Received: 28 April 2021 | Revised: 31 August 2021 | Accepted: 1 September 2021

DOI: 10.1002/wcms.1574

ADVANCED REVIEW

WIREs COMPUTATIONAL MOLECULAR SCIENCE WILEY

Density functional theory calculations of large systems: Interplay between fragments, observables, and computational complexity

William Dawson[1] | **Augustin Degomme**[2] | **Martina Stella**[3] | **Takahito Nakajima**[1] | **Laura E. Ratcliff**[3] | **Luigi Genovese**[2]

[1]RIKEN Center for Computational Science, Kobe, Japan

[2]Université Grenoble Alpes, INAC-MEM, L_Sim, Grenoble, France

[3]Department of Materials, Imperial College London, London, UK

Correspondence
Luigi Genovese, Université Grenoble Alpes, INAC-MEM, L_Sim, F-38000 Grenoble, France.
Email: luigi.genovese@cea.fr

Funding information
Engineering and Physical Sciences Research Council, Grant/Award Numbers: EP/P033253/1, TYC-101; Thomas Young Centre

Edited by: Modesto Orozco, Associate Editor

Abstract

In the past decade, developments of computational technology around density functional theory (DFT) calculations have considerably increased the system sizes which can be practically simulated. The advent of robust high performance computing algorithms which scale linearly with system size has unlocked numerous opportunities for researchers. This fact enables computational physicists and chemists to investigate systems of sizes which are comparable to systems routinely considered by experimentalists, leading to collaborations with a wide range of techniques and communities. This has important consequences for the investigation paradigms which should be applied to reduce the intrinsic complexity of quantum mechanical calculations of many thousand atoms. It becomes important to consider portions of the full system in the analysis, which have to be identified, analyzed, and employed as building-blocks from which decomposed physico-chemical observables can be derived. After introducing the state-of-the-art in the large scale DFT community, we will illustrate the emerging research practices in this rapidly expanding field, and the knowledge gaps which need to be bridged to face the stimulating challenge of the simulation of increasingly realistic systems.

This article is categorized under:

Electronic Structure Theory > Density Functional Theory

Software > Simulation Methods

Structure and Mechanism > Computational Materials Science

KEYWORDS

biomaterials, biomolecules, density functional theory, fragment molecular orbitals, large scale QM methods, macromolecular systems

1 | INTRODUCTION

Kohn–Sham (KS) density functional theory (DFT)[1,2] is a popular quantum mechanical (QM) framework for the modeling of molecules and materials. As the computing power available from high performance computing (HPC) technology continues to grow and novel algorithms are developed,[3] the prospect of research paradigms based on the application of

DFT to large systems (10,000 atoms or more) is becoming closer and closer to realization. Nonetheless, despite many articles demonstrating benchmark calculations on large systems,[4] applications which go beyond prototypes or demonstrators remain relatively uncommon.

With this in mind, in this review we will deviate from the many previously existing works which cover the algorithmic aspects of large DFT calculations, in order to focus on the practical work of applying such methods. We will consider what new paradigms must be employed, such as how to analyze calculated properties of large systems, the interplay between calculations and experimental knowledge, and the potential for large-scale DFT calculations being employed in other communities. We will show how the QM level of description which is available nowadays for researchers interested in large systems is largely sufficient for providing numerous insights that can complement characterization methods employed in other disciplines. There are, therefore, numerous opportunities for novel research at the intersection of different communities. The price to pay is to conceive the simulations differently than the common practice for "small scale" systems.

We will begin by briefly describing the computational techniques used for large systems. We will show how locality not only improves computational efficiency, but also leads to the concept of a "quasiobservable," which is a key tool for productive large scale calculations. We will then present recent applications of QM methods to materials, where the research paradigms for small systems can be readily deployed to larger length scales. Then we will turn to the study of biological systems, first looking at calculations based on the fragment molecular orbital (FMO) method and then with DFT. We will find that many of the lessons learned from FMO calculations will provide inspiration for the new research paradigms required to apply DFT to biological systems. We will also review large scale excited states calculations, and the opportunities such calculations present. Finally, we will discuss major differences in how large systems are prepared for DFT analysis, and describe types of "know how" and tools that must be developed for the DFT community in order to collaborate with other disciplines.

2 | DFT AT LARGE SCALE: THE IMPACT OF LOCALITY

In a previous advanced review article[4] we gave an overview of several mature, and in many cases open-source, programs which implement state-of-the-art theories for performing atomistic simulations and exploiting the power of supercomputers to speed up challenging calculations of large systems. The techniques employed in each code to accomplish this differ in the choice of basis set for the discretization of the KS orbitals, as well as the algorithms employed to search for the ground state solution. However, all such computational approaches deeply rely on Walter Kohn's nearsightedness principle,[5] which, in a nutshell, states that in a "well-behaved" system, the information about the electrons of a system's subregions can be completely described by considering a finite domain, localized around such a region, regardless of the actual size of the entire system. The rigorous formulation of this principle can be stated from the exponentially localized behavior of the single particle density matrix in systems which have a finite electronic gap (or metals at high temperature).

This nearsightedness can be seen both in a density matrix and localized orbital picture. In KS DFT, the KS orbitals $|\psi\rangle$ are discretized in terms of some set of M basis functions $|\phi\rangle$:

$$|\psi_i\rangle = \sum_{\alpha}^{M} c_i^{\alpha} |\phi_{\alpha}\rangle. \tag{1}$$

Such functions can have a localized support, for this reason they are also known as "support functions." This leads to the generalized eigenvalue problem:

$$\sum_{\beta} H_{\alpha\beta} c_i^{\beta} = E_i \sum_{\beta} S_{\alpha\beta} c_i^{\beta}, \tag{2}$$

where, $H_{\alpha\beta}$ are the matrix elements of the KS Hamiltonian in the basis set described by the support functions, $H_{\alpha\beta} \equiv \langle \phi_{\alpha} | \widehat{H} | \phi_{\beta} \rangle$, E_i the orbital energies, and $S_{\alpha\beta} \equiv \langle \phi_{\alpha} | \phi_{\beta} \rangle$ is the overlap matrix of the support functions. The single particle density operator $\widehat{F}$ may then be defined from the density kernel matrix, K, constructed from the KS coefficients:

$$K^{\alpha\beta} = \sum_{i}^{N} f_i c_i^{\alpha} c_i^{\beta} \Rightarrow \widehat{F} = \sum_{\alpha\beta} |\phi_\alpha\rangle K^{\alpha\beta} \langle\phi_\beta|, \tag{3}$$

where, f_i is the occupation number of the ith KS orbital. The KS orbitals themselves may be distributed across the entire domain. The nearsightedness principle may then be realized in the sparsity of the matrices S, H, and K. We note that nearsightedness is a property of the system, not of the basis set.

The representation of the electronic structure in terms of such matrices is a less common picture for basis sets like plane-waves, real space, or wavelets, where the overall number of degrees of freedom is such that those matrices become unmanageably large. In this case, usually, one directly manipulates the expression of the KS orbitals in the basis set, and may potentially lose track of the locality of the $\widehat{F}$ operator. However, even in this case, it is possible to benefit from the nearsightedness principle by extracting a posteriori information about the system's localized quantities. The simplest approach involves selecting k well-chosen columns of the density matrix to produce a set of nonorthogonal, localized orbitals, where k is the number of occupied orbitals. The choice of those columns can be guided by the pivot vector of the QR decomposition.[6] A pivoted Cholesky decomposition can also be employed to produce a set of local, orthogonal orbitals.[7] Localized basis sets may also be produced by performing various unitary transformations of the KS orbitals. For example, in the construction of Maximally Localized Wannier Functions[8] (or Boys orbitals[9]), a transformation is constructed to minimize the spread of the resulting orbitals. In the Recursive Subspace Bisection method,[10] a transformation is created using the CS decomposition[11] which localizes the orbitals on a hierarchical set of sub-domains. An orbital is considered localized on a subset of those sub-domains if the norm of the orbital outside those domains is below some threshold. For example, if 95% of an orbital is contained in only a single sub-domain at the lowest level after recursively bisecting each direction (x, y, z) four times, it would have a domain size of $\left(\frac{1}{2^4}\right)^3 \times 100\% = 0.024\%$ with a threshold of 0.05 or greater. In Figure 1 we plot some localization measures of the previously mentioned methods, by calculating the electronic structure with both a localized basis set description and a plane-wave based approach.

Such considerations have two fundamental consequences. On the one hand, it states that it should be possible to design an algorithm for QM of large systems whose computational workload is linearly proportional to the number of the atoms of the system. Secondly, once the size of a given system becomes large enough, the nearsightedness principle is intrinsically suggesting that, from a QM perspective, the system can be considered as a juxtaposition of smaller QM

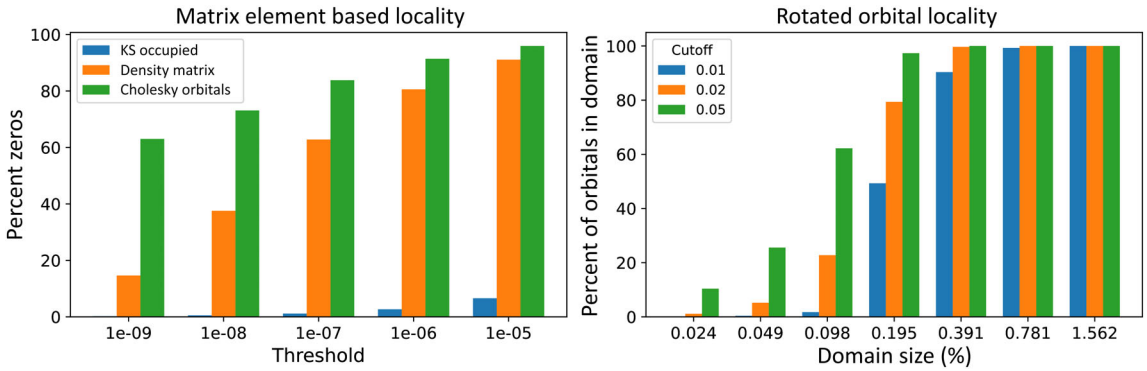

FIGURE 1 Measures of locality of a system composed of a peptide (the N3 inhibitor molecule in a X-ray crystal structure with SARS-CoV-2 Mpro from protein data Bank[PDB]: 6 LU7) in salt water (2536 total atoms) computed with the PBE functional.[12] The density matrix and Cholesky orbitals (Lowdin orthogonalized) were computed with the support functions of the BigDFT code[13] (a program based on in-situ optimized support functions) and the recursive subspace bisection (RSB) orbitals with the Qbox code[14] (a plane-wave program). Both codes employ a frozen-core approximation based on pseudopotentials.[15–17] *Left*: The percentage of matrix elements with an absolute value below a given threshold for the density matrix, KS orbitals, and Cholesky localized orbitals. When a suitable transformation is applied to the KS degrees of freedom, the matrices increase their level of sparsity. *Right*: The percentage of the domain on which the RSB orbitals are localized at different cutoff values. The RSB orbitals were localized by recursively subdividing the real space domain four times in each direction (4 4 4) and using various cutoffs (0.05, 0.02, 0.01). This figure illustrates that the system's degree of locality is a basis-independent quantity, and can be inspected even when nonlocalized computational degrees of freedom are employed

units, which of course are mutually entangled. This is a picture substantially different from the usual representation of a QM calculation for a traditional DFT use-case: a small system should always be considered as a unique QM region, whereas for a large system it should be possible to produce a "divide-and-conquer" strategy, not just for computational efficiency, but also to interpret the QM observables.

We know that the main physical content of a QM simulation is expressed in terms of the system's observables. For large systems it may become useful (under some suitable conditions) to interpret the observables of the system as a whole in terms of a sum of "quasiobservables," which are in turn associated to localized regions. Stated otherwise, in a large system, a QM observable can be decomposed into contributions which are associated with the system's fragments, and the concept of a fragment quasiobservable emerges, as illustrated by Figure 2. These quasiobservables, and applications that have exploited them, will be the main focus of this review.

3 | STANDARD PARADIGMS: MATERIALS SCIENCE APPLICATIONS

The aim of this review is to explore the use of large scale DFT in the context of systems which require a nontraditional approach, beyond the view of treating a system and its observables as a single QM identity. Nonetheless, while many (potential or existing) applications of large scale DFT require this new way of thinking, there are also a range of cases where DFT may be applied as-is. That is, given an approach which is able to treat a sufficient number of atoms, there are many new applications which open up without the need to substantially modify the workflows which are typically used to treat smaller systems. Many of these applications may be loosely described as falling into the area of materials science. Before moving on to the main topic of this review, here we, therefore, briefly highlight some recent examples of large scale DFT applications within this context.

The treatment of solid state systems based on the traditional approach of simulating a small unit cell using plane-wave DFT has been extremely successful. Nonetheless, there are many examples where it becomes necessary to treat a large number of atoms, either in order to generate a more realistic model for the material itself, or to include important environmental effects which may be highly relevant for use in a particular context. This includes the treatment of defects, disorder or domains, which prevent the system from being mapped onto a small unit cell.[19–25] Similarly, amorphous materials intrinsically require the treatment of many atoms in order to effectively capture their structure.[26,27] Another application area which has seen a number of examples of large scale simulations is in the treatment of two-dimensional (2D) materials. Examples include finite nanoflakes,[28,29] defect or adsorbate calculations,[30–32] and the investigation of heterostructures, where different twist angles between layers result in large unit cells.[33–36]

Large scale DFT is also particularly applicable for simulating nanomaterials. This includes nanoparticles (NPs) and quantum dots (QDs),[37–41] which often require the treatment of many atoms in order to approach realistic particle sizes. Furthermore, in other cases it is important to model interactions between NPs and adsorbed molecules or surfaces,[42–47] or for QDs embedded in other systems,[48,49] further increasing the system size. Other examples of large scale nanomaterial simulations include one-dimensional materials such as nanotubes,[50–55] nanowires,[56,57] and nanorods.[58–60]

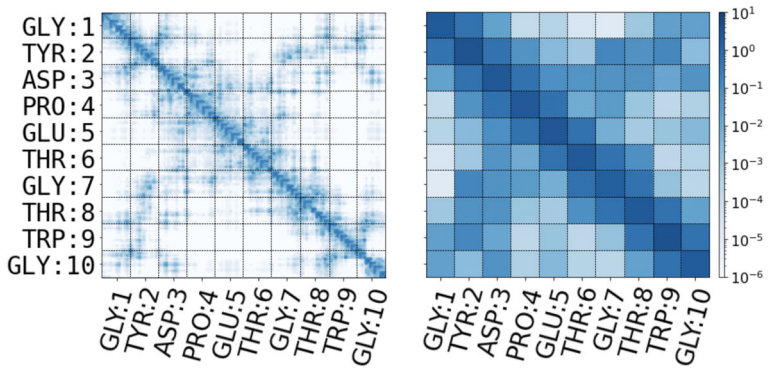

FIGURE 2 The density matrix (Lowdin orthogonalized) of a protein (1UAO[18]) computed with the BigDFT code using the PBE functional and HGH pseudopotentials.[15,16] *Left*: The decay of the density as measured by the (absolute) value of each matrix element. *Right*: The decay as measured by the norm of matrix blocks associated with each residue of the protein. "Quasiobservables" are expectation values taken over matrix blocks (both on and off diagonal)

Finally, there are various examples of applications to supramolecular systems, including donor–acceptor complexes and molecular crystals where the size and number of molecules involved leads to a large system size.[61–65] However, while such supramolecular systems do not *require* any special treatment beyond the ability to simulate at least several hundred atoms, they nonetheless may benefit from nonstandard applications of DFT, including fragment-based analysis, as we will describe in the forthcoming sections.

A number of common themes occur among the above applications concerning the types of properties which are calculated. This includes properties which are also common for small scale DFT calculations, such as geometry optimizations and structural comparisons, energetics such as adsorption and defect energies, and electronic properties such as band gaps, band structures, and densities of states. Spectral and excited state properties also arise, as is discussed further in Section 6. Finally, there are a number of examples where DFT is used to validate or compare with parameterized models.[29,31,32,41] While the use of DFT for small systems is already well established for parameterizing or validating a range of approaches, from force fields to model Hamiltonians, an important advantage of large scale DFT is that such approaches can be assessed at a *representative* length scale of the actual system of interest.

In summary, there already exist a number of applications of large systems which are, somehow, conceived with an investigation paradigm that is very close to traditional studies employing DFT for relatively few atoms. Thanks to recent advances in both computer codes and HPC technology, this sector of applications is rich in new opportunities. However, the considerations which are detailed in the forthcoming sections also offer the potential for conducting research with an increasingly interdisciplinary character, thanks to the increased realism in terms of atomic structure provided by large scale DFT.

4 | FRAGMENTATION OF LARGE SYSTEMS

Moving beyond applications based on a more traditional view of DFT, in this section we provide a first collection of examples in which the scientific investigation paradigm substantially differs from the established protocols for DFT at a small scale. Before examining DFT calculations, we first look at recent applications of the FMO method.[66] FMO is a well-established method, which has enabled the application of QM methods to large systems for many years. FMO is just one method in a whole family of fragment approaches which have been surveyed elsewhere,[67,68] however in this review we will limit our focus to FMO. By examining recent FMO applications, we will find hints about the types of workflows which are needed to apply DFT to large systems.

4.1 | FMO methodology

The first consideration which drives the FMO approach is the interpretation of the most relevant QM observable: the system's energy. In the FMO method, a system is partitioned a priori into a set of N nonoverlapping fragments. Then, the energy of the system, E, is written as a many body expansion:

$$E = \sum_{i}^{N} E_i + \sum_{i}^{N} \sum_{j}^{N} \left(E_{ij} - E_i - E_j\right)....$$

The most commonly employed FMO2 method truncates this expansion at the two body term. The three-body FMO3[69,70] and even four-body FMO4 methods[71] can also be employed for improved accuracy. To efficiently employ the above ansatz, the FMO procedure first self-consistently computes the energy of each lone fragment in the Coulomb bath of the other fragments. The dimer energies are then computed in the frozen environment of the other monomers. The electrostatic embedding employed by FMO plays a key role in the rapid convergence of the many-body expansions (see the early work of Stoll and Preuss[72]). By exploiting this fragment ansatz, FMO reduces the computational scaling of the employed QM method, and can be efficiently parallelized. The FMO method can be used to compute gradients[73,74] and full system molecular orbitals can be recovered.[75,76]

One benefit of the FMO method is that it generates as a byproduct a matrix of quasiobservables, $\Delta E_{ij} = E_{ij} - E_i - E_j$, which are called the pair interaction energies. These quasiobservables are a measure of the interaction strength between the different fragments of a system. When combined with the Kitaura–Morokuma energy decomposition analysis,[77] we

arrive at the pair interaction energy decomposition analysis (PIEDA).[78] This analysis allows for a detailed decomposition of the interactions between the system's fragments:

$$\Delta E_{ij} = \Delta E_{ij}^{ES} + E_{ij}^{EX} + E_{ij}^{DI} + E_{ij}^{CT+MIX}.$$

The terms in order are: electrostatic (ES), exchange-repulsion (EX), dispersion (DI), charge-transfer (CT), and mixed (MIX) terms which are the remainder. The decomposition may also be extended to include solvation effects and basis set superposition error. Recently, this analysis has been expanded to groups of fragments through a subsystem analysis framework.[79] Remarkably, PIEDA is able to take a fundamental weakness of the FMO method (writing the energy in terms of monomers and dimers) and transforms it into a strength. Now when an FMO calculation is performed on a system, detailed insight into the interactions between fragments can be obtained.

We note that FMO is not the only fragment method for which this type of interaction analysis can be performed. Recently, a fragment interaction analysis for large systems has also been presented as part of the Molecules-in-Molecules framework.[80] X-Pol[81] is another fragment based method which was developed in parallel to FMO (and indeed offers many improvements). The X-Pol method has been combined with symmetry adapted perturbation theory (SAPT)[82] to create the XSAPT[83] method. This combination can potentially improve on an FMO + PIEDA approach as SAPT offers an improved description of intermolecular interactions over the standard MP2 level of theory employed in FMO calculations.

4.2 | FMO applications

As a demonstration of the utility of this type of analysis, we can look at a recent article which utilizes the FMO + PIEDA method to study G-Protein Coupled Receptors (GPCRs).[84] GPCRs are a group of signaling proteins which are the target of about 40% of all drugs on the market. In this study, the authors performed FMO calculations on a database of 18 GPCR-ligand crystal structures, confirming that the total pair interaction energies between ligands correlated well with experimental affinity. Beyond this, they were able to identify which amino acids played an important role in binding, including some interactions that were previously unknown. For each interaction, they were able to measure the contributions of electrostatics, dispersion, and charge transfer to the protein–ligand interactions, resulting in detailed interaction maps.

Remarkably, in this study the FMO calculation was performed on only a cluster system defined by the amino acids within a 4.5 Å distance from the target ligand. Similar protocols have been recently used by other groups to develop antibiotics,[85] design antidiabetic drugs,[86] and to study proteins used in clinical diagonistics.[87] We will highlight another study which used this protocol to study interleukin-2 inducible T-cell kinase inhibitors (ITK).[88] Inhibition of ITK activity is an important target for the control of allergic asthma. This article describes the discovery of novel inhibitors of ITK using FMO as a guide. Starting with a known inhibitor, FMO was used to reveal the key interactions. During the refinement process, FMO revealed unexpected new interactions, and thereby design rules. Thus while FMO began as a method to decrease the computational costs of computing large systems, it has shown its value through its ability to take complex systems and extract detailed descriptions of their interactions, as illustrated in Figure 3. Systems are no longer seen as merely a collection of atoms, but instead can be described at the biologically relevant level of detail (amino acids).

This is not to say that FMO has been limited in application only to small systems. In a more recent study of GPCRs,[89] full protein systems were computed so that interaction strengths between different transmembranes could be quantified. By applying FMO to a set of 35 structures representing different branches of the evolutionary tree, the authors were able to construct a consensus network of interactions that are preserved across the dataset. Such a network can shed insight into how mutations far from the active site can have an effect on ligand binding. Many of the interactions identified are those which cannot be described well by classical force fields. In a similar spirit, Sladek et al.[90] used the PIEDA interaction energies to define protein residue networks. Protein residue networks (see Estrada[91] for a review) are a commonly used technique to analyze interactions in proteins. However, standard practice relies on distance criteria or force field energies. Using the FMO method, a more accurate description of protein interactions can be obtained. FMO can thus provide improved descriptions of system interactions that match the research paradigms of computational biology.

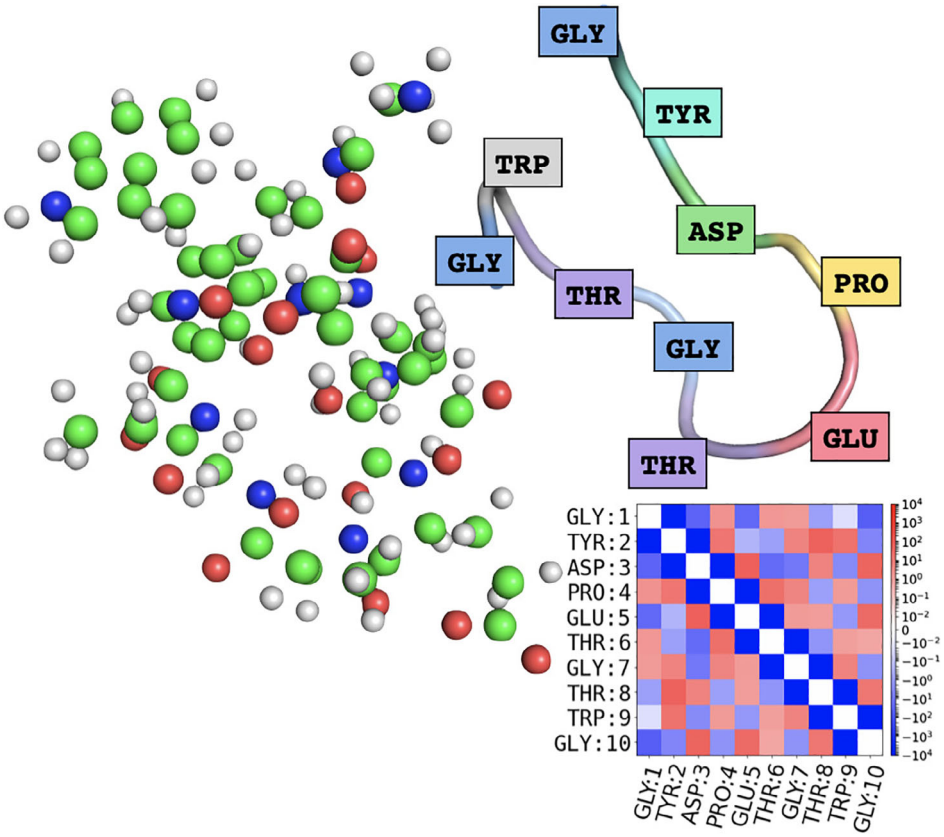

FIGURE 3 The FMO method transforms a complex collection of atoms into the language of biology by quantifying the interactions between protein residues. The heat map shows the fragment interaction energies (kcal/mol) of the pictured protein (1UAO) computed with the GAMESS program at the HF//6-31G* FMO2 level of theory

4.3 | The maturity of the FMO method

From the examples presented above, we can see that FMO is a mature computational method for studying large biological systems. A recent textbook[92] highlights a number of applications of FMO to the problem of drug design. As evidence of the maturity of the FMO method, we also highlight a recent study by Hatada et al.[93] In this study, the FMO + PIEDA approach was applied to study the main protease of the SARS-CoV-2 virus. This study was completed within only a few months of the release of crystal structures of SARS-CoV-2 proteins, demonstrating the readiness of the FMO method for practical application.

As we will discuss later in this review, the application of QM methods to biological systems requires new workflows for preprocessing/postprocessing. These kinds of new workflows have already been a subject of study for the FMO community. For the FMO method, it is particularly important to have tools which can sensibly partition a system into fragments. The FU program[94] is one such tool, providing a GUI application that can be used to prepare crystal structure files, fix structure errors, create input files for FMO calculations, and visualize the results. FU is a python program with the ability to run arbitrary, user-defined python scripts, which anticipates the complex workflows that may be required.

However, manual preparation requirements would limit the size of dataset to which FMO might be applied. To address this issue, an automatic protocol for performing FMO calculations (called auto-FMO) has been developed.[95] Recently a new FMO database has been created[96] which contains over 13,000 entries. Many of these entries were prepared using the auto-FMO program. The availability of automatic protocols and databases of the quasiobservables generated by FMO may be a promising source of interaction descriptors for future studies.

In this section, we have reviewed a number of applications of the FMO method. We have seen the advantages of FMO as a tool for studying large systems, bringing with it new paradigms which differ from typical DFT calculations at a small scale. The division of systems into fragments, and detailed analysis of those interactions, is the basis for FMO's application to drug design. Yet, there are also drawbacks to this approach, as an a priori partitioning of the system must

be employed. This type of approach relies heavily on chemical intuition, and does not provide opportunities for discovering new building blocks of the system. In the next section, we will turn to applications of DFT to large biological systems, where we will see how these previous FMO studies can provide guidance for new DFT-based paradigms.

5 | NEW PARADIGMS FOR DFT

In this section, we explore applications of linear scaling DFT to large molecular systems. We begin by discussing "traditional" applications which, like the application to materials presented earlier, focus on simply the energy and forces of a system. We will then continue to the use of DFT to derive novel system descriptors. Much like in the FMO section, we find that these novel applications provide an exciting blueprint for future QM paradigms.

5.1 | Energy and forces

The improved description of molecular systems offered by DFT over classical approaches make it an appealing method for computing the energies and forces of challenging systems. DFT in these cases acts as a type of improved force field which can handle a diverse array of systems. In an early application of large scale DFT, Otsuka et al.[97] performed geometry optimizations of the FK506 binding protein and three different ligands. It is promising to apply such an approach to the abundance of protein–ligand X-ray structures to overcome limitations in experimental resolution. They found that the geometries of the flexible protein side chains are significantly modified by the optimization procedure, with a correspondingly significant impact on binding energy estimates.

However, proteins in the body have a very different structure than those measured by crystallography due the effects of both the solute environment and finite temperature. When attempting to predict the affinity of proteins and ligands, it has become established scientific practice to include finite temperature sampling. One popular approach is the molecular mechanics Poisson–Boltzmann surface area (MM-PBSA) method,[98] which involves postprocessing molecular dynamics (MD) trajectories in order to compute a free energy of binding. While DFT is certainly too computationally expensive to perform such extensive sampling, using DFT to postprocess trajectories generated by classical force fields is one possible mixed workflow.

The MM-PBSA approach was modified by Fox et al.[99] to include enthalpic terms computed using linear scaling DFT. In that study, MD trajectories of T4 lysozyme in complex with various ligands were run using a classical force field, with DFT calculations then applied to snapshots (1000 snapshots taken over 19 ns of simulation time). They found that using DFT energies significantly improved the binding estimate over MM energies when compared with experiment. They noted that the energy values of the force field and DFT agreed well in vacuum, and that it was the more accurate QM solvation energy that contributed most to that improvement. Of course, such postprocessing might be possible using QM/MM with a well-chosen QM region. However, such a region can grow quickly when we are considering protein–protein binding, as was demonstrated by Cole et al.,[100] who used such an approach to study the RAD51-BRC4 protein complex. In this case, modeling of the interaction region required a cluster system of around 2800 atoms to converge the QM correction to the free energy of binding to less than 1 kcal/mol.

Geometry optimization is perhaps most useful when combined with transition state searches. While commonly employed classical force fields are well parameterized for ground state geometries, transition states are more challenging. Recently, Svensson et al.[101] performed transition state searches of inhibitors of insulin-regulated aminopeptidase (IRAP). Targeting the IRAP enzyme is a promising approach for treating cognitive disorders or dementia. Studying the transition state of a ligand bound to an enzyme can lead to design rules for drug discovery, as novel inhibitors can be designed to resemble the interactions of the transition state. Using the optimized DFT geometries they searched for similar structures to the transition state. Importantly, they could score candidate structures by docking them with the enzyme's transition geometry, allowing them to take into account conformation changes of the enzyme when influenced by ligands similar to the known transition state.

A significant limitation of the previous study was its use of small cluster systems. With recent developments in linear scaling DFT, however, modeling of the entire protein is now tractable, as demonstrated by Lever et al.[102] In this study, they computed the activation energy barrier and transition states of the classic Claisen rearrangement with Chorismate as the catalyst. They found that good agreement with experiment could be obtained, but only after including a QM region of just under 1000 atoms. Large scale DFT may therefore prove to be a crucial tool for future studies of enzymes.

Frequently, proteins contain metal atoms, which play an important role in the protein's function. Such systems are particularly difficult to model for classical force fields. In some cases, even DFT can fail to accurately model these multireference effects. In a study of myoglobin, Weber et al.[103] used a combination of linear scaling DFT and Dynamic Mean Field Theory to understand in detail the nature of myoglobin's binding to O_2 and CO. This example suggests that one important role which linear scaling DFT might play in the future is as an embedding environment for higher level QM methods.

5.2 | Quantum observables

DFT includes the explicit modeling of electronic degrees of freedom, which can lead to novel applications beyond its use as a more accurate and flexible force field for computing energy and forces. The KS orbitals can be used to shed insight into electronic interactions in a system. For example, Feliciano et al.[104] used orbitals around the HOMO-LUMO (highest occupied molecular orbital-lowest unoccupied molecular orbital) gap to understand electron transfer between *Geobacter sulfurreducens* (GS) and Fe(III) oxides. In this work, MD snapshots of the GS pilin (∼1000 atoms) were post-processed by DFT calculations in order to understand the long range electron transfer in this system. Visualization of the frontier orbitals revealed important amino acids for facilitating charge transfer, results which correlated well with amino acids which are evolutionarily unique to GS.

Transformations of the KS orbitals such as through the construction of natural bond orbitals (NBOs)[105] are another way the electronic degrees of freedom can be used to analyze the results of large scale calculations. In the previously discussed study of Claisen rearrangement,[102] NBOs were constructed[106] in order to understand the contribution of individual residues to the stabilization of the transition state. The computed NBOs can be combined with second order perturbation theory to estimate a stabilization energy coming from hydrogen bonding. This allowed the authors to identify important amino acids in the active site.

In a similar spirit to the PIEDA analysis mentioned in the previous section, energy decomposition analysis techniques[107] have been integrated into large scale DFT calculations.[108,109] In the ONETEP code, a hybrid scheme based on the ALMO EDA[110] and a frozen density analysis based on the LMO EDA method[111] has been implemented, decomposing the energy into: electrostatic, exchange, Pauli repulsion, correlation, polarization, and charge transfer. Solvation effects may be calculated and visualization of density differences is also available. This approach was applied to the analysis of inhibitors of Thrombrin proteins, which is a target for the treatment of thrombosis. By studying truncated systems of increasing size, it was found that exchange, Pauli repulsion, and correlation converged at 9 Å, while electrostatics did not converge until 15 Å, which shows the importance of large scale calculations for performing reliable analysis of protein–ligand interactions. By analyzing the different components of the EDA, similar ligands could be compared, and differences in the bonding nature could be understood even when overall interaction energies are similar. Analysis can also be performed on ligand functional groups, though care must be taken to understand the errors introduced by such a bond partitioning process.

The most fundamental observable of DFT calculations that might be used to analyze systems is the electron density itself. The electron density serves as the basis for topological based analysis such as the quantum theory of atoms in molecules.[112] Recently, the noncovalent interaction technique presented by Johnson et al.[113] has become a popular way to analyze interactions in large protein systems. The popularity of this method is in part because the density can be efficiently estimated using a super position of atomic densities. However, recently this approximation has been improved using a fragment based technique to construct the total density.[114] To the best of our knowledge, this analysis has not yet been paired with large scale DFT calculations. However, this kind of pairing represents a promising future development for large scale DFT.

5.3 | QM-based indicators

So far in this section we have provided some examples of how ideas which can be mutated from energy partitioning, pioneered by techniques like FMO, can be applied to DFT calculations of large systems. By now, it should be clear to the reader that this operation is by no means easy to handle. When a system becomes large, one cannot anymore rely on visual inspection of the entire critical regions of the system. It is therefore important to have the possibility of instead relying on *indicators* which would enable the user to understand if a particular region of the system requires a specialized treatment.

Recently, we have been exploring the connection between density based analysis techniques and fragmentation. As we have seen, fragmentation provides a powerful framework for analyzing interactions at a coarse grained level. However, the downside of fragment approaches is that a system must be decomposed a priori. In the previously presented cases of DFT calculations, the amino acids of proteins were chosen as the level of detail for analysis. However, it is not clear that this is the ideal fragmentation, as the role an amino acid might play can be heavily influenced by its environment. Furthermore, such an approach is not applicable to protein–ligand binding (beyond treating a ligand as a single fragment), metalloproteins, solid state calculations, and so on.

When postprocessing a DFT calculation, it is in theory possible to perform an analysis on an arbitrarily defined set of fragments. The only challenge is to define those fragments in a chemically meaningful way. To study the choice of fragmentation scheme, in a previous study we introduced a measure of fragment quality called the purity indicator.[115] This measure is directly based on the density, being computed as the deviation from idempotency of the density matrix block associated with a given fragment, as illustrated in Figure 4. By combining this approach with optimization algorithms,[116] it is possible to define a chemically meaningful set of de novo fragments. Similar to the subsystem analysis derived for FMO calculations,[79] fragments may be combined in order to understand interactions in a system at different length scales. The interactions of a system may be described in the same framework by computing a quasiobservable called the "fragment bond order" as a measure of the off-diagonal contributions of the density matrix.

As seen in Figure 4, the purity indicator and fragment bond order are computed directly from the density matrix. We also plot the purity indicator values of a small protein (1UAO). In this case we see one amino acid residue (nonterminal glycine) is evaluated to be a much poorer fragment than the others, which matches the know-how of the FMO community.[117] The Purity Indicator and Fragment Bond Order together measure the competition between a fragment's internal and external interactions. This approach has recently been applied to generating graph views of proteins as well as to understand the interaction between proteins and solvent molecules.[13]

Returning now to the topic of extended systems, we find that the locality of a system can be used to gain insight for these kinds of systems as well. For example, in a large scale study of vacancies in bulk silicon,[19] maximally localized Wannier functions[8] were constructed and visualized to gain a chemically intuitive understanding of the vacancy site. The recursive subspace bisection method[10] is another method which is able to localize orbitals into arbitrary subregions. While this approach was originally suggested for speeding up the computation of Hartree–Fock exchange,[118,119] it has also been suggested as a standalone analysis tool, able to reveal how surface effects can modify locality, or the locality that can exist even in metallic systems.[120] The challenge for constructing locality based, quasiobservables of

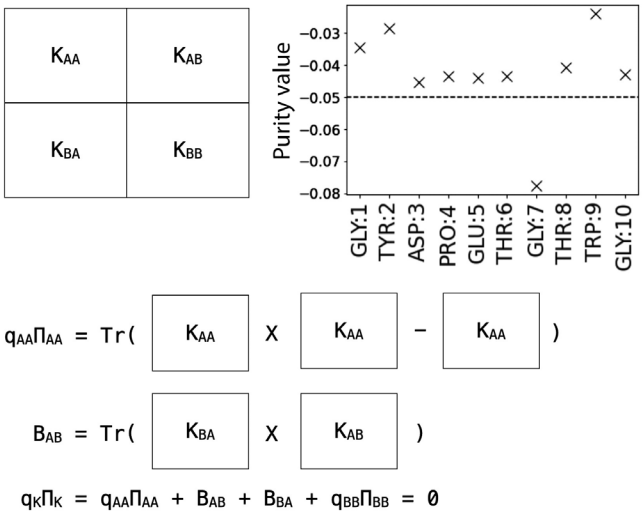

FIGURE 4 Diagrammatic representation of the calculation of the fragment purity (π) and bond order (B), where q is the number of electrons in a fragment. The overlap matrix can be incorporated either through a Mulliken or Lowdin type projection. The density matrix of the full system is idempotent, and thus has a π value of zero. Fragments have a negative, non-zero purity indicator value, which comes from truncating the B terms in the full expansion. Purity values of the 1UAO protein computed with BigDFT using the PBE functional are also plotted. (absolute) purity values below 0.05 (see previous work[115,116] for a justification of this threshold) indicate that the fragments are sufficiently pure and thus represent a good decomposition of the system

extended systems is in defining a set of chemically meaningful fragments. In particular it is desirable to have sets of fragments which are transferable both between systems and even within a system. In the context of further accelerating linear scaling calculations and constrained DFT calculations, a (pseudo-)fragment approach has been introduced,[121–123] wherein the transferability of a fragment can be assessed through rigid-body transformations.

In this section, we have seen ways in which DFT can be applied to gain insight into the properties of large systems. We have found that exploiting locality is a key ingredient in the new workflows that are emerging. Locality is particularly powerful when quasiobservables of well-defined fragments can be obtained. Thus, much like in the case of the FMO method, features of large systems that were originally of interest primarily for computational performance, can be exploited to gain intuitive understandings about system interactions.

6 | BEYOND THE GROUND STATE

So far, we have discussed calculations associated with ground state quantities. While DFT represents, in most cases, the best compromise between accuracy and computational complexity for ground state calculations, the availability of new tools and approaches to study large systems makes it also interesting for studying excited-state quantities. Excitations in molecular systems as well as in materials find application in many areas within physics, chemistry, biology and technology, such as in materials for energy conversion processes (e.g., organic solar cells[124]), molecular rotors,[125] and optical probes in biological systems (e.g., fluorophores in proteins).[126] In general, QM calculations of excited-state properties (including emission phenomena), particularly those based on DFT, are still considered a challenge. As opposed to their ground state counterparts, excited-states are short-lived and highly reactive making modeling their electronic structures a crucial but extremely involved task.[127] Furthermore, their description requires an accurate method to properly include environment effects.[128] As a result, the large-scale nature of molecular materials and biomolecular systems greatly challenges the use of conventional quantum-chemical methods because of the steep scaling of computational cost with system size.

Time-dependent density functional theory (TDDFT)[129] is the most widely used QM approach for computing transition energies and excited-state properties. While some recently developed algorithms[130,131] have made TDDFT, with its standard computational complexity, applicable to medium-sized systems, the simulation of larger systems, for example, biomolecules and nanostructures, requires linear scaling (LS) implementations (LS-TDDFT). Linear scaling TDDFT is implemented in various packages, for example, in ONETEP, CONQUEST,[132] and Siesta. In the ONETEP code,[133] linear-response TDDFT (both full and with the Tamm–Dancoff approximation) can be used to compute low-lying excited states and predict optical properties of large-scale systems (up to around 2000 atoms[134] with the PBE functional) ranging from nanostructured materials[135] to protein complexes.[136] In recent ONETEP applications, TDDFT has been used to study excitations of dyes in explicit solvent clusters,[137,138] the computation of excitons in multichromophore pigment–protein complexes,[139] and to model implicit and explicit host effects on excitons in pentacene derivatives.[140] The Siesta package has two main ways of obtaining the optical properties of finite systems: real-time TDDFT propagation and by computing the noninteracting dielectric function. PySCF-NAO is an open-source implementation of linear-response TDDFT which uses the KS orbitals from Siesta as a starting point.[141] Siesta has been employed for the computation of optical properties of compact metallic systems containing up to several hundreds of atoms.[142,143]

A widespread method for predicting electronic excitations is the GW approach, which is based on many-body perturbation theory.[144] Thanks to its favorable scaling with respect to system size compared with quantum chemistry methods, it has been historically applied to solids (materials, semiconductors, and insulators).[145,146] However, the robustness of its performance has motivated developments in the direction of large-scale GW simulations (where supercells can contain up to hundreds of atoms) of surfaces, interfaces and 2D materials, as well as molecular systems.[146,147] For example, Govoni et al.[148] have proposed a formulation of the GW method for large scale calculations which allows them to tackle systems of unprecedented size, including water/semiconductor interfaces with thousands of electrons and nanoparticles up to a diameter of 2.4 nm. As a future perspective, it is worth highlighting that the Bethe–Salpeter equation (BSE) formalism is establishing itself as a new efficient tool in the ensemble of computational methods available for the prediction of optical excitations. In particular, in combination with the GW approximation, it offers a similar computational cost to TDDFT with hybrid functionals, while providing an accuracy comparable with range-separated hybrid functionals. As discussed by Blase et al.,[149,150] the GW@BSE formalism has the potential to be used for systems containing several hundred atoms.

As explored in previous sections, the decomposition of systems into subsystems is a promising approach for modeling the properties of large systems. This has led to the development of multiscale and embedding methods which are nowadays largely employed for modeling complex systems, for example, transition metal complexes and biological systems. Among these, the quantum mechanics/molecular mechanics (QM/MM) formalism[151,152] is certainly the most used; it enables the description of a small part of the system, where reactions take place (active region) using QM, while the majority of the thousands of atoms that comprise the system's environment are handled through MM. However, extending the QM/MM formalism to excited states is still no trivial task as it requires high accuracy and therefore costly methods for the active region, while accurate polarization effects need to be taken into account.[153]

The FMO approach has been coupled to TDDFT within the GAMESS package to describe delocalized excitonic interactions in multichromophore systems such as pigment–protein complexes. In one study,[154] the approach was applied to a protein trimer (from the Fenna–Matthews–Olson complex) of about 6000 atoms. In this study, the FMO1 protocol was used, and significant analysis was required in order to construct large enough fragments to treat the delocalized states. The newly developed model expands the FMO method's capability to a wider range of applications in molecular crystals and photosynthetic complexes. In addition, the description accounts for self-consistent polarization of the environment that is overlooked with the standard nonpolarizable QM/MM description. Fragmentation can, hence, become a very powerful tool to improve and inform QM/MM setups for excited states, allowing for accurate modeling of the QM region (where the excitation takes place) and inclusion of environment effects. Furthermore, fragmentation approaches can be used to analyze in depth the excitations computed with TDDFT. One example is the Theodore interface[155] which performs three main functionalities: a fragment-based analysis for assigning state character, the computation of exciton sizes for measuring charge transfer, and generation of the natural transition orbitals used for visualization and quantification of multiconfigurational character.

In the direction of recent developments for the computation of spectroscopic properties of large and complex molecular systems, it is worth mentioning the latest release of the Dalton project (DP). DP is a python-based platform for accessing the quantum chemistry codes Dalton and LSDalton.[156] DP is capable of managing computer resources, handling user interactivity, steering computation, and performing data processing of results (themes we will revisit in Section 8). Recently, the capabilities of the package have been expanded for the modeling of spectroscopic properties of large systems through fragment-based approaches; the developers introduced the polarizable embedding (PE)[157] model which is a fragment-based scheme designed for the inclusion of environment effects in calculations of spectroscopic properties of large molecular systems and the PyFraME package which is used to automatize the generation of the embedding-potential parameters. The combination of these two modules allows for setting up workflows for the analysis of spectroscopic properties; an example of a possible workflow can be found in a recent review of the package[156] in an application to the Nile red chromophore.

In the context of fragment approaches, a natural playground can be found in supramolecular morphologies. For instance, in a recent article,[122] some of us have studied the impact of the disorder imposed by realistic solid-state morphologies on the quantities which govern electron transfer in OLEDs. These calculations employed constrained DFT, which allows a charge to be associated with a particular region of the material, for a system composed of a number of fragments. Thanks to the simplicity of its framework and its ability to model charge transfer mechanisms, constrained DFT has great potential for being employed within frameworks for studying large systems. Variations of its implementation can, thus, be found in a number of large scale DFT codes, for example, ONETEP,[62] Conquest,[158] and BigDFT.[121]

In this section we provided an overview of the variety of methodologies which are nowadays available for the description of excited states in large systems. Such diversity highlights that a meaningful modeling of excited states requires a certain minimal accuracy which poses a limit to the size of the investigated systems; for this reason there is still no unique answer on how to best model such large systems. However, thanks to advances in HPC capability, we have shown that modeling larger systems beyond the ground state is becoming more and more feasible, while remaining one of the most continuously evolving fields of computational physical chemistry.

7 | INTERPLAY BETWEEN EXPERIMENTAL AND SIMULATED DATA

In the previous sections, we have seen examples of calculations of systems of sizes compatible with many experimental apparatuses. Contrary to more traditional applications with conventional, small-scale DFT investigations, we are not referring here to model systems made of a few atoms, which require particular (sometimes peculiar) experimental

setups to be realized, or that are associated with theoretical simplifications of the realistic experiment. Rather, we are referring to realistic structures which can *directly* come from experimental data. This fact has a number of implications for people working on atomistic models, since all the above-mentioned studies have as a fundamental input the *structural* representation of the system.

To construct a "computational sample," and in particular to relate such a system to a realistic model of a particular experimental condition, there are numerous questions which have to be answered. We employ here the term "sample" because the typical length scales, of the order of $\simeq$10 nm, are identical to those accessible by a large range of experimental techniques. The first question to answer is whether a full QM treatment can be used alone, or should be employed in conjunction with other methods. In a previous review article,[4] we have already provided some criteria which can help in answering points such as this one. However, even for situations where full QM approaches may help, other important questions arise. For instance, how many conformations of the sample have to be considered? Would a traditional MD equilibration performed in a small region of the configuration space be enough to inspect in a satisfactory manner the variation of the quantities which are under investigation? Or, on the contrary, are we in a situation where large conformational samplings have to be inspected to study the system? Questions like these do not have a clear answer, nonetheless sometimes even a partial investigation of the problem may inform experimentalists working on the system. For these reasons, we believe it is important for computational physicists and chemists to work in close contact with other communities, in order to understand the complexities which are related to the retrieval of information which may appear simple from the atomistic simulation's viewpoint.

In the context of structural studies of macromolecular systems, X-ray, and neutron protein crystallography are important, complementary techniques. X-ray crystallography represents one of the main techniques of structural biology, and can benefit from recent technological advances as well the availability of numerous instruments worldwide.[159] A large number of PDB files from X-ray experiments are available in the RCSB Protein Databank.[160] However, the scattering cross section of X-rays is proportional to the electron density, therefore light atoms like hydrogens are poorly resolved by these techniques. In a biological system like DNA or a protein, this fact implies that some 25% of the atomic positions are not easily identified by X-ray crystallography. This is an important limitation, especially for cases where the positions of the hydrogens are important to determine biologically relevant information. For instance, knowledge about the protonation state of relevant residues of a protein is potentially of great importance in understanding the mechanisms which govern enzymatic reactions.

On the other hand, neutron-based crystallography is less frequently performed. The requirements for sample preparation (crystal size, sample deuteration) and conditioning make this technique more delicate, therefore, only a small percentage of the PDB database has been resolved by neutrons. Yet, with this technique it is possible to visualize the positions of hydrogen atoms (mostly substituted by deuterium), which gives information about the network of hydrogen-related bonding in such H-rich systems.[161] Furthermore, neutron protein crystallography can be performed at ambient temperatures, avoiding any artifacts that can emerge when cooling for X-ray analysis. When it comes to the study of such systems in solution, there are numerous techniques, like for instance small angle scattering, which can used to probe atomic structure;[162] however, their overview is beyond the scope of this article. Rather, the main message we want to highlight is that comparison with experimental data brings an additional set of "knowledge gaps" which need to be bridged in order to make different communities work together. Much of this knowledge is well known in other communities or among experts of large scale DFT simulations in biology, and we hope that highlighting it here will be helpful to DFT practitioners looking to scale up their simulations.

7.1 | DFT modeling of a crystallographic structure

To illustrate the challenges that exist for modeling large systems, we will consider example cases, extracted from the topic of drug design of potential molecules which target SARS-CoV-2. At the time of writing this review article, the Covid-19 pandemic represents a major issue that the world is facing. A number of crystallographic structures have been produced worldwide to offer to scientists of different communities information which may become useful to help mitigate the pandemic. In what follows, we will briefly discuss how such structures should be manipulated and examined in view of a DFT treatment. The rationale is not identical, depending on whether the structure comes from X-rays or from neutrons crystallography.

7.1.1 | Identifying the input structure from X-ray crystallography

Recently, Hoffman et al.[163] proposed a new ketone-based inhibitor of the SARS-CoV-2 main protease (MPro), which is one of the main targets for the design of drugs to treat the Covid-19 disease. Experimental crystal structures of this inhibitor in complex with this enzyme, as well as with the original SARS-CoV-1 main protease (which only differs by less than 4% in its aminoacidic sequence from is CoV-2 counterpart) are available (PDB code: 6XHM and 6XHN, respectively). However, even in our case where we have an experimental structure, performing a DFT calculation is not straightforward.

The PDB file contains multiple structures that must be separated out. In this case, the structure is a dimer containing two proteases which can either be separated out or computed together. Within a single strand of a dimer, certain atoms can crystallize in different conformations (denoted by the occupancy) and one should understand whether this is a consequence of the experimental resolution of a single average crystal position, or if this is a signal of the coexistence of multiple strands in crystalline form. Beyond this extraction step, many PDB files contain missing residues and atoms. This is related to experimental resolution, and also to the fact that atoms which have high mobility are more difficult to identify. Clearly, a DFT calculation would become unreliable if those atoms were missing. There exist a number of software programs which can handle the task of filling in the missing atoms of the system (e.g., CHARMM-GUI,[164] PDBFixer,[165] or Ambertools[166]). This step is usually followed by an equilibration process, where atomic force fields are employed to produce a more physically accurate set of atomic positions. Another important point is related to the total charge of the system. The protein's amino acids exchange protons with the environment, given a particular value of the pH of the solvent. This fact influences the total number of protons present in the structure as well as the total charge of the system, which of course must be explicitly set when performing DFT calculations. All of these considerations of course assume that the structure has been reliably prepared, however mistakes can be present, as shown by the recent work of the Coronavirus Structural Task Force.[167] In Figure 5 we provide a cartoon which illustrates the problem.

We see how points like these make the preparation of the structure completely different from the equivalent procedure which has to be performed in the case of smaller systems, for which global optimization can be readily performed, and equilibration done in a few tens of picoseconds. In this case, even after extracting a starting structure, it is appealing to relax or even equilibrate the structure of our protein using classical force fields. However, the presence of a covalently bonded ligand means that standard force fields cannot be applied without parameterization. This is a problem that, on one hand, adds further complexity to the preparation of the computational sample, and on the other hand represents another added value associated with the availability of QM methods for such systems, as the outcome of the DFT calculation may inform force field energy panoramas.

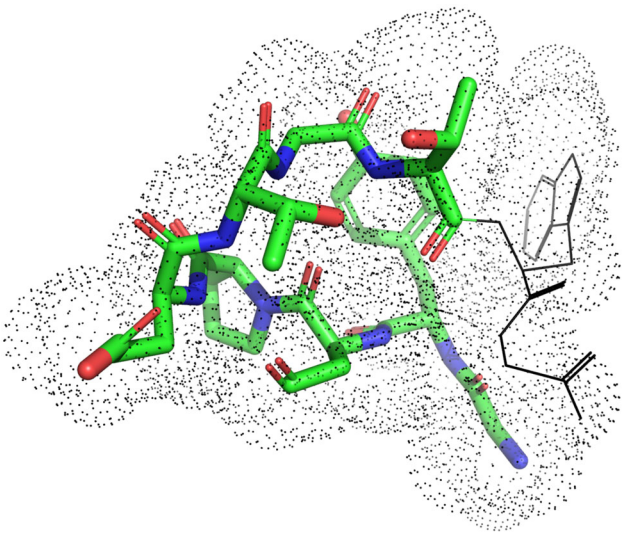

FIGURE 5 Proteins are challenging to model with DFT, as significant uncertainty exists about their structures. In data coming from X-rays, hydrogen atoms (and some heavy elements) are not present, whole residues might be missing (black lines), and the ensemble of conformations the target system might take in a realistic environment is unknown (black dots)

A treatment which follows the above-described guidelines has been also performed for systems like the X-ray crystal structure with PDB code 6LU7, which has been already mentioned in the context of the FMO + PIEDA approach[93] in Section 4.3, where the MPro protein is in complex with the N3 inhibitor (which in turn has been considered for the test of Figure 1). The 6LU7 system has also been employed in one of our recent articles[13] as an example for the fragmentation procedure described in Section 5.3. In that example, fragment charges as well as interaction graphs of the N3 inhibitor with the enzyme amino acids have been presented.

7.1.2 | Neutron crystallography and the position of light atoms

Here we will consider, in the same scientific topic, the possibility of generating a computational sample from a structure which comes from neutron crystallography. In a recent article,[168] the neutron crystallographic structure of the SARS-CoV-2 main protease has been analyzed in conjunction with another alpha-ketoamide inhibitor, the Hepatitis C antiretroviral drug Telaprevir (PDB code: 7LB7). The aim of this study was to identify the relationship between the inhibitor binding network with the protonation state of some key histidines of the active site. An investigation of the protonation state of histidines, especially the ones which are close to the enzyme's active sites, may help in clarifying catalytic mechanisms which govern enzyme behavior.

In this case as well, the positions need to be carefully analyzed before starting DFT calculations. Neutron crystallography PDB files provide also the coordinates of deuterium atoms (exchanged by hydrogen), and they appear in the PDB file with an occupancy factor which quantifies the amount of D-H exchange during the experiment. Therefore, in view of a DFT calculation, all the D atoms have to be removed in favor of their H counterpart. Since this structure was resolved at room temperature (22°C), analysis may be performed directly on this structure, and for some studies a force field equilibration step can be avoided (it may still be desired, although, to increase the number of configurations examined, like in the case of X-ray equilibrated structures).

A structure crystallized with the same technique has been also released by the same group[169] for the MPro without inhibitor (PDB code: 7JUN). Here also, the procedure to create a structure which is compatible for DFT calculations can be employed with the same guidelines as above.

7.1.3 | DFT as a postprocessing tool for interpreting crystallographic positions

As an illustrative example, we performed DFT calculations using the BigDFT code and the PBE functional on three sets of data: the neutron resolved structures (7LB7 and 7JUN) as well as the 6LU7 X-ray structure. For the X-ray structure, hydrogen atoms were added and the structure relaxed using OpenMM[165] and the AMBER FF14SB force field[170] (as discussed earlier, a real study would require an additional equilibration step). For the neutron structures, the actual crystal position were employed, without external modification. We focus on two types of observables: charges and forces. In Figure 6, we plot those quantities for all three structures. We see for almost all amino acids, the charges are in good agreement between these three data sources, with a difference existing for the 7LB7 structure in residue Glu288, which has a protonation state affected by the binding with the ligand, as discussed in the associated article.[168]

In the same figure, atomic and amino acid net forces are represented. For the forces on the amino acids, we see reasonable agreement here in terms of the range of force values, with the benefit of the neutron structures being more representative of actual room temperature conformations. However, if one focuses on the atomic forces, some hydrogen and sulfur atoms present unusually large force values for the neutron-resolved structures. A closer inspection reveals that the main deviation comes from the hydrogens bonded to sulfur in the cysteine residues, which, for both the neutron structures, appear to have been placed with a bond length that is unusually short. Refining of these issues is something that can be done by collaborations between DFT practitioners and experimentalists. Nonetheless, it is interesting to notice that despite these structure–model incompatibilities, the overall picture at the protein residues level appears to have not been affected, for the observables considered here.

From the discussion and calculations presented above, we will summarize with the following three points: (i) applying DFT to large systems requires processing different kinds of input data than at the small scale; (ii) analysis of this data should be done with the goal of computing coarse grained quantities that are evaluated in light of the many different conformations a system can take on; and (iii) electronic structure locality can ensure that modeling errors of one part of a system do not have an overwhelming effect on the total system, which means that useful quasiobservables

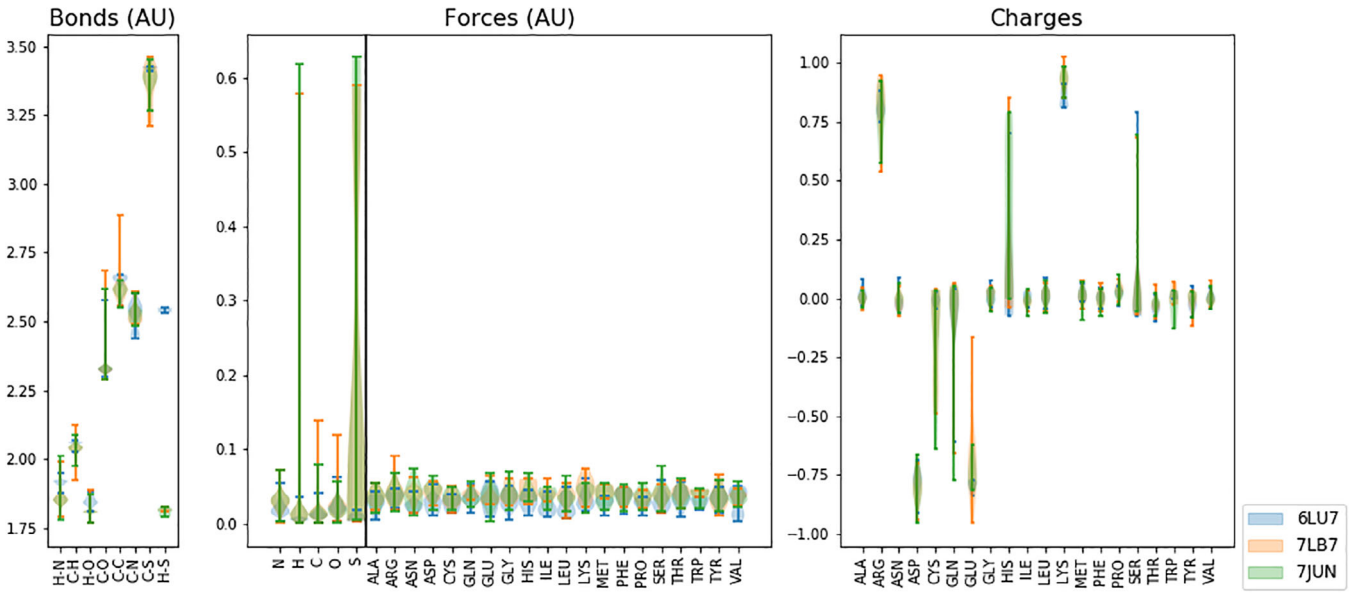

FIGURE 6 Violin plots of the bond lengths, atomic and amino acid forces and charges computed using the X-ray (6LU7) and neutron structures (7LB7 and 7JUN) described in the text

may be extracted from target regions without the need for refining the model with a high level of accuracy in the entire domain.

8 | BRINGING DFT TO OTHER COMMUNITIES

A general problem of atomistic simulations, and one particularly challenging aspect of large scale DFT codes, is that they require expert knowledge (both scientific and technical) to be employed. The codes have been developed by scientists and for scientists within the same community, and aspects like user-friendliness, graphical interfaces, or collaborative tools have not always been a priority. Such points are of paramount importance in order to enable collaborations between different communities. Additionally, performing atomistic simulations requires the use of large HPC resources. Whereas these prerequisites are ubiquitous in academia, they are not a given in industry, and best practices can vary between academic communities. There exist already commercial solutions on the market that target the industrial use of atomistic simulations. They provide proprietary software packages that simplify the usage of DFT codes and make them available for other communities. However, they follow a "traditional" on-premises strategy, where the client has to buy the software and install it on their own computational resources, requiring large investment in both software licenses and hardware.

By taking the most advanced codes for atomistic simulations and lowering their adoption barriers, it will be possible to "democratize" atomistic simulations and to open them up to a much broader community. To this end, solutions like Software-as-a-Service (SaaS) platforms represent an interesting opportunity to accelerate research of systems at the nanoscale. SaaS combines the most advanced simulation codes, predefined workflows, easy-to-use high-level tools, an intuitive and collaborative user interface, and flexible computing resources in the cloud to simplify the use of atomistic simulations.

Jupyter notebooks are one such tool which have been trending in recent years, as they allow for the writing of reproducible scientific workflows and gather in the same place preprocessing, running calculations, and post-processing/analysis of the results. Lots of codes from various scientific communities provide tools to interact with such notebooks. For example, in a recent summary of a special edition of the Journal of Chemical Physics on electronic structure software, native python interfaces were found to be present in: ABINIT, BERTHA, BigDFT, Dalton Project, Molpro, MPQC, NWChem, OpenMolcas, Psi4, PySCF, QMCPACK, Quantum ESPRESSO, Siesta, TeraChem, TheoDORE, and TurboRVB.[171] Introspection capacities make it also possible for notebooks to interact with specific subsets of the application, and create advanced workflows in higher level languages, without the need to interact with the

low level code. Such a separation of concerns makes it possible to see the codes as a service, providing a client part, python-based, for example, and a server part, running said computations on local or HPC resources. These kind of high level tools are increasingly a necessity as the scale of the experiments goes up to help the end users keep track of everything needed to run, understand, and reproduce experiments.

8.1 | Managing complicated workflows

HPC workflows have specific needs that make them hard to use from a single notebook engine. The experiments are not run locally but on distant supercomputers and generate large amounts of data that sometimes cannot be managed locally. In recent years, several tools have been developed to help reduce the complexity of these workflows and handle interactions with HPC resources in a code-agnostic manner, significantly reducing the amount of user tasks, as shown in Figure 7. Examples of such popular python-based tools are Parsl,[172] Signac,[173] or GC3Pie.[174] In the domain of materials science, several more focused tools have also emerged to provide common utilities to interact with the domain databases or file formats (ASE,[175] AiiDA,[176] or AFLOW[177] for example).

To illustrate such capabilities, we will discuss the example of Automated Interactive Infrastructure and Database for Computational Science (AiiDA), first developed by the Swiss Marvel NCCR and also part of the MaX CoE. Its engine allows one to build and generate workflows involving sometimes several codes and subworkflows. Developers of these codes or AiiDA developers provide and maintain plugins to interact with many materials science codes and utilities, handle their inputs and outputs, and provide representative workflows for their usage. The user can then decide to submit the computation blocks to local or remote HPC systems, as the engine can interact with most existing schedulers, generate submission scripts, submit them, monitor the execution, and retrieve the results the user wants to analyze.

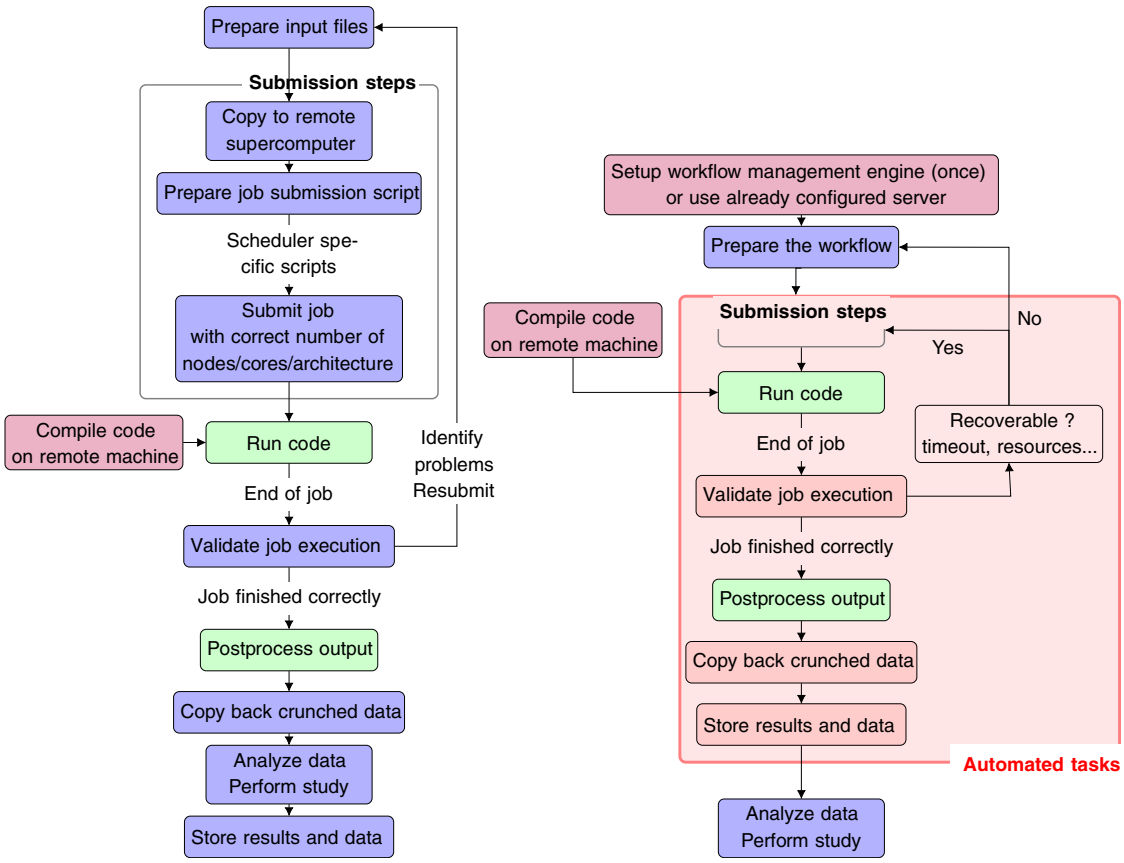

FIGURE 7 Depiction of a typical HPC workflow on the left, and how a workflow-management engine helps on the right. Blue nodes are user tasks, maroon ones are setup tasks that can be performed either by an administrator (code developers or cluster administrators, or the user themselves if they have the knowledge), green ones are computationally intense kernels run on the remote machine, and red ones are the tasks handled automatically by the workflow management engine

Everything from the inputs to the results is stored in a database, allowing safekeeping and querying for later use, as well as sharing the data with other users or caching to avoid running computations for which results are already known. Managing and keeping the provenance of each input or output is one of the main focus points of this particular workflow management system, to ensure the reproducibility of experiments.

Another effort from the AiiDA community is called AiiDAlab, which is a graphical user interface for AiiDA applications, running on a Jupyter server. The goal of this interface is to provide common graphical blocks for preparing and launching computations, for instance by getting structures directly from several materials science online databases from a user friendly interface. Most of the parameters for each code can then be selected from a dedicated graphical "app," installed through an application store, and then launched on a HPC platform through AiiDA. This would make DFT available for users without even the need to learn a high level language or write a notebook.

In this review, we have presented several ways in which calculations of large systems require new research paradigms and workflows. Along with these changes come increasingly complex preprocessing, large computational resource requirements, increases in the amount of data generated, and more computationally demanding post-processing operations. Fortunately, there exist a number of emerging tools, including SaaS, Jupyter notebooks, and workload managers like AiiDA, which can cover these new requirements and will be sure to enable productive scientific research using large scale DFT in the coming years.

9 | LARGE SCALE DFT CODES

Before concluding, we briefly outline some of the codes which were employed in the studies presented in the review. The purpose is not to give a detailed comparison of performance or capabilities, but to highlight their key selling-points and how they are differentiated.

CONQUEST

CONQUEST,[178] which is released under the MIT license, has been developed since 2002. Its formalism enables the user to choose the level of flexibility of the degrees of freedom employed, ranging from fixed localized degrees of freedom (blip functions or pseudoatomic orbitals) up to in-situ optimized basis functions. The code is capable of treating systems of hundreds of thousands atoms as well as first-principles molecular dynamics simulations.

CP2K

CP2K[179] is a GPL-licensed software package, which has been developed since 1999. The computational implementation is based on Gaussian basis functions combined with plane-waves. Code developments have been motivated by first-principles molecular dynamics simulations within a massively parallel and linear scaling approach.

GAMESS

GAMESS (US)[180] is a free, source-available quantum chemistry code developed since 1977. It is based on Gaussian basis functions and implements a large variety of ab-initio methods. For computing large systems, GAMESS includes implementations the Divide-and-Conquer approach as well as a variety of fragment methods, including the FMO method.

ONETEP

ONETEP[181] is released under various licenses, and has been developed since 2005. The approach is based on nonorthogonal generalized Wannier functions (NGWFs) expressed in terms of periodic cardinal sine functions, equivalent to a basis of plane-waves, with the advantage of being localized and, therefore, prone to linear scaling calculations.

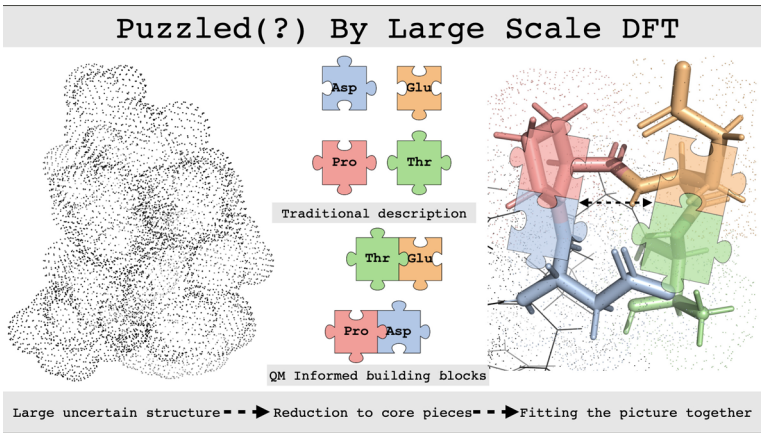

FIGURE 8 The challenge of applying DFT to large systems. The systems being studied are large and complex, with significant uncertainty about their actual conformation. By using DFT, we can create coarse grained descriptions that describe the interactions of large fragments in order to gain insight

In this regard, ONETEP provides a link between the formalisms of the solid-state community, where plane-waves are the basis set of choice, and calculations typical of quantum chemistry, which are based on localized atomic degrees of freedom.

SIESTA

SIESTA[182] is a GPL code, developed since 2000. SIESTA's efficiency stems from the use of a basis set of strictly-localized atomic orbitals. An important feature of the code is that its accuracy and cost can be tuned across a wide range, from quick exploratory calculations to highly accurate simulations.

BigDFT

BigDFT[13] is a GPL code, developed since 2008. The properties of BigDFT come from its flexible, systematic, and precise basis set: Daubechies wavelets. The code provides both cubic-scaling and linear-scaling approaches, and very large systems containing thousands of atoms can be simulated within a systematically improvable description. The authors of this review article are active developers of this code.

Other codes which provide similar features are available in the community and can be employed to perform studies which are similar to the ones presented here. This includes publicly available, linear scaling KS DFT codes like: ErgoSCF,[183] FHI-AIMS,[184] HONPAS,[185] LSDalton,[156,186] OpenMX,[187] SQDFT,[188] and FMO codes such as: ABINIT-MP[189] and PAICS.[190]

10 | CONCLUSION

When employing standard algorithms for performing DFT calculations, the computational cost grows with the third power of the system size. Thus even a one thousand fold increase in computational power would only enable the treatment of systems ten times larger. With this in mind, it is easy to see that the typical sizes of systems treated by DFT might remain relatively stable over the years, with research methodologies maturing around a common length scale. However, as linear scaling methods for DFT become more and more widely used, we can expect that the size of systems a DFT practitioner will be able to treat will change rapidly. This is both a blessing and a curse: while new problems can be solved, DFT experts will have to adapt to understand the unique challenges at each length scale. It is thus essential to establish research practices that can be transferred to larger and larger systems, and can quickly incorporate collaborations with new fields.

In this review, we have suggested a framework for thinking about large scale DFT, as illustrated in Figure 8. First, additional attention must be paid to the preparation of a "computational sample," which adequately handles the uncertainty about the configuration of a system in its natural environment. This is particularly evident in biological applications, where structures obtained from experiment cannot simply be used as-is. Second, DFT should be used as a guide for creating localized descriptions of a system, for which "quasiobservables" that lead to intuitive understanding can be derived. In this framework, there is a close interplay between: (a) locality leading to more efficient algorithms and (b) locality leading to clear chemical insights about increasingly complex systems. This kind of approach has been understood well by pioneers of the fields of linear scaling DFT and the FMO method, and can be seen in practice by the articles we have described here.

We have seen that DFT can be used in many ways at large scale, which enables new opportunities for researchers. One can use energy and forces of DFT to assess the quality of force fields in unknown situations, or employ DFT description as embedding environments for higher QM level of theories. Insights about interaction networks between systems' fragments can be described and inspected, and a system can be decomposed in to QM-consistent building blocks. A physico-chemical observable can be reinterpreted in terms of fragment observables that acquire a coherent meaning, especially when combined with consistency indicators. In addition to that, DFT can be employed with a "historical" mindset in larger systems, both for ground-state quantities and as a framework for electronic and optical excited states.

In the discussion presented here, we have largely ignored important questions about the choice of density functional, basis sets, solvation models, etc. Clearly, these considerations should not be neglected when treating systems at a large scale. However, what we have seen is that even with traditional, well-established DFT functionals (e.g., PBE + empirical dispersion), results coming from linear scaling codes reveal useful information, especially for coarse grained quantities and trends. Furthermore, we have seen that a true evaluation of these choices should be done in the context of structural uncertainty and with the goal of reliable quasiobservables, a validation framework that differs from DFT at the small scale. In the next few years, it will be very interesting to compare results at various levels of theory to understand the impact of such points on the overall outcome of the investigations.

In the coming decades, new computational techniques such as machine learning and the use of quantum computers are likely to have a significant impact on the state of practice of the QM modeling community. Joint workflows will be especially appealing for large systems. However, if these workflows only improve energies and forces, it will be a missed opportunity. Much like how DFT practitioners have struggled with the "inverse problem" for decades, machine learning experts struggle with the "black box" of artificial intelligence. Even in that field, though, considerations about computational performance can lead to new insights, such as a recent study using deep neural networks for which hyperparameter tuning revealed insight into the intrinsic length scales of materials.[191] Thus even if the underlying modeling changes, we expect that the issues discussed here about the interplay between fragments, observables, and computational complexity will continue to play a role in the study of large systems.

ACKNOWLEDGMENTS

We thank Viviana Cristiglio, Michel Masella, Bun Chan, Marco Zaccaria, and Babak Momeni for useful discussions and comments. Luigi Genovese, Takahito Nakajima, and William Dawson gratefully acknowledge the joint CEA-RIKEN collaboration action. Augustin Degomme and Luigi Genovese also acknowledge support from the MaX Center of Excellence. This work used computational resources of the supercomputer Fugaku provided by RIKEN through the HPCI System Research Project (Project ID: hp200179). Some calculations were performed using the Hokusai supercomputer system at RIKEN. LER and MS acknowledge support from an EPSRC Early Career Research Fellowship (EP/P033253/1) and the Thomas Young Centre under grant number TYC-101. 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.

CONFLICT OF INTEREST

The authors have declared no conflicting interests.

AUTHOR CONTRIBUTIONS

William Dawson: Conceptualization (equal); data curation (equal); methodology (equal); project administration (equal); resources (equal); supervision (equal); validation (equal); writing – original draft (equal); writing – review and

editing (equal). **Augustin Degomme:** Data curation (equal); investigation (equal); writing – original draft (equal); writing – review and editing (equal). **Martina Stella:** Formal analysis (equal); investigation (equal); resources (equal); validation (equal); writing – original draft (equal); writing – review and editing (equal). **Takahito Nakajima:** Supervision (equal); validation (equal); writing – review and editing (equal). **Laura Ratcliff:** Conceptualization (equal); data curation (equal); methodology (equal); project administration (equal); resources (equal); supervision (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Luigi Genovese:** Conceptualization (equal); data curation (equal); methodology (equal); project administration (equal); resources (equal); supervision (equal); validation (equal); writing – original draft (equal); writing – review and editing (equal).

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID


Luigi Genovese https://orcid.org/0000-0003-1747-0247

RELATED WIREs ARTICLE

Challenges in large scale quantum mechanical calculations

REFERENCES

1. Hohenberg P, Kohn W. Inhomogeneous electron gas. Phys Rev. 1964;136:B864–71. https://doi.org/10.1103/PhysRev.136.B864
2. Kohn W, Sham LJ. Self-consistent equations including exchange and correlation effects. Phys Rev. 1965;140:A1133–8. https://doi.org/10.1103/PhysRev.140.A1133
3. Bowler DR, Miyazaki T. O(N) methods in electronic structure calculations. Rep Prog Phys. 2012;75(3):036503. https://doi.org/10.1088/0034-4885/75/3/036503
4. Ratcliff LE, 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(1):e1290. https://doi.org/10.1002/wcms.1290
5. Kohn W. Density functional and density matrix method scaling linearly with the number of atoms. Phys Rev Lett. 1996;76:3168–71. https://doi.org/10.1103/PhysRevLett.76.3168
6. Damle A, Lin L, Ying L. Compressed representation of kohn–sham orbitals via selected columns of the density matrix. J Chem Theory Comput. 2015;11(4):1463–9. https://doi.org/10.1021/ct500985f
7. Aquilante F, Pedersen TB, de Merás AS, Koch H. Fast noniterative orbital localization for large molecules. J Chem Phys. 2006;125(17):174101. https://doi.org/10.1063/1.2360264
8. Marzari N, Vanderbilt D. Maximally localized generalized wannier functions for composite energy bands. Phys Rev B. 1997;56:12847–65. https://doi.org/10.1103/PhysRevB.56.12847
9. Boys SF. Construction of some molecular orbitals to be approximately invariant for changes from one molecule to another. Rev Mod Phys. 1960;32:296–9. https://doi.org/10.1103/RevModPhys.32.296
10. Gygi F. Compact representations of kohn-sham invariant subspaces. Phys Rev Lett. 2009;102:166406. https://doi.org/10.1103/PhysRevLett.102.166406
11. Stewart GW. Computing thecs decomposition of a partitioned orthonormal matrix. Numer Math. 1982;40(3):297–306.
12. Perdew JP, Burke K, Ernzerhof M. Generalized gradient approximation made simple. Phys Rev Lett. 1996;77:3865–8. https://doi.org/10.1103/PhysRevLett.77.3865
13. Ratcliff LE, Dawson W, Fisicaro G, Caliste D, Mohr S, Degomme A, et al. Flexibilities of wavelets as a computational basis set for large-scale electronic structure calculations. J Chem Phys. 2020;152(19):194110. https://doi.org/10.1063/5.0004792
14. Gygi F. Architecture of qbox: a scalable first-principles molecular dynamics code. IBM J Res Dev. 2008;52(1.2):137–44. https://doi.org/10.1147/rd.521.0137
15. Hartwigsen C, Goedecker S, Hutter J. Relativistic separable dual-space gaussian pseudopotentials from h to rn. Phys Rev B. 1998;58:3641–62. https://doi.org/10.1103/PhysRevB.58.3641
16. Willand A, Kvashnin YO, Genovese L, Vázquez-Mayagoitia Á, Deb AK, Sadeghi A, et al. Norm-conserving pseudopotentials with chemical accuracy compared to all-electron calculations. J Chem Phys. 2013;138(10):104109. https://doi.org/10.1063/1.4793260
17. Schlipf M, Gygi F. Optimization algorithm for the generation of oncv pseudopotentials. Comput Phys Commun. 2015;196:36–44. https://doi.org/10.1016/j.cpc
18. Honda S, Yamasaki K, Sawada Y, Morii H. 10 residue folded peptide designed by segment statistics. Structure. 2004;12(8):1507–18.
19. Corsetti F, Mostofi AA. System-size convergence of point defect properties: the case of the silicon vacancy. Phys Rev B. 2011;84:035209. https://doi.org/10.1103/PhysRevB.84.035209
20. Aguado-Puente P, Junquera J. Structural and energetic properties of domains in $pbtio_3$/$srtio_3$ superlattices from first principles. Phys Rev B. 2012;85:184105. https://doi.org/10.1103/PhysRevB.85.184105

21. Zhang W, Thiess A, Zalden P, Zeller R, Dederichs PH, Raty J-Y, et al. Role of vacancies in metal–insulator transitions of crystalline phase-change materials. Nat Mater. 2012;11(11):952–6. https://doi.org/10.1038/nmat3456
22. Tait EW, Ratcliff LE, Payne MC, Haynes PD, Hine NDM. Simulation of electron energy loss spectra of nanomaterials with linear-scaling density functional theory. J Phys: Condens Matter. 2016;28(19):195202. https://doi.org/10.1088/0953-8984/28/19/195202
23. Carnio EG, Hine NDM, Römer RA. Resolution of the exponent puzzle for the Anderson transition in doped semiconductors. Phys Rev B. 2019;99:081201. https://doi.org/10.1103/PhysRevB.99.081201
24. Das S, Motamarri P, Gavini V, Turcksin B, Li YW, Leback B. Fast, scalable and accurate finite-element based ab initio calculations using mixed precision computing: 46 pflops simulation of a metallic dislocation system. In Proceedings of the international conference for high performance computing, networking, storage and analysis, SC '19, Association for Computing Machinery, New York, NY; 2019.
25. Baker JS, Bowler DR. Polar morphologies from first principles: Pbtio3 films on srtio3 substrates and the p(2×λ) surface reconstruction. Adv Theory Simul. 2020;3(11):2000154. https://doi.org/10.1002/adts.202000154
26. Ronneberger I, Zhang W, Eshet H, Mazzarello R. Crystallization properties of the ge2sb2te5 phase-change compound from advanced simulations. Adv Funct Mater. 2015;25(40):6407–13. https://doi.org/10.1002/adfm.201500849
27. Hirata A, Kohara S, Asada T, Arao M, Yogi C, Imai H, et al. Atomic-scale disproportionation in amorphous silicon monoxide. Nat Commun. 2016;7(1):11591. https://doi.org/10.1038/ncomms11591
28. Hu W, Lin L, Yang C, Yang J. Electronic structure and aromaticity of large-scale hexagonal graphene nanoflakes. J Chem Phys. 2014;141(21):214704. https://doi.org/10.1063/1.4902806
29. Hu W, Huang Y, Qin X, Lin L, Kan E, Li X, et al. Room-temperature magnetism and tunable energy gaps in edge-passivated zigzag graphene quantum dots. npj 2D Mater Appl. 2019;3(1):17. https://doi.org/10.1038/s41699-019-0098-2
30. Cun H, Iannuzzi M, Hemmi A, Osterwalder J, Greber T. Two-nanometer voids in single-layer hexagonal boron nitride: formation via the "can-opener" effect and annihilation by self-healing. ACS Nano. 2014;8(7):7423–31. https://doi.org/10.1021/nn502645w
31. Wong D, Corsetti F, Yang W, Brar VW, Tsai H-Z, Wu Q, et al. Spatially resolving density-dependent screening around a single charged atom in graphene. Phys Rev B. 2017;95:205419. https://doi.org/10.1103/PhysRevB.95.205419
32. Corsetti F, Mostofi AA, Lischner J. First-principles multiscale modelling of charged adsorbates on doped graphene. 2D Mater. 2017;4(2):025070. https://doi.org/10.1088/2053-1583/aa6811
33. Constantinescu GC, Hine NDM. Energy landscape and band-structure tuning in realistic mos_2/$mose_2$ heterostructures. Phys Rev B. 2015;91:195416. https://doi.org/10.1103/PhysRevB.91.195416
34. Constantinescu GC, Hine NDM. Multipurpose black-phosphorus/hbn heterostructures. Nano Lett. 2016;16(4):2586–94. https://doi.org/10.1021/acs.nanolett.6b00154
35. Hu W, Lin L, Yang C, Dai J, Yang J. Edge-modified phosphorene nanoflake heterojunctions as highly efficient solar cells. Nano Lett. 2016;16(3):1675–82. https://doi.org/10.1021/acs.nanolett.5b04593
36. Wilson NR, Nguyen PV, Seyler K, Rivera P, Marsden AJ, Laker ZPL, et al. Determination of band offsets, hybridization, and exciton binding in 2d semiconductor heterostructures. Sci Adv. 2017;3(2):1–7. https://doi.org/10.1126/sciadv.1601832
37. Ruiz-Serrano Á, Skylaris C-K. A variational method for density functional theory calculations on metallic systems with thousands of atoms. J Chem Phys. 2013;139(5):054107. https://doi.org/10.1063/1.4817001
38. Bozyigit D, Yazdani N, Yarema M, Yarema O, Lin WMM, Volk S, et al. Soft surfaces of nanomaterials enable strong phonon interactions. Nature. 2016;531(7596):618–22. https://doi.org/10.1038/nature16977
39. Lamiel-Garcia O, Ko KC, Lee JY, Bromley ST, Illas F. When anatase nanoparticles become bulklike: properties of realistic tio2 nanoparticles in the 1–6 nm size range from all electron relativistic density functional theory based calculations. J Chem Theory Comput. 2017;13(4):1785–93. https://doi.org/10.1021/acs.jctc.7b00085
40. Morales-García Á, Valero R, Illas F. Reliable and computationally affordable prediction of the energy gap of (tio2)n ($10 \leq n \leq 563$) nanoparticles from density functional theory. Phys Chem Chem Phys. 2018;20:18907–11. https://doi.org/10.1039/C8CP03582B
41. Ellaby T, Aarons J, Varambhia A, Jones L, Nellist P, Ozkaya D, et al. Ideal versus real: simulated annealing of experimentally derived and geometric platinum nanoparticles. J Phys: Condens Matter. 2018;30(15):155301. https://doi.org/10.1088/1361-648x/aab251
42. Lin L, Larsen AH, Romero NA, Morozov VA, Glinsvad C, Abild-Pedersen F, et al. Investigation of catalytic finite-size-effects of platinum metal clusters. J Phys Chem Lett. 2013;4(1):222–6. https://doi.org/10.1021/jz3018286
43. Voznyy O, Sargent EH. Atomistic model of fluorescence intermittency of colloidal quantum dots. Phys Rev Lett. 2014;112:157401. https://doi.org/10.1103/PhysRevLett.112.157401
44. Wang Y-G, Mei D, Glezakou V-A, Li J, Rousseau R. Dynamic formation of single-atom catalytic active sites on ceria-supported gold nanoparticles. Nat Commun. 2015;6(1):6511. https://doi.org/10.1038/ncomms7511
45. Verga LG, Aarons J, Sarwar M, Thompsett D, Russell AE, Skylaris C-K. Effect of graphene support on large pt nanoparticles. Phys Chem Chem Phys. 2016;18:32713–22. https://doi.org/10.1039/C6CP07334D
46. Aarons J, Jones L, Varambhia A, MacArthur KE, Ozkaya D, Sarwar M, et al. Predicting the oxygen-binding properties of platinum nanoparticle ensembles by combining high-precision electron microscopy and density functional theory. Nano Lett. 2017;17(7):4003–12. https://doi.org/10.1021/acs.nanolett.6b04799
47. Verga LG, Aarons J, Sarwar M, Thompsett D, Russell AE, Skylaris C-K. Dft calculation of oxygen adsorption on platinum nanoparticles: coverage and size effects. Faraday Discuss. 2018;208:497–522. https://doi.org/10.1039/C7FD00218A
48. Heiss M, Fontana Y, Gustafsson A, Wüst G, Magen C, O'Regan DD, et al. Self-assembled quantum dots in a nanowire system for quantum photonics. Nat Mater. 2013;12:439–44. https://doi.org/10.1038/nmat3557

49. Ning Z, Gong X, Comin R, Walters G, Fan F, Voznyy O, et al. Quantum-dot-in-perovskite solids. Nature. 2015;523(7560):324–8. https://doi.org/10.1038/nature14563
50. Janssen JL, Beaudin J, Hine NDM, Haynes PD, Côté M. Bromophenyl functionalization of carbon nanotubes: an ab initio study. Nanotechnology. 2013;24(37):375702. https://doi.org/10.1088/0957-4484/24/37/375702
51. Bell RA, Payne MC, Mostofi AA. Does water dope carbon nanotubes? J Chem Phys. 2014;141(16):164703. https://doi.org/10.1063/1.4898712
52. Bell RA, Dubois SM-M, Payne MC, Mostofi AA. Electronic transport calculations in the onetep code: implementation and applications. Comput Phys Commun. 2015;193:78–88. https://doi.org/10.1016/j.cpc
53. Bouilly D, Janssen JL, Cabana J, Côté M, Martel R. Graft-induced midgap states in functionalized carbon nanotubes. ACS Nano. 2015;9(3):2626–34. https://doi.org/10.1021/nn506297z
54. Poli E, Elliott JD, Ratcliff LE, Andrinopoulos L, Dziedzic J, Hine NDM, et al. The potential of imogolite nanotubes as (co-)photocatalysts: a linear-scaling density functional theory study. J Phys: Condens Matter. 2016;28(7):074003. https://doi.org/10.1088/0953-8984/28/7/074003
55. Elliott JD, Poli E, Scivetti I, Ratcliff LE, Andrinopoulos L, Dziedzic J, et al. Chemically selective alternatives to photoferroelectrics for polarization-enhanced photocatalysis: the untapped potential of hybrid inorganic nanotubes. Adv Sci. 2017;4(2):1600153. https://doi.org/10.1002/advs.201600153
56. O'Rourke C, Mujahed SY, Kumarasinghe C, Miyazaki T, Bowler DR. Structural properties of silicon–germanium and germanium–silicon core–shell nanowires. J Phys: Condens Matter. 2018;30(46):465303. https://doi.org/10.1088/1361-648x/aae617
57. Kumarasinghe C, Bowler DR. DFT study of undoped and as-doped si nanowires approaching the bulk limit. J Phys: Condens Matter. 2019;32(3):035304. https://doi.org/10.1088/1361-648x/ab4b3c
58. Avraam PW, Hine NDM, Tangney P, Haynes PD. Factors influencing the distribution of charge in polar nanocrystals. Phys Rev B. 2011;83:241402. https://doi.org/10.1103/PhysRevB.83.241402
59. Avraam PW, Hine NDM, Tangney P, Haynes PD. Fermi-level pinning can determine polarity in semiconductor nanorods. Phys Rev B. 2012;85:115404. https://doi.org/10.1103/PhysRevB.85.115404
60. Hine NDM, Avraam PW, Tangney P, Haynes PD. Linear-scaling density functional theory simulations of polar semiconductor nanorods. J Phys: Conf Ser. 2012;367:012002. https://doi.org/10.1088/1742-6596/367/1/012002
61. Ratcliff LE, Haynes PD. Ab initio calculations of the optical absorption spectra of c60-conjugated polymer hybrids. Phys Chem Chem Phys. 2013;15:13024–31. https://doi.org/10.1039/C3CP52043A
62. Turban DHP, Teobaldi G, O'Regan DD, Hine NDM. Supercell convergence of charge-transfer energies in pentacene molecular crystals from constrained dft. Phys Rev B. 2016;93:165102. https://doi.org/10.1103/PhysRevB.93.165102
63. Xue H-T, Boschetto G, Krompiec M, Morse GE, Tang F-L, Skylaris C-K. Linear-scaling density functional simulations of the effect of crystallographic structure on the electronic and optical properties of fullerene solvates. Phys Chem Chem Phys. 2017;19:5617–28. https://doi.org/10.1039/C6CP08165G
64. Boschetto G, Xue H-T, Dziedzic J, Krompiec M, Skylaris C-K. Effect of polymerization statistics on the electronic properties of copolymers for organic photovoltaics. J Phys Chem C. 2017;121(5):2529–38. https://doi.org/10.1021/acs.jpcc.6b10851
65. Mondelli P, Boschetto G, Horton PN, Tiwana P, Skylaris C-K, Coles SJ, et al. Meta-analysis: the molecular organization of non-fullerene acceptors. Mater Horiz. 2020;7:1062–72. https://doi.org/10.1039/C9MH01439J
66. 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(3):701–6.
67. Gordon MS, Fedorov DG, Pruitt SR, Slipchenko LV. Fragmentation methods: a route to accurate calculations on large systems. Chem Rev. 2012;112(1):632–72. https://doi.org/10.1021/cr200093j
68. Herbert JM. Fantasy versus reality in fragment-based quantum chemistry. J Chem Phys. 2019;151(17):170901. https://doi.org/10.1063/1.5126216
69. Fedorov DG, Kitaura K. The importance of three-body terms in the fragment molecular orbital method. J Chem Phys. 2004;120(15):6832–40. https://doi.org/10.1063/1.1687334
70. Fedorov DG, Kitaura K. On the accuracy of the 3-body fragment molecular orbital method (fmo) applied to density functional theory. Chem Phys Lett. 2004;389(1):129–34. https://doi.org/10.1016/j.cplett.2004.03.072
71. Nakano T, Mochizuki Y, Yamashita K, Watanabe C, Fukuzawa K, Segawa K, et al. Development of the four-body corrected fragment molecular orbital (fmo4) method. Chem Phys Lett. 2012;523:128–33.
72. Stoll H, Preuß H. On the direct calculation of localized hf orbitals in molecule clusters, layers and solids. Theor Chim Acta. 1977;46(1):11–21. https://doi.org/10.1007/BF00551649
73. Kitaura K, Sugiki S-I, Nakano T, Komeiji Y, Uebayasi M. Fragment molecular orbital method: analytical energy gradients. Chem Phys Lett. 2001;336(1):163–70. https://doi.org/10.1016/S0009-2614(01)00099-9
74. Nagata T, Brorsen K, Fedorov DG, Kitaura K, Gordon MS. Fully analytic energy gradient in the fragment molecular orbital method. J Chem Phys. 2011;134(12):124115. https://doi.org/10.1063/1.3568010
75. Inadomi Y, Nakano T, Kitaura K, Nagashima U. Definition of molecular orbitals in fragment molecular orbital method. Chem Phys Lett. 2002;364(1):139–43. https://doi.org/10.1016/S0009-2614(02)01291-5
76. Tsuneyuki S, Kobori T, Akagi K, Sodeyama K, Terakura K, Fukuyama H. Molecular orbital calculation of biomolecules with fragment molecular orbitals. Chem Phys Lett. 2009;476(1):104–8.

77. Kitaura K, Morokuma K. A new energy decomposition scheme for molecular interactions within the hartree-fock approximation. Int J Quant Chem. 1976;10(2):325–40. https://doi.org/10.1002/qua.560100211
78. Fedorov DG, Kitaura K. Pair interaction energy decomposition analysis. J Comput Chem. 2007;28(1):222–37. https://doi.org/10.1002/jcc.20496
79. Fedorov DG, Kitaura K. Subsystem analysis for the fragment molecular orbital method and its application to protein–ligand binding in solution. J Phys Chem A. 2016;120(14):2218–31. https://doi.org/10.1021/acs.jpca.6b00163
80. Thapa B, Raghavachari K. Energy decomposition analysis of protein–ligand interactions using molecules-in-molecules fragmentation-based method. J Chem Inf Model. 2019;59(8):3474–84. https://doi.org/10.1021/acs.jcim.9b00432
81. Gao J, Truhlar DG, Wang Y, Mazack MJM, Löffler P, Provorse MR, et al. Explicit polarization: a quantum mechanical framework for developing next generation force fields. Acc Chem Res. 2014;47(9):2837–45. https://doi.org/10.1021/ar5002186
82. Szalewicz K. Symmetry-adapted perturbation theory of intermolecular forces. WIREs Comput Mol Sci. 2012;2(2):254–72. https://doi.org/10.1002/wcms.86
83. Lao KU, Herbert JM. Accurate and efficient quantum chemistry calculations for noncovalent interactions in many-body systems: the xsapt family of methods. J Phys Chem A. 2015;119(2):235–52. https://doi.org/10.1021/jp5098603
84. Heifetz A, Chudyk EI, Gleave L, Aldeghi M, Cherezov V, Fedorov DG, et al. The fragment molecular orbital method reveals new insight into the chemical nature of gpcr–ligand interactions. J Chem Inf Model. 2016;56(1):159–72. https://doi.org/10.1021/acs.jcim.5b00644
85. Zhang Q, Yu C, Min J, Wang Y, He J, Yu Z. Rational questing for potential novel inhibitors of fabk from *streptococcus pneumoniae* by combining fmo calculation, comfa 3d-qsar modeling and virtual screening. J Mol Model. 2011;17(6):1483–92. https://doi.org/10.1007/s00894-010-0847-9
86. Li S, Qin C, Cui S, Xu H, Wu F, Wang J, et al. Discovery of a natural-product-derived preclinical candidate for once-weekly treatment of type 2 diabetes. J Med Chem. 2019;62(5):2348–61. https://doi.org/10.1021/acs.jmedchem.8b01491
87. Abe Y, Shoji M, Nishiya Y, Aiba H, Kishimoto T, Kitaura K. The reaction mechanism of sarcosine oxidase elucidated using fmo and qm/mm methods. Phys Chem Chem Phys. 2017;19:9811–22. https://doi.org/10.1039/C6CP08172J
88. Heifetz A, Trani G, Aldeghi M, MacKinnon CH, McEwan PA, Brookfield FA, et al. Fragment molecular orbital method applied to lead optimization of novel interleukin-2 inducible t-cell kinase (itk) inhibitors. J Med Chem. 2016;59(9):4352–63. https://doi.org/10.1021/acs.jmedchem.6b00045
89. Heifetz A, Morao I, Babu MM, James T, Southey MWY, Fedorov DG, et al. Characterizing interhelical interactions of g-protein coupled receptors with the fragment molecular orbital method. J Chem Theory Comput. 2020;16(4):2814–24. https://doi.org/10.1021/acs.jctc.9b01136
90. 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(12):6623–31. https://doi.org/10.1021/acs.jctc.8b00733
91. Estrada E. The structure of complex networks: theory and applications. Oxford: Oxford University Press; 2012.
92. Heifetz A. Quantum mechanics in drug discovery. New York: Springer; 2020.
93. Hatada R, Okuwaki K, Mochizuki Y, Handa Y, Fukuzawa K, Komeiji Y, et al. Fragment molecular orbital based interaction analyses on covid-19 main protease - inhibitor n3 complex (pdb id: 6lu7). J Chem Inf Model. 2020;60(7):3593–602. https://doi.org/10.1021/acs.jcim.0c00283
94. Fedorov DG, Kitaura K. Modeling and visualization for the fragment molecular orbital method with the graphical user interface fu, and analyses of protein–ligand binding. In: Gordon MS, editor. Fragmentation: toward accurate calculations on complex molecular systems. New Jersey: Wiley; 2017. p. 119–40.
95. Watanabe C, Watanabe H, Okiyama Y, Takaya D, Fukuzawa K, Tanaka S, et al. Development of an automated fragment molecular orbital (fmo) calculation protocol toward construction of quantum mechanical calculation database for large biomolecules. Chem-Bio Inform J. 2019;19:5–18.
96. Takaya D, Watanabe C, Nagase S, Kamisaka K, Okiyama Y, Moriwaki H, et al. Fmodb: the world's first database of quantum mechanical calculations for biomacromolecules based on the fragment molecular orbital method. J Chem Inform Model. 2021;61:777–94. https://doi.org/10.1021/acs.jcim.0c01062
97. Otsuka T, Okimoto N, Taiji M, Bowler DR, Miyazaki T. Structural relaxation and binding energy calculations of FK506 binding protein complexes using the large-scale DFT code CONQUEST. J Phys: Conf Ser. 2013;454:012057. https://doi.org/10.1088/1742-6596/454/1/012057
98. Srinivasan J, Cheatham TE, Cieplak P, Kollman PA, Case DA. Continuum solvent studies of the stability of dna, rna, and phosphoramidate-dna helices. J Am Chem Soc. 1998;120(37):9401–9. https://doi.org/10.1021/ja981844
99. Fox SJ, Dziedzic J, Fox T, Tautermann CS, Skylaris C-K. Density functional theory calculations on entire proteins for free energies of binding: application to a model polar binding site. Proteins: Struct Funct Genet. 2014;82(12):3335–46.
100. Cole DJ, Skylaris C-K, Rajendra E, Venkitaraman AR, Payne MC. Protein-protein interactions from linear-scaling first-principles quantum-mechanical calculations. Europhys Lett. 2010;91(3):37004. https://doi.org/10.1209/0295-5075/91/37004
101. Svensson F, Engen K, Lundbäck T, Larhed M, Sköld C. Virtual screening for transition state analogue inhibitors of irap based on quantum mechanically derived reaction coordinates. J Chem Inf Model. 2015;55(9):1984–93. https://doi.org/10.1021/acs.jcim.5b00359
102. Lever G, Cole DJ, Lonsdale R, Ranaghan KE, Wales DJ, Mulholland AJ, et al. Large-scale density functional theory transition state searching in enzymes. J Phys Chem Lett. 2014;5(21):3614–9. https://doi.org/10.1021/jz5018703

103. Weber C, Cole DJ, O'Regan DD, Payne MC. Renormalization of myoglobin–ligand binding energetics by quantum many-body effects. Proc Natl Acad Sci USA. 2014;111(16):5790–5. https://doi.org/10.1073/pnas.1322966111
104. Feliciano GT, da Silva AJR, Reguera G, Artacho E. Molecular and electronic structure of the peptide subunit of *geobacter sulfurreducens* conductive pili from first principles. J Phys Chem A. 2012;116(30):8023–30. https://doi.org/10.1021/jp302232p
105. Reed AE, Curtiss LA, Weinhold F. Intermolecular interactions from a natural bond orbital, donor-acceptor viewpoint. Chem Rev. 1988; 88(6):899–926. https://doi.org/10.1021/cr00088a005
106. Lee LP, Cole DJ, Payne MC, Skylaris C-K. Natural bond orbital analysis in the onetep code: applications to large protein systems. J Comput Chem. 2013;34(6):429–44. https://doi.org/10.1002/jcc.23150
107. Phipps MJS, Fox T, Tautermann CS, Skylaris C-K. Energy decomposition analysis approaches and their evaluation on prototypical protein–drug interaction patterns. Chem Soc Rev. 2015;44:3177–211. https://doi.org/10.1039/C4CS00375F
108. Phipps MJS, Fox T, Tautermann CS, Skylaris C-K. Energy decomposition analysis based on absolutely localized molecular orbitals for large-scale density functional theory calculations in drug design. J Chem Theory Comput. 2016;12(7):3135–48. https://doi.org/10.1021/acs.jctc.6b00272
109. Phipps MJS, Fox T, Tautermann CS, Skylaris C-K. Intuitive density functional theory-based energy decomposition analysis for protein–ligand interactions. J Chem Theory Comput. 2017;13(4):1837–50. https://doi.org/10.1021/acs.jctc.6b01230
110. Khaliullin RZ, Cobar EA, Lochan RC, Bell AT, Head-Gordon M. Unravelling the origin of intermolecular interactions using absolutely localized molecular orbitals. Chem A Eur J. 2007;111(36):8753–65. https://doi.org/10.1021/jp073685z
111. Peifeng S, Li H. Energy decomposition analysis of covalent bonds and intermolecular interactions. J Chem Phys. 2009;131(1):014102. https://doi.org/10.1063/1.3159673
112. Richard F, Bader W. A quantum theory of molecular structure and its applications. Chem Rev. 1991;91(5):893–928. https://doi.org/10.1021/cr00005a013
113. Johnson ER, Keinan S, Mori-Sánchez P, Contreras-García J, Cohen AJ, Yang W. Revealing noncovalent interactions. J Am Chem Soc. 2010;132(18):6498–506. https://doi.org/10.1021/ja100936w
114. Arias-Olivares D, Wieduwilt EK, Contreras-García J, Genoni A. Nci-Elmo: a new method to quickly and accurately detect noncovalent interactions in biosystems. J Chem Theory Comput. 2019;15(11):6456–70. https://doi.org/10.1021/acs.jctc.9b00658
115. Mohr S, Masella M, Ratcliff LE, 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(9):4079–88. https://doi.org/10.1021/acs.jctc.7b00291
116. Dawson W, Mohr S, Ratcliff LE, Nakajima T, Genovese L. Complexity reduction in density functional theory calculations of large systems: system partitioning and fragment embedding. J Chem Theory Comput. 2020;16(5):2952–64. https://doi.org/10.1021/acs.jctc.9b01152
117. Fedorov D, Kitaura K. The fragment molecular orbital method: practical applications to large molecular systems. Cleveland, Ohio: CRC Press; 2009.
118. Gygi F, Duchemin I. Efficient computation of hartree–fock exchange using recursive subspace bisection. J Chem Theory Comput. 2013; 9(1):582–7. https://doi.org/10.1021/ct3007088
119. Dawson W, Gygi F. Optimized scheduling strategies for hybrid density functional theory electronic structure calculations. Proceedings of the international conference for high performance computing, networking, storage and analysis, SC '14. Piscataway, New Jersey: IEEE Press; 2014. p. 685–92.
120. Dawson W, Gygi F. Performance and accuracy of recursive subspace bisection for hybrid dft calculations in inhomogeneous systems. J Chem Theory Comput. 2015;11(10):4655–63. https://doi.org/10.1021/acs.jctc.5b00826
121. Ratcliff LE, Genovese L, Mohr S, Deutsch T. Fragment approach to constrained density functional theory calculations using daubechies wavelets. J Chem Phys. 2015;142(23):234105. https://doi.org/10.1063/1.4922378
122. Ratcliff LE, Grisanti L, Genovese L, Deutsch T, Neumann T, Danilov D, et al. Toward fast and accurate evaluation of charge on-site energies and transfer integrals in supramolecular architectures using linear constrained density functional theory (cdft)-based methods. J Chem Theory Comput. 2015;11(5):2077–86. https://doi.org/10.1021/acs.jctc.5b00057
123. Ratcliff LE, Genovese L. Pseudo-fragment approach for extended systems derived from linear-scaling DFT. J Phys: Condens Matter. 2019;31(28):285901. https://doi.org/10.1088/1361-648x/ab1664
124. Gúnes S, Neugebauer H, Sariciftci NS. Conjugated polymer-based organic solar cells. Chem Rev. 2007;107(4):1324–38. https://doi.org/10.1021/cr050149z
125. Koumura N, Zijlstra RWJ, van Delden RA, Harada N, Feringa BL. Light-driven monodirectional molecular rotor. Nature. 1999; 401(6749):152–5. https://doi.org/10.1038/43646
126. Pawlicki M, Collins HA, Denning RG, Anderson HL. Two-photon absorption and the design of two-photon dyes. Angew Chem Int Ed. 2009;48(18):3244–66. https://doi.org/10.1002/anie.200805257
127. Jacquemin D, Mennucci B, Adamo C. Excited-state calculations with td-dft: from benchmarks to simulations in complex environments. Phys Chem Chem Phys. 2011;13:16987–98. https://doi.org/10.1039/C1CP22144B
128. Mennucci B, Curutchet C. The role of the environment in electronic energy transfer: a molecular modeling perspective. Phys Chem Chem Phys. 2011;13:11538–50. https://doi.org/10.1039/C1CP20601J
129. Runge E, Gross EKU. Density-functional theory for time-dependent systems. Phys Rev Lett. 1984;52:997–1000. https://doi.org/10.1103/PhysRevLett.52.997
130. Rocca D, Gebauer R, Saad Y, Baroni S. Turbo charging time-dependent density-functional theory with lanczos chains. J Chem Phys. 2008;128(15):154105. https://doi.org/10.1063/1.2899649

131. Hübener H, Giustino F. Time-dependent density functional theory using atomic orbitals and the self-consistent sternheimer equation. Phys Rev B. 2014;89:085129. https://doi.org/10.1103/PhysRevB.89
132. O'Rourke C, Bowler DR. Linear scaling density matrix real time tddft: propagator unitarity and matrix truncation. J Chem Phys. 2015; 143(10):102801. https://doi.org/10.1063/1.4919128
133. Zuehlsdorff TJ, Hine NDM, Payne MC, Haynes PD. Linear-scaling time-dependent density-functional theory beyond the tamm-dancoff approximation: Obtaining efficiency and accuracy with in situ optimised local orbitals. J Chem Phys. 2015;143:204107. https://doi.org/10.1063/1.4936280
134. Zuehlsdorff TJ, Hine NDM, Spencer JS, Harrison NM, Riley DJ, Haynes PD. Linear-scaling time-dependent density-functional theory in the linear response formalism. J Chem Phys. 2013;139(6):064104. https://doi.org/10.1063/1.4817330
135. Corsini NRC, Hine NDM, Haynes PD, Molteni C. Unravelling the roles of size, ligands, and pressure in the piezochromic properties of cds nanocrystals. Nano Lett. 2017;17(2):1042–8. https://doi.org/10.1021/acs.nanolett.6b04461
136. Cole DJ, Chin AW, Hine NDM, Haynes PD, Payne MC. Toward ab initio optical spectroscopy of the Fenna-Matthews-Olson complex. J Phys Chem Lett. 2013;4(24):4206–12. https://doi.org/10.1021/jz402000c
137. Zuehlsdorff TJ, Haynes PD, Hanke F, Payne MC, Hine NDM. Solvent effects on electronic excitations of an organic chromophore. J Chem Theory Comput. 2016;12(4):1853–61. https://doi.org/10.1021/acs.jctc.5b01014
138. Zuehlsdorff TJ, Haynes PD, Payne MC, Hine NDM. Predicting solvatochromic shifts and colours of a solvated organic dye: the example of Nile red. J Chem Phys. 2017;146(12):124504. https://doi.org/10.1063/1.4979196
139. Zuehlsdorff TJ. Computing the optical properties of large systems. Switzerland: Springer International Publishing; 2015.
140. Charlton RJ, Fogarty RM, Bogatko S, Zuehlsdorff TJ, Hine NDM, Heeney M, et al. Implicit and explicit host effects on excitons in pentacene derivatives. J Chem Phys. 2018;148(10):104108. https://doi.org/10.1063/1.5017285
141. Koval P, Barbry M, Sánchez-Portal D. Pyscf-nao: an efficient and flexible implementation of linear response time-dependent density functional theory with numerical atomic orbitals. Comput Phys Commun. 2019;236:188–204.
142. Barbry M, Koval P, Marchesin F, Esteban R, Borisov AG, Aizpurua J, et al. Atomistic near-field nanoplasmonics: reaching atomic-scale resolution in nanooptics. Nano Lett. 2015;15(5):3410–9. https://doi.org/10.1021/acs.nanolett.5b00759
143. Marchesin F, Koval P, Barbry M, Aizpurua J, Sánchez-Portal D. Plasmonic response of metallic nanojunctions driven by single atom motion: quantum transport revealed in optics. ACS Photon. 2016;3(2):269–77. https://doi.org/10.1021/acsphotonics.5b00609
144. Aryasetiawan F, Gunnarsson O. The GW method. Rep Prog Phys. 1998;61(3):237–312. https://doi.org/10.1088/0034-4885/61/3/002
145. Hüser F, Olsen T, Thygesen KS. Quasiparticle gw calculations for solids, molecules, and two-dimensional materials. Phys Rev B. 2013; 87:235132. https://doi.org/10.1103/PhysRevB.87.235132
146. Golze D, Dvorak M, Rinke P. The gw compendium: a practical guide to theoretical photoemission spectroscopy. Front Chem. 2019;7: 377. https://doi.org/10.3389/fchem.2019.00377
147. van Setten MJ, Caruso F, Sharifzadeh S, Ren X, Scheffler M, Liu F, et al. Gw100: benchmarking g0w0 for molecular systems. J Chem Theory Comput. 2015;11(12):5665–87. https://doi.org/10.1021/acs.jctc.5b00453
148. Govoni M, Galli G. Large scale gw calculations. J Chem Theory Comput. 2015;11(6):2680–96. https://doi.org/10.1021/ct500958p
149. Blase X, Duchemin I, Jacquemin D, Loos P-F. The bethe-salpeter equation formalism: from physics to chemistry. J Phys Chem Lett. 2020;11(17):7371–82. https://doi.org/10.1021/acs.jpclett.0c01875
150. Duchemin I, Blase X. Cubic-scaling all-electron gw calculations with a separable density-fitting space-time approach. J Chem Theory Comput. 2021;(4):2383–2393. https://doi.org/10.1021/acs.jctc.1c00101
151. Magalhães RP, Fernandes HS, Sousa SF. Modelling enzymatic mechanisms with qm/mm approaches: current status and future challenges. Israel J Chem. 2020;60(7):655–66. https://doi.org/10.1002/ijch.202000014
152. Sun Q, Chan GK-L. Quantum embedding theories. Acc Chem Res. 2016;49(12):2705–12. https://doi.org/10.1021/acs.accounts.6b00356
153. Sneskov K, Schwabe T, Christiansen O, Kongsted J. Scrutinizing the effects of polarization in qm/mm excited state calculations. Phys Chem Chem Phys. 2011;13:18551–60. https://doi.org/10.1039/C1CP22067E
154. Kaliakin DS, Nakata H, Kim Y, Chen Q, Fedorov DG, Slipchenko LV. Fmoxfmo: elucidating excitonic interactions in the Fenna–Matthews–Olson complex with the fragment molecular orbital method. J Chem Theory Comput. 2020;16(2):1175–87. https://doi.org/10.1021/acs.jctc.9b00621
155. Plasser F. Theodore: a toolbox for a detailed and automated analysis of electronic excited state computations. J Chem Phys. 2020; 152(8):084108. https://doi.org/10.1063/1.5143076
156. Olsen JMH, Reine S, Vahtras O, Kjellgren E, Reinholdt P, Dundas KOH, et al. Dalton project: a python platform for molecular- and electronic-structure simulations of complex systems. J Chem Phys. 2020;152(21):214115. https://doi.org/10.1063/1.5144298
157. Olsen JM, Aidas K, Kongsted J. Excited states in solution through polarizable embedding. J Chem Theory Comput. 2010;6(12):3721–34. https://doi.org/10.1021/ct1003803
158. Sena AMP, Miyazaki T, Bowler DR. Linear scaling constrained density functional theory in conquest. J Chem Theory Comput. 2011; 7(4):884–9. https://doi.org/10.1021/ct100601n
159. Hoogerheide DP, Forsyth VT, Brown KA. Neutron scattering for structural biology. Phys Today. 2020;73(6):36–42. https://doi.org/10.1063/PT.3.4498
160. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res. 2000;28(1):235–42. https://doi.org/10.1093/nar/28.1.235

161. Zaccai NR, Coquelle N. Opportunities and challenges in neutron crystallography. EPJ Web Conf. 2020;236:02001. https://doi.org/10.1051/epjconf/202023602001
162. Zaccaria M, Dawson W, Cristiglio V, Reverberi M, Ratcliff LE, Nakajima T, et al. Designing a bioremediator: mechanistic models guide cellular and molecular specialization. Curr Opin Biotechnol. 2020;62:98–105.
163. Hoffman RL, Kania RS, Brothers MA, Davies JF, Ferre RA, Gajiwala KS, et al. Discovery of ketone-based covalent inhibitors of coronavirus 3cl proteases for the potential therapeutic treatment of covid-19. J Med Chem. 2020;63(21):12725–47. https://doi.org/10.1021/acs.jmedchem.0c01063
164. Jo S, Kim T, Iyer VG, Im W. Charmm-gui: a web-based graphical user interface for charmm. J Comput Chem. 2008;29(11):1859–65.
165. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, et al. Openmm 7: rapid development of high performance algorithms for molecular dynamics. PLOS Comput Biol. 2017;13(7):1–17. https://doi.org/10.1371/journal.pcbi.1005659
166. Salomon-Ferrer R, Case DA, Walker RC. An overview of the amber biomolecular simulation package. WIREs Comput Mol Sci. 2013;3(2):198–210.
167. Coronavirus Structural Task Force. https://insidecorona.net/. Accessed April 1, 2021.
168. Kneller DW, Phillips G, Weiss KL, Zhang Q, Coates L, Kovalevsky A. Direct observation of protonation state modulation in sars-cov-2 main protease upon inhibitor binding with neutron crystallography. J Med Chem. 2021;64:4991–5000. https://doi.org/10.1021/acs.jmedchem.1c00058
169. Kneller DW, Phillips G, Weiss KL, Pant S, Zhang Q, O'Neill HM, et al. Unusual zwitterionic catalytic site of sars-cov-2 main protease revealed by neutron crystallography. J Biol Chem. 2020;295(50):17365–17373. https://doi.org/10.1074/jbc.AC120.016154
170. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. J Chem Theory Comput. 2015;11(8):3696–713. https://doi.org/10.1021/acs.jctc.5b00255
171. Sherrill CD, Manolopoulos DE, Martínez TJ, Michaelides A. Electronic structure software. J Chem Phys. 2020;153(7):070401. https://doi.org/10.1063/5.0023185
172. Babuji Y, Woodard A, Li Z, Katz DS, Clifford B, Kumar R, et al. Parsl: pervasive parallel programming in python. In Proceedings of the 28th international symposium on high-performance parallel and distributed computing, HPDC '19, pp. 25–36. Association for Computing Machinery, New York, NY; 2019.
173. Adorf CS, Dodd PM, Ramasubramani V, Glotzer SC. Simple data and workflow management with the signac framework. Comput Mater Sci. 2018;146:220–9. https://doi.org/10.1016/j.commatsci.2018.01.035
174. Maffioletti S, Murri R. GC3Pie: a python framework for high-throughput computing. PoS. 2012;162(4):1–6. https://doi.org/10.22323/1.1z62.0143
175. Larsen AH, Mortensen JJ, Blomqvist J, Castelli IE, Christensen R, Dułak M, et al. The atomic simulation environment—a python library for working with atoms. J Phys: Condens Matter. 2017;29(27):273002. https://doi.org/10.1088/1361-648x/aa680e
176. Huber SP, Zoupanos S, Uhrin M, Talirz L, Kahle L, Häuselmann R, et al. Aiida 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. Sci Data. 2020;7(1):1–18. https://doi.org/10.1038/s41597-020-00638-4
177. Curtarolo S, Setyawan W, Hart GLW, Jahnatek M, Chepulskii RV, Taylor RH, et al. Aflow: an automatic framework for high-throughput materials discovery. Comput Mater Sci. 2012;58:218–26.
178. Nakata A, Baker JS, Mujahed SY, Poulton JTL, Arapan S, Lin J, et al. Large scale and linear scaling dft with the conquest code. J Chem Phys. 2020;152(16):164112. https://doi.org/10.1063/5.0005074
179. Kühne TD, Iannuzzi M, Del Ben M, Rybkin VV, Seewald P, Stein F, et al. Cp2k: an electronic structure and molecular dynamics software package - quickstep: efficient and accurate electronic structure calculations. J Chem Phys. 2020;152(19):194103. https://doi.org/10.1063/5.0007045
180. Barca GMJ, Bertoni C, Carrington L, Datta D, De Silva N, Deustua JE, et al. Recent developments in the general atomic and molecular electronic structure system. J Chem Phys. 2020;152(15):154102. https://doi.org/10.1063/5.0005188
181. Prentice JCA, Aarons J, Womack JC, Allen AEA, Andrinopoulos L, Anton L, et al. The onetep linear-scaling density functional theory program. J Chem Phys. 2020;152(17):174111. https://doi.org/10.1063/5.0004445
182. García A, Papior N, Akhtar A, Artacho E, Blum V, Bosoni E, et al. Siesta: recent developments and applications. J Chem Phys. 2020;152(20):204108. https://doi.org/10.1063/5.0005077
183. Rudberg E, Rubensson EH, Sałek P, Kruchinina A. Ergo: an open-source program for linear-scaling electronic structure calculations. SoftwareX. 2018;7:107–11. https://doi.org/10.1016/j.softx.2018.03.005
184. Blum V, Gehrke R, Hanke F, Havu P, Havu V, Ren X, et al. Ab initio molecular simulations with numeric atom-centered orbitals. Comput Phys Commun. 2009;180(11):2175–96.
185. Qin X, Shang H, Xiang H, Li Z, Yang J. Honpas: a linear scaling open-source solution for large system simulations. Int J Quant Chem. 2015;115(10):647–55. https://doi.org/10.1002/qua.24837
186. Aidas K, Angeli C, Bak KL, Bakken V, Bast R, Boman L, et al. The Dalton quantum chemistry program system. WIREs Comput Mol Sci. 2014;4(3):269–84. https://doi.org/10.1002/wcms.1172
187. Ozaki T. $O(n)$ krylov-subspace method for large-scale ab initio electronic structure calculations. Phys Rev B. 2006;74:245101.
188. Suryanarayana P, Pratapa PP, Sharma A, Pask JE. Sqdft: spectral quadrature method for large-scale parallel o(n) Kohn–Sham calculations at high temperature. Comput Phys Commun. 2018;224:288–98.

189. Nakano T, Mochizuki Y, Fukuzawa K, Amari S, Tanaka S. Chapter 2 - developments and applications of abinit-mp software based on the fragment molecular orbital method. In: Starikov EB, Lewis JP, Tanaka S, editors. Modern methods for theoretical physical chemistry of biopolymers. Amsterdam: Elsevier Science; 2006. p. 39–52.
190. Ishikawa T. PAICS: development of an open-source software of fragment molecular orbital method for biomolecule. Singapore: Springer Singapore; 2021.p. 69–76.
191. Mills K, Ryczko K, Luchak I, Domurad A, Beeler C, Tamblyn I. Extensive deep neural networks for transferring small scale learning to large scale systems. Chem Sci. 2019;10:4129–40. https://doi.org/10.1039/C8SC04578J

How to cite this article: Dawson W, Degomme A, Stella M, Nakajima T, Ratcliff LE, Genovese L. Density functional theory calculations of large systems: Interplay between fragments, observables, and computational complexity. WIREs Comput Mol Sci. 2022;12:e1574. https://doi.org/10.1002/wcms.1574