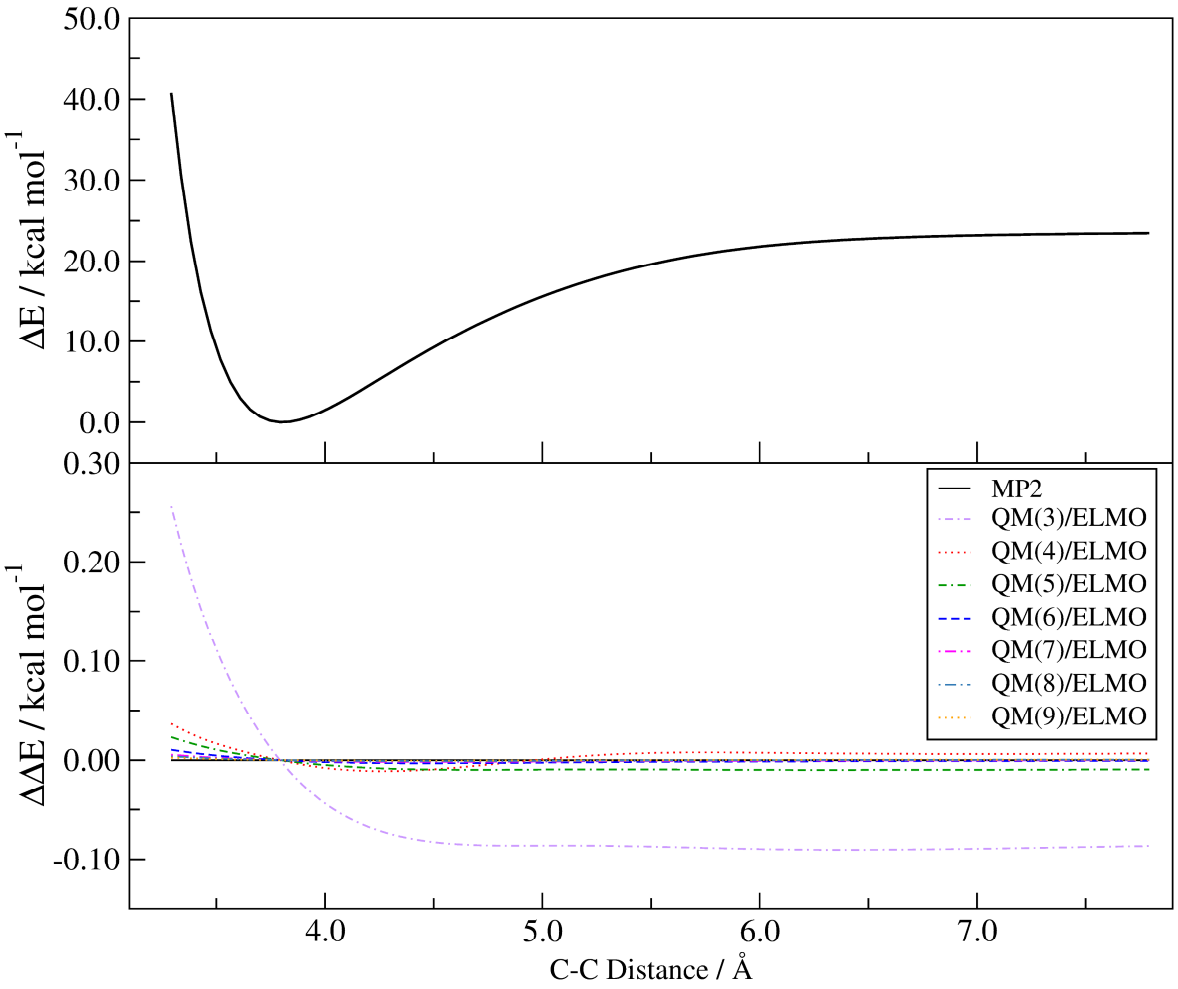


**Figure S7.** Interaction energy profile for the formic acid - decanoic acid dimer at MP2 level (top panel) and deviations from it when MP2/ELMO calculations are performed with QM regions of different size (bottom panel); all the curves refer to the cc-pVDZ basis-set.

**Table S1.** Number of (frozen and active) occupied molecular orbitals ($N_{occ}$), number of virtual molecular orbitals ($N_{virt}$) and timings associated with the CCSD/ELMO and CCSD calculations (cc-pVDZ basis-set) performed on the minimum energy structure of sorbitol.[(a)]

| Calculations | $N_{occ}$ | | $N_{virt}$ | CPU time (s) | % |
|---|---|---|---|---|---|
| | Frozen | Active | | | |
| QM(2)/ELMO | 37 | 12 | 69 | 1258.5 | 2.7 |
| QM(3)/ELMO | 31 | 18 | 101 | 3122.9 | 6.7 |
| QM(4)/ELMO | 25 | 24 | 133 | 8165.2 | 17.4 |
| QM(5)/ELMO | 19 | 30 | 165 | 18464.4 | 39.4 |
| Full QM (6) | 12 | 37 | 201 | 46901.0 | 100.0 |

[(a)] The recorded timings were obtained by performing parallel calculations on 16 Intel Xeon Gold 6130 2.1 GHz processors.

**Table S2.** Number of (frozen and active) occupied molecular orbitals ($N_{occ}$), number of virtual molecular orbitals ($N_{virt}$) and timings associated with the CCSD(T)/ELMO and CCSD(T) calculations (cc-pVDZ basis-set) performed on the minimum energy structure of the formic acid – decanoic acid dimer.[(a)]

| Calculations | $N_{occ}$ | | $N_{virt}$ | CPU time (s) | % |
|---|---|---|---|---|---|
| | Frozen | Active | | | |
| QM(3)/ELMO | 40 | 20 | 103 | 8521.9 | 1.8 |
| QM(4)/ELMO | 37 | 23 | 124 | 13541.0 | 2.8 |
| QM(5)/ELMO | 34 | 26 | 145 | 22536.8 | 4.6 |
| QM(6)/ELMO | 31 | 29 | 166 | 42836.4 | 8.8 |
| QM(7)/ELMO | 28 | 32 | 187 | 77277.4 | 15.9 |
| QM(8)/ELMO | 25 | 35 | 208 | 153539.2 | 31.6 |
| QM(9)/ELMO | 22 | 38 | 229 | 226973.7 | 46.8 |
| Full QM (11) | 15 | 45 | 275 | 485305.4 | 100.0 |

[(a)] The recorded timings were obtained by performing parallel calculations on 16 Intel Xeon Gold 6130 2.1 GHz processors.

## S3. References for the Supporting Information


[1] Stoll, H.; Wagenblast, G.; Preuss, H. On the Use of Local Basis Sets for Localized Molecular Orbitals. *Theor. Chim. Acta* **1980**, *57*, 169−178.

[2] Sironi, M.; Genoni, A.; Civera, M.; Pieraccini, S.; Ghitti, M. Extremely Localized Molecular Orbitals: Theory and Applications. *Theor. Chem. Acc.* **2007**, *117*, 685-698.

[3] Meyer, B.; Guillot, B.; Ruiz-Lopez, M. F.; Genoni, A. Libraries of Extremely Localized Molecular Orbitals. 1. Model Molecules Approximation and Molecular Orbitals Transferability. *J. Chem. Theory. Comput.* **2016**, *12*, 1052-1067.

[4] Meyer, B.; Guillot, B.; Ruiz-Lopez, M. F.; Jelsch, C.; Genoni, A. Libraries of Extremely Localized Molecular Orbitals. 2. Comparison with the Pseudoatoms Transferability. *J. Chem. Theory. Comput.* **2016**, *12*, 1068-1081.

[5] Philipp, D. M.; Friesner, R. A. Mixed Ab Initio QM/MM Modeling Using Frozen Orbitals and Tests with Alanine Dipeptide and Tetrapeptide. *J. Comput. Chem.* **1999**, *20*, 1468-1494.

[6] Ferré, N.; Assfeld, A.; Rivail, J.-L. Specific Force Field Parameters Determination for the Hybrid Ab Initio QM/MM LSCF Method. *J. Comput. Chem.* **2002**, *23*, 610-624.

[7] Meyer, B.; Genoni, A. Libraries of Extremely Localized Molecular Orbitals. 3. Construction and Preliminary Assessment of the New Databanks. *J. Phys. Chem. A* **2018**, *122*, 8965-8981.