Complexity Reduction in Large Quantum Systems: Fragment Identification and Population Analysis via a Local Optimized Minimal Basis

Stephan Mohr,[1] Michel Masella,[2] Laura E. Ratcliff,[3] and Luigi Genovese[4, 5]

[1] *Barcelona Supercomputing Center (BSC)*[*]
[2] *Laboratoire de Biologie Structurale et Radiologie,*
Service de Bioénergétique, Biologie Structurale et Mécanisme,
Institut de Biologie et de Technologie de Saclay,
CEA Saclay, F-91191 Gif-sur-Yvette Cedex, France
[3] *Argonne Leadership Computing Facility, Argonne National Laboratory, Illinois 60439, USA*
[4] *Univ. Grenoble Alpes, INAC-MEM, L_Sim, F-38000 Grenoble, France*
[5] *CEA, INAC-MEM, L_Sim, F-38000 Grenoble, France*[†]

(Dated: July 11, 2017)

We present, within Kohn-Sham Density Functional Theory calculations, a quantitative method to identify and assess the partitioning of a large quantum mechanical system into fragments. We then show how within this framework simple generalizations of other well-known population analyses can be used to extract, from first principles, reliable electrostatic multipoles for the identified fragments. Our approach reduces arbitrariness in the fragmentation procedure, and enables the possibility to assess, quantitatively, whether the corresponding fragment multipoles can be interpreted as observable quantities associated to a system's moiety. By applying our formalism within the code BigDFT, we show that the use of a minimal set of in-situ optimized basis functions allows at the same time a proper fragment definition and an accurate description of the electronic structure.

I. INTRODUCTION

First-principles computational quantum mechanical (QM) approaches are nowadays able to provide reasonably accurate modelizations for a wide variety of systems. In particular Density Functional Theory (DFT) approaches based on the Kohn-Sham (KS) formalism[1,2] are probably the most popular, usually presenting a good compromise between accuracy and computational complexity. Nonetheless, even when a DFT approach gives an accurate description of a microscopic system, it is advantageous in certain situations to consider an effective complexity reduction (ECR), allowing one to get the same level of accuracy by explicitly considering fewer degrees of freedom.

The fundamental principle of an ECR lies in the *identification* of the essential moieties (i.e. "fragments") of a system out of an atomic description. These fragments should then, in turn, be treated with an adequate methodology depending on the specific needs. A less complex description of a system might contribute to decreasing the computational cost of the calculation; however, this is not the sole advantage of an ECR. Within such a scheme the observable quantities that can be extracted for the system as a whole can also be assigned *separately* to each of the fragments. Such a procedure allows a better understanding of the relevant mechanisms which govern the interactions among the constituents of the system, together with the design and validation of coarse-graining models, that are adapted for systems of length scales for which atomistic QM models would be unnecessarily costly or even out of reach[3].

A great variety of fragmentation methods has been developed; an exhaustive overview can be found in Refs. 4,5. In all of these methods the fragments are chosen based on pre-defined conditions, such as geometrical criteria or basic chemical intuition, and there is no possibility for verifying *a posteriori* whether the chosen fragmentation is sensible or not for the actual setup of the simulation. A typical observable which is then determined for each moiety is the electronic charge, extracted from the charge density of the QM calculation of the entire system, partitioned among the fragments. Typically, the attention is focused on the atoms composing the system, and a large number of *atomic* charge population analyses have been developed.

All of these population analyses have their advantages and shortcomings, and applied to the very same system they might even give considerably different results[6,7]. However, from a conceptual point of view, all of them suffer from the same problem: the electrostatic multipoles of the atoms, considered separately, are *not* observable quantities of a QM system. The only electrostatic quantities that are truly QM observables are the charge multipoles of the whole system, which are of course well defined and independent of the basis set as they are a function of the charge density of the system, which should not alter under changes of the basis; all the methods should yield the *same* values, provided of course an adequate level of completeness. For a portion of the system like an atom or a fragment, electrostatic multipoles become "pseudo-observables", whose pertinence depends on the method chosen, the basis set used, and the definition of the subsystem itself. In the context of ECR methods based on electrostatic multipoles of a subsystem, this is a crucial fact that has to be taken into account. In other words, such methods suffer from two (somehow related) shortcomings: Firstly it is not possible to systematically validate the pertinence of the chosen fragmentation scheme, and secondly they do not allow one to quantify whether the electrostatic fragment multipoles extracted from the QM simulation can be considered as physical “pseudo-observables”, i.e. with a meaningful physico-chemical interpretation.

In this paper we propose a general theoretical scheme to identify subsystems (i.e. fragments) out of a large QM system, accounting for the aforementioned problems. Our method, which we will denote as “purity indicator”, allows one to assess quantitatively the suitability of the employed basis set for the chosen population method; thanks to this information we can therefore verify, in a quantitative way, whether a given fragmentation of a QM system is compatible with the employed *combination* of the basis set and the population method. We show that in situations where this is the case, the electrostatic quantities calculated on the pre-defined fragment moieties have the reliability of QM observables and can be interpreted as such. On the other hand, the same technique might also be employed to determine *a posteriori*, i.e. based on a QM calculation of the entire system, which are the essential moieties that can be considered as well defined entities for the actual fragmentation method and basis set.

Our approach is based on the *density matrix* of the system, which is a well-defined QM entity; this is in contrast to other popular QM-based fragmentation schemes, such as the Fragment Molecular Orbital (FMO) approach[8–10] or X-Pol[11–16], where only the pre-selected fragments are treated on a stringent ab-initio level. Like all methods based on the density matrix, this intrinsically only gives access to integrated quantities such as the charge or the total energy. This is in contrast to methods that explicitly calculate the wave functions within a fragmentation approach, which have also the advantage that they can be applied to excited states[17–20]. Within our framework we further point out the competitive advantages of a *minimal and optimized* basis set in the context of ECR methods. Using the purity indicator it can be shown that such a computational setup considerably simplifies both the fragment identification and multipole assignment. We additionally demonstrate that within this setup, even straightforward generalizations of pioneering approaches like Mulliken and Löwdin population analyses provide high quality and chemically sound results, whose reliability can be assessed in a quantitative way.

The outline of the paper is as follows. We first present in Sec. II the basic ideas of the identification and assessment of the system fragments and the calculation of the associated multipoles. In Sec. III we then discuss the important relation between the fragment definition and the employed basis set, by defining *a priori* a specific fragment definition and population scheme and then searching for the optimal basis for this setup. In Sec. IV we then inverse the focus and identify — within the setup of a given basis — the fragments for a large complex molecule in solvation.

II. METHODOLOGY

A. Fragment identification and assessment

Let us assume that a QM system can be split into M different fragments. This means that, in a “QM sense”, the wave function can be approximated by a *separable* wave function, i.e.

$$|\Psi\rangle \simeq |\Psi^1\rangle \otimes |\Psi^2\rangle \otimes \cdots \otimes |\Psi^M\rangle \;, \tag{1}$$

where each of the states $|\Psi^{\mathfrak{F}}\rangle$ is associated to the quantum description of the fragment $\mathfrak{F}$.

The simplest case where the above assumption is valid is the *cluster decomposition*, which also implies (the opposite is not necessarily true) that a spatial separation can be readily defined between the system elements and their respective wave functions do not overlap. In addition the Hilbert spaces of the different subsystems can be factorized in different subspaces where QM observables are correctly defined. To define pseudo-observables like the electrostatic multipoles of a system element, we are interested in a suitable realization of the above situation for a KS-DFT computation.

Let us suppose that we can express the (one-body) density operator of the system in a finite set of localized basis functions $|\phi_\alpha\rangle$ as follows:

$$\hat{F} = \sum_{\alpha,\beta} |\phi_\alpha\rangle K_{\alpha\beta} \langle\phi_\beta| \;. \tag{2}$$

This is a common ansatz for large scale DFT calculations[21–23]. In the following, the basis functions $|\phi_\alpha\rangle$ will be called *support functions*, and the matrix $\mathbf{K}$ will denote the *kernel.* If $\hat{F}$ is obtained from a many-body wave function $|\Psi\rangle$ expressed via a single Slater determinant the above density matrix is idempotent, i.e. $\hat{F}^2 = \hat{F}$, and the kernel is pure, i.e. it obeys $\mathbf{KSK} = \mathbf{K}$, where $S_{\alpha\beta} = \langle\phi_\alpha|\phi_\beta\rangle$ is the overlap matrix among the support functions.

When a QM system is genuinely separable, it should be possible to define a projector operator $\hat{W}^{\mathfrak{F}}$ associated with each fragment $\mathfrak{F}$ such that $\hat{W}^{\mathfrak{F}}|\Psi\rangle = |\Psi^{\mathfrak{F}}\rangle$. For such a separable system, the QM measure of any observable $\hat{O}$ may also be associated with the fragment, by evaluating $\mathrm{Tr}\left(\hat{F}\hat{W}^{\mathfrak{F}}\hat{O}\right)$. The quantity $\hat{F}^{\mathfrak{F}} = \hat{F}\hat{W}^{\mathfrak{F}}$ may thus be referred to as the “fragment density matrix”. For a separable system such a density operator is idempotent by construction. Separability of the associated many-body wavefunctions $|\Psi^{\mathfrak{F}}\rangle$ also implies that different fragments are orthogonal, i.e. $\hat{F}^{\mathfrak{F}}\hat{F}^{\mathfrak{G}} = \hat{F}^{\mathfrak{F}}\delta_{\mathfrak{F}\mathfrak{G}}$. For a reasonable fragment definition we should require that the complete set of fragments represents a partitioning of the system, i.e.

$$\sum_{\mathfrak{F}} \hat{F}^{\mathfrak{F}} = \hat{F} \;. \tag{3}$$

To proceed further we assume that $\hat{W}^{\mathfrak{F}}$ can be provided in the same basis set as that used to describe the density matrix:

$$\hat{W}^{\mathfrak{F}} = \sum_{\mu,\nu} \left|\phi_\mu\right\rangle R^{\mathfrak{F}}_{\mu\nu} \left\langle\phi_\nu\right| \ , \tag{4}$$

where the (still to be defined) matrix $\mathbf{R}^{\mathfrak{F}}$ determines the character of the fragment projection operator; several examples will be given later.

For a QM system that is not genuinely separable, a "fragment quantity" is *not* a well-defined quantum observable. Of course there is no universal recipe to define the fragment partitioning, which leads to the question of the *pertinence* of the operator $\hat{W}^{\mathfrak{F}}$. We would like then to *quantify* the reliability of the identification of $\mathfrak{F}$ as a system's moiety by the projector defined from $\mathbf{R}^{\mathfrak{F}}$. If such a fragment restriction makes sense, the operator $\hat{F}^{\mathfrak{F}} \equiv \hat{F}\hat{W}^{\mathfrak{F}}$ should — following the above discussion — be idempotent, i.e. $\left(\hat{F}^{\mathfrak{F}}\right)^2 = \hat{F}^{\mathfrak{F}}$. Hence, the quantity

$$\mathrm{Tr}\left(\left(\hat{F}^{\mathfrak{F}}\right)^2 - \hat{F}^{\mathfrak{F}}\right) = \mathrm{Tr}\left(\left(\mathbf{K}\mathbf{S}^{\mathfrak{F}}\right)^2 - \mathbf{K}\mathbf{S}^{\mathfrak{F}}\right) \ , \tag{5}$$

with $\mathbf{S}^{\mathfrak{F}} \equiv \mathbf{S}\mathbf{R}^{\mathfrak{F}}\mathbf{S}$, is well suited to quantify the pertinence of fragment $\mathfrak{F}$ being considered as a genuine fragment of the full system. We will call this quantity from now on the *purity indicator*; the closer this index is to zero the more properly the fragment $\mathfrak{F}$ is identified. In order to define an intensive quantity we may additionally normalize the purity indicator and thus consider the quantity

$$\Pi = \frac{1}{q}\,\mathrm{Tr}\left(\left(\mathbf{K}\mathbf{S}^{\mathfrak{F}}\right)^2 - \mathbf{K}\mathbf{S}^{\mathfrak{F}}\right) \ , \tag{6}$$

where we indicate with q the total number of electrons of the fragment in gas phase.

The above derivation makes apparent that the purity indicator is an explicit functional of the matrix $\mathbf{R}^{\mathfrak{F}}$ and the basis set $\{\phi_\mu\}$. Consequently it is evident that this quantity is *not* a QM observable. Rather, it has to be interpreted as a *necessary* condition for the matrix $\mathbf{R}^{\mathfrak{F}}$ to be meaningful for the identification of a fragment within a given basis. If this condition is not fulfilled and the purity indicator is high, it is unlikely that the value of $\mathrm{Tr}\left(\hat{F}^{\mathfrak{F}}\hat{O}\right)$ can be associated with an observable quantity of the fragment $\mathfrak{F}$.

It is important to stress here that these criteria are less stringent than a simple spatial separation between the fragments, as they are defined in terms of entries of the density matrix operator in the employed basis set. As an illustrative example for a proper fragmentation, we can choose any operator that selects one (or more) KS orbitals,

$$\hat{W}^{j} = \left|\psi_j\right\rangle \left\langle\psi_j\right| \ . \tag{7}$$

Indeed this is a suitable definition: Due to the orthonormality of the KS orbitals the trace in Eq. (5) is exactly zero, and $\sum_j \hat{W}^j = \hat{F}$, thus also fulfilling Eq. (3). This is consistent with the obvious consideration that it makes sense to project density matrix-related quantities onto a subset of the KS orbitals.

B. Atomic charge population analyses

Traditionally the most common choice for the fragments are the individual atoms. We therefore want to briefly revisit some popular atomic charge population analyses. A pioneering example is provided by the Mulliken approach[24], which directly uses the atomically localized basis functions in which the QM molecular orbitals are expressed, and is thus conceptually very simple. On the other hand the outcome of the Mulliken analysis depends strongly on the used basis set (see e.g. Refs. 25,26 and references therein) and a bad choice might yield completely misleading results. The Löwdin population analysis[27,28] is akin in spirit, with the difference that it works with a set of orthonormalized orbitals. The strong sensitivity with respect to the basis set is considerably reduced by an approach like the natural population analysis (NPA)[25], which evaluates the atomic charges as the occupancies of a set of special "Natural Atomic Orbitals" (NAO). The advantage of NPA over Mulliken and Löwdin is that the first one is built upon "wavefunction-based" physical concepts, like the definition of the Natural Atomic Orbitals, whereas the latter ones rely on a partitioning scheme that considers *all* the basis functions on the atom on an equal footing.

Now we want to see how this connects to our general framework, by applying it to KS-DFT calculations and comparing with the aforementioned well-established methods. If a fragment is a well defined and independent subsystem, there exists a set of "fragment states" $|\psi^{\mathfrak{F}}_\mu\rangle$ (which are eigenfunctions of the projector $\hat{W}^{\mathfrak{F}}$), together with their dual functions $\langle\tilde{\psi}^{\mathfrak{F}}_\mu|$, thus fulfilling $\langle\tilde{\psi}^{\mathfrak{F}}_\mu|\psi^{\mathfrak{F}}_\nu\rangle = \delta_{\mu\nu}$. As we are here dealing with fragments formed by the individual atoms, we can in the same way assume atomic states $|\psi^A_\mu\rangle$ and define the projector $\hat{W}^A$ onto that atom by summing over them:

$$\hat{W}^A \equiv \sum_\mu |\psi^A_\mu\rangle\langle\tilde{\psi}^A_\mu| \ . \tag{8}$$

The most straightforward approach to identify fragments out of a system described by localized basis functions is to *associate* a set of basis functions with a given atom A. These atoms can then also eventually be combined to form a fragment $\mathfrak{F}$ constituted by this group of atoms, as will be discussed later. The restriction to an atom A can be implemented by the diagonal matrix $T^A_{\mu\nu} = \delta_{\mu\nu}\theta(A,\mu)$, where $\theta(A,\mu)$ is defined as

$$\theta(A,\mu) = \begin{cases} 1 & \text{if } \mu \text{ is associated with atom } A \\ 0 & \text{otherwise} \end{cases} \ . \tag{9}$$

Such an association is clearly arbitrary and is based on considerations about the (presumed) center of the associated basis function. Information about the basis extensions are often neglected and might lead to unreliable partitionings, as the clear association of a basis function with an atom is not obvious any more. When adopting this approach of fragment selection it is important to remember the previously mentioned bi-orthogonality and to distinguish between direct and dual "fragment states". Suppose we define $|\psi_\mu^A\rangle = \sum_\beta T^A_{\mu\beta}|\phi_\beta\rangle$ as an atomic state. The orthonormality constraint then imposes that $\langle\tilde{\psi}_\mu^A| = \sum_\beta \langle\phi_\beta|S^{-1}_{\beta\mu}$. By plugging this into Eq. (8) and comparing with Eq. (4) it follows that the projector matrix reads

$$\mathbf{R}^A_M = \mathbf{T}^A\mathbf{S}^{-1}\,. \tag{10}$$

As will be shown later, such a definition corresponds to nothing other than the traditional Mulliken population analysis.

Proceeding in an analogous way, we can also define the fragment states in terms of a basis which is first orthogonalized, giving $|\psi_\mu^A\rangle = \sum_{\beta\gamma} T^A_{\mu\beta}S^{-1/2}_{\beta\gamma}|\phi_\gamma\rangle$, and therefore $\langle\tilde{\psi}_\mu^A| = \langle\psi_\mu^A|$. This leads to the projector matrix

$$\mathbf{R}^A_L = \mathbf{S}^{-1/2}\mathbf{T}^A\mathbf{S}^{-1/2}\,, \tag{11}$$

which corresponds, as will be demonstrated later, to the definition of the Löwdin population analysis.

We may also revisit the NPA method under this light. Here the degrees of freedom of the subsystem are defined in the basis of Natural Atomic Orbitals (NAO) which are by construction orthonormal. These are generated in a procedure involving several steps, resulting in an expression that can be written as linear combinations (with coefficients $\mathbf{B}^A$) of the original basis functions projected on the atoms A (see Ref. 25):

$$|\psi_\mu^A\rangle = \sum_\beta B^A_{\mu\beta}|\phi_\beta\rangle\,. \tag{12}$$

Within this scheme the NAO projector operator is defined as

$$\mathbf{R}^A_{NAO} = \mathbf{B}^{A^T}\mathbf{T}^A\mathbf{B}^A\,. \tag{13}$$

The transformation matrix $\mathbf{B}^A$ is defined in such a way to ensure that the NAO are eigenstates of the density operator for a given atom, that can thus directly be written in terms of the sum over the NAO:

$$\hat{F}^A = \sum_\mu \theta(A,\mu)|\psi_\mu^A\rangle N_\mu^A\langle\psi_\mu^A|\,. \tag{14}$$

The NPA method is considered to be more robust than the Mulliken and Löwdin approaches, since it removes the strong dependence of the results on the basis set. This superiority is related to the fact that basis sets with diffuse degrees of freedom often contain components that considerably contribute to the description of empty states. In the NPA scheme their contribution is weighted by the eigenvalue N_α^A, whereas in the Mulliken or Löwdin scheme all the atomic components have the same weight. A similar approach to NPA is the use of so-called AOIMs (atomic orbitals in molecular environments)[29], which as well have the goal of providing a reliable and stable population analysis for variable (and in particular large) basis sets. The AOIMs are defined as the solution of the single-electron Schrödinger equation with an effective potential given by the spherical average of the molecular potential centered on the given atom. Once the AOIMs have been determined, a standard population scheme such as the Mulliken approach yields reliable and robust results.

For the Mulliken projector (Eq. (10)), the condition of Eq. (5) corresponds to the idempotency of the matrix $\mathbf{KST}^A$, i.e. the block of the $\mathbf{KS}$ matrix associated with the indices of atom A. The Löwdin projector (Eq. (11)), on the other hand, can be considered meaningful if the atomic block matrix of $\mathbf{S}^{1/2}\mathbf{KS}^{1/2}$ is close to idempotency. By orthogonality of the NAO, the NPA approach is idempotent if all the NPA eigenvalues N_α^A associated with the atom A are 0 or 1.

A situation in which Mulliken and Löwdin are unreliable corresponds to a setup that yields a non-pure atomic (or more general fragment) kernel. The above considerations show that this non-purity is not only a consequence of a inappropriate fragment choice, but also related to the basis. This is an important point, as it means that even simple population schemes might lead to unbiased and reliable results if the basis employed leads to pure fragment kernels. Indeed we show later in Sec. III that, whenever it is possible to identify a sensible fragment, a *minimal* basis leads — for both Mulliken and Löwdin — to such a favorable situation.

C. Generalized multipole decomposition

To analyze the features of the density matrix of a system, the most intuitive objects to use are the multipoles of the charge density $\rho(\mathbf{r})$. These read

$$\begin{aligned} Q^R_{\ell m} &\equiv \sqrt{\frac{4\pi}{2\ell+1}}\int \mathcal{S}_{\ell m}(\mathbf{r}-\mathbf{r}_R)\rho(\mathbf{r})\,\mathrm{d}\mathbf{r} \\ &= \sqrt{\frac{4\pi}{2\ell+1}}\,\mathrm{Tr}\left(\hat{F}\hat{\mathcal{S}}^R_{\ell m}\right) = \mathrm{Tr}\left(\mathbf{K}\mathbf{P}^R_{\ell m}\right)\,, \end{aligned} \tag{15}$$

where we have defined the multipole matrices $\mathbf{P}^R_{\ell m}$ as

$$P^R_{\ell m;\alpha\beta} = \sqrt{\frac{4\pi}{2\ell+1}}\,\langle\phi_\alpha|\hat{\mathcal{S}}^R_{\ell m}|\phi_\beta\rangle\,. \tag{16}$$

In the above equation the superscript R indicates that the solid harmonic operators $\hat{\mathcal{S}}^R_{\ell m}(\mathbf{r}) \equiv \mathcal{S}_{\ell m}(\mathbf{r}-\mathbf{r}_R)$ are centered on the reference position $\mathbf{r}_R$; their proper definitions are presented in Appendix A for completeness. We may therefore say that the electrostatic multipoles are functions of the density matrix and the center $\mathbf{r}_R$ of the reference system.

The resulting $Q^R_{\ell m}$ can however also be used for the calculation of multipoles with respect to a different origin $\mathbf{r}_{R'}$. As is shown in more detail in Appendix A we obtain the relation

$$Q^{R'}_{\ell m} = \sum_{\ell'=0}^{\ell} \sum_{m'=-\ell'}^{\ell'} Q^R_{\ell' m'} \mathcal{C}^{\ell m}_{\ell' m'}(\mathbf{r}_{R'} - \mathbf{r}_R) \ , \tag{17}$$

where the functions $\mathcal{C}^{\ell m}_{\ell' m'}(\mathbf{r})$ can be expressed in terms of the $\mathcal{S}_{\ell-\ell' m''}(\mathbf{r})$. For the important cases of the monopole and dipole components these equations are very simple and provide

$$Q^{R'}_{00} = Q^R_{00} \ , \tag{18a}$$

$$Q^{R'}_{1m} = \sqrt{\frac{3}{4\pi}} \mathcal{S}_{1m}(\mathbf{r}_{R'} - \mathbf{r}_R) Q^R_{00} + Q^R_{1m} \ . \tag{18b}$$

As the electrostatic multipoles are functions of $\mathbf{K}$ and $\mathbf{r}_R$, we can also obtain these quantities for a fragment of a system. All we have to do is to associate the fragment with a "fragment kernel" $\mathbf{K}^{\mathfrak{F}} \equiv \mathbf{K}\mathbf{S}\mathbf{R}^{\mathfrak{F}}$, by following the considerations of Sec. II A. The above definitions must therefore be generalized. Again restricting ourselves to the case of atomic fragments, this leads to the following definition of the atomic multipoles:

$$Q^A_{\ell m} \equiv \mathrm{Tr}(\mathbf{K}\mathbf{S}\mathbf{R}^A \mathbf{P}^A_{\ell m}) \ . \tag{19}$$

With this definition we can now also briefly revisit the projector matrices introduced in Sec. II B. As the monopole matrix is given by $\mathbf{P}_{00} = \mathbf{S}$, the monopole term for the Mulliken approach (Eq. (10)) reads $Q^A_{00} = \mathrm{Tr}(\mathbf{K}\mathbf{S}\mathbf{T}^A)$, i.e. the trace of $\mathbf{KS}$ evaluated only for those elements belonging to atom A, which is indeed nothing other than the well-known Mulliken charge population analysis. For the Löwdin approach (Eq. (11)) we obtain $Q^A_{00} = \mathrm{Tr}(\mathbf{S}^{1/2}\mathbf{K}\mathbf{S}^{1/2}\mathbf{T}^A)$, which indeed corresponds to the Löwdin charge population analysis. As $\sum_A \mathbf{T}^A = \mathbb{1}$, both the definitions satisfy the property of Eq. (3), which is important to ensure the preservation of the total monopole of the system. In other terms, we always have $\sum_A Q^A_{00} = \mathrm{Tr}\,(\mathbf{KS})$. In the NPA approach the self-duality of the NAO gives $Q^A_{00} = \mathrm{Tr}\left(\mathbf{N}^A\right)$.

If the fragment is not a single atom, but rather an ensemble of atoms, the projector $\hat{W}^{\mathfrak{F}}$ onto that fragment can then simply be defined as the sum over the projectors onto the atoms constituting the fragment, i.e. $\hat{W}^{\mathfrak{F}} = \sum_{A \in \mathfrak{F}} \hat{W}^A$. By linearity and by employing Eq. (17), we can obtain the fragment's multipoles in terms of their atomic counterparts:

$$Q^{\mathfrak{F}}_{\ell m} = \sum_{\ell'=0}^{\ell} \sum_{m'=-\ell'}^{\ell'} \sum_{A \in \mathfrak{F}} Q^A_{\ell' m'} \mathcal{C}^{\ell m}_{\ell' m'}(\mathbf{r}_{\mathfrak{F}} - \mathbf{r}_A) \ . \tag{20}$$

With a fragment projector defined as a sum of atomic projectors we can provide the *atomic* contribution to the electrostatic description of a given fragment. However, such an "atoms-in-molecule" description of the fragment must be taken with great care: Indeed, even if the fragment $\mathfrak{F}$ is reliable in the sense described by Eqs. (3) and (5), these conditions are in general *not* fulfilled for the atoms $A \in \mathfrak{F}$. If this is the case the atomic multipoles $Q^A_{\ell m}$ must *not* be considered as (pseudo-) observables as only the fragment as a whole is a reasonable partition of the system.

The above consideration is very important. A charge population analysis may be meaningful for a *molecule*, but not for the atoms belonging to the molecule; this is due to the fact that the atoms themselves are *not* separable entities of the molecule. Our criteria allow us to quantify this separability in the basis set used for the population analysis, thereby giving the possibility of associating (or not) such pseudo-observables with well-identified portions of the QM system.

III. RELATION BETWEEN FRAGMENT DEFINITION AND BASIS SET

We have presented a *quantitative* criterion to identify a fragment within a large system, and we pointed out that its fulfillment does not only depend on the actual fragmentation choice, but also on the nature of the support function basis. In other terms, the possibility of "splitting" a system into fragments is not only an intrinsic property of the system, but also of the set of support functions used to describe it.

Indeed, we have so far avoided any discussion about the specific localized basis set that we use — rather we simply assumed that a suitable choice exists. In principle there is no constraint on the exact form of the support functions — they can either be a contraction of an underlying basis set, which is the case for example in BigDFT[30,31] or ONETEP[22,32–34], or predefined atomic basis functions, either numerical or analytic, as for instance in Conquest[23,35,36], Quickstep[37] or SIESTA[38,39]; an overview over popular electronic structure codes and the basis sets they use can be found in Ref. 3.

In this section we want to discuss this important relation between fragment definition and basis set in more detail. More precisely, we define *a priori* the fragments and the projector matrix, and we discuss the impact of different basis sets in fragments identifications within this setup. As an illustrative test we take a system where the fragments can readily be identified by chemical intuition, namely a droplet of 100 water molecules, extracted manually from a larger bulk liquid water system; as it will only serve as a playground, no particular thermalization/relaxation was performed.

a. Basis Set Setups: Optimized Molecular Orbitals vs. *Atomic Orbitals* In Fig. 1 we present three quantities — density of states (DoS), purity indicator and molecular dipoles — for different basis sets. We compare a setup where we use the optimized quasi-orthogonal support functions of BigDFT (Fig. 1a) with a setup where we use non-optimized atomic orbitals (AO, Fig. 1b). The optimized support functions are obtained by minimizing — within the underlying Daubechies wavelet[40] basis of BigDFT — a target function that ensures both accuracy and locality, and it has been demonstrated that they are capable of representing the KS orbitals and derived quantities well[30,31]. The AO, on the other hand, are obtained by solving — again within the wavelet basis — the KS equation for the isolated atom using HGH pseudopotentials[41] including a nonlinear core correction[42]. Both the AO and the optimized basis were confined in localization regions centered around the atoms with a radius of 3.7 Å for H and 4.0 Å for O, and the PBE functional[43] was used.

For each setup we varied the number of support functions per O/H atom, namely (following the nomenclature of atomic orbitals) of type sp/s, spd/sp and spdf/spd. Note that in the augmented setups we did *not* alter the localization regions of the basis, we only included more components. All setups are compared with a reference calculation done using the cubic scaling version of the BigDFT code[44], which does not use any localization constraints. For the calculation of the purity indicator and the molecular dipoles we show results for both the Mulliken and Löwdin approaches, in order to also investigate the effect of the particular choice of the projector matrix.

b. Description of the Electronic Structure As can be seen from the uppermost panel of Fig. 1, all calculations with the optimized functions reproduce the reference DoS. The atomic orbitals, on the other hand, exhibit serious deviations for the minimal sp/s basis sets; reasonably accurate results can only be obtained for the larger spd/sp and, even better, spdf/spd setups. In other words, the basis must be larger, compared to the optimized case, to describe the electronic structure precisely — a fact which is well known from codes which use fixed atomic orbitals[45].

c. Purity indicator for the H_2O molecules Next we investigate the influence of these different basis sets on the fragment definitions, using the purity indicator derived above. According to the definition in Eq. (6), a value below a "level of confidence" of the order of a few percent seems to be low enough to consider the fragment as a subsystem. We set from now on our criterion to 5%; in other terms, we consider the subsystem as a fragment if the projection operator modifies the value of the fragment monopole by no more than 5%.

As can be seen from the values in the second row of Fig. 1, the setups using a small basis yield almost pure fragment kernels, whereas those using a larger basis lead to considerable deviations from zero.

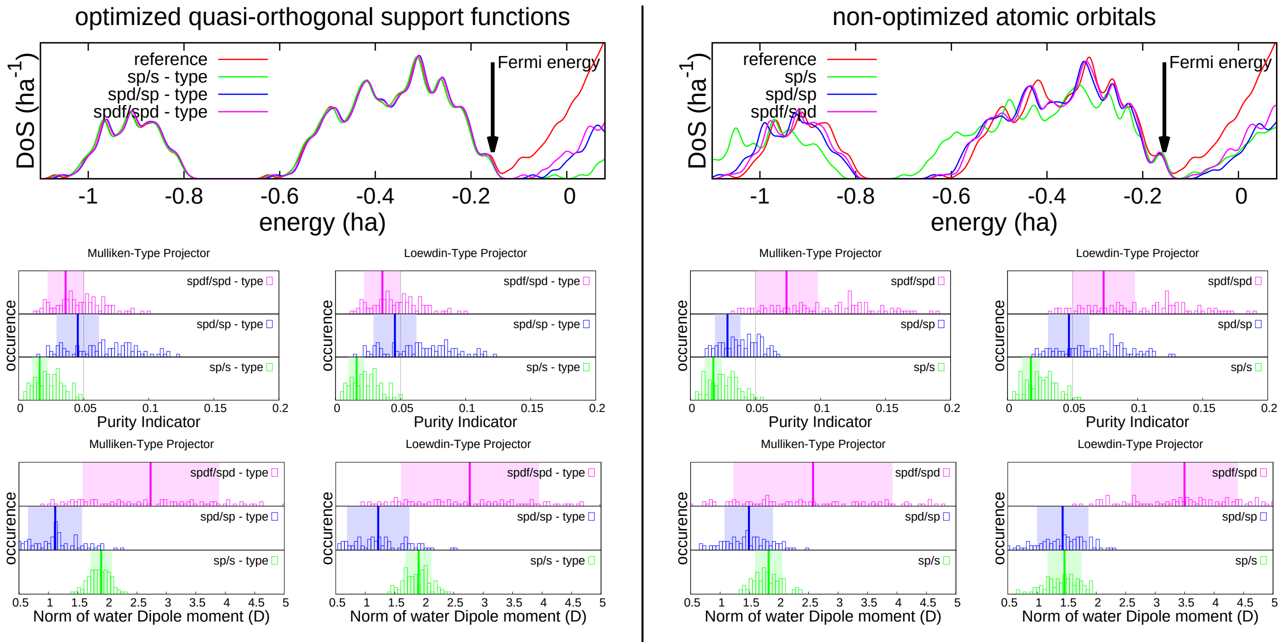
d. Influence on the measure of the molecular dipoles Let us now discuss how this translates into the calculation of the pseudo-observables of the fragments. In the third panel of Fig. 1, we plot the distribution of the individual water dipoles within the droplet, calculated as described in Sec. II C We have found the well-known result that the multipole values depend *strongly* on the basis set, even though the DoS is correctly reproduced. More precisely, we see that — in an apparently counterintuitive way — the more "complete" the basis set is, the less sound the results for these quantities are. However, taking into account the results from the second panel, these outcomes become understandable: For those setups yielding large values for the purity indicator we lose the interpretation of molecular dipoles as (pseudo-)observables. For the non-minimal basis sets, a Mulliken or Löwdin analysis appears therefore unjustified — in contrast to the minimal setup, where we get sound values of the molecular dipoles within this (unthermalized) toy droplet.

We should however recall that the purity indicator does not reflect the information about the completeness of the basis set, but only the suitability of the fragment identification *within* the basis. Indeed, for the sp/s setup, the purity indicators are equally good for the optimized and AO setups, as shown in Tab. I. Nevertheless, for the AO sp/s setup we still have — as pointed out before — a too crude representation of the electronic structure of the droplet, as the KS orbitals are badly expressed in this small AO basis.

e. Unreliability of atomic multipoles In Tab. I we also present the purity indicators considering only the individual atoms as fragments. Compared to the molecules, those values are substantially higher, indicating that atomic multipoles within a water molecule can *not* be considered as physical observables. Rather it is necessary to consider a water molecule as a single non-splittable unit, and only the multipole values for this unit can be considered as meaningful and allow a physical interpretation. We will give another demonstration of the unreliability of atomic multipoles in Sec. IV.

f. Reliability of Mulliken vs. *Loewdin* Additionally we also want to emphasize that all results for the optimized support functions are almost invariant under the choice between Mulliken and Löwdin, whereas the numbers obtained from the atomic orbitals change noticeably. This is a direct consequence of the quasi-orthogonality of the BigDFT support functions, in contrast to the non-orthogonality of the atomic orbitals. Indeed we see in Tab. II, showing the mean molecular dipole moments, that the AO Löwdin results are considerably worse than the three other setups. This can be explained by the fact that the Löwdin approach increases the support of the basis whilst orthogonalizing them, thereby losing the correspondence between orbital and atom.

g. Advantages of Optimized and Minimal (Molecular) basis In summary, the purity indicators suggest that *only* the minimal basis setup is meaningful within a Mulliken or Löwdin approach. We thus see a clear advantage of using a basis set which is optimized in situ, as indicated by the summary in Tab. III. Such an optimized minimal basis is *complete enough* for an accurate description of the electronic structure, but also *small enough* for



(a) optimized quasi-orthogonal support functions

(b) non-optimized atomic orbitals

FIG. 1. Comparison of the density of states (first row), purity indicator (second row), and molecular dipoles (third row), for a non-relaxed water droplet consisting of 100 molecules. For the density of states the curves were shifted such that the Fermi energies always coincide with that of the reference calculation, and a Gaussian smearing with $\sigma = 0.27\,\mathrm{eV}$ was applied. For the purity indicator and the dipole moments we present the result for both the Mulliken and Löwdin approaches. The vertical bar at 0.05 in the second panel indicates the "level of confidence", i.e. we consider a fragment to be reasonable for values below this threshold. Fig. 1a shows the outcome for the optimized quasi-orthogonal BigDFT support functions, whereas Fig. 1b shows the situation when the support functions are replaced by (unoptimized) atomic orbitals. As can be seen the first case is rather insensitive to the choice of the approach (Mulliken or Löwdin), whereas the non-orthogonal atomic orbitals show strong deviations. Moreover and most important for this study, the only setup which yields a good result for all measurements is that using the *minimal* set of optimized support functions.

	sp/s optimized			sp/s atomic orbitals		
	H_2O	O	H	H_2O	O	H
Mulliken	0.02(1)	0.16(1)	0.45(0)	0.03(1)	0.16(1)	0.46(1)
Löwdin	0.03(1)	0.16(1)	0.45(0)	0.03(1)	0.17(1)	0.48(0)
quality	✔	✘	✘	✔	✘	✘

TABLE I. Purity indicator of the droplet constituents for the sp/s setup, using the definition of Eq. (6). The values for the atoms are considerably larger than those for the entire molecules, indicating that the atoms alone should not be considered as independent fragments.

	sp/s optimized		sp/s atomic orbitals	
	Mulliken	Löwdin	Mulliken	Löwdin
H_2O dipole (D)	1.89(18)	1.90(22)	1.83(23)	1.46(29)
quality	✔	✔	✔	✘

TABLE II. Mean value of the molecular dipole moment of the droplet molecules, for the sp/s setups of Fig. 1.

	optimized			atomic orbitals		
	sp/s	spd/sp	spdf/spd	sp/s	spd/sp	spdf/spd
DoS	✔	✔	✔	✘	✔	✔
non-purity	✔	✘	✘	✔	✘	✘
H_2O dipole	✔	✘	✘	✔/✘	✘	✘

TABLE III. Summary of the quality of the description provided by the different setups, highlighting how the quality of the results potentially depends on the basis setup. Overall only the optimized minimal basis is able to provide reliability in all the categories.

an accurate description of atomic charges and molecular dipoles.

This advantage of a smaller basis for the characterization of the atomic charges and dipoles might appear counterintuitive. However we have to recall that the richer the basis is the more Rydberg states it contains, making the fragment kernel less pure since both Mulliken and Löwdin treat all basis functions on an equal footing. An approach aiming at coping as well with such larger basis sets should thus be able to filter out those basis functions which mainly contribute to the representation of virtual states. Since neither Mulliken nor Löwdin have this ability, this implies that these approaches work best – if not exclusively – for a minimal basis, which in turn means that it is indispensable to use an optimized basis set in order to reach a high precision. The other way around, we see that the use of a minimal and optimized basis allows the usage of simple projectors like Mulliken and Löwdin, without the need to resort to more involved approaches.

IV. APPLICATION TO A COMPLEX HETEROGENEOUS SYSTEM — SOLVATED DNA

In Sec. III we have seen that the use of a minimal and optimized basis allows the use of simple population schemes like Mulliken or Löwdin while still yielding a precise description of the electronic structure. We now want to apply the developed concepts to a complex heterogeneous system where the fragments are not immediately identifiable. We present results for a rather large system, namely an 11 base pair DNA fragment (made only of Guanine and Cytosine nucleotides) which is embedded into a sodium-water solution, giving in total 15,613 atoms. To get a realistic setup we took one snapshot from an extended MD simulation, run with Amber 11[46,47] and the ff99SB force field[48]; the system is shown in Fig. 2.

In spite of the large dimensions, the linear scaling approach of BigDFT[30,31] can easily perform a full QM calculation of the entire system. Following the considerations of Sec. III a minimal set of basis functions was employed. As a first step we took as candidates for the fragments just the individual atoms; in Fig. 3a we show the atomic charges that we get from such a fragment definition using the Mulliken projector.

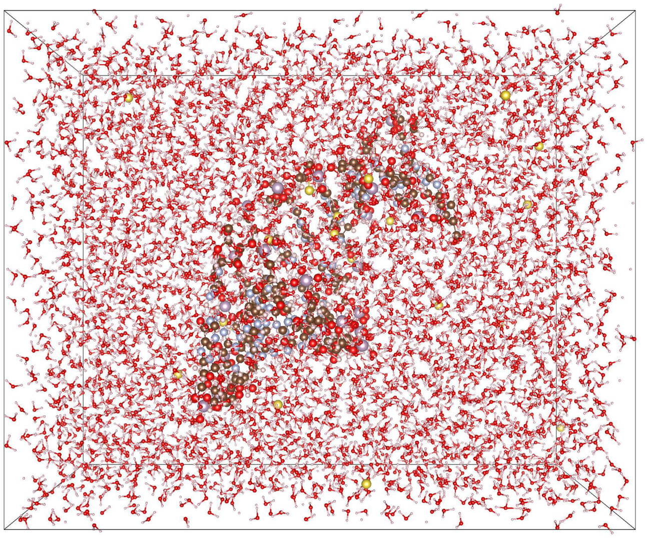
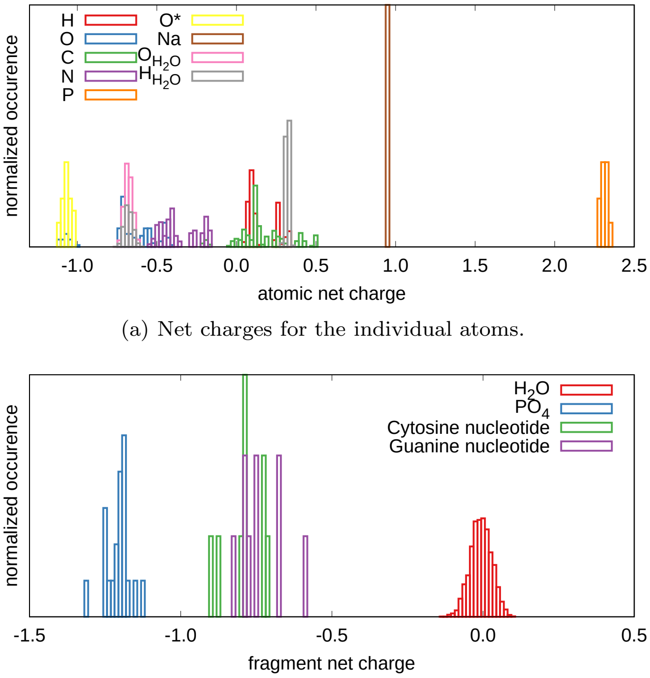


FIG. 2. Visualization[49] of the used DNA fragment (11 base pairs) in Na-water solution, consisting in total of 15,613 atoms.



(a) Net charges for the individual atoms.

(b) Net charges for some reasonably selected fragments.

FIG. 3. Fragment net charges for the system shown in Fig. 2 for various fragment definitions. In Fig. 3a we chose as fragments the individual atoms, whereas in Fig. 3b we chose as fragments groups of atoms, following chemical intuition.

a. Identification of systems' moieties To verify whether this fragment choice was sensible, we show in Tab. IVa the purity indicator for each atom type. As can be seen there are considerable differences, ranging from 4% for Na to 48% for H and C. Once again, this means that for such population methods in this basis care should be taken when extracting atomic charges and multipoles, as in general the atoms cannot be considered as "independent". As specific examples we focus on the two species which have a large positive net charge, namely Na and P. The purity indicator for Na is very small and thus confirms that *the basis functions employed* are in line with the chemical intuition that these Na atoms can be considered as "independent fragments", assuming a fragment selection provided by Mulliken-like projectors. This also agrees with their chemically sound atomic net charge, which is close to 1. The purity indicator for P, on the other hand, is considerably larger, together with a

	H	C	N	O	Os	Na	P
purity indicator	0.48	0.48	0.32	0.15	0.12	0.04	0.34
quality	✘	✘	✘	✘	✘	✔	✘

(a) Non-purity for the individual atoms.

	PO_4	Cyt	Gua	H_2O
purity indicator	0.05	0.01	0.01	0.01
quality	✔	✔	✔	✔

(b) Non-purity for some reasonably chosen fragments.

TABLE IV. Purity indicator according to Eq. (6) (using the Mulliken projector matrix of Eq. (10)), where the fragment is either a single atom (Tab. IVa) or — following chemical intuition — composed of several atoms (Tab. IVb).

surprisingly high value for its net charge. This indicates that P alone is not an optimal definition of a fragment in this case. Indeed the phosphorus atoms are part of a phosphate group PO_4, and if we take this unit as a fragment definition the purity indicator decreases considerably and is close to that of Na, as shown in Tab. IVb. The same scenario also applies for the other atoms of the system. If we consider each full nucleotide within the DNA as a fragment, we can see that the purity indicator decreases even further. The same order of magnitude can be observed for the water molecules, which — not surprisingly — again form a reliable fragment. In Fig. 3b we show the fragment charges for these more reasonable fragment choices. In summary, the purity indicator allows us — for a given choice of the basis and projector — to select fragments in an unbiased and reliable way.

b. Charge population analysis of the DNA nucleotides The above charge analysis also allows us to determine how much of the Na charge has gone to the DNA. The 20 Na atoms have lost 19.2 electrons (corresponding to an average ionization of 0.96), out of which 3.6 have gone to the water and the other 15.6 to the DNA. Considering the aforementioned purity indicators, the charge transfer appears to be chemically reliable as it corresponds to a transfer between well-defined fragments.

V. CONCLUSIONS AND OUTLOOK

The scope of this paper was to discuss the identification and representation of fragments within large quantum systems. In particular we aimed to answer the question of under which circumstances the properties of such fragments can be considered as meaningful (pseudo-)observables. As a basic criterion for the suitability of a fragment definition we identified the purity of the density matrix belonging to the fragment. This so-called purity indicator is a functional of the fragment projector chosen, the basis set employed, and the fragment considered.

If the value of the purity indicator is small, there is only a little coupling between the density operators of the fragment and the system, and the fragment can be considered as a meaningful sub-unit. In this case it is likely that the characteristics of the fragments can be considered as meaningful observables with a physical interpretation. Moreover, the reverse conclusion is even more important: Since a low value of the purity indicator is a *necessary* condition, it will be very difficult to describe the electronic structure of a fragment with meaningful observables — like for example electrostatic multipoles or partial DoS — if it does not fulfill this requirement within the given computational setup.

In addition we demonstrated that the use of an *optimized and minimal* localized basis set is of great advantage. This allows, on the one hand, to correctly identify the fragment even for simple projection methods like Mulliken and Loewdin, and on the other hand to describe the electronic structure with a high precision. Using a larger basis leads to considerably less pure fragment kernels, even for chemically sound fragments such as water molecules within a droplet, and thus renders the entire fragmentation procedure questionable; using a non-optimized basis requires the use of a large set of functions in order to correctly describe the electronic structure, which in turn leads to the aforementioned fragmentation issues and the need to use more delicate and involved fragment projection methods. Only the combination of a minimal and optimized basis thus provides satisfying results with respect to both aspects.

Concerning the observables, we focused in this paper on the multipoles of the fragments. Our formalism allows the calculation of multipoles of any order, which is important to provide an accurate description of the fragment's electrostatic potential[50], in line with established results based on atomic descriptions[51–57]. They might thus be used in the context of an electrostatic embedding, thereby reducing the complexity of a QM calculation and paving the way towards powerful multiscale calculations[3]. The use of such electrostatic observables in the context of embedded QM calculations will be considered in a forthcoming publication[58].

VI. ACKNOWLEDGMENTS

We would like to thank Thierry Deutsch for valuable discussions and Fátima Lucas for providing various test systems and helpful discussions. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. SM acknowledges support from the European Centre of Excellence MaX (project ID 676598). LG acknowledges support from the EU ExtMOS project (project ID 646176) and the European Centre of Excellence EoCoE (project ID 676629).

Appendix A: Definition of the spherical harmonics and translation relations

The (real) solid spherical harmonics are defined in terms of the corresponding complex functions as (using $\mathbf{r} \equiv (r, \Omega)$):

$$
\mathcal{S}_{\ell m}(r,\Omega) \equiv
\begin{cases}
\frac{1}{\sqrt{2}} r^{\ell} \left((-1)^m Y_{\ell m}(\Omega) + Y_{\ell,-m}(\Omega) \right) & m > 0 \ , \\
r^{\ell} Y_{\ell 0}(\Omega) & m = 0 \ , \\
\frac{1}{\sqrt{2}\mathrm{i}} r^{\ell} \left((-1)^m Y_{\ell |m|}(\Omega) - Y_{\ell,-|m|}(\Omega) \right) & m < 0 \ .
\end{cases}
\tag{A1}
$$

With these conventions they satisfy the orthogonality relation

$$
\int \frac{S_{\ell m}(r,\Omega) S_{\ell' m'}(r,\Omega)}{r^{2\ell}} \, \mathrm{d}\Omega = \delta_{\ell \ell'} \delta_{m m'} \ , \tag{A2}
$$

for any radial value $r > 0$. The real spherical harmonics satisfy the relation[59]

$$
\mathcal{S}_{\ell m}(\mathbf{r} + \mathbf{\Delta}) = \sum_{\ell'=0}^{\ell} \sum_{m'=-\ell'}^{\ell'} \mathcal{S}_{\ell' m'}(\mathbf{r}) \times \sqrt{\frac{4\pi}{2\ell'+1}} \mathcal{C}^{\ell m}_{\ell' m'}(\mathbf{\Delta}) \ , \tag{A3}
$$

where the functions

$$
\mathcal{C}^{\ell m}_{\ell' m'}(\mathbf{r}) = \sqrt{\frac{2\ell'+1}{4\pi}} \sum_{m''=\ell'-\ell}^{\ell-\ell'} \mathcal{S}_{\ell-\ell' m''}(\mathbf{r}) C^{\ell m}_{\ell' m' m''} \tag{A4}
$$

are described in terms of the coeffcients $C^{\ell m}_{\ell' m' m''}$ given by (see also Supplementary Information of Ref. 59):

$$
\begin{aligned}
C^{\ell,m}_{0,0,m''} &= \sqrt{4\pi} \delta_{m m''} \ , \\
C^{\ell,m}_{\ell,m',m''} &= \sqrt{4\pi} \delta_{m' m} \delta_{0 m''} \ , \\
C^{2,-2}_{1 m' m''} &= \sqrt{4\pi} \sqrt{\frac{5}{3}} \left(\delta_{m',-1} \delta_{m'',1} + \delta_{m',1} \delta_{m'',-1} \right) \ , \\
C^{2,-1}_{1 m' m''} &= \sqrt{4\pi} \sqrt{\frac{5}{3}} \left(\delta_{m',-1} \delta_{m'',0} + \delta_{m',0} \delta_{m'',-1} \right) \ , \\
C^{2,0}_{1 m' m''} &= \sqrt{4\pi} \frac{\sqrt{5}}{3} \left(-\delta_{m',-1} \delta_{m'',-1} \right. \\
&\qquad \left. + \ 2\delta_{m',0} \delta_{m'',0} - \delta_{m',1} \delta_{m'',1} \right) \ , \\
C^{2,1}_{1 m' m''} &= \sqrt{4\pi} \sqrt{\frac{5}{3}} \left(\delta_{m',0} \delta_{m'',1} + \delta_{m',1} \delta_{m'',0} \right) \ , \\
C^{2,2}_{1 m' m''} &= \sqrt{4\pi} \sqrt{\frac{5}{3}} \left(-\delta_{m',-1} \delta_{m'',-1} + \delta_{m',1} \delta_{m'',1} \right) \ .
\end{aligned}
\tag{A5}
$$

[*] stephan.mohr@bsc.es

[†] luigi.genovese@cea.fr

[1] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," Phys. Rev. **136**, B864 (1964).

[2] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," Phys. Rev. **140**, A1133 (1965).

[3] Laura E. Ratcliff, Stephan Mohr, Georg Huhs, Thierry Deutsch, Michel Masella, and Luigi Genovese, "Challenges in large scale quantum mechanical calculations," Wiley Interdiscip. Rev.-Comput. Mol. Sci. **7**, e1290 (2017), e1290.

[4] Mark S. Gordon, Dmitri G. Fedorov, Spencer R. Pruitt, and Lyudmila V. Slipchenko, "Fragmentation methods: A route to accurate calculations on large systems," Chem. Rev. **112**, 632 (2012), pMID: 21866983, http://dx.doi.org/10.1021/cr200093j.

[5] Michael A. Collins and Ryan P. A. Bettens, "Energy-based molecular fragmentation methods," Chem. Rev. **115**, 5607 (2015), pMID: 25843427, http://dx.doi.org/10.1021/cr500455b.

[6] Kenneth B. Wiberg and Paul R. Rablen, "Comparison of atomic charges derived via different procedures," J. Comput. Chem. **14**, 1504 (1993).

[7] Célia Fonseca Guerra, Jan-Willem Handgraaf, Evert Jan Baerends, and F. Matthias Bickelhaupt, "Voronoi deformation density (vdd) charges: Assessment of the mulliken, bader, hirshfeld, weinhold, and vdd methods for charge

analysis," J. Comput. Chem. **25**, 189 (2004).

[8] Cheol Ho Choi and Dmitri G. Fedorov, "Reducing the scaling of the fragment molecular orbital method using the multipole method," Chem. Phys. Lett. **543**, 159 (2012).

[9] Kazuo Kitaura, Eiji Ikeo, Toshio Asada, Tatsuya Nakano, and Masami Uebayasi, "Fragment molecular orbital method: an approximate computational method for large molecules," Chem. Phys. Lett. **313**, 701 (1999).

[10] Dmitri G. Fedorov and Kazuo Kitaura, "Extending the power of quantum chemistry to large systems with the fragment molecular orbital method," J. Phys. Chem. A **111**, 6904 (2007), pMID: 17511437, http://dx.doi.org/10.1021/jp0716740.

[11] Jiali Gao, "Toward a molecular orbital derived empirical potential for liquid simulations," J. Phys. Chem. B **101**, 657 (1997), http://dx.doi.org/10.1021/jp962833a.

[12] Jiali Gao, "A molecular-orbital derived polarization potential for liquid water," J. Chem. Phys. **109**, 2346 (1998), http://dx.doi.org/10.1063/1.476802.

[13] Scott J. Wierzchowski, David A. Kofke, and Jiali Gao, "Hydrogen fluoride phase behavior and molecular structure: A qm/mm potential model approach," J. Chem. Phys. **119**, 7365 (2003), http://dx.doi.org/10.1063/1.1607919.

[14] Wangshen Xie and Jiali Gao, "Design of a next generation force field: the x-pol potential," J. Chem. Theory Comput. **3**, 1890 (2007), pMID: 18985172, http://dx.doi.org/10.1021/ct700167b.

[15] Wangshen Xie, Lingchun Song, Donald G. Truhlar, and Jiali Gao, "The variational explicit polarization potential and analytical first derivative of energy: Towards a next generation force field," J. Chem. Phys. **128**, 234108 (2008), http://dx.doi.org/10.1063/1.2936122.

[16] Wangshen Xie, Lingchun Song, Donald G. Truhlar, and Jiali Gao, "Incorporation of a qm/mm buffer zone in the variational double self-consistent field method," J. Phys. Chem. B **112**, 14124 (2008), pMID: 18937511, http://dx.doi.org/10.1021/jp804512f.

[17] Fangqin Wu, Wenjian Liu, Yong Zhang, and Zhendong Li, "Linear-scaling time-dependent density functional theory based on the idea of from fragments to molecule," J. Chem. Theory Comput. **7**, 3643 (2011), pMID: 26598260, http://dx.doi.org/10.1021/ct200225v.

[18] Junzi Liu, Yong Zhang, and Wenjian Liu, "Photoexcitation of light-harvesting cpc60 triads: A flmo-td-dft study," J. Chem. Theory Comput. **10**, 2436 (2014), pMID: 26580764, http://dx.doi.org/10.1021/ct500066t.

[19] Zhendong Li, Hongyang Li, Bingbing Suo, and Wenjian Liu, "Localization of molecular orbitals: From fragments to molecule," Acc. Chem. Res. **47**, 2758 (2014), pMID: 25019464, http://dx.doi.org/10.1021/ar500082t.

[20] Hongyang Li, Wenjian Liu, and Bingbing Suo, "Localization of open-shell molecular orbitals via least change from fragments to molecule," J. Chem. Phys. **146**, 104104 (2017), http://dx.doi.org/10.1063/1.4977929.

[21] E. Hernández and M. J. Gillan, "Self-consistent first-principles technique with linear scaling," Phys. Rev. B **51**, 10157 (1995).

[22] Chris-Kriton Skylaris, Peter D. Haynes, Arash A. Mostofi, and Mike C. Payne, "Introducing onetep: Linear-scaling density functional simulations on parallel computers," J. Chem. Phys. **122**, 084119 (2005).

[23] D. R. Bowler, R. Choudhury, M. J. Gillan, and T. Miyazaki, "Recent progress with large-scale ab initio calculations: the conquest code," Phys. Status Solidi B **243**, 989 (2006).

[24] "Electronic population analysis on lcao-mo molecular wave functions. i," J. Chem. Phys. **23**, 1833 (1955), http://dx.doi.org/10.1063/1.1740588.

[25] Alan E. Reed, Robert B. Weinstock, and Frank Weinhold, "Natural population analysis," J. Chem. Phys. **83**, 735 (1985).

[26] Peter Politzer and Robert S. Mulliken, "Comparison of two atomic charge definitions, as applied to the hydrogen fluoride molecule," J. Chem. Phys. **55**, 5135 (1971).

[27] Per-Olov Löwdin, "On the non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals," J. Chem. Phys. **18**, 365 (1950).

[28] Per-Olov Löwdin, "On the nonorthogonality problem," Adv. Quantum Chem. **5**, 185 (1970).

[29] Wenjian Liu and Lemin Li, "A method for population and bonding analyses in calculations with extended basis sets," Theor. Chim. Acta **95**, 81 (1997).

[30] Stephan Mohr, Laura E. Ratcliff, Paul Boulanger, Luigi Genovese, Damien Caliste, Thierry Deutsch, and Stefan Goedecker, "Daubechies wavelets for linear scaling density functional theory," J. Chem. Phys. **140**, 204110 (2014).

[31] Stephan Mohr, Laura E. Ratcliff, Luigi Genovese, Damien Caliste, Paul Boulanger, Stefan Goedecker, and Thierry Deutsch, "Accurate and efficient linear scaling dft calculations with universal applicability," Phys. Chem. Chem. Phys. **17**, 31360 (2015).

[32] Peter D. Haynes, Chris-Kriton Skylaris, Arash A. Mostofi, and Mike C. Payne, "Onetep: linear-scaling density-functional theory with local orbitals and plane waves," Phys. Status Solidi B **243**, 2489 (2006).

[33] A. A. Mostofi, P. D. Haynes, C. K. Skylaris, and M. C. Payne, "Onetep: linear-scaling density-functional theory with plane-waves," Mol. Simul. **33**, 551 (2007), http://dx.doi.org/10.1080/08927020600932801.

[34] Chris-Kriton Skylaris, Peter D Haynes, Arash A Mostofi, and Mike C Payne, "Recent progress in linear-scaling density functional calculations with plane waves and pseudopotentials: the onetep code," J. Phys.: Condens. Matter **20**, 064209 (2008).

[35] D. R. Bowler, I. J. Bush, and M. J. Gillan, "Practical methods for ab initio calculations on thousands of atoms," Int. J. Quantum Chem. **77**, 831 (2000).

[36] D R Bowler and T Miyazaki, "Calculations for millions of atoms with density functional theory: linear scaling shows its potential." J. Phys.: Condens. Matter **22**, 074207 (2010).

[37] Joost VandeVondele, Matthias Krack, Fawzi Mohamed, Michele Parrinello, Thomas Chassaing, and Jürg Hutter, "Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach," Comput. Phys. Commun. **167**, 103 (2005).

[38] José M Soler, Emilio Artacho, Julian D Gale, Alberto García, Javier Junquera, Pablo Ordejón, and Daniel Sánchez-Portal, "The siesta method for ab initio order- n materials simulation," J. Phys.: Condens. Matter **14**, 2745 (2002).

[39] Emilio Artacho, E Anglada, O Diéguez, J D Gale, A García, J Junquera, R M Martin, P Ordejón, J M Pruneda, D Sánchez-Portal, and J M Soler,

"The siesta method; developments and applicability," J. Phys.: Condens. Matter **20**, 064208 (2008).

[40] Ingrid Daubechies, *Ten lectures on wavelets* (Society for Industrial and Applied Mathematics, Philadelphia, 1992).

[41] C. Hartwigsen, S. Goedecker, and J. Hutter, "Relativistic separable dual-space gaussian pseudopotentials from h to rn," Phys. Rev. B **58**, 3641 (1998).

[42] Alex Willand, Yaroslav O Kvashnin, Luigi Genovese, Álvaro Vázquez-Mayagoitia, Arpan Krishna Deb, Ali Sadeghi, Thierry Deutsch, and Stefan Goedecker, "Norm-conserving pseudopotentials with chemical accuracy compared to all-electron calculations," J. Chem. Phys. **138**, 104109 (2013).

[43] John P. Perdew, Kieron Burke, and Matthias Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett. **77**, 3865 (1996).

[44] Luigi Genovese, Alexey Neelov, Stefan Goedecker, Thierry Deutsch, Seyed Alireza Ghasemi, Alexander Willand, Damien Caliste, Oded Zilberberg, Mark Rayson, Anders Bergman, and Reinhold Schneider, "Daubechies wavelets as a basis set for density functional pseudopotential calculations." J. Chem. Phys. **129**, 014109 (2008).

[45] Frank Jensen, "Atomic orbital basis sets," Wiley Interdiscip. Rev.-Comput. Mol. Sci. **3**, 273 (2013).

[46] David A. Case, Thomas E. Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M. Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J. Woods, "The Amber biomolecular simulation programs," J. Comput. Chem. **26**, 1668 (2005), arXiv:NIHMS150003.

[47] David A. Case, T. A. Darden, T. E. Cheatham, Carlos L. Simmerling, J. Wang, Robert E. Duke, Ray Luo, Michael Crowley, Ross C. Walker, W. Zhang, K. M. Merz, B. Wang, S. Hayik, Adrian Roitberg, Gustavo Seabra, I. Kolossváry, K. F. Wong, F. Paesani, J. Vanicek, X. Wu, Scott R. Brozell, Tom Steinbrecher, Holger Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D. H. Mathews, M. G. Seetin, C. Sagui, V. Babin, and Peter A. Kollman, *Amber 11*, University of California, San Francisco.

[48] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling, "Comparison of multiple amber force fields and development of improved protein backbone parameters," Proteins Struct. Funct. Bioinf. **65**, 712 (2006).

[49] Koichi Momma and Fujio Izumi, "VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data," J. Appl. Crystallogr. **44**, 1272 (2011).

[50] A. J. Stone, "Distributed multipole analysis, or how to describe a molecular charge distribution," Chem. Phys. Lett. **83**, 233 (1981).

[51] W. Andrzej Sokalski and R. A. Poirier, "Cumulative atomic multipole representation of the molecular charge distribution and its basis set dependence," Chem. Phys. Lett. **98**, 86 (1983).

[52] Donald E. Williams, "Representation of the molecular electrostatic potential by atomic multipole and bond dipole models," J. Comput. Chem. **9**, 745 (1988).

[53] W.A. Sokalski, M. Shibata, R. Rein, and R.L. Ornstein, "Cumulative atomic multipole moments complement any atomic charge model to obtain more accurate electrostatic properties," J. Comput. Chem. **13**, 883 (1992).

[54] C. E. Whitehead, C. M. Breneman, N. Sukumar, and M. D. Ryan, "Transferable atom equivalent multicentered multipole expansion method," J. Comput. Chem. **24**, 512 (2003).

[55] Graeme M. Day, W. D. Sam Motherwell, and William Jones, "Beyond the Isotropic Atom Model in Crystal Structure Prediction of Rigid Molecules: Atomic Multipoles versus Point Charges," Cryst. Growth Des. **5**, 1023 (2005).

[56] Nuria Plattner and Markus Meuwly, "Higher order multipole moments for molecular dynamics simulations," J. Mol. Model. **15**, 687 (2009).

[57] Christian Kramer, Tristan Bereau, Alexander Spinn, Klaus R. Liedl, Peter Gedeck, and Markus Meuwly, "Deriving static atomic multipoles from the electrostatic potential," J. Chem. Inf. Model. **53**, 3410 (2013).

[58] Stephan Mohr, Michel Masella, Laura E. Ratcliff, and Luigi Genovese, "Complexity reduction in large quantum systems: Reliable electrostatic embedding for multiscale approaches via optimized minimal basis functions," (to be submitted).

[59] Jaime Fernández Rico, Rafael López, Ignacio Ema, and Guillermo Ramírez, "Translation of real solid spherical harmonics," Int. J. Quantum Chem. **113**, 1544 (2013).