Complexity Reduction in Density Functional Theory Calculations of Large Systems: System Partitioning and Fragment Embedding

William Dawson,*,† Stephan Mohr,‡ Laura E. Ratcliff,¶ Takahito Nakajima,† and Luigi Genovese*,§

†*RIKEN Center for Computational Science, Kobe, Japan*
‡*Barcelona Supercomputing Center (BSC)*
¶*Department of Materials, Imperial College London, London SW7 2AZ, United Kingdom*
§*Univ. Grenoble Alpes, INAC-MEM, L_Sim, F-38000 Grenoble, France*
‖*CEA, INAC-MEM, L_Sim, F-38000 Grenoble, France*

E-mail: william.dawson@riken.jp; luigi.genovese@cea.fr

Abstract

With the development of low order scaling methods for performing Kohn-Sham Density Functional Theory, it is now possible to perform fully quantum mechanical calculations of systems containing tens of thousands of atoms. However, with an increase in the size of system treated comes an increase in complexity, making it challenging to analyze such large systems and determine the cause of emergent properties. To address this issue, in this paper we present a systematic complexity reduction methodology which can break down large systems into their constituent fragments, and quantify inter-fragment interactions. The methodology proposed here requires no a priori information or user interaction, allowing a single workflow to be automatically applied to any system of interest. We apply this approach to a variety of different systems, and show how it allows for the derivation of new system descriptors, the design of QM/MM partitioning schemes, and the novel application of graph metrics to molecules and materials.

1 Introduction

Linear scaling algorithms for Kohn-Sham (KS) Density Functional Theory (DFT),[1,2] developed already some time ago,[3,4] have recently become accessible to a broader community thanks to the introduction of reliable and robust approaches (see e.g. Ref. 5 and references therein). This fact has important consequences for the interpretation and design of first-principle approaches, as the possibility of tackling systems of unconventionally large sizes allows for the addressing of new scientific questions. However, when treating heterogeneous systems, an increase in system size leads to an increase in complexity, making the interpretation of computational results challenging.

For a system containing many thousand atoms, it is likely that the fundamental constituents (or "moieties") of the system are of $\mathcal{O}(1)$, i.e. their size does not increase with the total number of atoms of the system. It appears therefore interesting to single out such moieties, and to try to model their mutual interactions with a less complex description. Thanks to linear scaling DFT techniques, the full quantum-mechanical (QM) calculation of the original system can be used as an assessment of the quality

of such simplified descriptions.

When linking together various length scales, such considerations are no longer optional, but they rather become compulsory. Performing a set of production QM simulations with an unnecessarily costly approach would result in a study of poor quality, as the simulation scheme entangles interactions with different length scales and couplings. In other terms, the dogma "the more complex the simulation the better" is not true in all situations. Taking these considerations into account allows one to focus on the regions of the system that *require* a high level of theory, leading to a better understanding of the fundamental mechanisms and avoiding an unnecessary waste of computational resources.

In this context — which we will from now on denote as "complexity reduction" — we briefly want to point out the important difference between *fragmentation* and *embedding*. In the first case, the full QM system is partitioned into several fragments, which are each individually treated at a full QM level, but which are mutually interacting in a simplified way. Fragmentation methods are conceived to simplify the full ab-initio calculation of a large QM system, i.e. they aim to treat the entire system at the same level of theory. Famous examples are, for instance, the Fragment Molecular Orbital approach,[6,7] the X-Pol method,[8–15] the Molecular Tailoring Approach,[16–19] or subsystem DFT.[20,21] Embedding methods, on the other hand, aim to split the system into a target region and an environment, each treated at different computational cost. Embedding approaches use various levels of theory within a single calculation, thus paving the way towards coarse grained models which can be used within multi-scale QM/MM simulations. Among others, we quote here the methods detailed in Refs. 22–36.

A problem common to both fragmentation and embedding methods is how to derive a general and reliable method for partitioning an arbitrary system into a set of fragments. As a matter of fact, the concept of fragmentation is to some extent an "ad-hoc" operation, based on the assumption that the system can be somehow partitioned into subsystems that mutually interact. In a previous publication,[37] we derived a simple method of determining in a quantitative way whether a chosen fragmentation is reasonable. If this is the case, the fragments become "independent" of each other and can be assigned "pseudo-observables" i.e. quantities with an interpretable physicochemical meaning.

In this paper, we build upon our previous work on evaluating fragments in order to develop a full methodology for complexity reduction. We will begin in Sec. 2 by reintroducing the *purity indicator* as a measure of fragment quality. Then in Sec. 3, we will define a new measure called the *fragment bond order*, which quantifies the interaction strength between fragments. We then will utilize the fragment bond order to determine the chemical significance of the purity indicator. In Sec. 4 we will further use the fragment bond order to define an embedding environment for fragments, and show how that can be used to build a graph like view of a molecular system. In Sec. 4.1, we will describe an automatic procedure that uses the fragment bond order to fragment a given system such that the purity indicator is close to zero for each fragment. Finally, we will conclude by demonstrating this methodology on a number of systems, and discuss how this methodology might bring together the concepts of fragmentation and embedding, enabling general multilayered schemes for both the calculation and interpretation of complex, heterogeneous systems.

2 Fragmentation and Interpretation of Observables

In a QM system, the expectation value of a one-body observable $\hat{O}$ can be expressed as $\langle \hat{O} \rangle \equiv \mathrm{tr}\left(\hat{F}\hat{O}\right)$, where we denote by $\hat{F} = |\Psi\rangle\langle\Psi| = \hat{F}^2$ the one-body density matrix of the system, that can be identified in terms of the ground-state wavefunction $|\Psi\rangle$. When a QM system is susceptible to be genuinely separable in to fragment states $|\Psi^{\mathcal{F}}\rangle$, it should be possible to define

a projection operator $\hat{W}^{\mathcal{F}}$ associated with each fragment $\mathcal{F}$ such that $\hat{W}^{\mathcal{F}}|\Psi\rangle = |\Psi^{\mathcal{F}}\rangle$. Performing such a fragmentation operation *a posteriori* is a procedure that presents, of course, some degrees of arbitrariness and is susceptible to provide, in the worst case, a system partitioning into physically meaningless moieties. The spirit of the fragmentation procedure described in[37] is to provide indicators that helps in assessing the *physical pertinence* of a given fragmentation. Let us briefly review this methodology here.

We assume that the density matrix of the system, as well as the projection operator, can be defined in a set of localized, not necessarily orthonormal, basis functions $|\phi_\alpha\rangle$ as follows:

$$\hat{F} = \sum_{\alpha,\beta} |\phi_\alpha\rangle K_{\alpha\beta} \langle\phi_\beta| \ , \tag{1}$$

$$\hat{W}^{\mathcal{F}} = \sum_{\mu,\nu} |\phi_\mu\rangle \, R^{\mathcal{F}}_{\mu\nu} \, \langle\phi_\nu| \ , \tag{2}$$

and that a generic one-body operator $\hat{O}$ can be expressed by the matrix elements $O_{\alpha\beta} = \langle\phi_\alpha| \, \hat{O} \, |\phi_\beta\rangle$. In this context the overlap matrix $S_{\alpha\beta} \equiv \langle\phi_\alpha|\phi_\beta\rangle$ can be seen as the matrix representation of the identity operator.

To be meaningful, the fragment projector should satisfy:

$$\hat{W}^{\mathcal{F}}\hat{W}^{\mathcal{G}} = \hat{W}^{\mathcal{F}}\delta_{\mathcal{F}\mathcal{G}} \Rightarrow \mathbf{R}^{\mathcal{F}}\mathbf{S}\mathbf{R}^{\mathcal{G}} = \mathbf{R}^{\mathcal{F}}\delta_{\mathcal{F}\mathcal{G}}, \tag{3}$$

$$\sum_{\mathcal{F}} \hat{W}^{\mathcal{F}} = \hat{\mathbb{I}} \Rightarrow \sum_{\mathcal{F}} \mathbf{S}\mathbf{R}^{\mathcal{F}}\mathbf{S} = \mathbf{S}, \tag{4}$$

which are the obvious orthogonality (including projection) and resolution-of-the-identity conditions that a reasonable fragmentation should implement. The latter condition, when combined with the idempotency of $\hat{F}$, provides $\hat{F}\left(\sum_{\mathcal{F}} \hat{W}^{\mathcal{F}}\right)\hat{F} = \hat{F}$, which would imply the interesting equation:

$$\hat{W}^{\mathcal{G}}\hat{F}\left(\sum_{\mathcal{F}} \hat{W}^{\mathcal{F}}\right)\hat{F}\hat{W}^{\mathcal{G}} = \hat{W}^{\mathcal{G}}\hat{F}\hat{W}^{\mathcal{G}}. \tag{5}$$

Nonetheless, when the system's fragmentation is exact, the fragment density matrices $|\psi^{\mathcal{F}}\rangle \langle\psi^{\mathcal{F}}| = \hat{W}^{\mathcal{F}}\hat{F}\hat{W}^{\mathcal{F}}$ should also be idempotent. Together with Eqs. (5) and (3) this would imply:

$$\hat{W}^{\mathcal{G}}\hat{F}\left(\sum_{\mathcal{F}\neq\mathcal{G}} \hat{W}^{\mathcal{F}}\right)\hat{F}\hat{W}^{\mathcal{G}} = 0 \ , \tag{6}$$

which is a condition that can be realized (excluding pathological situations) by assuming that *in a meaningful fragmentation, the fragment representation of the density matrix is block-diagonal*, i.e. $\hat{W}^{\mathcal{G}}\hat{F}\hat{W}^{\mathcal{F}} \equiv \hat{F}W^{\mathcal{F}}\delta_{\mathcal{F}G} \equiv \hat{W}^{\mathcal{F}}\hat{F}\delta_{\mathcal{F}G}$ in the span of the basis set chosen. We may therefore rephrase a meaningful fragmentation as the *purity condition* $(\hat{F}^{\mathcal{F}})^2 = \hat{F}^{\mathcal{F}}$ where we have defined the fragment density matrix as $\hat{F}^{\mathcal{F}} \equiv \hat{F}W^{\mathcal{F}}$. Such a condition depends on the *combination* of the basis set ϕ_α and the projection $\mathbf{R}^{\mathcal{F}}$, and cannot be guaranteed a priori, nor is it a sufficient condition for fragmentation. Simply, when this condition holds, a system is susceptible to be fragmented by the set of projections identified by the operators $\hat{W}^{\mathcal{F}}$. However, we emphasize that the purity condition above is more stringent than the idempotency of the operator $|\psi^{\mathcal{F}}\rangle \langle\psi^{\mathcal{F}}|$ as the latter would impose the block-like behaviour of $\hat{F}$.

At the same time, an operator can be projected onto the fragment subspace by defining $\hat{O}^{\mathcal{F}} \equiv \hat{W}^{\mathcal{F}}\hat{O}\hat{W}^{\mathcal{F}}$, which would provide, in the basis set representation, $\mathbf{O}^{\mathcal{F}} = \mathbf{S}\mathbf{R}^{\mathcal{F}}\mathbf{O}\mathbf{R}^{\mathcal{F}}\mathbf{S}$.

The purity condition is itself represented in the basis set by the expression:

$$\mathbf{K}\mathbf{S}\mathbf{R}^{\mathcal{F}}\mathbf{S}\mathbf{K}\mathbf{S}\mathbf{R}^{\mathcal{F}} = \mathbf{K}\mathbf{S}\mathbf{R}^{\mathcal{F}}, \tag{7}$$

whose trace enables us to introduce the *purity indicator*, defined by:

$$\Pi_{\mathcal{F}} = \frac{1}{q_{\mathcal{F}}} \mathrm{Tr}\left(\left(\mathbf{K}\mathbf{S}^{\mathcal{F}}\right)^2 - \mathbf{K}\mathbf{S}^{\mathcal{F}}\right) \ , \tag{8}$$

where $q_{\mathcal{F}}$ is the total number of electrons of the isolated fragment in gas phase and $\mathbf{S}^{\mathcal{F}} \equiv \mathbf{S}\mathbf{R}^{\mathcal{F}}\mathbf{S}$. We note that $\Pi \leq 0$ and call *pure* a fragment whose projection satisfies the condition $\Pi \simeq 0$.

Such a condition, which we emphasize to be *non-linear* in the projector matrix elements $R^{\mathcal{F}}$, when fulfilled, enables one to interpret the

fragment-expectation value:

$$\langle \hat{O} \rangle_{\mathcal{F}} \equiv \mathrm{tr}\left(\hat{F}^{\mathcal{F}}\hat{O}\right) = \mathrm{tr}\left(\mathbf{KSR}^{\mathcal{F}}\mathbf{O}\right) \quad , \qquad (9)$$

as a pseudo-observable of the fragment $\mathcal{F}$. Indeed, by resolution-of-the-identity, we may decompose the expectation value in to fragment-wise values, namely $\langle \hat{O} \rangle = \sum_{\mathcal{F}} \langle \hat{O} \rangle_{\mathcal{F}}$. We retrieve here the *extensivity* of the expectation values: as this condition is linear in the fragment projection operator, a collection of fragments is itself a fragment and their expectation value is the sum of the separate contributions. More importantly, thanks to this property a fragment pseudo-observable can be *decomposed* into different contributions. Let $\mathbf{R}^{\mathcal{F}} = \mathbf{R}^{\mathcal{F}}_{\mathbf{1}} + \mathbf{R}^{\mathcal{F}}_{\mathbf{2}}$. Even if the fragments $\mathcal{F}_{1,2}$ were not pure, still we would have $\langle \hat{O} \rangle_{\mathcal{F}} = \langle \hat{O} \rangle_{\mathcal{F}_1} + \langle \hat{O} \rangle_{\mathcal{F}_2}$. This fact enables us to define the fragment projection matrix from, for example, atomic projectors, even when, as in most of the cases, the atoms cannot be considered as pure system moieties.

Instead of Eq. (9), we could have defined the fragment expectation value by the equation:

$$\langle \hat{O} \rangle_{\mathcal{F}} = \mathrm{tr}\left(\hat{W}^{\mathcal{F}}\hat{F}\hat{W}^{\mathcal{F}}\hat{O}\right) = \mathrm{tr}\left(\hat{F}\hat{O}^{\mathcal{F}}\right) \ , \quad (10)$$

which we know for a pure fragment would have lead to the same result. This shows that, even in the case of an operator that is not fragment-block diagonal, for a pure fragment only the diagonal term contributes to the expectation value, which is a natural result of the use of Hermitian operators.

2.1 Population Analysis of Fragments

Within this framework, traditional population analysis schemes might be extended to a system's fragments. In Ref.[37] we introduced expressions for the Mulliken (M) and Löwdin (L) projectors, which in the basis representation are:

$$\mathbf{R}^{\mathcal{F}}_{M} \equiv \mathbf{T}^{\mathcal{F}}\mathbf{S}^{-1} \ , \quad \mathbf{R}^{\mathcal{F}}_{M} \equiv \mathbf{S}^{-1/2}\mathbf{T}^{\mathcal{F}}\mathbf{S}^{-1/2} \ , \ (11)$$

where $T^{\mathcal{F}}$ is a diagonal matrix which has a value of one for the indices $\alpha \in \mathcal{F}$ that are associated to the fragment $\mathcal{F}$. Such an association is somehow arbitrary, in the sense that it is based on simple geometric considerations on the domain of the basis functions. The value of $\Pi^{M,L}_{\mathcal{F}}$ enables one to assess whether the fragmentation is reliable within the chosen population scheme. Also, the matrix $T^{\mathcal{F}}$ may be expressed as:

$$T^{\mathcal{F}} = \sum_{a \in \mathcal{F}} T^{a} \ , \qquad (12)$$

where we define the matrices T_a by associating each index α to one atom a of the system. We retrieve in this way the traditional Mulliken and Löwdin atomic projections. The well-known unreliability of these population methods for atoms may the be ascribed to the fact that, in general, the atoms cannot be associated to pure fragments: in most of the cases Π_a would be significantly different from zero.

3 Significance of the Purity Indicator

We may give to the purity indicator a chemical significance. Indeed, given a basis set and a projection method, the orbital population of a fragment can be defined as follows:

$$q_{\mathcal{F}}\Pi_{\mathcal{F}} = \langle \hat{F}^{\mathcal{F}} \rangle_{\mathcal{F}} - \langle \hat{\mathbb{I}} \rangle_{\mathcal{F}} = \mathcal{B}_{\mathcal{F}\mathcal{F}} - \langle \hat{\mathbb{I}} \rangle_{\mathcal{F}} \ . \quad (13)$$

In the above definitions we have employed the *orbital population* of the fragment $\mathcal{F}$, defined as $\langle \hat{\mathbb{I}} \rangle_{\mathcal{F}} = \mathrm{tr}\left(\mathbf{KS}^{\mathcal{F}}\right)$, as well as the *fragment bond order*, which is a quantity that in general involves two fragments:

$$\mathcal{B}_{\mathcal{F}\mathcal{G}} = \mathrm{Tr}\left(\mathbf{KS}^{\mathcal{F}}\mathbf{KS}^{\mathcal{G}}\right) = \langle \hat{F}^{\mathcal{G}} \rangle_{\mathcal{F}} \ . \qquad (14)$$

Such a quantity is associated to the overall bonding ability of the two fragments $\mathcal{F}$ and $\mathcal{G}$ *with respect to the chosen basis set and population scheme*. This quantity is similar to the Wiberg index,[38] and in the case of the Mulliken representation with atomic fragments corresponds to the Mayer bond order.[39] In this case, we have defined a more general fragment

bond order, which describes the interaction between two arbitrary fragments.

The purity condition defined in this way strongly resembles the concept of chemical valence,[40,41] which measures the ability of an atom to form chemical bonds in its current environment, but in this case we include off diagonal contributions and scale by the number of electrons. Indeed it is enough to notice that, in the Mulliken population scheme, for a fragment made only of atom a the purity indicator is the opposite of the atomic valence:

$$q_a \Pi_a = -V_a \equiv= \mathrm{tr}\left((\mathbf{KST}^a)^2 - \mathbf{KST}^a\right). \quad (15)$$

Following this interpretation we can rephrase the purity condition with a chemical meaning: a fragment is pure if it has a "zero-valence" condition - i.e. the value of the fragment bond order with itself equals the fragment orbital population. Despite its physico-chemical interpretation, such a zero-valence condition is a property of the computational setup and of the projection method, and it is not a chemical property *per se*; however, when the basis set and the projection scheme are suitably chosen, it enables the splitting of the system's observables into fragments.

As mentioned, the purity indicator has a non-linear behaviour with respect to the combination of fragments. It is easy to verify that, for two fragments $\mathcal{F}$ and $\mathcal{G}$ we can expand the purity indicator in terms of the fragment bond order as follows:

$$q_{\mathcal{F}+\mathcal{G}} \Pi_{\mathcal{F}+\mathcal{G}} = q_{\mathcal{F}} \Pi_{\mathcal{F}} + q_{\mathcal{G}} \Pi_{\mathcal{G}} + \mathcal{B}_{\mathcal{F}\mathcal{G}} + \mathcal{B}_{\mathcal{G}\mathcal{F}}, \quad (16)$$

such a result will turn out to be useful in the forthcoming section.

4 Fragmentation and Subsystems

We have seen that the fragmentation operators are useful to identify pseudo-observables that can be associated to a system's moieties. This is clearly helpful in characterizing a system, providing information on the impact a given fragmentation will have on the reliability of a fragment's expectation value. However, for certain observables, it would be nonetheless interesting to rely on moieties which are defined beforehand, and analyse their mutual interaction in order to characterize the system's building blocks from an electronic point of view.

Let us consider an example scenario where a given target set of atoms $\mathcal{T}$ is chosen a priori, and the goal is to compute its properties using only a subset of the full system. Associated with that target fragment is a purity indicator $\Pi_{\mathcal{T}}$ with a absolute value that may be higher than some desired threshold ϵ. We have seen that this implies that the density matrix is not assumed block-diagonal in the fragmentation provided by $\mathcal{T}$. Let us define the embedded purity indicator $\Pi_{\mathcal{T}:\mathcal{E}}$ as the purity indicator of the joint T and E system, but without considering the contribution associated to the environment alone.

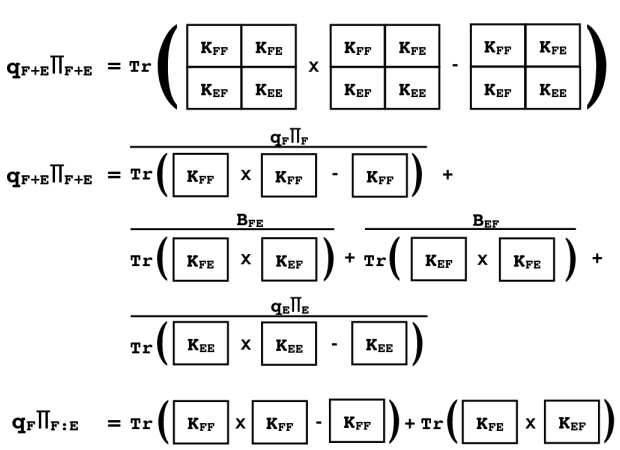


Figure 1: Summary of an expansion of the purity indicator of two fragments $\mathcal{F}$ and $\mathcal{E}$ in terms of the fragment bond order. For simplicity, we assume the Mulliken (or Löwdin) population scheme with a overlap matrix S that is unitary, but in the general case the above diagram is the same with the matrix K replaced by KS.

We can define the embedded purity indicator as follows:

$$q_{\mathcal{T}} \Pi_{\mathcal{T}:\mathcal{E}} \equiv q_{\mathcal{T}} \Pi_{\mathcal{T}} + \mathcal{B}_{\mathcal{T}\mathcal{E}}\,. \quad (17)$$

To clarify the interpretation of this quantity, in Fig. 1 we provide a matricial representation

of this block view of the purity indicator.

We note that the correction term $\mathcal{B}_{\mathcal{TE}}$ represents the "strength" of the "electronic" interaction between the two fragments. Crucially, the value of $\Pi_{\mathcal{E}}$ is not included, with the trace only running along the $\mathcal{FF}$ block. Thus, a good environment need not satisfy the purity condition itself. A suitable embedding environment is one such that the sum of the fragment bond order values of all fragments excluded from the environment is below some cutoff. In general, this environment might also be split into a number of different fragments.

By defining a fragmentation procedure and embedding scheme, we see that a graph like view of a system emerges. In this representation, fragments are nodes, and edges are drawn between fragments in the same embedding environment. This representation can be efficiently computed using the results of a calculation of the full system. Through judicious choices of a fragmentation and embedding cutoffs, a coarse grained view of large complex systems can be achieved.

4.1 Automatic Fragmentation

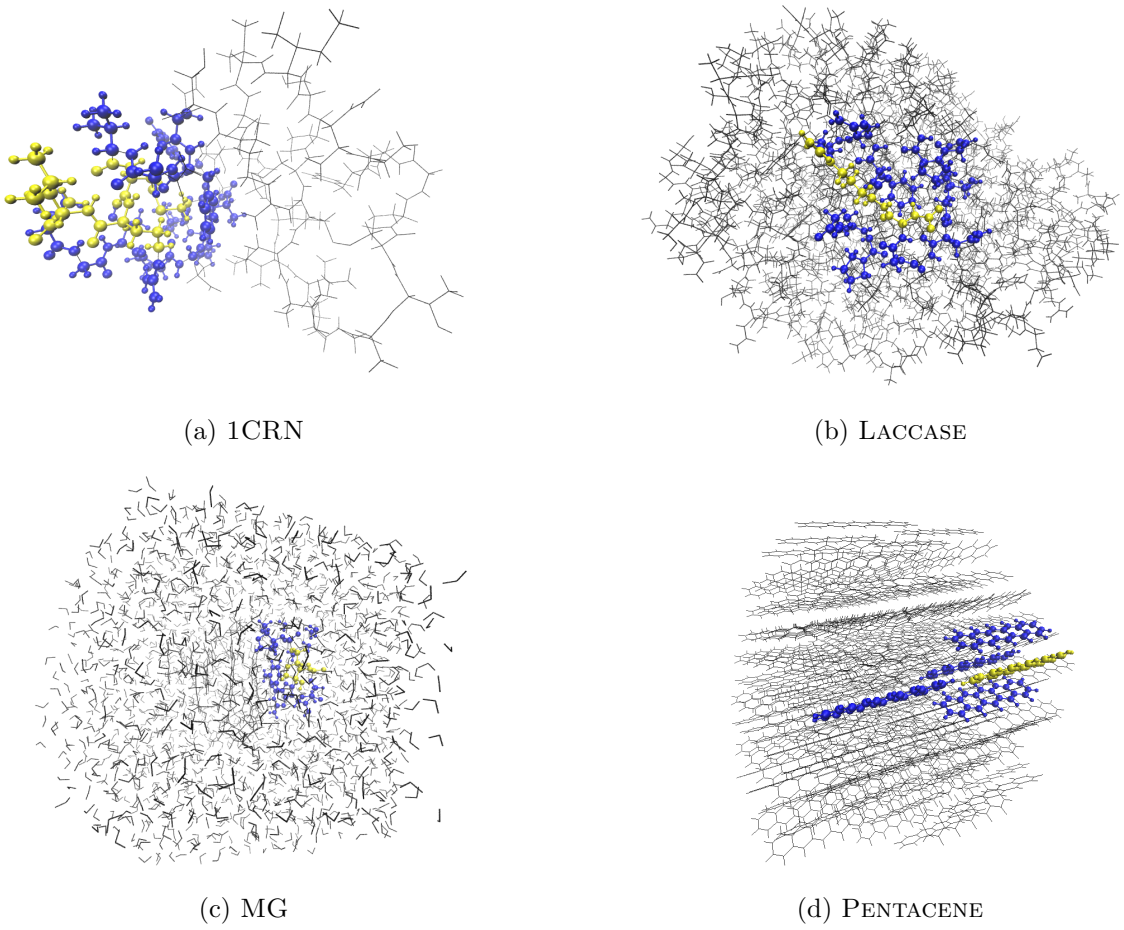
In Sec. 2, we established the purity indicator $\Pi_{\mathcal{F}}$ as a means of quantifying the choice of a given fragment $\mathcal{F}$. With a figure of merit established, we now consider how to partition a system such that each fragment fulfills that criteria. Determining the best fragmentation of a system is ill defined as presented so far, as several different fragmentations of the same system can fulfill the purity condition. Additional constraints must be introduced, such as locality in space, similarity to other fragmentation schemes, uniformity in fragment size, or maximizing the total number of fragments.

For the purposes of this paper, we will consider a simple greedy, spatially motivated algorithm for fragmenting the system. We begin by treating each atom as its own separate fragment. Then, we select the fragment with the lowest purity value to be merged. The fragment bond order between this fragment and its neighbors within a 10 Bohr radius are computed, and we merge it with the fragment with the largest bond order. This process is repeated until all fragments satisfy the purity condition $\Pi_{\mathcal{F}} > \epsilon$. While this fragmentation is not guaranteed to maximize the number of fragments, it is efficient to compute, and the spatially local fragments will help with subsequent analysis.

5 Reliability of the Approach

We will now demonstrate the previously presented tools on a number of example systems. We consider four example systems: a cambrin protein (1CRN),[42] a Laccase enzyme from Trametes versicolor (Laccase),[43] a cluster of pentacene molecules (Pentacene), and an RNA molecule binding magnesium (based on PDB 1I7J[44]) in solution (MG) (see Fig. 2). Details of the DFT calculations performed are presented in Sec. A of the Appendix. These systems each represent different challenges for complexity reduction. 1CRN is a well studied model, and coarse graining of the system might be achieved by simply decomposing the fragments based on the amino acid sequence. Laccase, on the other hand, has four copper atoms in it, making it not possible to decompose it purely using the amino acids. For the Pentacene system, the fragmentation guidance is somehow obvious, and it is interesting to decompose the observables into bulk-like and surface fragments. For the MG system, while partitioning of water molecules is an obvious start, whether the RNA molecule can be partitioned remains uncertain, as is determining a suitable fragmentation for the magnesium ions. For all of these systems, even once a decomposition has been established, the choice of an embedding environment for each target fragment remains challenging.

The tools established in the preceding sections require no a priori information about the system to be applied, and can generate an unbiased coarse graining of each type of system. In this section, we will systematically fragment and compute embedding environments for these example systems, and evaluate these reduced models with a number of different metrics.



(a) 1CRN (b) Laccase

(c) MG (d) Pentacene

Figure 2: Embeddings of target fragments in the four sample systems. The target regions are in yellow, and the embedding environment (using a bond order cutoff of 0.01) are in blue. Atoms in black are those which belong to the full system but are excluded from the subsystem calculations.

5.1 Choice of Purity Indicator Cutoff

We begin by exploring the choice of purity indicator cutoff's effect on the number of fragments in a given system. For each system, the auto fragmentation procedure described in Sec. 4.1 is applied. The number of fragments for each system at various cutoffs are plotted in Fig. 3. We have also analyzed the number of fragments of just the RNA molecule in the MG system.

One point of interest in the data of Fig. 3 is that the two proteins (1CRN, Laccase) follow an extremely similar trend when comparing the relative number of fragments at a given cutoff value. This suggests that the average size of a fragment is similar when systems are composed of similar building blocks. This is in contrast to the Pentacene system which has similarly

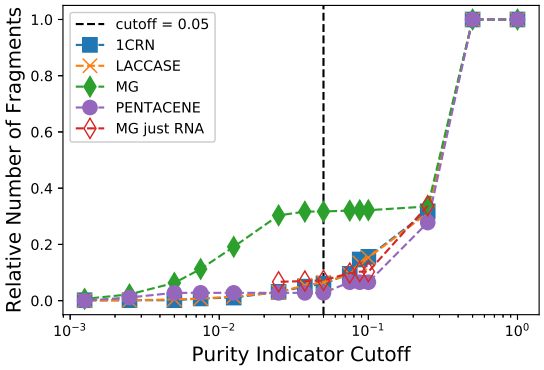


Figure 3: The relative number of fragments in each system compared to an atomic fragmentation at various purity indicator cutoff (absolute) values.

sized fragments, but different building blocks with different behavior. As a further point of contrast, the MG system has relatively more fragments at high cutoff values than all of the other systems. However, when separately examining the RNA molecule fragments in MG, we see a different picture, with a stricter cutoff leading to fewer fragments.

There appears to exist a region between $\Pi = -0.01$ and -0.05 where the number of fragments is relatively stable, and a coarse grained view of the system is possible. We note that this regime matches a similar finding as a study using localized orbitals to partition domains for the purpose of accelerating exact exchange calculations.[45] In that study, the cutoff was defined in terms of the norm of truncated localized orbital, which corresponds closely with the purity indicator.

For a given cutoff value, there exists some freedom based on how coarse grained a view of a system is desired. For example, with the MG system, a cutoff in this range may be too fine grained a view of the system, leading to a tighter cutoff value for fragmenting the solution. This is due to the large number of water molecules in the system. Water molecules have very low (absolute) purity values, meaning that a fragmentation of a solution is stable even as the cutoff value is tightened.

For the remainder of this paper, we will use a cutoff of $\Pi = -0.05$ for automatically fragmenting systems. Further analysis will be performed on this choice of cutoff in Sec. 6.

5.2 Choice of Fragment Bond Order Cutoff and Reliability of Fragment Observables

We now consider the appropriate threshold for defining an embedding environment. As a figure of merit, we focus our attention to the electrostatic dipole as the chosen fragment observable (the operator $\hat{O}$ of the equations in Sec. 2). We have chosen the dipole since a faithful representation of such observable would demonstrate the reliability of the electronic density as well as a good approximation for one of the main quantities needed for the long-range potential that a fragment would generate.

To do this, we begin by fragmenting each of the example systems, with a cutoff of $\Pi = -0.05$. Next, for each system we select the fragment with the largest dipole value (to increase the signal to noise of subsequent calculations) and define it as the target fragment for embedding. We then define embedding environments based on various threshold values. We also compared this approach with an environment computed by the nearest neighbor distance between fragments.

Calculations were then performed from scratch on the target and embedding environment, and observables were recomputed. The dipoles of each target fragment in the various embedding environments were computed from the atomic dipoles according to the equations in our previous publication.[37] We emphasize that no external potential from outside the embedding environment was included except through the net charge which was rounded to the nearest electron.

Images of the various target regions inside an embedding region with a bond order cutoff of 0.01 are shown in Fig. 2. Errors in the dipole values are plotted in Fig. 4.

We note that it is remarkable that these calculations are accurate at all given that in many places we have cut covalent bonds without using a capping procedure. Calculations performed on these systems also smoothly converged to the ground state. The lack of a need for capping can be attributed to the $\Pi = -0.05$ purity value of the fragments, which already limits the amount of charge being leaked. In a sense, the low (absolute) purity value of the embedding fragments enable them to act as a general type of cap on the target fragment. By adding the embedding environment, it is possible to significantly reduce the errors, until improvement stagnates in general with a bond order cutoff of between 0.01 and 0.001. We also note that while a conservative distance criteria can define a suitable embedding environment, the converged distance value is significantly affected by the specific system geometry. Using the auto fragmentation procedure and bond or-

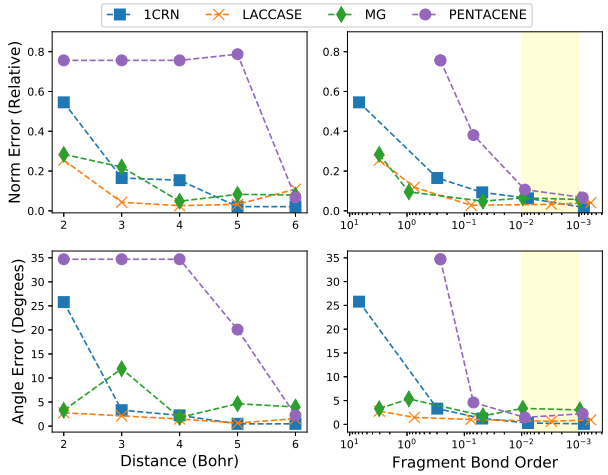


Figure 4: Error in the dipole in various embedding environments. The relative error in norm is defined as $\frac{||d'-d||_2}{||d||_2}$ where d is the dipole computed from the full calculation, and d' the dipole computed from the subsystem calculation. The angle error is the angle between the dipole computed from the full and subsystem calculations. The region between a fragment bond order cutoff of 0.01 and 0.001 has been highlighted to emphasize the converged observables.

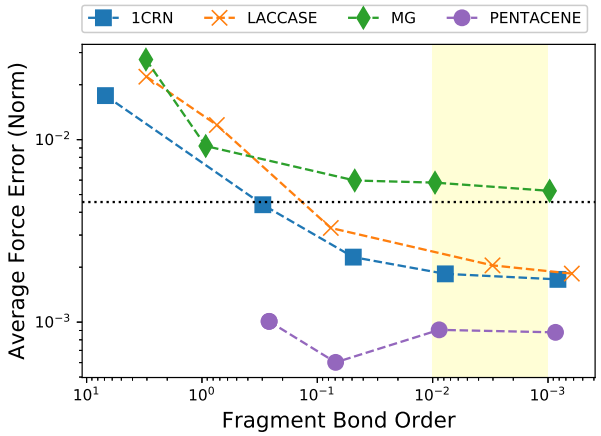


Figure 5: Average error in the atomic forces (a.u.) in various embedding environments. The dotted horizontal line is the estimate of the noise in the force from the calculation of the full 1CRN system.

der tools together, one can automatically define an embedding of all system fragments which accurately reproduces desired observables.

We also consider the error in the atomic forces. We know that, as the atoms cannot be associated to pure fragments, an embedding environment is needed to guarantee that atomic forces can become reliable. We have computed the average error in the forces inside the target fragment and plotted those values in Fig. 5. Here we see a similar convergence trend, with the exception being the PENTACENE system, which has very low forces on any given fragment. To put these errors into context, we define an estimate of noise in the forces as the standard deviation of the forces with mean zero. The error in the forces presented becomes of the same order of magnitude as the noise in the forces from calculations of the full system, and are similar to the errors in forces that come from using the linear scaling version of BigDFT.[46] The stagnation in force error reduction with an increasing environment size is further evidence that we have captured the essential fragment environment for all systems using a bond order cutoff between 0.01 and 0.001.

For a fragment identified with a $|\Pi| = 0.05$, and using a bond order cutoff value of 0.01, the number of atoms in the four system target regions are 228, 191, 97, and 180 for 1CRN, LACCASE, MG, and PENTACENE respectively. The embedding environments contain 172, 159, 78, and 144 atoms. If these regions were used for production QM/MM calculations, they would represent larger QM regions than are usually treated, though recent studies have favored bigger QM regions.[47] A reduction of the QM region size is possible when the MM region realistically mimics the external region, such as through the inclusion of capping atoms or the use of well tuned MM potentials.

The procedure presented here requires no direct user interaction, and is instead a general workflow for studying any kind of system. This generality is further shown in the supplementary information, as the calculations on each system can be performed with the same script by only changing the input geometry file. The generality of this scheme makes it a promising approach for high-throughput calculations aimed at complexity reduction.

6 Evaluating The Coarse Grained View of the System

We now continue our analysis on these systems by performing a more information centric analysis of the system fragments. We will begin by studying the transferability of fragments, and comparing them to the amino acids of proteins. We will then generate graph like views of each system, and see how choices of fragment purity and embedding environment affect various graph metrics.

6.1 Fragmentation Comparison of Proteins

For 1CRN, a different natural fragmentation might be to use the amino acid sequence of the protein instead of the auto fragmentation procedure. We have computed the purity values of those fragments as generated by the FU program,[48] and plotted them in Fig. 6. We see that by our purity criteria of $\Pi > -0.05$, the amino acids are a reasonable system fragmentation, which is not surprising for this kind of model system. Nonetheless, the auto fragmentation procedure requires no a priori fragment information, making it applicable to a wider class of systems. When additional fragmentation guidance is available, the two approaches can be combined if a coarser grained view of the system is desired. For example, with the 1CRN system, tightening the threshold from $\Pi = -0.05$ to -0.025 to -0.01 reduces the number of fragments from 39 to 18 to 5. When starting from the amino acids for the Laccase system, the drop is from 452 to 215 to 77 fragments.

This result further demonstrates that some arbitrariness exists in the choice of system fragmentation. This might hint that there exist a broader set of descriptors for describing biological systems than just the amino acid sequence. Using the open babel code,[49,50] we can for each fragment compute a molecular fingerprint, and then compute a similarity score between each pair of fingerprints. For this study, we will use the FP2 fingerprint, which creates

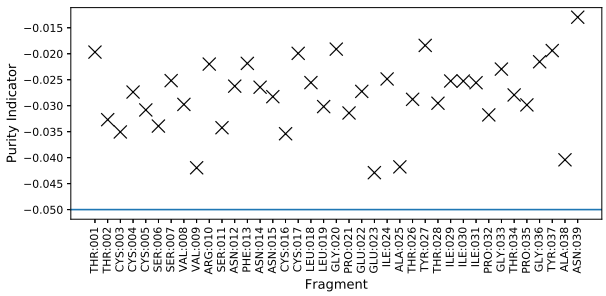


Figure 6: Purity indicator values for the 1CRN system when fragmented by amino acid using the FU program.

a binary string representation of a fragment based on short linear and ring molecule substructures, and evaluate the similarity between those strings using the Tanimoto coefficient (see Willet[51] for an overview of this approach). In Fig. 7, we demonstrate this approach by first comparing the fragments of 1CRN using the auto fragmentation tool and the amino acid sequence. We see that these fragments are indeed significantly different, despite the fact that both the amino acid partitioning and the auto fragmentation procedure result in the same number of fragments (39). Next, we investigate the transferability of fragments by comparing the fragments of 1CRN with Laccase. For Laccase, it is difficult to determine an appropriate fragmentation for the copper atoms without a tool like the auto fragmentation procedure, but for this comparison we have by hand merged all the copper atoms with their neighboring cystine amino acids. When comparing the amino acids of 1CRN and Laccase, we unsurprisingly identify many similar fragments. However, we also compare the fragments of 1CRN generated with the auto fragment tool, and find that there are also many similar fragments shared between the two systems. Thus, while the auto fragmentation tool identifies new kinds of fragments, these fragments remain transferable, making them a promising source of new descriptors that can more adequately be put in relation with QM calculations in a given computational setup.

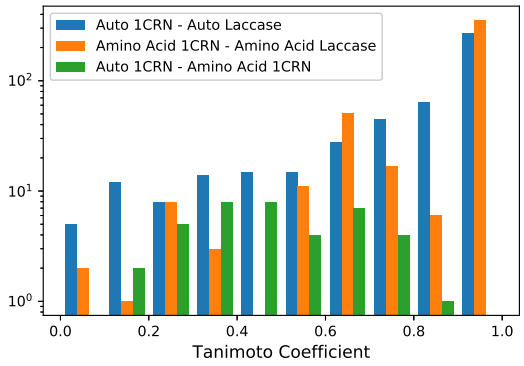


Figure 7: Histogram of Tanimoto coefficients when comparing the fragments of 1CRN with LACCASE. Note that Tanimoto coefficient values range from 0 to 1, with values closer to 1 being more similar.

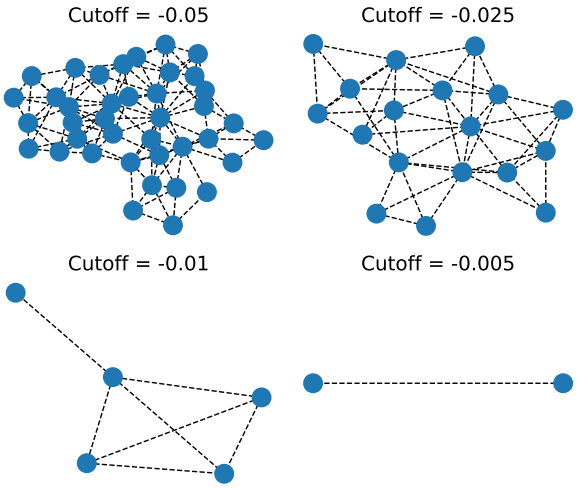


Figure 8: Coarse graining of the 1CRN graph structure starting with the amino acids fragments. The diagrams here are generated using the Kamada-Kawai algorithm[52] for visual clarity, so node locations are not related to atomic positions in space.

6.2 Graph Metrics on General Systems

We finish this demonstration by turning to the generation of graph like views of a system. For each of the example systems, we once again perform auto fragmentation with a $\Pi = -0.05$ cutoff and use this fragmentation to define the graph's nodes. Then, for each fragment, we compute its embedding environment at various thresholds, and use that environment to define the edges of the graph.

We may examine the graph characteristics of a system with a change in purity indicator cutoff while keeping the bond order cutoff fixed. By increasing the purity cutoff closer and closer to zero, we can generate low resolution views of a system's connectivity. This process is demonstrated in Fig. 8. In this example, we begin with the connectivity of the 1CRN system with the fragments defined by the amino acids and the connectivity with a 0.01 bond order cutoff. As we push the cutoff closer to zero, the shape of the graph changes significantly, resulting in a simpler and simpler picture of the system.

From this representation, we compute some sample graph metrics: the average shortest path length and the average clustering coefficient.[53] These metrics have been applied to proteins in the past, as reviewed by Estrada.[54] In most previous studies, however, the focus was on long range van der Waals interactions. Other authors have used interaction energies such as those derived from forcefield calculations.[55] Sladek and coworkers recently utilized pair interaction energies[56] to define network edges, and showed how the properties analyzed using an energy based model differ from standard distance based analysis.[57]

Values of the average shortest path length metric are reported in Fig. 9. Note that in the analysis presented here, the average shortest path length is defined in terms of the number of edges traversed, without consideration to physical inter-fragment distances. From this figure, we find additional supporting evidence for a bond order cutoff of 0.01 for the embedding environment. When a smaller value is used, the graphs of these systems are no longer fully connected. Even with a bond order cutoff value of 0.1, the MG system is disconnected, reflecting how pure the water molecule fragments are. The two proteins are connected as soon as any bond order is considered, but the average shortest path length quickly decreases with a decrease in fragment bond order cutoff, leading to a very different description of the system.

The average clustering coefficients are plot-

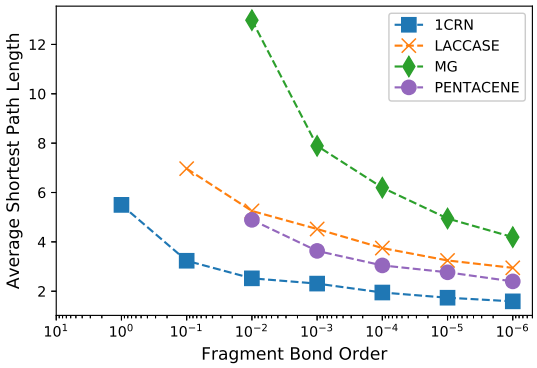


Figure 9: Average shortest path length for each system at various cutoff values. The path length of a disconnected system is ∞, so those values are not shown here.

ted in Fig. 10. The low average clustering value for the MG system reflects the general lack of structure of the water molecules, while the other systems are significantly more connected. For both of these metrics, we find inflection points with a bond order cutoff of around 0.01 or 0.001, after which the measures increases/decrease linearly with the fragment bond order. However, the slope of the linear region depends on the system. Thus, while networks might be generated with large distance cutoffs to incorporate long range interactions, the fundamental structure of graphs can be understood by looking at the short ranged covalent interactions using the fragment bond order tool.

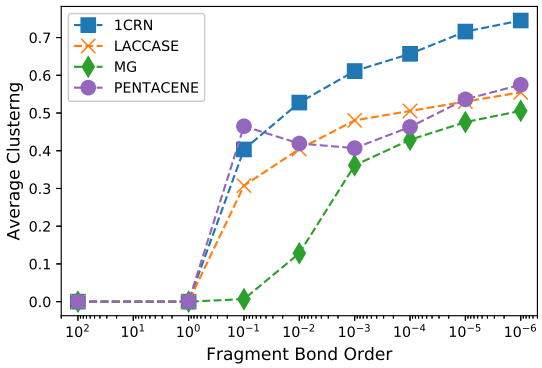


Figure 10: Average clustering coefficients for each system at various cutoff values.

Next, we apply this approach to all four example systems using the fragments defined by the auto fragmentation procedure. For each of these systems, we compute the average shortest path length at various purity cutoffs, and plot it against the log of the number of fragments in Fig. 11. For the MG system, certain values of the network were disconnected, so an average was taken over each subgraph. For a network with small world characteristics, the average shortest path length should grow logarithmically with the number of nodes.[58] Intriguingly, we do see such growth for the two protein molecules, though with an inflection point around a purity indicator cutoff of $\Pi = -0.025$. For the other two systems, there also appears to be two distinct patterns centered at a purity value of $\Pi = -0.025$. This suggests that there is a cross over point at which a view of the local structure is lost and the global structure dominates the description of the molecular system. In the complexity reduction framework proposed here, this information can be extracted using the BigDFT code, enabling insight into system properties at the desired level of detail.

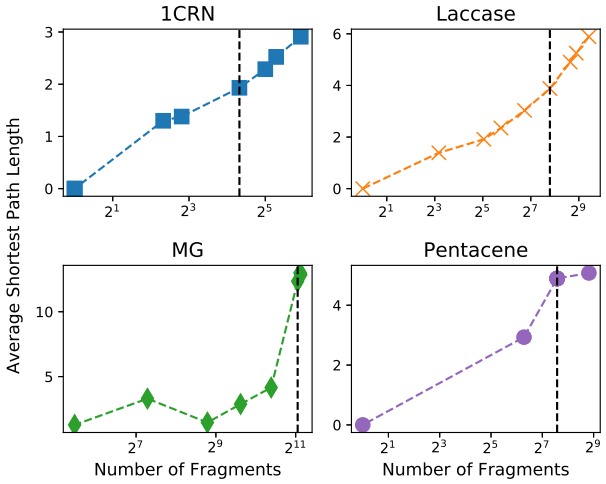


Figure 11: The average shortest path length vs. the number of fragments for each system at various purity value levels. The dotted vertical line represents the number of nodes when a purity value of $\Pi = -0.025$ is used.

7 Fragmentation vs. Embedding

In the framework introduced thus far, users have two degrees of freedom to consider for building a coarse grained model of a system. The first is the coarseness of the fragments, as determined by the purity indicator. The second is the choice of embedding environment as defined by the fragment bond order. The balance between these two variables depends on the choice of observable one wishes to compute, as we will demonstrate through the following case study: computing the density of states (DoS) of the LACCASE system.

First, we consider the problem of computing the density of states of a system projected on to a given fragment (PDoS). In the framework of fragmentation, the projected density of states can be computed by the formula

$$\rho_{\mathcal{F}}(\omega) = \sum_i \mathrm{tr}\left(\hat{W}^{\mathcal{F}} \left|\psi_i\right\rangle \left\langle\psi_i\right|\right) \delta(\epsilon_i - \omega) \ , \quad (18)$$

where $|\psi_i\rangle$ are the Kohn-Sham orbitals of energy ϵ_i. It is indeed possible to compute the PDoS for any arbitrary fragment. However, we know that the DoS of a given fragment will be influenced by its environment, with the degree of influence determined by the purity indicator of that fragment. When a fragment is not pure, if we recompute that fragment in isolation, we can't expect to reliably reproduce the PDoS embedded in the full system. However, by including more and more environment in a buffer region using the fragment bond order as a guide, we can eventually reach a converged result, as shown in Fig. 12. Thus, the bond order tool allows us to make up for the lack of purity of a given fragment by performing embedding calculations.

Now we turn to computing the full DoS of the entire systems. Here we might be tempted to use the same approach (i.e. for each fragment, we compute its embedding environment, perform an embedded calculation, compute the projected DoS, and finally sum up the values). However, an alternative approach would be to simply use the auto fragmentation procedure

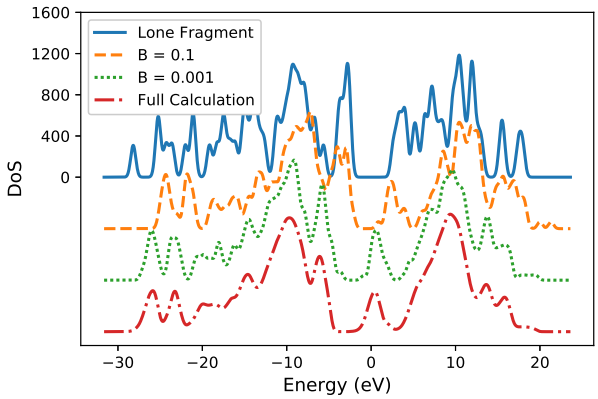


Figure 12: The projected density of states of the target region of the LACCASE protein computed with different embedding environments. Plot was obtained using a smearing parameter of 0.3 eV and each line has been shifted for visual clarity.

with a stricter cutoff, and use no embedding environment, as shown in Fig. 13. The benefit of an embedding environment is that the target region has an effective purity value that is closer to zero than by itself. However, by building the embedding environment, we also improve the purity value of each of the embedding fragments, resulting in a total purity value for a joint target-embedding fragment that is much closer to zero than the lone target. Thus, a non-buffer approach can reduce the amount of repeated work from overlapping environments, with the trade-off depending on the scaling of the computational cost with the system size.

The difference between computing the DoS and the PDoS is that in one case we care about a *system level observable* and in the other a *fragment level pseudo-observable*. For each calculation quantity, we should consider the level of detail it is computed at. The fragment dipole and the PDoS are fragment level quantities, whereas the DoS or the total energy are system level quantities. The purity indicator gives us a measure that informs whether such quantities can be computed independent of the environment. In cases where it is not possible, the bond order tool can define a suitable embedding such that one reproduces the desired value.

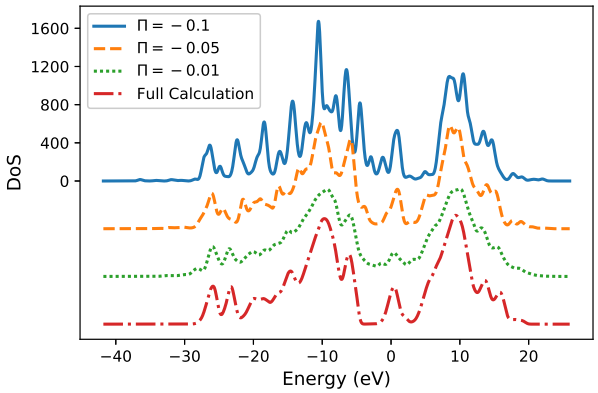


Figure 13: The density of states of the full LACCASE protein computed as a composite from subsystems computed at different purity values. Plot was obtained using a smearing parameter of 0.3 eV and each line has been shifted for visual clarity.

8 Discussion

In this work, we have presented a complexity reduction scheme which takes large, heterogeneous systems, and uses the results of linear scaling DFT calculations to generate coarse grained models. We have demonstrated how this approach can be applied without bias to different classes of systems with no a priori user information. Furthermore, by applying this method to generate model systems of a target in an environment, and using these models to accurately compute fragment observables, we have shown that this approach provides chemically meaningful descriptions of system interactions.

Fragments form the basis for many low order scaling methods for computing large systems. Those methods compute the properties of a system from the ground up, using either predefined fragments (for example, using the amino acid structure of proteins[30,59]), fragments refined using distances between fragment elements,[60–62] cheminformatics,[63] or bonding information in combination with chemically motivated rules[64–67] to define the partitioning. A recent review by Collins and Bettens[27] describes many of these types of methods. The approach presented here differs in that it is instead works from the top down, using the results of linear scaling calculations to determine system fragments. This work is thus more focused on post-processing systems for chemical understanding than on fast calculations. Nonetheless, our approach can serve as a complement for such methods by defining initial partitioning and embedding systems, which then can be treated at a higher level or theory, have their geometry optimized, or used to perform molecular dynamics with such fragment methods.

The interaction between fragments also has been a topic of many studies, in particular when trying to determine intermolecular forces for studying reactions.[56,68,69] This work goes back to the pioneering development of the theory of atoms in molecules,[70] as the critical points in the electron density can be used to define which atoms interact.[71–77] It has been continued by recent work in the framework of partition density functional theory[78] with a focus on describing chemical reactivity.[79,80] Similar to the methodology presented here, these works also can describe both covalent and non-covalent interactions between fragments, though the focus is on an atomic level. The methodology we have presented here is instead density matrix based, and works at a coarser, fragment level view. Partitioning is achieved through a fall off of the density matrix in the linear scaling regime.

In this work, three classical ideas have re-emerged: valence, population analysis, and bond order (see Mayer[81] for a review). At the time when these ideas were developed, calculations on systems with even hundreds of atoms remained out of reach, allowing for careful analysis of individual atomic contributions. In this work, we have taken those ideas which were defined at the atomic level of granularity and redefined them for molecular fragments. By moving to the fragment level, not only is it possible to derive more chemically meaningful observables, but also to enable coarse grained analysis of large systems.

In the following few years, the next generation of exascale class supercomputers promises to enable the routine application of fully quantum mechanical methods to systems with tens of thousands of atoms. With this, information derived from the electronic structure will begin

to have an impact on entirely new disciplines. One piece of information from these calculations is the locality of the electronic structure, which we have used to partition systems and describe interactions between fragments. The novel fragments generated by this approach and the graph structures that tie them together are promising new tools for theoretical studies. Our future work will focus on applying this methodology to an even wider class of systems, in hopes of generating novel design rules and insights into large, heterogeneous systems.

9 Acknowledgements

This work was supported by the Next-Generation Supercomputer project (the K computer) and the FLAGSHIP2020 project (Supercomputer Fugaku) within the priority study5 (Development of new fundamental technologies for high-efficiency energy creation, conversion/storage and use) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan. Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr). Calculations were performed using the Hokusai supercomputer system at RIKEN. 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. S.M. and L.G. acknowledge support from the European Centre of Excellence MaX, funded within the Horizon2020 Framework of the European Union under project ID 676598. L.G. acknowledges support from the EU ExtMOS project and the European Centre of Excellence EoCoE, funded within the Horizon2020 Framework of the European Union under project IDs 646176 and 676629, respectively. L.G., T.N., and W.D. also gratefully acknowledge the joint CEA-RIKEN collaboration action. L.E.R. acknowledges support from an EPSRC Early Career Research Fellowship (EP/P033253/1).

We would also like to acknowledge Kazuo Kitaura for his assistance using the FU program [48] for generating the amino acid partitioning of the two protein systems and preprocessing of the input geometry. We would like to thank Thierry Deutsch for valuable discussions and Fátima Lucas for providing various test systems and helpful discussions. Images of molecular systems were generated using VMD. [82]

A DFT Calculation Details

Calculations of each system were performed with the BigDFT code [83] using density functional theory in the linear scaling mode [46,84] with the PBE [85] exchange and correlation functional and free boundary conditions. Hartwigsen-Goedecker-Hutter (HGH) [86,87] pseudopotentials were used with 11 and 2 valence electrons for copper and magnesium respectively. Fragmentation and bond order calculations have been implemented in a new python based pre/post-processing library called PyBigDFT. These calculations may be run using python notebooks, as have been included in the supplementary materials along with all geometry files.

In the linear scaling mode of BigDFT, finite distance based cutoffs for kernel values are employed to maintain the sparsity of the hamiltonian and density matrix. [88] We note that these distances are much larger than the embedding region sizes tested in the preceding sections, and as such these a priori cutoffs should not affect the resulting analysis. One of the key steps for this analysis is computing the product of the density matrix and overlap matrix. As an extra precaution, we compute this matrix with no distance cutoff, instead filtering values of magnitude below 1×10^{-6} using the NTPoly library. [89]

B Supporting Information

The python notebooks used to setup calculations, perform analysis, and generate figures

have been included in the supporting information.

- CR2.IPYNB a python notebook for performing a complexity reduction analysis on any of the systems used in this paper.
- CR2-DOS.IPYNB a python notebook for performing the density of states case study.
- SUMMARY.IPYNB a python notebook for generating summarizing figures.

In addition to the actual notebooks, static websites generated by these notebooks have been included (see the .HTML files in the directory STATIC/) for each of the system considered.

Geometry files in the XYZ format, as well as BDA files used to identify amino acid fragments (only of the proteins systems) have also been included in the GEOMETRIES/ and BDA/ directories, respectively.

References

(1) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864–B871.

(2) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133–A1138.

(3) Goedecker, S. Linear scaling electronic structure methods. *Rev. Mod. Phys.* **1999**, *71*, 1085–1123.

(4) Bowler, D. R.; Miyazaki, T. O(N) methods in electronic structure calculations. *Rep. Prog. Phys.* **2012**, *75*, 036503.

(5) Ratcliff, L. E.; Mohr, S.; Huhs, G.; Deutsch, T.; Masella, M.; Genovese, L. Challenges in large scale quantum mechanical calculations. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2017**, *7*, e1290.

(6) 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.

(7) Fedorov, D. G.; Kitaura, K. Extending the Power of Quantum Chemistry to Large Systems with the Fragment Molecular Orbital Method. *J. Phys. Chem. A* **2007**, *111*, 6904–6914.

(8) Gao, J. Toward a Molecular Orbital Derived Empirical Potential for Liquid Simulations. *J. Phys. Chem. B* **1997**, *101*, 657–663.

(9) Gao, J. A molecular-orbital derived polarization potential for liquid water. *J. Chem. Phys.* **1998**, *109*, 2346–2354.

(10) Wierzchowski, S. J.; Kofke, D. A.; Gao, J. Hydrogen fluoride phase behavior and molecular structure: A QM/MM potential model approach. *J. Chem. Phys.* **2003**, *119*, 7365–7371.

(11) Xie, W.; Gao, J. Design of a Next Generation Force Field: The X-POL Potential. *J. Chem. Theory Comput.* **2007**, *3*, 1890–1900.

(12) Xie, W.; Song, L.; Truhlar, D. G.; Gao, J. The variational explicit polarization potential and analytical first derivative of energy: Towards a next generation force field. *J. Chem. Phys.* **2008**, *128*, 234108.

(13) Xie, W.; Song, L.; Truhlar, D. G.; Gao, J. Incorporation of a QM/MM Buffer Zone in the Variational Double Self-Consistent Field Method. *J. Phys. Chem. B* **2008**, *112*, 14124–14131.

(14) Wang, Y.; Sosa, C. P.; Cembran, A.; Truhlar, D. G.; Gao, J. Multilevel X-Pol: A Fragment-Based Method with Mixed Quantum Mechanical Representations of Different Fragments. *J. Phys. Chem. B* **2012**, *116*, 6781–6788.

(15) Gao, J.; Wang, Y. Communication: Variational many-body expansion: Accounting

for exchange repulsion, charge delocalization, and dispersion in the fragment-based explicit polarization method. *J. Chem. Phys.* **2012**, *136*, 071101.

(16) Gadre, S. R.; Shirsat, R. N.; Limaye, A. C. Molecular Tailoring Approach for Simulation of Electrostatic Properties. *J. Phys. Chem.* **1994**, *98*, 9165–9169.

(17) 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.

(18) Gadre, S. R.; Ganesh, V. Molecular tailoring approach: towards PC-based ab initio treatment of large molecules. *J. Theor. Comput. Chem.* **2006**, *5*, 835–855.

(19) Sahu, N.; Gadre, S. R. Molecular Tailoring Approach: A Route for ab Initio Treatment of Large Clusters. *Acc. Chem. Res.* **2014**, *47*, 2739–2747.

(20) Jacob, C. R.; Neugebauer, J. Subsystem density-functional theory. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2014**, *4*, 325–362.

(21) Krishtal, A.; Sinha, D.; Genova, A.; Pavanello, M. Subsystem density-functional theory as an effective tool for modeling ground and excited states, their dynamics and many-body interactions. *J. Phys.: Condens. Matter* **2015**, *27*, 183202.

(22) Bakowies, D.; Thiel, W. Hybrid Models for Combined Quantum Mechanical and Molecular Mechanical Approaches. *J. Phys. Chem.* **1996**, *100*, 10580–10594.

(23) Ferré, N.; Ángyán, J. G. Approximate electrostatic interaction operator for QM/MM calculations. *Chem. Phys. Lett.* **2002**, *356*, 331–339.

(24) others,, et al. The ONIOM method and its applications. *Chem. Rev.* **2015**, *115*, 5678–5796.

(25) QM/MM: What have we learned, where are we, and where do we go from here? *Theor Chem. Acc.* **2007**, *117*, 185.

(26) Senn, H. M.; Thiel, W. QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed.* **2009**, *48*, 1198–1229.

(27) Collins, M. A.; Bettens, R. P. Energy-based molecular fragmentation methods. *Chem. Rev.* **2015**, *115*, 5607–5642.

(28) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation methods: A route to accurate calculations on large systems. *Chem. Rev.* **2011**, *112*, 632–672.

(29) Raghavachari, K.; Saha, A. Accurate composite and fragment-based quantum chemical models for large molecules. *Chem. Rev.* **2015**, *115*, 5643–5677.

(30) He, X.; Zhu, T.; Wang, X.; Liu, J.; Zhang, J. Z. Fragment quantum mechanical calculation of proteins and its applications. *Acc. Chem. Res.* **2014**, *47*, 2748–2757.

(31) Li, S.; Li, W.; Ma, J. Generalized energy-based fragmentation approach and its applications to macromolecules and molecular aggregates. *Acc. Chem. Res.* **2014**, *47*, 2712–2720.

(32) Hirata, S.; Gilliard, K.; He, X.; Li, J.; Sode, O. Ab initio molecular crystal structures, spectra, and phase diagrams. *Acc. Chem. Res.* **2014**, *47*, 2721–2730.

(33) Sahu, N.; Gadre, S. R. Molecular tailoring approach: a route for ab initio treatment of large clusters. *Acc. Chem. Res.* **2014**, *47*, 2739–2747.

(34) Pruitt, S. R.; Bertoni, C.; Brorsen, K. R.; Gordon, M. S. Efficient and accurate fragmentation methods. *Acc. Chem. Res.* **2014**, *47*, 2786–2794.

(35) Mezey, P. G. Fuzzy electron density fragments in macromolecular quantum chemistry, combinatorial quantum chemistry,

functional group analysis, and shape–activity relations. *Acc. Chem. Res.* **2014**, *47*, 2821–2827.

(36) Herbert, J. M. Fantasy versus reality in fragment-based quantum chemistry. *J. Chem. Phys.* **2019**, *151*, 170901.

(37) Mohr, S.; Masella, M.; Ratcliff, L. E.; Genovese, L. Complexity Reduction in Large Quantum Systems: Fragment Identification and Population Analysis via a Local Optimized Minimal Basis. *J. Chem. Theory Comput.* **2017**, *13*, 4079–4088.

(38) Wiberg, K. B. Application of the pople-santry-segal CNDO method to the cyclopropylcarbinyl and cyclobutyl cation and to bicyclobutane. *Tetrahedron* **1968**, *24*, 1083–1096.

(39) Mayer, I. Bond order and valence: Relations to Mulliken's population analysis. *Int. J. Quantum Chem.* **1984**, *26*, 151–154.

(40) Borisova, N.; Semenov, S. The molecular-orbital determination of the chemical bond order. *Vestn. Leningr. Univ., Ser. 4: Fiz., Khim.* **1973**, 119–124.

(41) Armstrong, D. R.; Perkins, P. G.; Stewart, J. J. Bond indices and valency. *J. Chem. Soc., Dalton Trans.* **1973**, 838–840.

(42) Teeter, M. Water structure of a hydrophobic protein at atomic resolution: Pentagon rings of water molecules in crystals of crambin. *Proc. Natl. Acad. Sci. U. S. A.* **1984**, *81*, 6014–6018.

(43) Dellafiora, L.; Galaverna, G.; Reverberi, M.; DallAsta, C. Degradation of aflatoxins by means of laccases from Trametes versicolor: An in silico insight. *Toxins* **2017**, *9*, 17.

(44) Adamiak, D. A.; Rypniewski, W. R.; Milecki, J.; Adamiak, R. W. The 1.19 Å X-ray structure of 2-O-Me (CGCGCG) 2 duplex shows dehydrated RNA with 2-methyl-2, 4-pentanediol in the minor groove. *Nucleic Acids Res.* **2001**, *29*, 4144–4153.

(45) Dawson, W.; Gygi, F. Performance and accuracy of recursive subspace bisection for hybrid DFT calculations in inhomogeneous systems. *J. Chem. Theory Comput.* **2015**, *11*, 4655–4663.

(46) Mohr, S.; Ratcliff, L. E.; Genovese, L.; Caliste, D.; Boulanger, P.; Goedecker, S.; Deutsch, T. Accurate and efficient linear scaling DFT calculations with universal applicability. *Phys. Chem. Chem. Phys.* **2015**, *17*, 31360–31370.

(47) Cole, D. J.; Hine, N. D. Applications of large-scale density functional theory in biology. *J. Phys.: Condens. Matter* **2016**, *28*, 393001.

(48) Fedorov, D. G.; Kitaura, K. Modeling and visualization for the fragment molecular orbital method with the graphical user interface FU, and analyses of protein–ligand binding. In *The fragment molecular orbital method: practical applications to large molecular systems*; John Wiley & Sons, 2017; Chapter 3, pp 119–140.

(49) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.

(50) O'Boyle, N. M.; Morley, C.; Hutchison, G. R. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.* **2008**, *2*, 1–7.

(51) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug discovery today* **2006**, *11*, 1046–1053.

(52) Kamada, T.; Kawai, S. An Algorithm for Drawing General Undirected Graphs. *Inf. Process. Lett.* **1989**, *31*, 7–15.

(53) Saramäki, J.; Kivelä, M.; Onnela, J.-P.; Kaski, K.; Kertesz, J. Generalizations of

the clustering coefficient to weighted complex networks. *Phys. Rev. E* **2007**, *75*, 027105.

(54) Estrada, E. *The structure of complex networks: theory and applications*; Oxford University Press, 2012; pp 277–295.

(55) Vijayabaskar, M.; Vishveshwara, S. Interaction energy based protein structure networks. *Biophys. J.* **2010**, *99*, 3704–3715.

(56) Fedorov, D. G.; Kitaura, K. Pair interaction energy decomposition analysis. *J. Comput. Chem.* **2007**, *28*, 222–237.

(57) Sladek, V.; Tokiwa, H.; Shimano, H.; Shigeta, Y. Protein Residue Networks from Energetic and Geometric Data: Are They Identical? *J. Chem. Theory Comput.* **2018**, *14*, 6623–6631.

(58) Barrat, A.; Weigt, M. On the properties of small-world network models. *Eur. Phys. J. B* **2000**, *13*, 547–560.

(59) Wang, X.; Liu, J.; Zhang, J. Z.; He, X. Electrostatically embedded generalized molecular fractionation with conjugate caps method for full quantum mechanical calculation of protein energy. *J. Phys. Chem. A* **2013**, *117*, 7149–7161.

(60) 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.

(61) Hua, S.; Hua, W.; Li, S. An efficient implementation of the generalized energy-based fragmentation approach for general large molecules. *J. Phys. Chem. A* **2010**, *114*, 8126–8134.

(62) Hua, S.; Li, W.; Li, S. The Generalized Energy-Based Fragmentation Approach with an Improved Fragmentation Scheme: Benchmark Results and Illustrative Applications. *ChemPhysChem* **2013**, *14*, 108–115.

(63) Steinmann, C.; Ibsen, M. W.; Hansen, A. S.; Jensen, J. H. FragIt: a tool to prepare input files for fragment based quantum chemical calculations. *PLoS One* **2012**, *7*, e44480.

(64) Deev, V.; Collins, M. A. Approximate ab initio energies by systematic molecular fragmentation. *J. Chem. Phys.* **2005**, *122*, 154102.

(65) Le, H.-A.; Tan, H.-J.; Ouyang, J. F.; Bettens, R. P. Combined fragmentation method: A simple method for fragmentation of large molecules. *J. Chem. Theory Comput.* **2012**, *8*, 469–478.

(66) Collins, M. A. Systematic fragmentation of large molecules by annihilation. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7744–7751.

(67) Collins, M. A.; Cvitkovic, M. W.; Bettens, R. P. The combined fragmentation and systematic molecular fragmentation methods. *Acc. Chem. Res.* **2014**, *47*, 2776–2785.

(68) Szalewicz, K. Symmetry-adapted perturbation theory of intermolecular forces. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 254–272.

(69) Parrish, R. M.; Parker, T. M.; Sherrill, C. D. Chemical assignment of symmetry-adapted perturbation theory interaction energy components: the functional-group SAPT partition. *J. Chem. Theory Comput.* **2014**, *10*, 4417–4431.

(70) Bader, R. F. A quantum theory of molecular structure and its applications. *Chem. Rev.* **1991**, *91*, 893–928.

(71) Contreras-García, J.; Johnson, E. R.; Keinan, S.; Chaudret, R.; Piquemal, J.-P.; Beratan, D. N.; Yang, W. NCIPLOT: a program for plotting noncovalent interaction regions. *J. Chem. Theory Comput.* **2011**, *7*, 625–632.

(72) Espinosa, E.; Souhassou, M.; Lachekar, H.; Lecomte, C. Topological analysis of the electron density in hydrogen bonds. *Acta Crystallogr., Sect. B: Struct. Sci.* **1999**, *55*, 563–572.

(73) Grabowski, S. J. Ab initio calculations on conventional and unconventional hydrogen bonds study of the hydrogen bond strength. *J. Phys. Chem. A* **2001**, *105*, 10739–10746.

(74) Becke, A. D.; Edgecombe, K. E. A simple measure of electron localization in atomic and molecular systems. *J. Chem. Phys.* **1990**, *92*, 5397–5403.

(75) Silvi, B.; Savin, A. Classification of chemical bonds based on topological analysis of electron localization functions. *Nature* **1994**, *371*, 683–686.

(76) Contreras-Garcia, J.; Recio, J. Electron delocalization and bond formation under the ELF framework. *Theor. Chem. Acc.* **2011**, *128*, 411–418.

(77) Johnson, E. R.; Keinan, S.; Mori-Sánchez, P.; Contreras-García, J.; Cohen, A. J.; Yang, W. Revealing noncovalent interactions. *J. Am. Chem. Soc.* **2010**, *132*, 6498–6506.

(78) Elliott, P.; Burke, K.; Cohen, M. H.; Wasserman, A. Partition density-functional theory. *Phys. Rev. A* **2010**, *82*, 024501.

(79) Cohen, M. H.; Wasserman, A. On hardness and electronegativity equalization in chemical reactivity theory. *J. Stat. Phys.* **2006**, *125*, 1121–1139.

(80) Cohen, M. H.; Wasserman, A. On the foundations of chemical reactivity theory. *J. Phys. Chem. A* **2007**, *111*, 2229–2242.

(81) Mayer, I. Bond order and valence indices: A personal account. *J. Comput. Chem.* **2007**, *28*, 204–221.

(82) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.

(83) Genovese, L.; Neelov, A.; Goedecker, S.; Deutsch, T.; Ghasemi, S. A.; Willand, A.; Caliste, D.; Zilberberg, O.; Rayson, M.; Bergman, A.; Schneider, R. Daubechies wavelets as a basis set for density functional pseudopotential calculations. *J. Chem. Phys.* **2008**, *129*, 014109.

(84) Mohr, S.; Ratcliff, L. E.; Boulanger, P.; Genovese, L.; Caliste, D.; Deutsch, T.; Goedecker, S. Daubechies wavelets for linear scaling density functional theory. *J. Chem. Phys.* **2014**, *140*, 204110.

(85) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(86) Hartwigsen, C.; Gœdecker, S.; Hutter, J. Relativistic separable dual-space Gaussian pseudopotentials from H to Rn. *Phys. Rev. B* **1998**, *58*, 3641–3662.

(87) Willand, A.; Kvashnin, Y. O.; Genovese, L.; Vázquez-Mayagoitia, Á.; Deb, A. K.; Sadeghi, A.; Deutsch, T.; Goedecker, S. Norm-conserving pseudopotentials with chemical accuracy compared to all-electron calculations. *J. Chem. Phys.* **2013**, *138*, 104109.

(88) Mohr, S.; Dawson, W.; Wagner, M.; Caliste, D.; Nakajima, T.; Genovese, L. Efficient computation of sparse matrix functions for large-scale electronic structure calculations: the CheSS library. *J. Chem. Theory Comput.* **2017**, *13*, 4684–4698.

(89) Dawson, W.; Nakajima, T. Massively parallel sparse matrix function calculations with NTPoly. *Comput. Phys. Commun.* **2018**, *225*, 154–165.

Graphical TOC Entry