Libraries of Extremely Localized Molecular Orbitals. 3. Construction and Preliminary Assessment of the New Databanks

Benjamin Meyer, Alessandro Genoni

Libraries of Extremely Localized Molecular Orbitals. 3. Construction and Preliminary Assessment of the New Databanks

Benjamin Meyer [(1)#], 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.

[#] Present Address:

Laboratory for Computational Molecular Design, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland.

* Correspondence to:

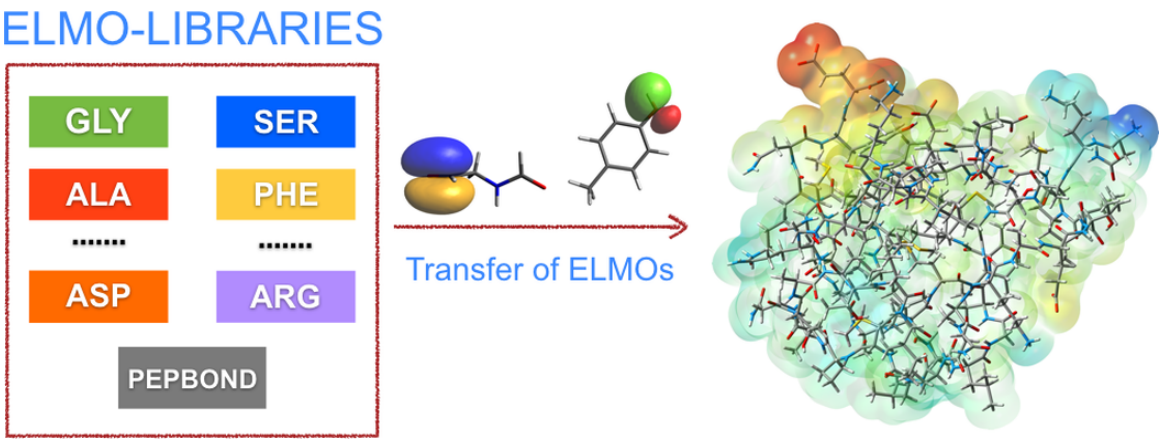
Alessandro Genoni, Université de Lorraine & CNRS, Laboratoire de Physique et Chimie Théoriques (LPCT), UMR CNRS 7019, 1 Boulevard Arago, F-57078 Metz, France. E-mail: Alessandro.Genoni@univ-lorraine.fr; Phone: +33 (0)3 72 74 91 70.

Abstract

The fast and reliable determination of wave functions and electron densities of macromolecules has been one of the goals of theoretical chemistry for a long time and, in this context, several linear scaling techniques have been successfully devised over the years. Different approaches have been adopted to tackle this problem and one of them exploits the fact that, according to the traditional chemical perception, molecules can be seen as constituted of recurring units (e.g., functional groups) with well-defined chemical features. This has led to the development of methods in which the global wave functions or electron densities of macromolecules are obtained by simply transferring density matrices or fuzzy electron densities associated with molecular fragments. In this context, we propose an alternative strategy that aims at quickly reconstructing wave functions and electron densities of proteins through the transfer of extremely localized molecular orbitals (ELMOs), which are orbitals strictly localized on small molecular units and, for this reason, easily transferable from molecule to molecule. To accomplish this task we have constructed original libraries of ELMOs that cover all the possible elementary fragments of the twenty natural amino acids in all their possible protonation states and forms. Our preliminary test calculations have shown that, compared to more traditional methods of quantum chemistry, the transfers from the novel ELMO-databanks allow to obtain wave function and electron densities of large polypeptides and proteins at a significantly reduced computational cost. Furthermore, notwithstanding expected discrepancies, the obtained electron distributions and electrostatic potentials are in very good agreement with those obtained at Hartree-Fock and density functional theory (DFT) levels. Therefore, the results encourage to use the new libraries as alternatives to the popular pseudoatom-databases of crystallography in the refinement of

crystallographic structures of macromolecules. In particular, in this context, we have already envisaged the coupling of the ELMO-databanks with the promising Hirshfeld atom refinement technique to extend the applicability of the latter to very large systems.

TOC Graphic


ELMO-LIBRARIES
GLY
SER
ALA
PHE
.....
.....
ASP
ARG
PEPBOND
Transfer of ELMOs

1. Introduction

One of the main goals of contemporary theoretical chemistry is the development of efficient computational methods able to provide reliable descriptions of macromolecular electronic structures in a reasonable lapse of time. In fact, although available computational resources are more and more powerful and quantum chemistry strategies are constantly improved in terms of time scaling, the current limit for basic (e.g., Hartree-Fock or density functional theory (DFT)) quantum chemistry calculations is represented by relatively large systems constituted of only very few hundreds of atoms.[1-5] In order to obtain a fully quantum mechanical (QM) description of larger systems (e.g., very large proteins as Hsp90), it is necessary to resort to linear scaling techniques (or, *O(N)* methods),[6-8] which are approximate strategies whose computational cost scales linearly with the dimensions of the system under exam. All these approaches strongly rely on the traditional chemical perception, according to which every molecule can be seen as composed of small and well-defined units (e.g. functional groups) that keep their main features in different systems and that are only partially influenced by their nearest chemical environment. Following this reasoning, the *O(N)* strategies basically consist in performing calculations on small molecular fragments and, afterwards, in assembling the results obtained on the subunits to recover the global wave function, electron density or physical properties for the large target system under investigation. In this context it is worth mentioning the "divide & conquer" techniques[9-14] and the molecular tailoring approach,[15-23] but also the fragment interaction strategies, such as all the current variants of the molecular fractionation with conjugated caps (MFCC) technique,[24-31] the well-known fragment molecular orbital (FMO) approach[32-36] and the kernel energy method (KEM).[37-47]

Another interesting class of linear scaling strategies comprises the LEGO-type approaches. They are based on the transferability principle and practically aim at reconstructing global electron densities, density matrices and wave functions of macromolecules by simply transferring fragment electron distributions, subunit density matrices or strictly localized molecular orbitals (MOs) previously determined on proper model molecules and afterwards stored in databanks. Prominent examples are the molecular electron density LEGO assembler (MEDLA) technique[48,49] and the adjustable density matrix assembler (ADMA) strategy[50-52] proposed by Mezey and coworkers, the transferable atom equivalents (TAE) method introduced by Breneman[53,54] and real-space approaches developed by Bader[55-57] and Matta.[58]

Among the approaches that we have just mentioned, it is also possible to include a technique that we have recently anticipated in two papers of ours[59,60] and that basically consists in reconstructing wave functions and electron densities of very large molecules (e.g., proteins) through the transfer of the so-called extremely localized molecular orbitals (ELMOs).[61] As we will briefly show in the introductory theory section, ELMOs are orbitals that are variationally determined by constraining them to expand in local basis-sets that result from the *a priori* definition of chemically meaningful fragments. This enables to obtain MOs that, unlike the traditional localized molecular orbitals of quantum chemistry,[62-66] do not show orthogonalization tails that extend beyond the main localization region and that, for this reason, are strictly associable with small molecular subunits and, in principle, easily transferable from one molecule to another. Therefore, ELMOs are orbitals characterized by an intrinsic connection with the usual chemical perception and allow the introduction of traditional chemical concepts in *ab initio* computations, as also recently testified by

their direct use[67-72] in the framework of Jayatilaka's X-ray constrained wave function approach.[73-81]

In the two previous papers of this series,[59,60] detailed studies on the transferability of the extremely localized molecular orbitals have been shown and discussed. The main conclusion was that, except for expected and unavoidable discrepancies, the electron densities resulting from the transfer of ELMOs are generally very similar *i)* to the corresponding Hartree-Fock ones (with the size of the observed differences comparable to the extent of the discrepancies between the DFT and Hartree-Fock electron distributions) and *ii)* to those obtained through the transfer of the multipole-model pseudoatoms[82-84] of crystallography. This confirmed the reliable exportability of the ELMOs, which had been previously investigated by only considering small target systems.[85-90] Furthermore, it decisively encouraged the construction of new libraries of extremely localized molecular orbitals with the final goal of refining crystallographic structures and computing approximate properties of very large molecules. Actually, a database of ELMOs (*DENPOL*) was already available and was constructed by Sironi and coworkers to rapidly obtain approximate electron densities of proteins.[91] Nevertheless, *DENPOL* is limited to the minimal basis-set STO-4G and, in the associated program for the transfer of the extremely localized molecular orbitals, the transfer protocol leads to final geometries of the target systems that always deviate from the input ones, which is clearly incompatible with our final objective of using the ELMO-databanks for the refinement of crystallographic structures.

The new ELMO-libraries, whose features and performances will be described in this paper, will clearly represent a valid alternative to the exiting databanks of the above-mentioned multipole-model pseudoatoms.[92-109] In fact, crystallographers have

successfully and widely used the pseudoatoms databanks not only to obtain instantaneous reconstructions of electron densities of large systems, but also to perform accurate structural refinements and to compute approximate electrostatic properties of various biomolecules. Four are the main libraries of pseudoatoms currently used: *i)* the "experimental" ELMAM2 database,[96,97] which stems from the pioneering ELMAM databank[92-95] and which has been constructed by averaging the values of multipole parameters corresponding to chemically equivalent pseudoatoms previously obtained from high-quality experimental charge density refinements; the theoretical *ii)* UBDB[98-102] (University at Buffalo Pseudoatom Databank) and *iii)* Invariom databases,[103-107] which have been built starting from charge density refinements against theoretical structure factors obtained from gas-phase *ab initio* calculations; and *iv)* the more recent SBFA (Synthon-Based Fragments Approach) libraries,[108,109] which are designed to specifically include intermolecular interactions and are mainly used in crystal engineering. From the comparisons conducted over the years,[110-112] it emerged that the different libraries have their own specific advantages and disadvantages and they generally lead to very similar results. Of course, their most important applications have been the refinements of crystallographic structures of small proteins and macromolecules, such as the well-know refinement of the crambin structure performed by Jelsch and coworkers,[113] the more recent ELMAM refinement of the diisopropylfluorophosphatase[114] or the Invariom refinement of the thiopeptide antibiotic thiostrepton.[115]

Notwithstanding the very good quality of the obtained crystallographic structures, one of the intrinsic shortcomings of the pseudoatoms databanks is that, after the refinements, only the reconstructed electron distributions (and not the wave functions) of the systems under exam are available. This obviously limits the number of

properties that can be determined since the exact functional relation between ground state wave functions and electron densities of many-electron systems is still unknown. On the contrary, this drawback will be completely overcome through the introduction of the new libraries of extremely localized molecular orbitals that will enable to get an approximate wave function at each refinement-step and that, for this reason, pave the way towards the development of a novel, fast and fully quantum mechanical approach to refine the crystallographic structures of macromolecules.

In particular, the new ELMO-libraries seem suitable for the extension of the Hirshfeld atom refinement (HAR) technique[116-121] to very large systems. HAR is an emerging refinement method of quantum crystallography[122-126] that, by only exploiting X-ray diffraction data, is able to locate the positions of the hydrogen atoms with the same precision and accuracy usually attained by means of neutron diffraction measurements,[118-121] mostly within a single standard deviation. Since this remains true also exploiting X-ray diffraction data at resolutions as low as 0.8 Å,[120] HAR might be exploited to successfully refine protein crystallographic structures at sub-atomic resolution, whose number will certainly increase in the next few years as a consequence of the recent large investments in the construction of facilities for the production of intense high-synchrotron radiations and X-ray free electron lasers. Nevertheless, HAR requires an *ab initio* calculation (usually at Hartree-Fock or DFT levels) at each step of the refinement in order to update the molecular electron density of the investigated system according to the geometry changes. It is obvious that, due to the scaling of the traditional QM methods, HAR cannot be efficiently applied to macromolecules. The only way to accomplish this task is to couple it with a linear scaling technique and, in this direction, the combination of HAR with the new ELMO-libraries might be a completely valid perspective to be considered.

In this paper we will present in detail the recently constructed libraries of extremely localized molecular orbitals. After reviewing the theoretical background of the ELMOs, we will show the structure and the organization of the new ELMO-databanks. Then we will describe the preliminary test calculations that we have performed to assess capabilities and performances of the libraries. In particular, we will analyze the scalability of the ELMOs transfer, we will evaluate the deviation of computed quantities from those obtained by means of traditional quantum chemistry calculations and we will also consider the first applications of the ELMO-databases to real proteins. Finally, in the last section of the paper, we will draw our final conclusions and we will discuss some possible future applications for the new databanks.

2. Theory

Over the years different strategies have been devised for the computation of molecular orbitals strictly localized on small molecular fragments.[61,127-136] In chronological order, the starting point is certainly the "group function method",[137,138] which was developed by McWeeny in 1960s and which has been the reference for all the later techniques belonging to this family. In this context, in 1980, Stoll and coworkers proposed a new set of eigenvalue equations that can be seen as a generalization of the canonical Hartree-Fock equations for the computation of extremely localized molecular orbitals.[61] Fornili *et al.*[86] have afterwards implemented these equations in a reliable program that has been fruitfully used to construct the ELMO-libraries presented in this paper. For this reason, in this section we will briefly focus on the fundamentals of the Stoll equations. Afterwards, we will also introduce

the technique proposed by Philipp and Friesner[139] for the rotation of strictly localized molecular orbitals.

The Stoll equations. Following the Stoll method,[61] the molecule under exam is subdivided into fragments, which, according to the traditional chemical intuition, generally correspond to atoms, bonds and functional groups. After this fragmentation, each pre-defined subunit is automatically associated with a local basis-set $\beta_i = \left\{ \left| \chi_{i\mu} \right\rangle \right\}_{\mu=1}^{M_i}$ that consists of the only basis functions centered on the atoms belonging to the fragment and that is used to expand only the ELMOs of the subunit. For example, the generic α-th ELMO for the *i*-th fragment can be expressed like this:

$$| \varphi_{i\alpha} \rangle = \sum_{\mu=1}^{M_i} C_{i\mu,i\alpha} \left| \chi_{i\mu} \right\rangle \qquad (1)$$

The other assumption of the Stoll method is that the system under exam is described by a single Slater determinant constructed with extremely localized molecular orbitals defined by equation (1) (from now on, ELMO wave function):

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

where $\hat{A}$ is the usual antisymmetrizer, n_i is the number of occupied ELMOs for the *i*-th fragment, $\varphi_{i\alpha}$ is a spin-orbital with spatial part $\varphi_{i\alpha}$ and spin part α and $\overline{\varphi}_{i\alpha}$ is a spin-orbital with spatial part $\varphi_{i\alpha}$ and spin part β. Furthermore, $det[\mathbf{S}]$ is the determinant of the overlap-matrix of the occupied ELMOs. In fact, ELMOs are intrinsically non-orthogonal orbitals because the pre-defined molecular subunits generally share part of their local basis-sets.

ELMOs are obtained by simply variationally minimizing the energy associated with the ELMO wave function. This is mathematically equivalent to solve self-consistently

these modified Hartree-Fock equations (namely, the Stoll equations) for each fragment:

$$\hat{F}_i|\varphi_{i\alpha}\rangle = \varepsilon_{i\alpha}\,|\varphi_{i\alpha}\rangle \qquad (3)$$

where $\hat{F}_i$ is the modified Fock operator for the *i*-th subunit, given by this expression:

$$\hat{F}_i = \left(1 - \hat{\rho} + \hat{\rho}_i^{\dagger}\right)\hat{F}\left(1 - \hat{\rho} + \hat{\rho}_i\right) \qquad (4)$$

with $\hat{F}$ as the usual Fock operator, $\hat{\rho}$ as the global density operator

$$\hat{\rho} = \sum_{i,j=1}^{f}\sum_{\alpha=1}^{n_i}\sum_{\beta=1}^{n_j} |\varphi_{i\alpha}\rangle\, S^{-1}_{i\alpha,j\beta}\, \langle\varphi_{j\beta}| \qquad (5)$$

and $\hat{\rho}_i$ as the density operator for the *i*-th fragment

$$\hat{\rho}_i = \sum_{j=1}^{f}\sum_{\alpha=1}^{n_i}\sum_{\beta=1}^{n_j} |\varphi_{i\alpha}\rangle\, S^{-1}_{i\alpha,j\beta}\, \langle\varphi_{j\beta}| \qquad (6)$$

For the sakes of clarity and completeness, in Figure 1 we have depicted the ELMOs that can be obtained for the ammonia molecule by imposing a localization scheme strictly corresponding to the Lewis structure of the system. It is easy to observe that, by using this localization pattern (Figure 1A), we can straightforwardly obtain ELMOs that describe core and lone-pair electrons on the nitrogen atoms (Figures 1B and 1C) and ELMOs associated with the three N-H bonds of the molecule (Figure 1D).

Finally, it is important to note point out that, due to the non-orthogonality of extremely localized molecular orbitals, convergence problems may sometimes arise when we try to directly solve equations (3). As done by Fornili and coworkers in their implementation of the Stoll strategy,[86] this drawback can be overcome by determining ELMOs through direct ELMO-energy minimizations that exploit approximate Hessians evaluated analytically only at the first iteration and afterwards updated using the Broyden-Fletcher-Goldfarb-Shanno formula.[140]

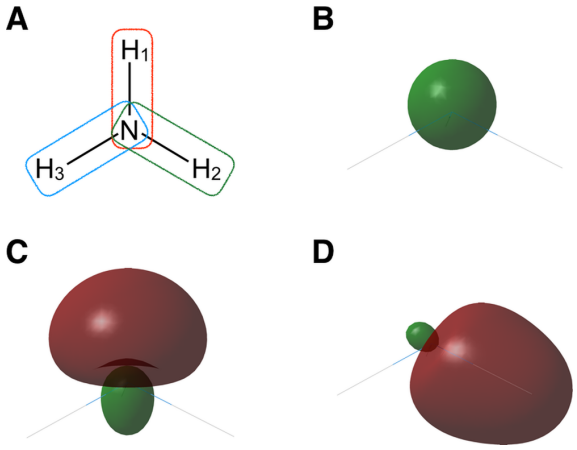


Figure 1. (A) Localization scheme associated with the Lewis structure of the ammonia molecule, with the three overlapping fragments N-H explicitly shown (the atomic fragment N is not shown for the sake of clarity); (B) ELMO describing the core electrons of the nitrogen atom, (C) ELMO describing the lone-pair electrons of the nitrogen atom, (D) ELMO describing one of the N-H bonds of the molecule (basis-set cc-pVDZ and 0.2 a.u. isosurfaces for all the orbitals).

Rotation of the ELMOs. In the Introduction, we have stressed that, due to their strict localization, ELMOs are orbitals reliably transferable from molecule to molecule. In order to transfer an ELMO from a model molecule (namely, the molecule on which it is originally determined) to a target system that we want to study, it is necessary to define a proper rotation matrix that transforms the ELMO coefficients originally determined on the geometry of the model molecule to new coefficients that are consistent with the geometry of the target system. The way of obtaining this rotation matrix has been suggested by Philipp and Friesner in the framework of QM/MM techniques with boundary regions described by strictly localized bond orbitals.[139] Since this strategy has been already reviewed in details in our first paper about the ELMOs transferability,[59] here we will only show its essential points.

Following Philipp and Friesner, the definition of a proper rotation matrix for the coefficients of an ELMO needs the definition of two reference frames (see Figure 2): one in the model molecule (here indicated as (**a**, **c**, **d**)) and one in the target molecule (here indicated as (**a**', **c**', **d**')), each of them resulting from the choice of a triad of atoms that guarantees the uniqueness of the rotation. In particular, given the two corresponding triads (A_1, A_2, A_3) and (A_1', A_2', A_3'), the reference frames (**a**, **c**, **d**) and (**a**', **c**', **d**') are defined like this: **a** (**a**') is the position vector of A_2 (A_2') relative to A_1 (A_1') (see Figure 2), while **c** (**c**') and **d** (**d**') are given by the following vector products:

$$\begin{cases} \mathbf{c} = \mathbf{a} \times \mathbf{b} & (\mathbf{c}' = \mathbf{a}' \times \mathbf{b}') \\ \mathbf{d} = \mathbf{c} \times \mathbf{a} & (\mathbf{d}' = \mathbf{c}' \times \mathbf{a}') \end{cases} \quad (7)$$

where **b** (**b'**) is the position vector of A_3 (A_3') with respect to A_1 (A_1') (see again Figure 2).

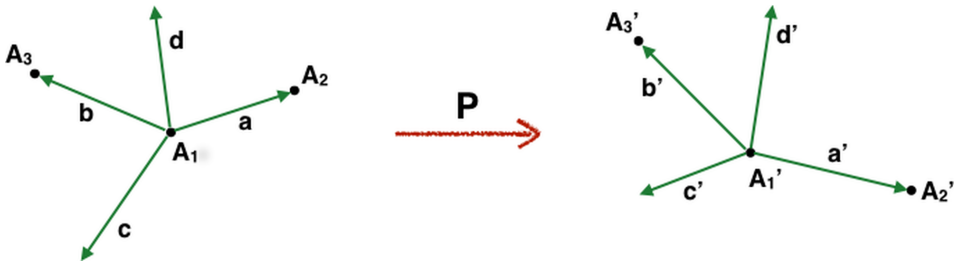


Figure 2. Reference frames and triads of atoms for the definition of the ELMO rotation from the geometry of the model molecule (left) to the geometry of the target system (right).

Concerning the choice of the triads of atoms, the following simple rules are followed. If an ELMOs is localized on only one atom (i.e., ELMO describing core or lone-pair electrons), the triad is given by the atom of interest and by two other atoms, which are generally those bonded to the one on which the ELMO is localized. When an ELMO describes a bond, the triad is obviously given by the two atoms involved in the bond

along with a third one that must represent the local dissymmetry of the bond under exam.[141] Finally, when we have three-center ELMOs, which are the most delocalized type of orbitals stored in our databanks (see Section 3), the triads of atoms are automatically defined. If we had an ELMO localized on more than three atoms, it would obviously impossible to define a triad (and, consequently, a reference frame) that simultaneously takes into account the orientation of all the atoms of the subunit. For this reason, as it will be described in the next section, all the ELMOs stored in the current libraries are localized at the largest on three atoms.

After defining the two reference frames, the fundamental step to obtain the ELMO rotation matrix consists in determining the matrix **P** associated with the rotation from reference frame (**a**, **c**, **d**) to reference frame (**a**', **c**', **d**'). Using this transformation matrix is actually possible to construct the matrices that rotate the different types of basis functions and the associated ELMO coefficients. In fact, excluding the *s*-type orbitals, which obviously remain unchanged due to their spherical symmetry, the *p*-type basis functions (and the related coefficients) transform exactly according to rotation matrix **P**, while basis functions (and corresponding coefficients) with angular momentum greater than 1 transform according to rotation matrices that can be easily expressed in terms of **P**.

Finally, for the sake of completeness, it is important to note that the method proposed by Philipp and Friesner works only if ELMOs are expanded in terms of Cartesian Gaussian basis functions. Furthermore, since the strategy does not automatically take into account the variations in the values of the overlap integrals of the basis functions due to changes of bond lengths and angles in the target molecule, suitable ELMOs renormalizations must be performed after each rotation.

3. Structure of the ELMO-libraries

General features. The current ELMO-libraries have been constructed with the goal of proposing a new tool for the instantaneous reconstruction of approximate wave functions and electron densities of molecules ranging from small polypeptides to very large proteins. For this reason, they cover all the possible fragments of the water molecule and of the twenty natural amino acids in all their possible protonation states and forms (namely, N-terminal, non-terminal and C-terminal forms). Most of the molecular orbitals in the databanks are localized on one-atom and two-atom fragments, which describe core/lone-pair electrons and bond electrons, respectively. Nevertheless, in order to properly describe situations in which it is very important to deal with the delocalized nature of the electronic structure, we have also considered molecular orbitals localized on three-atom subunits. In particular, in the databases we have included: *i)* ELMOs localized on O-C-O fragments to correctly describe σ and π electrons of the carboxylate groups in all the C-terminal residues and in glutamate (see Figure 3A); *ii)* ELMOs localized on O-C-N subunits to properly treat the eight electrons of the peptide bonds and of the amide bonds in asparagine and glutamine (i.e., the electrons associated with the two σ bonds, with the C-O π bond and with the delocalized lone-pair of the nitrogen atom; see Figure 3B); *iii)* ELMOs localized on C-C-C, C-N-C, C-C-N, N-C-N and N-C-C fragments to describe each delocalized π electron pair in the aromatic rings of the phenylalanine, histidine, tryptophan and tyrosine residues (see Figure 3C); *iv)* ELMOs localized on fragment N-C-N of arginine to deal with the eight delocalized electrons corresponding to the guanidino group of that amino acid (see Figure 3D).

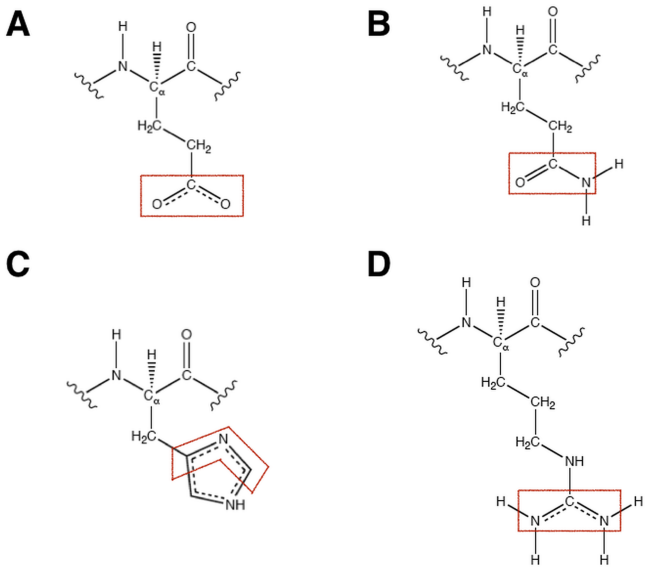


Figure 3. Examples of three-atom fragments considered in the ELMO-libraries: (A) O-C-O subunit to describe the σ and π electrons of the carboxylate group in glutamate; (B) O-C-N fragment to describe the eight electrons involved in the amide bond of glutamine; (C) C-N-C subunit to describe one of the π electron pairs of the imidazole ring of histidine; (D) N-C-N fragment to describe the eight (σ and π) electrons of the guanidino group of arginine.

It is important to point out that, if, on the one hand, determining ELMOs localized on more than three atoms would be quite straightforward, on the other hand, for the reasons explained in the Theory section, their exact rotation/transfer to the target system would not be possible, which is fundamental for our final goal of refining crystallographic structures. In case of a fragment constituted by four or more atoms, the only possibility to circumvent the problem would consist in storing (in the libraries) ELMOs for each possible combination of the values (not only the minima values) of all the dihedral angles of the fragment in exam, which is not practically feasible. Furthermore, in relation to the choice of describing the π electron systems of the aromatic rings by means of ELMOs localized on three-atom fragments, it is worth noting that, although the most preferable option would be to treat them through orbitals completely delocalized on the whole aromatic unit, one has also to bear in

mind that in experimental crystallographic geometries (which are the main targets of the present libraries) aromatic rings are not completely planar. Therefore, if we exploited ELMOs "delocalized" over the whole rings and previously determined on planar (theoretically optimized) molecular geometries, the rotations/transfers would not be optimal. For all these reasons a compromise has been found and, for all the aromatic rings of the twenty natural amino acids, the ELMOs describing the π electron pairs have been localized on three atoms.

At the moment, the ELMO-libraries are available in five standard basis-sets of quantum chemistry: 6-31G, 6-311G, 6-31G(d,p), 6-311G(d,p) and cc-pVDZ. We are currently extending the databanks to other traditional sets of basis functions of theoretical chemistry.

Construction of the databanks. The ELMOs stored in the databases have been previously obtained by performing traditional ELMO calculations on model molecules properly designed according to the Nearest Functional Group Approximation (NFGA), which resulted as the most reliable among all the model-molecules approximations that we have examined in our preliminary study on the ELMOs transferability.[59] The only exception was for the computation of the ELMOs describing the peptide bonds, for which we have considered a model molecule (see model molecule A in Figure S1 of the Supporting Information) obtained with the simpler Nearest Bond Approximation (NBA).

For the sake of completeness, we remind that, given a fragment for which we want to compute the corresponding extremely localized molecular orbitals, the NFGA model molecule is obtained by considering the fragment of interest and its nearest neighbor functional groups properly capped with hydrogen atoms, while the NBA model

molecule results from the fragment of interest and its nearest neighbor bonds always saturated with hydrogen atoms.

Overall, for the construction of our new ELMO libraries, we have considered 115 model molecules (see Figure S1 and XYZ files in the Supporting Information), whose geometries have been preliminarily optimized at B3LYP/6-311++G(d,p) level using the Gaussian 09 quantum chemistry package.[142] The optimized geometries have been afterwards used to compute the ELMOs that we stored in the current databanks. All the ELMO calculations have been carried out by exploiting version 8 of the GAMESS-UK suite of programs[143] that has been properly modified to introduce the implementation of the ELMO method proposed by Stoll.

Organization of the libraries. In our ELMO-libraries we have decided to resort to the AMBER[144] nomenclature for residues. This allowed us to easily assign a well-defined and characteristic label to all the amino acids in all their possible protonation states and bonding situations (e.g., cysteine involved in disulfide bridges). For example, for histidine, instead of using the only label HIS, which is usually found in PDB files, we have considered three different labels, each of them corresponding to a specific protonation state of the residue (see Figure 4): *i)* HID, to indicate a histidine residue with hydrogen on nitrogen δ; *ii)* HIE, to indicate a histidine residue with hydrogen on nitrogen ε and *iii)* HIP, to indicate a positively charged histidine residue with hydrogens on both the nitrogen atoms. Furthermore, always according to the AMBER convention, we have added an "N" and "C" prefix to the labels for the non-terminal amino acids in order to define the labels for the N-terminal and C-terminal residues, respectively. For instance, the label for N-terminal alanine is NALA, while the label for C-terminal alanine is CALA. Concerning the names for the different atom-types of the standard amino acids, we have adopted the same convention of the PDB files,

which is also the one used in the AMBER package.[144] The adopted atom-names have been consequently exploited to set up the labels for the ELMO-fragments of each residue.

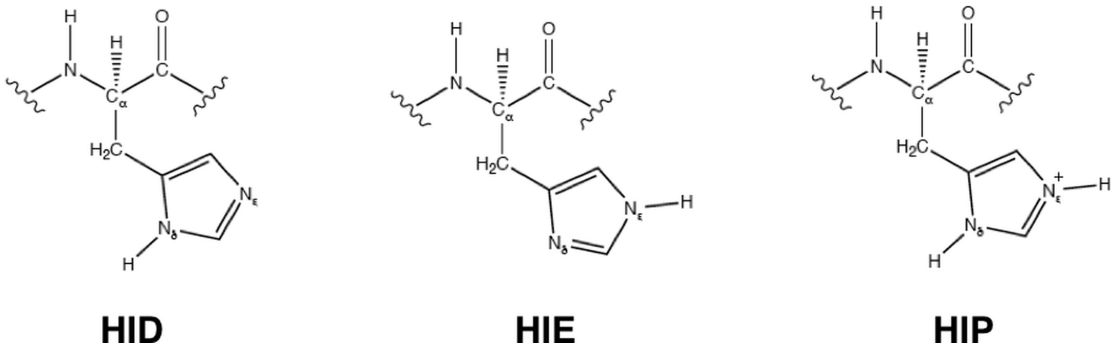


Figure 4. Histidine in its three possible protonation states, each of them characterized by a specific label.

To better explain the organization of the ELMO-databases, in Table 1 we have reported all the considered fragments for the arginine residue in its non-terminal form (ARG). First of all, it is worth noting that one-, two- and three-atom fragments have been considered: the one-atom subunits correspond to ELMOs that describe all the core and lone-pair electrons in the arginine residue, the two-atom fragments are associated with ELMOs that describe the ordinary bonds of the amino acid, while the three-atom fragments are related to ELMOs that describe the peptide bonds in which the residue is involved (fragment C_O_N) and to ELMOs that deal with the guanidino group of the amino acid (fragment NH1_CZ_NH2). Each subunit is also associated with a capital letter that simply indicates the model molecule on which the ELMOs of the subunit have been previously computed. Furthermore, we can observe that, for each fragment, we have two triads of atoms that allow the definition of the matrix for the transfer/rotation of the ELMOs localized on the fragment. As explained in Section 2, one triad is composed of three atoms in the model molecule (for instance, for

fragment CA_CB, atoms C6, C13 and C16 in model molecule E), while the other triad consists of three atoms in the target molecule (for example, always for fragment CA_CB, atoms CA, CB and CG in each arginine residue of the target polypeptide/protein). In this regard it is important to observe that, in the new ELMO-libraries, for each fragment we have not stored only the corresponding extremely localized molecular orbitals, but also the coordinates of the atoms that constitute the corresponding atomic triad in the model molecule. For the sake of completeness, the complete lists of fragments for the water molecule and for all the amino acids in all their possible protonation states and bonding situations is given in the Supporting Information, where, as in Table 1, for each subunit we have indicated both the two triads of atoms that enable to define the ELMO rotation matrix and the model molecules on which the corresponding ELMOs have been previously determined.

The ELMOdb program. The automatic transfer of ELMOs from the new libraries to target protein/polypeptide structures has been efficiently implemented in our in-house program *ELMOdb*. In its current version, *ELMOdb* requires an input PDB file suitably pre-processed through the *tleap* or *xleap* modules of the AMBER Molecular Dynamics package.[144] In particular, given a deposited PDB file (e.g., a PDB file stored in the Protein Data Bank), before performing the real ELMOs transfer, it is necessary *i)* to select only one of the possible conformers for the disordered parts of the polypeptide/protein, *ii)* to specify the protonation states of the residues by assigning them the proper AMBER residue-labels (as indicated in the previous subsection) and, finally, *iii)* to run *tleap* or *xleap* in order to add missing hydrogen atoms (only if necessary) and to prepare the PDB file in a suitable format to be read from *ELMOdb*.

Table 1. List of the all the fragments for the arginine residue in its non-terminal form, each of them with the label of the model molecule on which the corresponding ELMOs have been computed and with the two triads of atoms for the definition of the ELMO rotation matrix.

Fragment	Model Molecule	Model Molecule Triad	Target Molecule Triad
CA	C	N1_C6_C8	N_CA_C
N	C	C6_N1_H2	CA_N_H
C	C	C6_C8_O9	CA_C_O
O	D	C5_O6_N7	C_O_N
CB	E	C6_C13_C16	CA_CB_CG
CG	F	C1_C5_C8	CB_CG_CD
CD	G	C7_C4_N1	CG_CD_NE
NE	H	C1_N4_C6	CD_NE_CZ
CZ	I	N1_C3_N4	NH1_CZ_NH2
NH1	I	C3_N1_H2	CZ_NH1_HH11
NH2	I	C3_N4_H6	CZ_NH2_HH21
CA_HA	C	C6_H7_C8	CA_HA_C
CA_N	C	C6_N1_H2	CA_N_H
N_H	C	N1_H2_C6	N_H_CA
CA_C	C	C6_C8_O9	CA_C_O
C_O_N	A	C1_O2_N3	C_O_N
CA_CB	E	C6_C13_C16	CA_CB_CG
CB_HB2	E	C13_H14_C16	CB_HB2_CG
CB_HB3	E	C13_H14_C16	CB_HB3_CG
CB_CG	J	C13_C16_C19	CB_CG_CD
CG_HG2	F	C5_H6_C8	CG_HG2_CD
CG_HG3	F	C5_H6_C8	CG_HG3_CD
CG_CD	K	C7_C4_N1	CG_CD_NE
CD_HD2	G	C4_H5_N1	CD_HD2_NE
CD_HD3	G	C4_H5_N1	CD_HD3_NE
CD_NE	H	C1_N4_C6	CD_NE_CZ
NE_HE	H	N4_H5_C6	NE_HE_CZ
NE_CZ	H	N4_C6_N8	NE_CZ_NH1
NH1_CZ_NH2	I	N1_C3_N4	NH1_CZ_NH2
NH1_HH11	I	N1_H2_C3	NH1_HH11_CZ
NH1_HH12	I	N1_H2_C3	NH1_HH12_CZ
NH2_HH21	I	N1_H2_C3	NH2_HH21_CZ
NH2_HH22	I	N1_H2_C3	NH2_HH22_CZ

Afterwards, the program analyzes the new PDB file residue by residue. For each of them, it systematically processes one fragment at a time by retrieving in the libraries both the stored ELMOs and the stored atomic coordinates of the model molecule triad. Then, after identifying the atomic coordinates of the target molecule triad in the PDB file, the rotation matrix for the ELMOs of the fragment under exam is constructed and the extremely localized molecular orbitals are instantaneously transferred to the target system. As already mentioned in the Theory section, in order to take into account the differences of bond lengths and angles between the model and target molecules, the transferred ELMOs are properly renormalized after the transfer.

Since the current ELMO libraries cover only fragments for the water molecule and the twenty natural amino acids, the *ELMOdb* program has been also conceived in order to possibly read tailor-made ELMOs for special fragments, ligands or solvent molecules that may be present in the PDB files of the systems under investigation. Of course, the customized ELMOs must be previously computed on suitable model molecules (or even on the same molecule of interest) and then properly stored (along with the geometries of the corresponding model molecules triads) in a dedicated folder from which the *ELMOdb* program can read them.

Finally, in output, other than producing a text file containing very general information, *ELMOdb* provides the binary file of the rotated ELMOs (in other words, the approximate wave function for the target molecule), the binary file of the ELMO one-electron density matrix (which is obviously related to the approximate electron density of the target system) and, above all, a complete Gaussian[142] formatted checkpoint file that can be used to perform subsequent analyses or calculations.

4. Test calculations

In order to test the capabilities and performances of the new ELMO-libraries and of the associated *ELMOdb* program, we have performed preliminary test calculations mainly aiming at *i)* evaluating the time-scaling of the ELMOs transfer and *ii)* comparing the results obtained through the transfer of ELMOs to those obtained by means of traditional quantum chemical methods. The outcomes of these investigations will be presented in this section. At first we will discuss the results of the tests that have been performed to determine the CPU times strictly associated with the transfer of extremely localized molecular orbitals to polypeptides of increasing sizes. Then we will analyze the comparisons of charges, electron densities, electrostatic potentials and density matrices obtained at ELMO, Hartree-Fock and DFT levels for a series of eight oligopeptides. Finally, we will show the first applications of the new ELMO-libraries to proteins.

Scaling of the ELMOs transfer. To evaluate the scalability associated with the transfer of extremely localized molecular orbitals, for each natural amino acid we have considered linear homopeptides constituted by a number of residues ranging from 5 to 50. The PDB files for all these linear polypeptides have been obtained by using the option *sequence* of the *tleap* module in AMBER and, afterwards, they have been used as input files for the *ELMOdb* program. In order to perform the most stringent stress-tests for the new databases, all the transfers have been performed considering the 6-311G(d,p) basis-set, which is the largest set of basis functions currently available for the ELMO-databanks. For all the computations we have evaluated the CPU time strictly associated with the transfer of extremely localized molecular orbitals.

The results of these test calculations are reported in Table 2. It is easy to observe that, for all the amino acids, the real transfer of ELMOs from the libraries is really instantaneous, with the overall CPU time that never exceeds 0.5 seconds, even for some of the largest systems considered in our investigation, such as the poly-tryptophan and the poly-tyrosine constituted by 50 amino acids, for which the number of basis functions is 16331 and 14131, respectively.

Table 2. CPU times (in s) associated with the transfer of ELMOs from the ELMO-library for basis-set 6-311G(d,p) to linear homopeptides of increasing size for each type of amino acid.

Amino Acid	Number of Residues									
	5	10	15	20	25	30	35	40	45	50
Ala	0.013	0.015	0.022	0.033	0.039	0.050	0.057	0.069	0.080	0.089
Arg	0.023	0.037	0.059	0.080	0.106	0.137	0.174	0.207	0.218	0.257
Asn	0.020	0.024	0.034	0.046	0.059	0.077	0.088	0.108	0.129	0.148
Asp	0.022	0.020	0.029	0.042	0.056	0.071	0.086	0.099	0.118	0.136
Cys	0.017	0.019	0.025	0.038	0.048	0.059	0.071	0.087	0.101	0.114
Gln	0.028	0.030	0.059	0.083	0.102	0.124	0.153	0.153	0.244	0.244
Glu	0.026	0.031	0.055	0.084	0.090	0.095	0.133	0.136	0.217	0.195
Gly	0.014	0.016	0.026	0.039	0.044	0.046	0.047	0.087	0.068	0.087
His	0.043	0.040	0.082	0.066	0.116	0.156	0.142	0.266	0.190	0.238
Ile	0.023	0.028	0.042	0.063	0.077	0.098	0.186	0.174	0.211	0.214
Leu	0.038	0.044	0.070	0.085	0.104	0.109	0.167	0.182	0.215	0.243
Lys	0.042	0.043	0.098	0.080	0.136	0.151	0.146	0.172	0.298	0.255
Met	0.032	0.036	0.042	0.079	0.098	0.101	0.163	0.184	0.196	0.215
Phe	0.038	0.052	0.081	0.108	0.135	0.229	0.210	0.247	0.227	0.256
Pro	0.023	0.037	0.040	0.048	0.061	0.076	0.079	0.093	0.110	0.148
Ser	0.015	0.018	0.027	0.039	0.048	0.063	0.072	0.084	0.092	0.110
Thr	0.020	0.022	0.036	0.050	0.063	0.077	0.093	0.113	0.128	0.125
Trp	0.046	0.064	0.086	0.130	0.194	0.204	0.246	0.285	0.342	0.330
Tyr	0.035	0.041	0.066	0.096	0.114	0.146	0.172	0.201	0.252	0.311
Val	0.018	0.020	0.037	0.046	0.062	0.092	0.109	0.105	0.127	0.167

Notwithstanding the impressive scaling of the ELMOs transfer, it is worth mentioning that, when ELMOs are used to compute electron densities (even when ELMOs are transferred), the most time-consuming step is the calculation of the one-electron density matrix. In fact, as mentioned in the Theory section, ELMOs are non-orthogonal orbitals and, due to this fact, the computation of the one-electron density matrix **P** requires the inversion of the overlap matrix **S** of the occupied molecular orbitals, as shown in the following equation:

$$P_{\mu\nu} = \sum_{i,j=1}^{N} C_{\mu j}\, S_{ji}^{-1}\, C_{\nu i} \quad (8)$$

where N is the number of occupied molecular orbitals, while $C_{\mu j}$ and $C_{\nu i}$ are coefficients of the molecular orbitals expansions in the adopted global basis-set. For very large systems, the inversion of matrix **S** might become computationally demanding. As we will show in one of the next subsections, the first applications of the new databanks to proteins have confirmed that the most expensive step in the reconstruction of wave functions and electron densities through the transfer of ELMOs is indeed the calculation of the one-electron density matrix. Nevertheless, we have also observed that its computational cost is not as unfavorable as we would expect. Furthermore, to further and decisively speed up the reconstruction of wave functions and electron densities by transferring ELMOs from the recently constructed databanks, we have already envisaged the introduction of one of the cost-effective and currently available linear scaling algorithms/techniques[145-152] for the quick inversion of large overlap matrices. These algorithms have been known for a long time and more and more efficient strategies are continuously proposed. Therefore, in reason of the algorithmic methodologies that are nowadays at our hands, the necessity of

inverting large overlap matrices of occupied ELMOs will not at all prevent the application of the current ELMO libraries to very large proteins.

Comparison to traditional quantum chemistry methods. To further evaluate the performances of the ELMO-libraries we have afterwards decided to compare charges, electron densities, electrostatic potentials and density matrices obtained through the transfer of ELMOs with those resulting from more traditional Hartree-Fock and DFT calculations. The reason for the comparison to Hartree-Fock and DFT results originates from the fact that, as mentioned in the Introduction, one of the possible future applications of the ELMO-databanks could be their coupling with the Hirshfeld atom refinement. In fact, HAR is a quantum crystallographic refinement technique that requires a quantum chemical calculation at each iteration, and, to the best of our knowledge, so far only Hartree-Fock and DFT computations have been carried out in this context.

To perform these comparisons we have considered eight oligopeptides (three pentapeptides and five hexapeptides), whose geometries have been extracted from PDB files downloaded from the Protein Data Bank: 1BC5, 1BXX, 1JW6, 2D5W, 3FG5, 3FOD, 3OW9 and 3WNE. In each of them we kept only the atomic coordinates of the polypeptide to be investigated, removing, when necessary, all the coordinates of all the other small molecules (water molecules, ions, other possible ligands, *etc.*) and those of the protein that interacts with the oligopeptide under exam. Afterwards, as already discussed in the subsection describing the *ELMOdb* program, the resulting PDB files have been further pre-processed *i)* by choosing only one of the possible conformers (chain C in 2D5W, chain H in 3FOD, chain B in 3OW9 and chain C in 3WNE), *ii)* by assigning the correct protonation states to the different residues according to the reported pH of crystallization and *iii)* by adding the missing

hydrogen atoms through *tleap*. After the pre-processing, the geometries in the final PDB files have been used as input structures to perform ELMOs transfers and to carry out Hartree-Fock and DFT (B3LYP functional) calculations with the 6-31G, cc-pVDZ and 6-311G(d,p) basis-sets. The results obtained with the different methods have been eventually compared exploiting global similarity indexes.

At first, our comparisons focused on different types of quantum chemistry-based point charges resulting from the calculations. In particular, we have considered both the Mulliken[153] and the Merz-Singh-Kollman charges,[154,155] the former somehow related to the electron density and the latter more connected with the electrostatic potential. Furthermore, since all the investigated polypeptides are constituted by a very large number of atoms, we have decided to resort to a descriptor that would allow us to globally evaluate the performances of the ELMOs transfer as compared to the two traditional quantum chemistry methods taken into account. For this reason, for each polypeptide and basis-set, we have computed root-mean-square deviations (RMSDs) between the different sets of charges, always using the Hartree-Fock (HF) one as reference. The RMSDs obtained for the Mulliken charges are reported in Table 3, while the RMSDs for the Merz-Singh-Kollman ones are shown in Table 4.

Table 3. Root-mean-square deviations of the Mulliken charges (in e) using the Hartree-Fock values as references.

Peptide	6-31G		cc-pVDZ		6-311G(d,p)	
	B3LYP	ELMO	B3LYP	ELMO	B3LYP	ELMO
1BC5	0.134	0.105	0.098	0.175	0.068	0.119
1BXX	0.136	0.103	0.093	0.181	0.069	0.118
1JW6	0.120	0.094	0.088	0.151	0.059	0.098
2D5W	0.116	0.115	0.082	0.142	0.062	0.111
3FG5	0.114	0.111	0.082	0.161	0.059	0.112
3FOD	0.124	0.103	0.088	0.158	0.072	0.111
3OW9	0.116	0.109	0.086	0.153	0.064	0.114
3WNE	0.137	0.101	0.095	0.176	0.071	0.116

Analyzing Table 3, we can observe that, for the 6-31G basis-set, the ELMO Mulliken charges are always globally closer to the to the Hartree-Fock ones than the Mulliken charges obtained from the B3LYP calculations, even if the RMSDs are roughly of the same order of magnitude. The trend is opposite for the cc-pVDZ and 6-311G(d,p) basis-sets, for which the agreement between Hartree-Fock and DFT Mulliken charges is definitely better than the agreement between Hartree-Fock and ELMO charges for all the examined polypeptides. Concerning the Merz-Singh-Kollman charges (see Table 4), for all the basis-sets, the B3LYP root-mean-square deviations are almost always lower than the ELMO ones. The only situation in which the charges obtained at ELMO level are in a better agreement with the Hartree-Fock ones is the case of polypeptide 3FOD when basis-set 6-31G is used. For the same polypeptide, the ELMO and DFT methods give very similar RMSDs when the cc-pVDZ basis-set is used.

Table 4. Root-mean-square deviations of the Merz-Singh-Kollman charges (in e) using the Hartree-Fock values as references.

Peptide	6-31G		cc-pVDZ		6-311G(d,p)	
	B3LYP	ELMO	B3LYP	ELMO	B3LYP	ELMO
1BC5	0.111	0.120	0.101	0.132	0.056	0.142
1BXX	0.098	0.122	0.087	0.134	0.063	0.143
1JW6	0.082	0.112	0.072	0.124	0.049	0.133
2D5W	0.080	0.094	0.068	0.100	0.051	0.105
3FG5	0.079	0.098	0.073	0.115	0.052	0.116
3FOD	0.100	0.091	0.092	0.095	0.072	0.096
3OW9	0.092	0.108	0.086	0.119	0.067	0.124
3WNE	0.080	0.095	0.069	0.107	0.053	0.110

Therefore, the data collected in Tables 3 and 4 seem to indicate that, as expected, the transfer of ELMOs is only an approximation and that the agreement between the results of Hartree-Fock and DFT calculations is generally better than the agreement between the results of Hartree-Fock computations and ELMOs transfers. Nevertheless, it is also worth noting that the obtained RMSDs values are relatively small. Furthermore, both Mulliken and Merz-Singh-Kollman charges are only point values and more global comparisons would be necessary to draw final and more general conclusions.

To accomplish this task we have thus decided to directly compare the electron densities and electrostatic potentials obtained with the different strategies and basis-sets taken into account. These comparisons have been carried out by exploiting three real-space similarity measures: *i)* the traditional Carbó similarity index,[156] *ii)* the Carbó Euclidean distance[156] and *iii)* the more traditional root-mean-square deviation.

The traditional Carbó similarity index between two molecular electron densities $\rho_I(\boldsymbol{r})$ and $\rho_J(\boldsymbol{r})$ is defined like this:

$$C_{IJ} = Z_{IJ}\left[Z_{II}Z_{JJ}\right]^{-1/2} \qquad (9)$$

where Z_{IJ} (and, in analogous way Z_{II} and Z_{JJ}) is given by the following integral:

$$Z_{IJ} = \int d\boldsymbol{r}\ \rho_I(\boldsymbol{r})\,\rho_J(\boldsymbol{r}) \qquad (10)$$

Of course, the similarity index C_{IJ} assumes values in the interval $[0, 1]$ and the complete similarity occurs when C_{IJ} is equal to 1.

On the other hand, the Carbó Euclidean distance between two electron densities is defined in this way

$$D_{IJ} = \left[Z_{II} + Z_{JJ} - 2\,Z_{IJ}\right]^{1/2} \qquad (11)$$

and its values can be found in the interval $[0, +\infty]$. Obviously, the higher the value of the index is, the less similar the two distributions are.

We also remind that the root-mean-square deviation between two charge distributions is defined as:

$$RMSD\left(\rho_I, \rho_J\right) = \left[\frac{\sum_{i=1}^{n_p}\left(\rho_I(\boldsymbol{r}_i) - \rho_J(\boldsymbol{r}_i)\right)^2}{n_p}\right]^{1/2} \qquad (12)$$

where n_p is the number of points of the grids on which the electron densities have been evaluated. The same similarity measures have been also used for the comparisons of the obtained molecular electrostatic potentials. The only difference was that, in those cases, only grid-points with corresponding electron density values lower than or equal to 0.001 e/bohr^3 have been taken into account in the comparison.

Furthermore, it is worth noting that all these similarity indexes have been evaluated by considering suitable three-dimensional grids (both for the electron densities and for the electrostatic potentials) characterized by a 0.08 bohr step-size for each direction

and by global sizes that allow to recover the correct numbers of electrons of the investigated systems when the electron densities are integrated numerically.

At a first stage we have considered the obtained electron densities and, always using the Hartree-Fock charge distributions as references, we have computed the values of the traditional Carbó similarity index given by equation (9). The results are reported in Table 5, where we can immediately observe that, for all the polypeptides and for all the basis-sets used in the calculations, the similarities among the charge distributions are very high. This is further confirmed by the values of the global RMSDs (always referred to the Hartree-Fock electron densities; see Table 6), which are mostly of the order of 10^{-4} e/bohr^3. Furthermore, and more importantly, in Tables 5 and 6 we can also note that in many cases, unlike what we have observed for the point charges, the similarity between Hartree-Fock and ELMO electron densities is higher than the similarity between Hartree-Fock and B3LYP charge distributions. The same trend can be noted also in Table S1 of the Supporting Information, where we have reported the values of the Euclidean Carbó distances D_{IJ}. This index can be rather considered as a dissimilarity measure and, for this reason, it also allowed to better highlight the global discrepancies between the examined electron densities.

Table 5. Values of the Carbó similarity index between the Hartree-Fock, B3LYP and ELMO electron densities. The Hartree-Fock electron distributions were used as references.

Peptide	6-31G		cc-pVDZ		6-311G(d,p)	
	B3LYP	ELMO	B3LYP	ELMO	B3LYP	ELMO
1BC5	0.999993	0.999993	0.999989	0.999991	0.999985	0.999991
1BXX	0.999994	0.999994	0.999990	0.999992	0.999985	0.999992
1JW6	0.999995	0.999995	0.999991	0.999994	0.999987	0.999994
2D5W	0.999994	0.999993	0.999989	0.999991	0.999985	0.999992
3FG5	0.999994	0.999991	0.999990	0.999989	0.999985	0.999989
3FOD	0.999993	0.999994	0.999989	0.999992	0.999984	0.999993
3OW9	0.999993	0.999993	0.999988	0.999991	0.999983	0.999992
3WNE	0.999994	0.999993	0.999990	0.999991	0.999985	0.999991

Table 6. Values of the RMSDs (in $\times 10^{-4}$ e/bohr^3) between the Hartree-Fock, B3LYP and ELMO electron densities. The Hartree-Fock electron distributions were used as references.

Peptide	6-31G		cc-pVDZ		6-311G(d,p)	
	B3LYP	ELMO	B3LYP	ELMO	B3LYP	ELMO
1BC5	8.116	7.909	11.008	9.140	11.921	8.836
1BXX	6.888	6.826	9.457	7.831	10.587	7.655
1JW6	7.117	6.643	9.962	7.564	11.047	7.377
2D5W	6.831	7.320	9.684	8.251	10.843	8.042
3FG5	7.088	8.418	10.051	9.464	11.143	9.330
3FOD	8.902	8.248	12.118	9.517	13.555	9.220
3OW9	9.036	8.329	12.238	9.528	13.397	9.292
3WNE	7.979	8.160	11.216	9.528	12.509	9.418

In analogous way, the values of the traditional Carbó similarity measure and of the root-mean-square deviations have been afterwards evaluated also to quantitatively compare the obtained molecular electrostatic potentials. The results are shown in Tables 7 and 8 and, also in this situation, they indicate high similarities among the considered quantities, although slightly lower compared to the case of the electron densities. Furthermore, also in this case it is important to note that the similarity between Hartree-Fock and ELMO electrostatic potentials is not necessarily worse than the one between Hartree-Fock and DFT electrostatic potentials. In Table S2 of the Supporting Information we have also reported the values for the Carbó Euclidean distances that better revealed the global dissimilarities between the examined distributions and that basically confirm the trends already observed in Tables 7 and 8.

Table 7. Values of the Carbó similarity index between the Hartree-Fock, B3LYP and ELMO molecular electrostatic potentials. The Hartree-Fock electrostatic potentials were used as references.

Peptide	6-31G		cc-pVDZ		6-311G(d,p)	
	B3LYP	ELMO	B3LYP	ELMO	B3LYP	ELMO
1BC5	0.993362	0.997077	0.991162	0.996588	0.999468	0.996575
1BXX	0.997979	0.991530	0.998135	0.990551	0.998546	0.989835
1JW6	0.999358	0.989149	0.999175	0.987393	0.999416	0.986332
2D5W	0.999871	0.998659	0.999876	0.998461	0.999759	0.998381
3FG5	0.999445	0.995265	0.999540	0.994326	0.999601	0.994042
3FOD	0.998195	0.999098	0.998079	0.998913	0.997862	0.99883
3OW9	0.986005	0.998787	0.982281	0.998619	0.982815	0.998544
3WNE	0.999699	0.996713	0.999737	0.996583	0.999751	0.996173

Table 8. Values of the RMSDs (in $\times 10^{-3}$ hartree) between the Hartree-Fock, B3LYP and ELMO molecular electrostatic potentials. The Hartree-Fock electrostatic potentials were used as references.

Peptide	6-31G		cc-pVDZ		6-311G(d,p)	
	B3LYP	ELMO	B3LYP	ELMO	B3LYP	ELMO
1BC5	10.302	7.362	11.617	7.921	2.848	7.989
1BXX	2.035	4.620	1.914	4.773	1.752	5.005
1JW6	4.215	7.174	5.198	7.615	1.845	7.910
2D5W	2.017	6.989	1.996	7.460	2.722	7.637
3FG5	2.036	6.199	1.820	6.639	1.697	6.835
3FOD	16.429	5.462	17.430	5.927	17.099	6.321
3OW9	18.797	6.393	20.315	6.772	19.864	6.983
3WNE	3.231	6.546	4.601	6.910	4.347	7.212

Finally, we have also compared the one-electron density matrices resulting from the three strategies taken into account. In particular, we have evaluated a sort of Euclidean distance between density matrices,[52] namely:

$$\Delta P = \frac{1}{M^2}\left[\sum_{i,j=1}^{M}\left(P_{ij}^{REF} - P_{ij}^{X}\right)^2\right]^{\frac{1}{2}} \quad (13),$$

where M is the number of basis functions, REF stands for "reference method" (in our case, Hartree-Fock) and X is the method to compare (in our case, B3LYP and ELMO).

The obtained values have been collected in Table 9. It is easy to observe that, for all the basis-sets and for all the polypeptides, the B3LYP density matrices are always closer to the Hartree-Fock ones. However, we can also note that, in general, the values of the ELMO/Hartree-Fock and B3LYP/Hartree-Fock distances are completely comparable and of the same order of magnitude.

Table 9. Values (multiplied by 10^3) of the distances between the Hartree-Fock, B3LYP and ELMO one-electron density matrices. The Hartree-Fock density matrices were used as references.

Peptide	6-31G		cc-pVDZ		6-311G(d,p)	
	B3LYP	ELMO	B3LYP	ELMO	B3LYP	ELMO
1BC5	3.697	5.950	2.064	2.485	0.775	2.027
1BXX	2.371	5.431	1.362	2.128	0.644	1.821
1JW6	2.846	4.655	1.683	1.944	0.618	1.612
2D5W	2.629	6.568	1.432	2.362	0.736	2.077
3FG5	2.764	5.642	1.589	2.237	0.669	1.879
3FOD	3.435	6.200	1.706	2.331	0.984	1.971
3OW9	3.425	4.982	1.747	1.929	0.933	1.625
3WNE	3.030	6.962	1.650	2.460	0.850	2.181

As expected, all the comparisons presented in this subsection confirm that the transfer of ELMOs provide approximate density matrices, electron densities, electrostatic potentials. Nevertheless, notwithstanding the unavoidable discrepancies, the obtained results also show quite good agreements among the different quantities that we have considered. For this reason we believe that the electron densities reconstructed through the transfer of ELMOs represent reliable approximations to the rigorous self-consistent charge distributions resulting from traditional quantum chemistry calculations. This result is especially important in view of using the new databanks for very large systems as proteins (see next subsection) and, in particular, in view of coupling them with HAR to perform structural refinements of macromolecules.

Application to proteins. The final preliminary tests on the ELMO-libraries consisted in their fist application to proteins. To accomplish this task we have considered

antifreeze protein RD1 (PDB code: 1UCS; 64 residues and 997 atoms) and the N-terminal domain of protein PEX14 (PDB code: 5L87; 62 residues and 1019 atoms). Also in this case, after keeping only the atomic coordinates of the proteins under exam, the PDB files have been further processed by selecting only the major conformers in the disordered parts of the proteins, by properly defining the protonation states of the different residues and by adding the coordinates of missing hydrogen atoms through *tleap*.

Using the geometries contained in the resulting PDB files, we have afterwards used the ELMO-libraries to reconstruct electron densities and electrostatic potential of the two proteins (see Figures 5 and 6). In Table 10 and 11, for each basis-set we have reported both the global CPU time associated with the calculation performed through the ELMO-databanks the and the CPU times associated with specific tasks of the *ELMOdb* program. For the sake of comparison, we have also shown the global CPU times of the corresponding serial Hartree-Fock calculations.

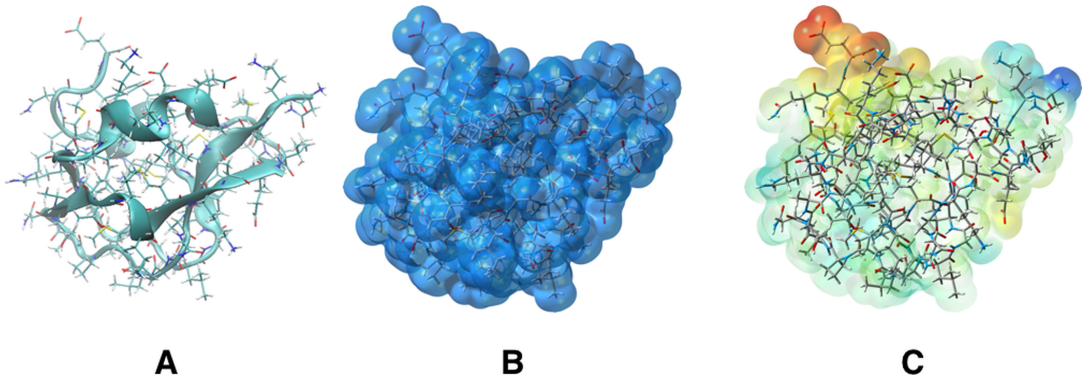


Figure 5. (A) structure of antifreeze protein RD1 from PDB file 1UCS with (B) its electron density (0.001 e/bohr^3 isosurface) and (C) its electrostatic potential (on the 0.001 e/bohr^3 electron density isosurface; red = -0.323 hartree and blue = 0.323 hartree) obtained through the transfer of extremely localized molecular orbitals from the libraries (cc-pVDZ basis-set).

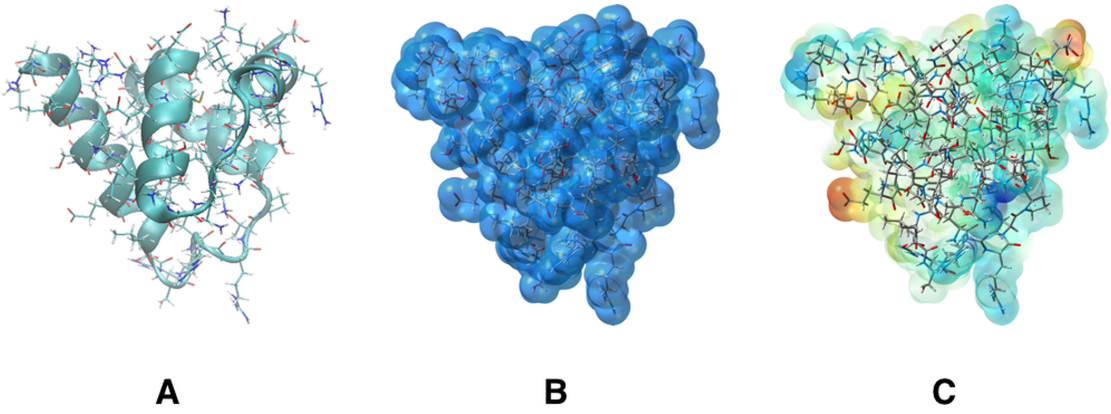


Figure 6. (A) structure of the N-terminal domain of protein PEX14 from PDB file 5L87 with (B) its electron density (0.001 e/bohr^3 isosurface) and (C) its electrostatic potential (on the 0.001 e/bohr^3 electron density isosurface; red = -0.100 hartree and blue = 0.168 hartree) obtained through the transfer of extremely localized molecular orbitals from the libraries (cc-pVDZ basis-set).

The results confirm that the transfer of ELMOs is really instantaneous and that, as anticipated above, the most time consuming step is always the computation of the one-electron density matrix. Nevertheless, it is worth noting that, although either linear scaling or parallel algorithms have been not introduced yet, the electron density reconstructions are quite fast, especially if compared with traditional quantum chemistry calculations at Hartree-Fock level. Again, this is obviously very promising and encouraging for the future application of the ELMO-libraries to the refinement of protein crystallographic structures and, in particular, for the possible coupling of the ELMO-databanks with the Hirshfeld atom refinement, for which an updated protein electron density would be necessary at each iteration of the procedure.

Table 10. Global and partial CPU times (in seconds) taken by the *ELMOdb* program in reconstructing approximate ELMO wave functions of antifreeze protein RD1 (PDB code: 1UCS) through transfers from the ELMO-libraries. For the sake of comparison, the global CPU times associated with the corresponding serial Hartree-Fock calculations are also shown.

Task of the *ELMOdb* program	Basis-set				
	6-31G	6-311G	6-31G(d,p)	6-311G(d,p)	cc-pVDZ
Overlap integrals computation	28.37	55.80	48.83	83.60	54.07
ELMOs transfer	0.21	0.26	0.31	0.24	0.26
Density matrix calculation	123.14	181.87	289.31	653.71	299.27
Formatted checkpoint file	18.15	40.26	65.11	102.68	89.02
Global *ELMOdb* CPU time	170.12	278.57	404.14	841.17	443.22
Hartree-Fock CPU time	21h 56m 8.2s	78h 32m 4.6s	112h 40m 49.2s	234h 11m 21.6s	150h 50m 5.9s

Table 11. Global and partial CPU times (in seconds) taken by the *ELMOdb* program in reconstructing approximate ELMO wave functions of the PEX14 N-terminal domain (PDB code: 5L87) through transfers from the ELMO-libraries. For the sake of comparison, the global CPU times associated with the corresponding serial Hartree-Fock calculations are also shown.

Task of the *ELMOdb* program	Basis-set				
	6-31G	6-311G	6-31G(d,p)	6-311G(d,p)	cc-pVDZ
Overlap integrals computation	30.11	64.84	52.96	91.79	60.34
ELMOs transfer	0.19	0.21	0.24	0.34	0.32
Density matrix calculation	145.89	257.77	386.91	601.82	441.05
Formatted checkpoint file	21.28	51.17	89.83	105.97	74.95
Global *ELMOdb* CPU time	197.66	374.54	530.58	800.98	577.26
Hartree-Fock CPU time	22h 9m 47.8s	70h 31m 29.5s	87h 42m 42.1s	195h 24m 21.4s	143h 12m 37.5s

In order to further test the capabilities of the new ELMO-databases, the protein electron densities, electrostatic potentials and density matrices obtained through the transfer of ELMOs have been compared to the corresponding ones obtained at Hartree-Fock level. As for the oligopeptides examined in the previous section, the similarities between electron densities and electrostatic potentials have been assessed through the evaluation of global similarity indexes (traditional Carbó index, Euclidean Carbó distance and global root-mean-square deviation) over three-dimensional grids. In this case, for all the electron densities we have still considered fine grids with a 0.08313 bohr step-size for each direction, while, due to the larger computational cost, for the electrostatic potentials we have increased the step-size to 0.166667 bohr. To compare the density matrices we have exploited again the sort of Euclidean distance defined in equation (13).

The results of the comparisons are reported in Tables 12, 13 and 14, respectively. Concerning the electron distributions, also for the considered protein structures the ELMO/Hartree-Fock agreement is quite high, although slightly lower than the one previously observed for the eight examined polypeptides. For instance, we can easily note that the root-mean-square deviations are of the order of 10^{-3} e/bohr^3, while, for the oligopeptides, the RMSDs were of the order of 10^{-4} e/bohr^3. Analyzing the results for the electrostatic potentials, although we have adopted coarser three-dimensional grids compared to the studies on the polypeptides and we cannot perform direct comparisons, we can anyway observe that the similarities between the Hartree-Fock and ELMO results remain quite high. Finally, considering the comparison of the density matrices, we can notice that the computed ELMO/Hartree-Fock distances are of the same order of magnitude of those previously obtained for the examined oligopeptides. In analogy with the previous results it is also possible to see that the

ELMO/Hartree-Fock discrepancies reduce when larger basis-sets are used. Therefore, the comparisons of electron densities, electrostatic potential and density matrices computed on proteins further confirm that the transfer of ELMOs from properly constructed libraries can be considered a useful strategy to rapidly obtain approximate but reliable electron densities, electrostatic potentials and density matrices of macromolecules.

Table 12. Values of the similarity indexes between the electron densities obtained at Hartree-Fock and ELMO-libraries levels for antifreeze protein RD1 (PDB code: 1UCS) and for the N-terminal domain of protein PEX14 (PDB code: 5L87).

Protein & Similarity indexes	Basis-set				
	6-31G	6-311G	6-31G(d,p)	6-311G(d,p)	cc-pVDZ
1UCS					
Traditional Carbó index	0.999989	0.999987	0.999989	0.999988	0.999988
Carbó Euclidean distance [(a)]	0.773	0.823	0.779	0.814	0.804
RMSD [(b)]	1.399	1.489	1.409	1.472	1.455
5L87					
Traditional Carbó index	0.999986	0.999985	0.999986	0.999984	0.999985
Carbó Euclidean distance [(a)]	0.800	0.846	0.806	0.839	0.833
RMSD [(b)]	1.593	1.684	1.605	1.671	1.658

(a) Values in e; (b) Values in 10^{-3} e/bohr^3.

Table 13. Values of the similarity indexes between the molecular electrostatic potentials obtained at Hartree-Fock and ELMO-libraries levels for antifreeze protein RD1 (PDB code: 1UCS) and for the N-terminal domain of protein PEX14 (PDB code: 5L87).

Protein & Similarity indexes	Basis-set				
	6-31G	6-311G	6-31G(d,p)	6-311G(d,p)	cc-pVDZ
1UCS					
Traditional Carbó index	0.993686	0.993284	0.993700	0.993819	0.993545
Carbó Euclidean distance [(a)]	3.544	3.710	3.658	3.813	3.796
RMSD [(b)]	6.403	6.705	6.611	6.889	6.859
5L87					
Traditional Carbó index	0.998609	0.998537	0.998558	0.998567	0.998532
Carbó Euclidean distance [(a)]	4.262	4.372	4.336	4.324	4.391
RMSD [(b)]	8.481	8.700	8.628	8.604	8.737

(a) Values in hartree×bohr^3; (b) Values in 10^{-3} hartree.

Table 14. Values (multiplied by 10^3) of the distances between the one-electron density matrices obtained at Hartree-Fock and ELMO-libraries levels for antifreeze protein RD1 (PDB code: 1UCS) and for the N-terminal domain of protein PEX14 (PDB code: 5L87).

Protein	Basis-set				
	6-31G	6-311G	6-31G(d,p)	6-311G(d,p)	cc-pVDZ
1UCS	1.771	1.431	0.704	0.583	0.735
5L87	1.749	1.415	0.704	0.569	0.749

5. Conclusions and perspectives

In this work we have introduced the new libraries of Extremely Localized Molecular Orbitals, whose construction has been encouraged by recent studies on the ELMOs

transferability.[59,60] At the moment the new ELMO-databanks cover all the possible fragments of the water molecule and of the twenty natural amino acids in all their possible protonation states and forms, thus allowing automatic and fast reconstructions of wave functions and electron densities of molecules ranging from small polypeptides to very large proteins.

The first preliminary tests have shown that the transfer of ELMOs is practically instantaneous and that the new ELMO-databases indeed allow to obtain approximate wave functions and electron densities of very large systems at a computational cost significantly lower than the computational cost associated with usual quantum chemistry calculations. This result is remarkable also in light of the fact either any linear scaling techniques or parallel algorithms have not been used yet for the computation of the ELMO one-electron density matrix, which, due the non-orthogonality of ELMOs, remains the most expensive step in the reconstruction of wave functions and electron densities by transferring extremely localized molecular orbitals from dedicated libraries. Therefore, the envisaged introduction of proper routines[145-152] for the efficient computation of the ELMO one-particle density matrix will further improve the time scaling of the strategy proposed in this paper.

The test calculations have also shown that, notwithstanding expected and unavoidable discrepancies, the electron densities and electrostatic potentials obtained through the transfer of ELMOs are generally in very good agreement with those obtained through more traditional methods. This clearly opens the possibility of applying the recently constructed ELMO-libraries in situations and computational techniques in which the rapid evaluations of electron densities and electrostatic potentials are crucial to obtain quick results and information.

In this context, the most direct and logic exploitation of the new databanks of extremely localized molecular orbitals will be their use in refinements of protein crystallographic structures, for which an approximate electron density is necessary at each step of the process. In particular, given the intrinsic quantum mechanical nature of the electron distributions resulting from the ELMOs transfers, the new ELMO-libraries seem really suitable to be coupled with the promising Hirshfeld atom refinement technique of crystallography, which requires a quantum mechanical calculation at each iteration and which, for this reason, at the moment cannot be straightforwardly applied to very large systems. This will allow us to exploit, also for macromolecules (particularly for proteins), the precision and accuracy of HAR in determining the positions of the hydrogen atoms by only using X-ray diffraction data. Finally, given the excellent time scaling of the ELMOs transfer, in the near future we also envisage to applying the new databases in the framework of molecular docking for the fast evaluation of quantum mechanically rigorous (although approximate) electrostatic potentials of protein target systems (or at least of their most crucial regions) at a moderate computational cost and without resorting to very large computational facilities.

Acknowledgments

A.G. acknowledges the French Research Agency (ANR) for financial support of the Young Researcher Project *QuMacroRef* through Grant No. ANR-17-CE29-0005.

Supporting Information Available

Table S1 showing the values of the Carbó Euclidean distances between the Hartree-Fock, B3LYP and ELMO electron densities, Table S2 showing the values of the

Carbó Euclidean distances between the Hartree-Fock, B3LYP and ELMO molecular electrostatic potentials and Figure S1 showing the model molecules used for the computations of the ELMOs stored in the libraries (PDF file).

Geometries of all the model molecules used for the calculations of the ELMOs stored in the libraries (XYZ files).

List of all the considered fragments considered in the construction of the ELMO-databanks, each of them with the specification of the model molecules on which the corresponding ELMOs have been computed and of the two atomic triads used for the definition of the associated ELMO rotation matrix (TXT files).

This material is available free of charge via the Internet at http://pubs.acs.org/.

References

1. Van Alsenoy, C.; Yu, C.-H.; Peeters, A.; Martin, J. M. L.; Schäfer, L. Ab Initio Geometry Determinations of Proteins.1. Crambin. *J. Phys. Chem. A* **1998**, *102*, 2246-2251.
2. Scuseria, G. E. Linear Scaling Density Functional Calculations with Gaussian Orbitals. *J. Phys. Chem. A*. **1999**, *103*, 4782-4790.
3. Sato, F.; Yoshihiro, T.; Era, M; Kashiwagi, H. Calculation of all-electron wavefunction of hemoprotein cytochrome c by density functional theory *Chem. Phys. Lett.* **2001**, *341*, 645-651.
4. Inaba, T.; Tahara, S.; Nisikawa, N.; Kashiwagi, H.; Sato, F. All-electron density functional calculation on insulin with quasi-canonical localized orbitals *J. Comput. Chem.* **2005**, *26*, 987-993.
5. Inaba, T.; Sato, F. Development of parallel density functional program using distributed matrix to calculate all-electron canonical wavefunction of large molecules. *J. Comput. Chem.* **2007**, *28*, 984-995.
6. Li, X.-P.; Nunes, R. W.; Vanderbilt, D. Density-matrix electronic-structure method with linear system-size scaling. *Phys. Rev. B* **1993**, *47*, 10891-10894.
7. Goedecker, S. Linear scaling electronic structure methods. *Rev. Mod. Phys.* **1999**, *71,* 1085-1123.
8. Merz, K. M., Jr. Using Quantum Mechanical Approaches to Study Biological Systems. *Acc. Chem. Res.* **2014**, *47*, 2804-2811.
9. Yang, W. Direct calculation of electron density in density-functional theory. *Phys. Rev. Lett.* **1991**, *66*, 1438-1441.
10. Yang, W. Direct calculation of electron density in density-functional theory: Implementation for benzene and a tetrapeptide. *Phys. Rev. A* **1991**, *44*, 7823-7826.

11. Yang, W.; Lee, T.-S. A density-matrix divide-and-conquer approach for electronic structure calculations of large molecules. *J. Chem. Phys.* **1995**, *103*, 5674-5678.
12. Dixon, S. L.; Merz, K. M., Jr. Semiempirical molecular orbital calculations with linear system size scaling. *J. Chem. Phys.* **1996**, *104*, 6643-6649.
13. Dixon, S. L.; Merz, K. M., Jr. Fast, accurate semiempirical molecular orbital calculations for macromolecules. *J. Chem. Phys.* **1997**, *107*, 879-893.
14. He, X.; Merz, K. M., Jr. Divide and Conquer Hartree-Fock Calculations on Proteins. *J. Chem. Theory Comput.* **2010**, *6*, 405-411.
15. Gadre, S. R.; Shirsat, R. N.; Limaye, A. C. Molecular Tailoring Approach for Simulation of Electrostatic Properties. *J. Phys. Chem.* **1994**, *98*, 9165-9169.
16. Babu, K.; Gadre, S. R. Ab initio quality one-electron properties of large molecules: Development and testing of molecular tailoring approach. *J. Comput. Chem.* **2003**, *24*, 484-495.
17. Babu, K.; Ganesh V.; Gadre S. R.; Ghermani N. E. Tailoring approach for exploring electron densities and electrostatic potentials of molecular crystals. *Theor. Chem. Acc.* **2004**, *111*, 255-263.
18. Ganesh, V.; Dongare, R. K.; Balanarayan, P.; Gadre, S. R. Molecular Tailoring Approach for Geometry Optimization of Large Molecules: Energy Evaluation and Parallelization Strategies. *J. Chem. Phys.* **2006**, *125*, 104109.
19. Gadre, S. R.; Ganesh, V. Molecular Tailoring Approach: Towards PC-based ab intio Treatment of Large Molecules. *J. Theor. Comput. Chem.* **2006**, *5*, 835-855.
20. Rahalkar, A. P.; Katuoda, M. Gadre, S. R.; Nagase, S. Molecualr Tailoring Approach in Conjunction with MP2 and RI-MP2 Codes: A Comparison with Fragment Molecular Orbital Method. *J. Comput. Chem.* **2010**, *31*, 2405-2418.

21. Sahu, N.; Gadre, S. R. Molecualr Tailoring Approach: A Route for *ab initio* Treatment of Large Clusters. *Acc. Chem. Res.* **2014**, *47*, 2739-2747.
22. Singh, G.; Nandi, A.; Gadre, S. R. Breaking the bottleneck: Use of molecular tailoring approach for the estimation of binding energies at MP2/CBS limit for large water clusters. *J. Chem. Phys.* **2016**, *144*, 104102.
23. Sahu, N.; Gadre, S. R. Vibrational infrared and Raman spectra of polypeptides: Fragments-in-fragments within molecular tailoring approach. *J. Chem. Phys.* **2016**, *144*, 114113.
24. Zhang, D. W.; Zhang, J. Z. H. Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein-molecule interaction energy. *J. Chem. Phys.* **2003**, *119*, 3599-3605.
25. Zhang, D. W.; Xiang, Y.; Zhang, J. Z. H. New Advance in Computational Chemistry: Full Quantum Mechanical ab Initio Computation of Streptavidin-Biotin Interaction Energy. *J. Phys. Chem. B* **2003**, *107*, 12039-12041.
26. Gao, A. M.; Zhang, D. W.; Zhang, J. Z. H.; Zhang, Y. An efficient linear scaling method for ab initio calculation of electron density of proteins. *Chem. Phys. Lett.* **2004**, *394*, 293-297.
27. Xiang, Y.; Zhang, D. W.; Zhang, J. Z. H. Fully quantum mechanical energy optimization for protein-ligand structure. *J. Comput. Chem.* **2004**, *25*, 1431-1437.
28. Mey, Y.; Zhang, D. W.; Zhang, J. Z. H. New method for direct linear-scaling calculation of electron density of proteins. *J. Phys. Chem. A* **2005**, *109*, 2-5.
29. He, X.; Zhang, J. Z. H. A new method for direct calculation of total energy of protein. *J. Chem. Phys.* **2005**, 122, 031103.
30. He, X.; Zhang, J. Z. H. The generalized molecular fractionation with conjugate caps/molecular mechanics method for direct calculation of protein energy. *J.*

Chem. Phys. **2006**, 124, 184703.

31. Li, S.; Li, W.; Fang, T. An efficient fragment-based approach for predicting the ground-state energies and structures of large molecules. *J. Am. Chem. Soc.* **2005**, *127*, 7251-7226.
32. Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. Fragment molecular orbital method: an approximate computational method for large molecules. *Chem. Phys. Lett.* **1999**, *313*, 701-706.
33. Nakano, T.; Kaminuma, T.; Sato, T.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. Fragment molecular orbital method: application to polypeptides. *Chem. Phys. Lett.* **2000**, *318*, 614-618.
34. Fedorov, D. G.; Kitaura, K. Theoretical development of the fragment molecular orbital (FMO) method. In *Modern Methods for Theoretical Physical Chemistry and Biopolymers*; Starikov, E. B., Lewis, J. P., Tanaka, S., Eds.; Elsevier: Amsterdam, 2006; Chapter 1, pp 3-38.
35. Nakano, T.; Mochizuki, Y.; Fukuzawa, K.; Amari, S.; Tanaka, S. Developments and applications of ABINIT-MP software based on the fragment molecular orbital method. In *Modern Methods for Theoretical Physical Chemistry and Biopolymers*; Starikov, E. B., Lewis, J. P., Tanaka, S., Eds.; Elsevier: Amsterdam, 2006; Chapter 2, pp 39-52.
36. Fedorov, D. G.; Kitaura K. Theoretical Background of the Fragment Molecular Orbital (FMO) Method and Its Implementation in GAMESS. In *The Fragment Molecular Orbital Method: Practical Applications to Large Molecular Systems*; Fedorov, D. G., Kitaura, K., Eds.; CRC Press - Taylor & Francis Group: Boca Raton, FL, 2009; Chapter 2, pp 5-36.
37. Huang L.; Massa, L.; Karle, J. Quantum kernels and quantum crystallography:

Applications in biochemistry. In *Quantum Biochemistry: Electronic Structure and Biological Activity*; Matta, C. F. Ed.; Wiley-VCH: Weinheim, 2010; Chapter 1, pp 3-60.

38. Huang, L.; Massa, L.; Karle, J. Kernel energy method applied to vesicular stomatitis virus nucleoprotein. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 1731-1736.
39. Huang, L.; Massa, L.; Karle, J. The Kernel Energy Method: Application to a tRNA. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 1233-1237.
40. Huang, L.; Massa, L.; Karle, J. Kernel energy method illustrated with peptides. *Int. J. Quantum Chem.* **2005**, *103*, 808-817.
41. Huang, L.; Massa, L.; Karle, J. Kernel energy method: Application to DNA. *Biochemistry* **2005**, *44*, 16747-16752.
42. Huang, L.; Massa, L.; Karle, J. Kernel energy method: Application to insulin. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 12690-12693.
43. Huang, L.; Bohorquez, H.; Matta, C. F.; Massa, L. The Kernel Energy Method: Application to Graphene and Extended Aromatics. *Int. J. Quantum Chem.* **2011**, *111*, 4150-4157.
44. Huang, L.; Massa, L.; Matta, C. F. A graphene flake under external electric fields reconstructed from field-perturbed kernels. *Carbon* **2014**, *76*, 310-320.
45. Timm, M. J.; Matta, C. F.; Massa, L.; Huang, L. The localization-delocalization matrix and the electron density-weighted connectivity matrix of a finite graphene flake reconstructed from kernel fragments. *J. Phys. Chem. A* **2014**, *118*, 11304-11316.
46. Huang, L.; Matta, C. F.; Massa, L. The kernel energy method (KEM) delivers fast and accurate QTAIM electrostatic charge for atoms in large molecules *Struct. Chem.* **2015**, *26*, 1433-1442.

47. Polkosnik, W.; Massa, L. Single determinant *N*-representability and the kernel energy method applied to water clusters. *J. Comp. Chem.* **2018**, *39*, 1038-1043.

48. Walker, P. D.; Mezey, P. G. Molecular electron density Lego approach to molecule building. *J. Am. Chem. Soc.* **1993**, *115*, 12423-12430.

49. Walker, P. D.; Mezey, P. G. Ab Initio Quality Electron Densities for Proteins: A MEDLA Approach. *J. Am. Chem. Soc.* **1994**, *116*, 12022-12032.

50. Exner, T. E.; Mezey, P. G. Ab initio-quality electrostatic potentials for proteins: An application of the ADMA approach. *J. Phys. Chem. A* **2002**, *106*, 11791-11800.

51. Exner, T. E.; Mezey, P. G. Ab initio quality properties for macromolecules using the ADMA approach. *J. Comput. Chem.* **2003**, *24*, 1980-1986.

52. Szekeres, Z.; Exner, T.; Mezey, P. G. Fuzzy Fragment Selection Strategies, Basis Set Dependence and HF-DFT Comparisons in the Applications of the ADMA Method of Macromolecular Quantum Chemistry. *Int. J. Quantum Chem.* **2005**, *104*, 847-860.

53. Breneman, C. M.; Thompson, T. R.; Rhem, M.; Dung, M. Electron density modeling of large systems using the transferable atom equivalent method. *Comput. Chem.* **1995**, *19*, 161-179.

54. Breneman, C. M.; Rhem, M. QSPR Analysis of HPLC column capacity factors for a set of high-energy materials using electronic van der Waals surface property descriptors computed by transferable atom equivalent method. *J. Comput. Chem.* **1997**, *18*, 182-197.

55. Chang, C.; Bader, R. F. W. Theoretical construction of a polypeptide. *J. Phys. Chem.* **1992**, *96*, 1654-1662.

56. Bader, R. F. W.; Martín, F. J. Interdeterminancy of basin and surface properties

of an open system. *Can. J. Chem.* **1998**, *76*, 284-291.

57. Martín, F. J. "Theoretical Synthesis of Macromolecules from Transferable Functional Groups". Ph.D. Thesis; McMaster University: Hamilton, 2001.
58. Matta, C. F. Theoretical reconstruction of the electron density of large molecules from fragments determined as proper open quantum systems: the properties of the oripavine PEO, enkephalins, and morphine. *J. Phys. Chem. A* **2001**, *105*, 11088-11101.
59. 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.
60. 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.
61. Stoll, H.; Wagenblast, G.; Preuss, H. On the Use of Local Basis Sets for Localized Molecular Orbitals. *Theor. Chim. Acta* **1980**, *57*, 169−178.
62. Boys, S. F. Construction of Some Molecular Orbitals to be Approximately Invariant for Changes from One Molecule to Another. *Rev. Mod. Phys.* **1960**, *32*, 296–299.
63. Foster, J. M.; Boys, S. F. Canonical Configurational Interaction Procedure. *Rev. Mod. Phys.* **1960**, *32*, 300–302.
64. Edmiston, C.; Ruedenberg, K. Localized Atomic and Molecular Orbitals. *Rev. Mod. Phys.* **1963**, *35*, 457–465.
65. Edmiston, C.; Ruedenberg, K. Localized Atomic and Molecular Orbitals. II. *J. Chem. Phys.* **1965**, *43*, S97–S116.
66. Pipek, J.; Mezey, P. G. A Fast Intrinsic Localization Procedure Applicable for Ab

Initio and Semiempirical Linear Combination of Atomic Orbital Wave Functions. *J. Chem. Phys.* **1989**, *90*, 4916–4926.

67. Genoni, A.; Molecular Orbitals Strictly Localized on Small Molecular Fragments from X-ray Diffraction Data. *J. Phys. Chem. Lett.* **2013**, *4*, 1093-1099.
68. Genoni, A.; X-ray Constrained Extremely Localized Molecular Orbitals: Theory and Critical Assessment of the New Technique. *J. Chem. Theory Comput.* **2013**, *9*, 3004-3019.
69. Dos Santos, L. H. R.; Genoni, A.; Macchi, P. Unconstrained and X-ray constrained extremely localized molecular orbitals: analysis of the reconstructed electron density. *Acta Crystallogr., Sect. A* **2014**, *70*, 532-551.
70. Genoni, A.; Meyer, B. X-Ray Constrained Wave Functions: Fundamentals and Effects of the Molecular Orbitals Localization. *Adv. Quantum Chem.* **2016**, *73*, 333-362.
71. Genoni, A. A first-prototype multi-determinant X-ray constrained wavefunction approach: the X-ray constrained extremely localized molecular orbital-valence bond method, *Acta Crystallogr., Sect. A* **2017**, *73*, 312-316.
72. Casati, N.; Genoni, A.; Meyer, B.; Krawczuk, A.; Macchi, P. Exploring charge density analysis in crystals at high pressure: data collection, data analysis and advanced modelling. *Acta Crystallogr., Sect. B* **2017**, *73*, 584-597.
73. Jayatilaka, D. Wave Function for Beryllium from X-ray Diffraction Data. *Phys. Rev. Lett.* **1998**, *80*, 798−801.
74. Jayatilaka, D.; Grimwood, D. J. Wavefunctions Derived from Experiment. I. Motivation and Theory. *Acta Crystallogr., Sect. A* **2001**, *57*, 76−86.
75. Grimwood, D. J.; Jayatilaka, D. Wavefunctions Derived from Experiment. II. A Wavefunction for Oxalic Acid Dihydrate. *Acta Crystallogr., Sect. A* **2001**, *57*,

87−100.

76. Bytheway, I.; Grimwood, D.; Jayatilaka, D. Wavefunctions Derived from Experiment. III. Topological Analysis of Crystal Fragments. *Acta Crystallogr., Sect. A* **2002**, *58*, 232−243.
77. Bytheway, I.; Grimwood, D. J.; Figgis, B. N.; Chandler, G. S.; Jayatiaka, D. Wavefunctions Derived from Experiment. IV. Investigation of the Crystal Environment of Ammonia. *Acta Crystallogr., Sect. A* **2002**, *58*, 244−251.
78. Grimwood, D. J.; Bytheway, I.; Jayatilaka, D. Wavefunctions Derived from Experiment. V. Investigation of Electron Densities, Electrostatic Potentials, and Electron Localization Functions for Noncentrosymmetric Crystals. *J. Comput. Chem.* **2003**, *24*, 470−483.
79. Hudák, M.; Jayatilaka, D.; Peraínova, L.; Biskupic, S.; Kozísek, J.; Bucinský, L. X-ray Constrained Unrestricted Hartree−Fock and Douglas−Kroll−Hess Wavefunctions. *Acta Crystallogr., Sect. A* **2010**, *66*, 78−92.
80. Genoni, A.; Dos Santos, L. H. R.; Meyer, B.; Macchi, P. Can X-ray constrained Hartree-Fock wavefunctions retrieve electron correlation?, *IUCrJ* **2017**, *4*, 136-146.
81. Genoni, A.; Franchini, D.; Pieraccini, S.; Sironi, M. X-ray Constrained Spin-Coupled Wavefunction: a New Tool to Extract Chemical Information from X-ray Diffraction Data. *Chem. Eur. J.* **2018**, DOI: 10.1002/chem.201803988.
82. Hirshfeld, F. L. Difference Densities by Least-Squares Refinement: Fumaramic Acid. *Acta Crystallogr., Sect. B* **1971**, *27*, 769−781.
83. Stewart, R. F. Electron Population Analysis with Rigid Pseudoatoms. *Acta Crystallogr., Sect. A* **1976**, *32*, 565−574.
84. Hansen, N. K.; Coppens, P. Testing Aspherical Atom Refinements on Small-

Molecule Data Sets. *Acta Crystallogr., Sect. A* **1978**, *34*, 909−921.

85. Sironi, M.; Famulari, A.; Raimondi, M.; Chiesa, S. The transferability of extremely localized molecular orbitals. *J. Mol. Struct. (THEOCHEM)* **2000**, *529*, 47-54.
86. Fornili, A.; Sironi, M.; Raimondi, M. Determination of extremely localized molecular orbitals and their application to quantum mechanics/molecular mechanics methods and to the study of intramolecular hydrogen bonding. *J. Mol. Struct. (THEOCHEM)* **2003**, *632*, 157–172.
87. Burresi, E.; Sironi, M. Determination of extremely localized molecular orbitals in the framework of density functional theory. *Theor. Chem. Acc.* **2004**, *112*, 247-253.
88. Genoni, A; Sironi, M. A Novel Approach to Relax Extremely Localized Molecular Orbitals: the Extremely Localized Molecular Orbital-Valence Bond Method. *Theor. Chem. Acc.* **2004**, *112*, 254-262.
89. Genoni, A.; Fornili, A.; Sironi, M. Optimal Virtual Orbitals to Relax Wave Functions Built Up with Transferred Extremely Localized Molecular Orbitals. *J. Comput. Chem.* **2005**, *26*, 827-835.
90. Sironi, M.; Genoni, A.; Civera, M.; Pieraccini, S.; Ghitti, M. Extremely Localized Molecular Orbitals: Theory and Applications. *Theor. Chem. Acc.* **2007**, *117*, 685-698.
91. Sironi, M.; Ghitti, M.; Genoni, A.; Saladino, G.; Pieraccini, S. DENPOL: A new program to determine electron densities of polypeptides using extremely localized molecular orbitals. *J. Mol. Struct. (THEOCHEM)* **2009**, *898*, 8-16.
92. Pichon-Pesme, V.; Lecomte, C.; Lachekar, H. On Building a Data Bank of Transferable Experimental Electron Density Parameters: Application to

Polypeptides. *J. Phys. Chem.* **1995**, *99*, 6242−6250.

93. Jelsch, C.; Pichon-Pesme, V.; Lecomte, C.; Aubry, A. Transferability of Multipole Charge-Density Parameters: Application to Very High Resolution Oligopeptide and Protein Structures. *Acta Crystallogr., Sect. D* **1998**, *54*, 1306-1318.
94. Zarychta, B.; Pichon-Pesme, V.; Guillot, B.; Lecomte, C.; Jelsch, C. On the Application of an Experimental Multipolar Pseudo-Atom Library for Accurate Refinement of Small-Molecule and Protein Crystal Structures. *Acta Crystallogr., Sect. A* **2007**, *63*, 108-125.
95. Lecomte, C.; Jelsch, C.; Guillot, B.; Fournier, B.; Lagoutte, A. Ultrahigh-Resolution Crystallography and Related Electron Density and Electrostatic Properties in Proteins. *J. Synchroton Rad.* **2008**, *15*, 202-203.
96. Domagała, S.; Munshi, P.; Ahmed, M.; Guillot, B.; Jelsch, C. Structural Analysis and Multipole Modelling of Quercetin Monohydrate. A Quantitative and Comparative Study. *Acta Crystallogr., Sect. B* **2011**, *67*, 63-78.
97. Domagała, S.; Fournier, B.; Liebschner, D.; Guillot, B.; Jelsch, C. An Improved Experimental Databank of Transferable Multipolar Atom Models – ELMAM2. Construction Details and Applications. *Acta Crystallogr., Sect. A* **2012**, *68*, 337-351.
98. Koritsanszky, T.; Volkov, A.; Coppens, P. Aspherical-Atom Scattering Factors from Molecular Wave Functions. 1. Transferability and Conformation Dependence of Atomic Electron Densities of Peptides within the Multipole Formalism. *Acta Crystallogr., Sect. A* **2002**, *58*, 464−472.
99. Volkov, A.; Li, X.; Koritsanszky, T.; Coppens, P. Ab Initio Quality Electrostatic Atomic and Molecular Properties Including Intermolecular Energies from a Transferable Theoretical Pseudoatom Databank. *J. Phys. Chem. A* **2004**, *108*,

4283−4300.

100. Li, X.; Volkov, A. V.; Szalewicz, K.; Coppens, P. Interaction Energies between Glycopeptide Antibiotics and Substrates in Complexes Determined by X-ray Crystallography: Application of a Theoretical Databank of Aspherical Atoms and a Symmetry-Adapted Perturbation Theory-Based Set of Interatomic Potentials. *Acta Crystallogr., Sect. D* **2006**, *62*, 639-647.

101. Dominiak, P. M.; Volkov, A.; Li, X.; Messerschmidt, M.; Coppens, P. A Theoretical Databank of Transferable Aspherical Atoms and Its Application to Electrostatic Interaction Energy Calculations of Macromolecules. *J. Chem. Theory Comput.* **2007**, *3*, 232−247.

102. Volkov, A.; Messerschmidt, M.; Coppens, P. Improving the Scattering-Factor Formalism in Protein Refinement: Application of the University at Buffalo Aspherical-Atom Databank to Polypeptide Structures. *Acta Crystallogr., Sect. D* **2007**, *63*, 160-170.

103. Dittrich, B.; Koritsanszky, T.; Luger, P. A Simple Approach to Nonspherical Electron Densities by Using Invarioms. *Angew. Chem., Int. Ed.* **2004**, *43*, 2718−2721.

104. Dittrich, B.; Hübschle, C. B.; Messerschmidt, M.; Kalinowski, R.; Grint, D.; Luger, P. The Invariom Model and Its Application: Refinement of D,L-Serine at Different Temperatures and Resolution. *Acta Crystallogr., Sect. A* **2005**, *61*, 314−320.

105. Dittrich, B.; Hübschle, C. B.; Luger, P.; Spackman, M. A. Introduction and Validation of an Invariom Database for Amino-Acid, Peptide and Protein Molecules. *Acta Crystallogr., Sect. D* **2006**, *62*, 1325−1335.

106. Dittrich, B; Hübschle, C. B.; Holstein, J. J.; Fabbiani, F. P. A. Towards

Extracting the Charge Density from Normal-Resolution Data. *J. Appl. Cryst.* **2009**, *42*, 1110-1121.

107. Dittrich, B; Hübschle, C. B.; Pröpper, K.; Dietrich, F.; Stolper, T.; Holstein, J. J. The Generalized Invariom Database (GID). *Acta Crystallogr., Sect. B* **2013**, *69*, 91-104.

108. Hathwar, V. R.; Thakur, T. S.; Dubey, R.; Pavan, M. S.; Row, T. N. G.; Desiraju, G. R. Extending the Supramolecular Synthon Based Fragment Approach (SBFA) for Transferability of Multipole Charge Density Parameters to Monofluorobenzoic Acids and their Cocrystals with Isonicotinamide: Importance of C-H···O, C-H···F, and F···F Intermolecular Regions. *J. Phys. Chem. A* **2011**, *115*, 12852-12863.

109. Hathwar, V. R.; Thakur, T. S.; Row, T. N. G.; Desiraju, G. R. Transferability of Multipole Charge Density Parameters for Supramolecular Synthons: A New Tool for Quantitative Crystal Engineering. *Cryst. Growth Des.* **2011**, *11*, 616-623.

110. Pichon-Pesme, V.; Jelsch, C.; Guillot, B.; Lecomte, C. A Comparison between Experimental and Theoretical Aspherical-Atom Scattering Factors for Charge-Density Refinement of Large Molecules. *Acta Crystallogr., Sect. A* **2004**, *60*, 204-208.

111. Volkov, A.; Koritsanszky, T.; Li, X.; Coppens, P. Response to the Paper A Comparison between Experimental and Theoretical Aspherical-Atom Scattering Factors for Charge-Density Refinement of Large Molecules, by Pichon-Pesme, Jelsch, Guillot & Lecomte (2004). *Acta Crystallogr., Sect. A* **2004**, *60*, 638-639.

112. Bąk, J. M.; Domagała, S.; Hübschle, C.; Jelsch, C.; Dittrich, B.; Dominiak, P. M. Verification of Structural and Electrostatic Properties Obtained by the Use of Different Pseudoatom Databases. *Acta Crystallogr., Sect. A* **2011**, *67*, 141-153.

113. Jelsch, C.; Teeter, M. M.; Lamzin, V.; Pichin-Pesme, V.; Blessing, R. H.; Lecomte, C. Accurate protein crystallography at ultra-high resolution: Valence electron distribution in crambin. *Proc. Natl. Acad. USA* **2000**, *97*, 3171-3176.

114. Elias, M.; Liebschner, D.; Koepke, J.; Lecomte, C.; Guillot, B.; Jelsch, C. Hydrogen atoms in protein structures: high-resolution X-ray diffraction structure of the DFPase. *BMC Research Notes* **2013**, *6*, 308.

115. Pröpper, K.; Holstein, J. J.; Hübschle, C. B.; Bond, C. S.; Dittrich, B. Invarion refinement of a new monoclinic solvate thiostrepton at 0.64 Å resolution. *Acta Crystallogr., Sect. D* **2013**, *69*, 1530-1539.

116. Jayatilaka, D.; Dittrich, B. X-ray structure refinement using aspherical atomic density functions obtained from quantum mechanical calculations. *Acta Crystallogr., Sect. A* **2008**, *64*, 383-393.

117. Capelli, S.C.; Bürgi, H.-B.; Dittrich, B.; Grabowsky, S.; Jayatilaka, D. Hirshfeld Atom Refinement. *IUCrJ* **2014**, *1*, 361-379.

118. Woińska, M.; Jayatilaka, D.; Spackman, M. A.; Edwards, A. J.; Dominiak, P. M.; Woźniak, K.; Nishibori, E.; Sugimoto, K.; Grabowsky, S. Hirshfeld atom refinement for modeling strong hydrogen bonds. *Acta Crystallogr., Sect. A* **2014**, *70*, 483-498.

119. Wall, M. E. Quantum crystallographic charge density of urea, *IUCrJ* **2016**, *3*, 237-246.

120. Woińska, M.; Grabowsky, S.; Dominiak, P. M.; Woźniak, K.; Jayatilaka, D. Hydrogen atoms can be located accurately and precisely by x-ray crystallography. *Sci. Adv.* **2016**, *2*, e1600192.

121. Dittrich, B.; Lübben, J.; Mebs, S.; Wagner, A.; Luger, P.; Flaig, R. Accurate Bond Lengths to Hydrogen Atoms from Single-Crystal X-ray Diffraction by

Including Estimated Hydrogen ADPs And Comparison to Neutron and QM/MM Benchmarks. *Chem. Eur. J.* **2017**, *23*, 4605-4614.

122. Genoni, A.; Bučinský, L.; Claiser, N.; Contreras-García, J.; Dittrich, B.; Dominiak, P. M.; Espinosa, E.; Gatti, C.; Giannozzi, P.; Gillet, J.-M.; Jayatilaka, D.; Macchi, P.; Madsen, A. Ø.; Massa, L. J.; Matta, C. F.; Merz, K. M., Jr.; Nakashima, P. N. H.; Ott, H.; Ryde. U.; Schwarz, K.; Sierka, M.; Grabowsky, S. Quantum Crystallography: Current Developments and Future Perspectives. *Chem. Eur. J.* **2018**, *24*, 10881-10905.

123. Novara, R. F.; Genoni, A.; Grabowsky, S. What is Quantum Crystallography? *ChemViews* **2018**, DOI:10.1002/chemv.201800066.

124. Grabowsky, S.; Genoni, A.; Bürgi, H.-B. Quantum Crystallography. *Chem. Sci.* **2017**, *8*, 4159-4176.

125. Massa, L.; Matta, C. F. Quantum Crystallography: A perspective. *J. Comput. Chem.* **2017**, *39*, 1021-1028.

126. Tsirelson, V. Early days of quantum crystallography: A personal account. *J. Comput. Chem.* **2017**, *39*, 1029-1037.

127. Adams, W. H. On the Solution of the Hartree-Fock Equations in Terms of Localized Orbitals *J. Chem. Phys.* **1961**, *34*, 89-102.

128. Huzinaga, S.; Cantu, A. A. Theory of Separability of Many-Electron Systems. *J. Chem. Phys.* **1971** *55*, 5543–5549.

129. Gilbert, T. L. Multiconfiguration self-consistent-field theory for localized orbitals. II. Overlap constraints, Lagrangian multipliers, and the screened interaction field *J. Chem. Phys.* **1974**, *60*, 3835-3844.

130. Matsuoka, O. Expansion methods for Adams–Gilbert equations. I. Modified Adams–Gilbert equation and common and fluctuating basis sets. *J. Chem. Phys.* **1977**, *66*, 1245-1254.

131. Smits, G. F.; Altona, C. Calculation and properties of non-orthogonal, strictly local molecular-orbitals. *Theor. Chim. Acta* **1985**, *67*, 461–475.

132. Francisco, E.; Martín Pendás, A.; Adams, W. H. Generalized Huzinaga Building-Block Equations for Nonorthogonal Electronic Groups – Relation to the Adams-Gilbert Theory. *J. Chem. Phys.* **1992**, *97*, 6504-6508.

133. Ordejón, P.; Drabold, D.; Grumbach, M.; Martin, R. Unconstrained Minimization Approach for Electronic Computations that Scales Linearly with System Size. *Phys. Rev. B* **1993**, *48*, 14646–14649.

134. Couty, M.; Bayse, C. A.; Hall, M. B. Extremely localized molecular orbitals (ELMO): a non-orthogonal Hartree-Fock method. *Theor. Chem. Acc.* **1997**, *97*, 96–109.

135. Sironi, M.; Famulari, A. An orthogonal approach to determine extremely localised molecular orbitals. *Theor. Chem. Acc.* **2000**, *103*, 417-422.

136. Szekeres, Z.; Surján, P. R. Direct determination of fragment localized molecular orbitals and the orthogonality constraint. *Chem. Phys. Lett.* **2003**, *369*, 125–130.

137. McWeeny, R. The Density Matrix in Many-Electron Quantum Mechanics. 1. Generalized Product Functions – Factorization and Physical Interpretation of the Density Matrices. *Proc. R. Soc. London Ser. A* **1959**, *253*, 242–259.

138. McWeeny, R. Some Recent Advances in Density Matrix Theory. *Rev. Mod. Phys.* **1960**, *32*, 335–369.

139. 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.

140. Press, W. H.; Teukolsky, S. A.; Vetterling,W. T.; Flannery, B. P. Numerical Recipes in Fortran 77: The Art of Scientific Computing, 2nd ed.; Cambridge University Press: New York, 1992; pp 387−448.

141. 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.

142. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision D.01; Gaussian, Inc., Wallingford, CT, USA, 2009.

143. Guest, M. F.; Bush, I. J.; van Dam, H. J. J.; Sherwood, P.; Thomas, J. M. H.; van Lenthe, J. H.; Havenith, R. W. A.; Kendrick, J. The GAMESS-UK Electronic Structure Package: Algorithms, Developments and Applications. *Mol. Phys.* **2005**, *103*, 719−747.

144. Case, D. A.; Ben-Shalom, I. Y.; Brozell, S. R.; Cerutti, D. S.; Cheatham, T. E., III; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Ghoreishi, D.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Greene, D.; Harris, R.; Homeyer, N.; Izadi, S.; Kovalenko, A.; Kurtzman, T.; Lee,T. S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Mermelstein, D. J.; Merz, K. M.; Miao, Y.; Monard, G.; Nguyen, C.; Nguyen, H.; Omelyan, I.; Onufriev, A.; Pan, F.; Qi, R.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shen, J.; Simmerling, C. L.; Smith, J.; Salomon-Ferrer, R.; Swails, J.; Walker, R. C.; Wang, J.; Wei, H.; Wolf, R. M.; Wu, X.; Xiao, L.; York, D. M.; Kollman, P. M. *AMBER 2018*; University of California San Francisco, San Francisco, CA, USA, 2018.

145. Mauri, F.; Galli, G.; Car, R. Orbital formulation for electronic-structure calculations with linear system-size scaling. *Phys. Rev. B* **1993**, *47*, 9973-9976.

146. Benzi, M.; Meyer, C.; Tuma, M. A sparse approximate inverse preconditioner for the conjugate gradient method. *SIAM J. Sci. Comput.* **1996**, *17*, 1135-1149.

147. Millam, J. M.; Scuserie, G. E. Linear scaling conjugate gradient density matrix search as an alternative to diagonalization for first principles electronic strucrue calculations. *J. Chem. Phys.* **1997**, *106*, 5569-5577.

148. Challacombe, M. A. A simplified density matrix minimization for linear scaling self-consistent field theory. *J. Chem. Phys.* **1999**, *110*, 2232-2242.

149. Ozaki, T. Efficient recursion method for inverting an overlap matrix. *Phys. Rev. B* **2001**, *64*, 195110.

150. Niklasson, A. M. N. Iterative refinement method for the approximate factorization of a matrix inverse. *Phys. Rev. B* **2004**, *70*, 193102.

151. Genoni, A.; Ghitti, M.; Pieraccini, S.; Sironi, M. A novel extremely localized molecular orbitals based technique for the one-electron density matrix

computation. *Chem. Phys. Lett.* **2005**, *415*, 256-260.

152. Negre, A. F. A.; Mniszewski, S. M.; Cawkwell, M. J.; Bock, N.; Wall, M.; Niklasson, A. M. N. Recursive Factorization of the Inverse Overlap Matrix in Linear-Scaling Quantum Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2016**, *12*, 3063-3073.

153. Mulliken, R. S. Electronic Population Analysis on LCAO-MO Molecular Wave Functions. I. *J Chem. Phys.* **1955**, *23*, 1833-1840.

154. Singh, U. C.; Kollman, P. A. An approach to computing electrostatic charges for molecules. *J. Comp. Chem.* **1984**, *5*, 129-145.

155. Besler, B. H.; Merz, K. M., Jr.; Kollman, P. A. Atomic Charges Derived from Semiempirical Methods. *J Comput. Chem.* **1990**, *11*, 431-439.

156. Carbó-Dorca, R.; Besalli, E.; Amat, L.; Fradera, X. In *Advances in Molecular Simialrity*, Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: London, U.K., 1996; Vol. 1, Chapter 1, pp 1-42.