Applications of Large-Scale Density Functional Theory in Biology

Daniel J. Cole[1,2] **and Nicholas D. M. Hine**[3]

[1] Theory of Condensed Matter group, Cavendish Laboratory, 19 JJ Thomson Ave, Cambridge CB3 0HE, U.K.

[2] School of Chemistry, Newcastle University, Newcastle upon Tyne NE1 7RU, U.K.

[3] Department of Physics, University of Warwick, Coventry CV4 7AL, U.K.

E-mail: `daniel.cole@ncl.ac.uk`

Abstract. Density functional theory (DFT) has become a routine tool for the computation of electronic structure in the physics, materials and chemistry fields. Yet the application of traditional DFT to problems in the biological sciences is hindered, to a large extent, by the unfavourable scaling of the computational effort with system size. Here, we review some of the major software and functionality advances that enable insightful electronic structure calculations to be performed on systems comprising many thousands of atoms. We describe some of the early applications of large-scale DFT to the computation of the electronic properties and structure of biomolecules, as well as to paradigmatic problems in enzymology, metalloproteins, photosynthesis and computer-aided drug design. With this review, we hope to demonstrate that first principles modelling of biological structure–function relationships are approaching a reality.

Submitted to: *J. Phys.: Condens. Matter*

1. Introduction

An important goal of molecular biology is an understanding of the relationship between structure and function. Often, it is not only the overall shape of the biological molecule (such as the arrangement of α-helices and β-sheets in proteins) and the identity of its building blocks (such as the sequence of nucleobases in DNA), but also its underlying *electronic structure* that determines the function of the molecule. Quantum mechanical (QM) simulations have played a crucial role in determining the electronic properties of the amino acids, nucleic acids, carbohydrates, lipids and so forth that go into building a biological organism. Yet, if first principles calculations are to become a truly predictive tool in the biological sciences, then we require the ability to study the properties of these molecules within their natural environment. For example, the bond breaking/forming reactions in enzymes, the absorption of light by optically-active pigment molecules, and the transport of electrons in DNA and proteins may all be studied using simplified models in vacuum. However, we will argue that it is only by studying the complex effects of the biological environment that the functional mechanism of the system may be accurately quantified from first principles.

In this review, we begin by motivating the need to construct realistic, large-scale QM models when studying biological systems. In particular, we find, by collecting together literature spanning a wide range of biological fields, that QM system sizes in excess of around 500 atoms are required for the accurate determination of many properties of interest. Therefore, to focus our review, we limit ourselves to the literature describing biomolecular simulations with a minimum of 500 atoms treated at the QM level of theory. This size regime is typically at the limit of that which can be studied using traditional QM calculations, which are at least three orders of magnitude more expensive than molecular mechanics (MM) force fields and which typically scale as the third (or greater) power of the number of atoms in the simulation. However, over the last two decades, the ability of density functional theory (DFT) to treat large system sizes has advanced dramatically, particularly with the advent of linear-scaling methods. When combined with increasing computational processing power, this enables the almost routine application of quantum mechanics to the description of the physics and chemistry of complex biological molecules.

There have been excellent general reviews of both DFT [1] and linear-scaling DFT [2, 3] in recent years. We will not attempt to re-visit the material covered in these previous works, but rather focus on the specific challenges associated with modelling biomolecules, summarise the recent successes, and assess how close we are to routine first-principles descriptions of biological processes. We will therefore review relevant methodology such as local orbital methods, linear-scaling algorithms, embedding, electrostatics, spectroscopy and the advanced treatment of dispersion and correlated electronic effects. Finally, in section 4, we will cover some of the main applications of large-scale DFT to the study of biological function, including the prediction of biomolecular structure and electronic properties, computational enzymology, metalloproteins, photosynthesis and computer-aided drug design. In general, the chosen applications are paradigmatic problems that have been widely studied, and are well-characterised through experimental measurements. However, a vast range of biological problems are much less straightforward to access experimentally. It is therefore an opportune time to reflect and consider if we are ready for truly predictive first principles modelling that can explain the functional interactions and chemistry of complex biological molecules.

1.1. Necessity for Large-Scale QM Simulations

Despite the huge complexity of biological organisms, many phenomena can be studied in a relatively localised region, such as an active site of an enzyme, or a binding pocket, or a chromophore. In these cases, quantum mechanical accuracy in the description of for example bond-breaking, transition metal chemistry, intermolecular binding or electronic excitations is both appealing and computationally feasible. Assuming that one requires a QM treatment, then the choice is between a QM cluster approach and a hybrid QM/MM simulation, in which a QM subsystem is embedded inside a larger region treated at a less expensive molecular mechanics level of theory [4]. Ideally the two methods should converge for sufficiently large QM regions. The advantage of the former approach is simplicity – there is no need to assign an MM force field or to model the boundary between the QM and the MM regions. The advantage of the latter is that, in theory, properties of interest should converge with a smaller QM region. Although the MM region does not explicitly account for electronic polarisation, this may be accounted for in an approximate manner by embedding the system in a polarisable continuum model [5, 6], or in a fully-atomistic polarisable medium [7, 8, 9]. As a cautionary note, QM/MM results can depend sensitively on the methods used [10] and one must be careful about how the boundary is treated. Shaw et al. give the example of solvating a single QM water molecule in a bath of MM water [11]. One widely used water model, despite performing well in purely MM simulations, undergoes significant structural distortions when used in a QM/MM scheme. In this section, we summarise the results of a number of studies that have explicitly considered the convergence of a range of different observables with respect to the size of the QM region. We give examples of both QM cluster and hybrid QM/MM studies of biological systems, and we will show evidence that, in both cases, large-scale QM regions in excess of 500 atoms are required to converge many properties of interest.

The first systematic study of the dependence of biological properties on the size of the QM region was performed by Solt et al. [12]. Here, the authors studied the force error on polar and apolar residues in the lysozyme protein using a QM/MM approach with the QM region described by the PM3 semiempirical potential. As the size of the QM region centred on the residue of interest is increased, the force error decreases. However, perhaps surprisingly, the forces at the centre of the QM region do not reach acceptable levels of convergence (< 0.1 eV/Å) until beyond 500 QM atoms (~ 9 Å radius) for the polar region and 300 QM atoms (~ 7.5 Å radius) for the apolar region. More relevantly for the biological modelling community, the free energy of a proton transfer reaction between a water molecule and a glutamic acid residue in the polar region showed similarly slow convergence with the size of the QM region. For example, increasing the radius of the QM region from 3 to 6 Å, decreases the free energy of proton transfer by 3 kcal/mol. These effects are attributed to inconsistencies between QM and MM treatment of long-ranged electrostatic interactions, and the neglect of explicit polarisation in the latter. Indeed, the same effects have been observed in solution, where it has been shown that spherical shells of QM water of $\sim 8 - 9$ Å radius are required to converge interaction energies between a series of small organic molecules and water [13, 14].

Sumowski et al. also studied the effect of the environment on proton transfer isomerisation energies, in this case to investigate the protonation states of a stacked arginine pair in adenovirus Ad11, a motif that has been shown to be important for binding to its receptor CD46 [15]. While small QM models favour proton transfer to neighbouring residues to form a neutral arginine pair, inclusion of a larger, more realistic computational model stabilises the zwitterionic form. As above, a relatively large QM region (6 Å, 437 atoms) is required to converge the QM/MM isomerisation energies to less than 1 kcal/mol. Although, interestingly, there is still a discrepancy in excess of 3 kcal/mol between the converged QM/MM result and the largest purely QM calculation (1000 atoms), indicating that long-ranged electrostatics are particularly crucial here.

Another widespread use of QM in biological modelling is in enzymology, and this is reflected in the number of studies that examine the dependence of computed structural and energetic data on the number of atoms that are treated quantum mechanically. Sadeghian et al. study base excision repair of oxidised guanine by the bacterial glycosylase, MutM, using QM/MM with large-scale QM regions [16]. The authors perform extensive convergence tests of the reaction pathways with the size of the QM region studied. Particularly noteworthy is the barrier height of the second stage of the proposed reaction (the deprotonation of a proline residue), which varies from 28 kcal/mol using 143 QM atoms, to 6 kcal/mol using 278 QM atoms, to the converged value of 14 kcal/mol using 493 atoms. All barrier heights and interaction energies are converged to within 2 kcal/mol using QM clusters of radius 8 Å (~ 600 atoms). These conclusions are broadly consistent with studies of nucleophilic attack and proton transfer in acetylene hydratase [17], and of a methyl transfer reaction in catechol-O-methyltransferase [18].

Spectroscopy is another field where charge transfer and polarisation are expected to strongly influence computed results, and where the size of the QM region may be crucial. Our own interests lie in computing the optical excited states of chlorophyll pigments embedded in protein complexes for electronic excitation energy transfer. We found that QM clusters of around 1500–2000 atoms were required to converge the observed excitation energies to within 10 wavenumbers [19]. Isborn et al. computed the UV/vis absorption spectra of the photoactive yellow protein (PYP) chromophore in aqueous and protein environments [20], and compared QM/MM calculations with a full QM treatment of the system. In solution, it was found that quantum mechanical treatment of 40-200 water molecules surrounding the chromophore are needed to converge the peak position and width of the computed absorption spectrum. In the protein, the effect of QM treatment is even more visible – upon extending the QM region from 104 to 723 atoms, the spectrum is red-shifted by 0.3 eV. It was confirmed that the 723 atom result was converged by comparing with a calculation in which the entire protein was included in the QM region. Similarly, Zuehlsdorff et al. show that around 380 QM water molecules are required to converge time-dependent DFT calculations of the magnitude of the solvatochromic shift of an alizarin dye in solution [21]. In this system, use of an implicit solvent model actually shows qualitatively wrong behaviour (a blue shift of the spectrum relative to vacuum) when compared with the full QM result (red shift). It was further conclusively demonstrated that the origin of the red shift is the partial delocalisation of the excitation onto a 7 Å volume of water surrounding the pigment, an effect which cannot be modelled without explicit QM treatment of the environment.

Even stronger sensitivity to the environment has been observed when studying NMR chemical shifts in proteins. For a DNA-enzyme complex, the computed ^{1}H and ^{13}C shifts of a particular DNA lesion were found to drop below acceptable thresholds (0.1 and 0.5 ppm respectively) only when a QM cluster of radius 9 Å (1400 atoms) was used [22]. Although this minimum radius reduces to around 8 Å when used in a QM/MM scheme, this is still a formidable system size for conventional DFT approaches. Similar results were obtained for the 46 amino acid fungal dockerin domain, which is small enough that the entire protein may be treated quantum mechanically [23]. Here a QM radius of at least 6 Å is recommended (albeit with less stringent thresholds on convergence).

In contrast, local spectroscopic measurements may converge very quickly with the size of the QM region. For example, manganese isotropic hyperfine coupling constants in the oxygen evolving complex of photosystem II converge already after including around 100 atoms in the QM region, if used as part of a QM/MM protocol [24]. However, when using QM-only clusters, more than 200 atoms are required due to structural artefacts of the model that propagate into the inorganic core.

Summary: A consensus is forming that large QM regions are required to ensure that spectroscopic and energetic observables are converged with system size. Although QM/MM shows faster convergence than purely QM approaches, still QM regions in excess of 500 atoms are required in many cases. Such system sizes are problematic for many conventional QM approaches and so, in what follows, we outline some of the methods that will in future enable these calculations to become routine.

1.2. Feasibility of Large-Scale Simulations

In this section we will summarise the electronic structure methodologies that are capable of simulations at the length scales that we have described, focusing particularly on those that have been most actively used for biological applications. In practice, this restricts us to codes based on the Kohn-Sham formulation of density functional theory [25, 26], and its linear-scaling reformulations, as these are to date the only methodologies offering the requisite balance of speed with accuracy. Traditional approaches to DFT are in general based on finding eigenstates $\psi_i(\mathbf{r})$ with eigenvalues ϵ_i of the Kohn-Sham equation:

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + V_{\text{ext}}(\mathbf{r}) + V_{\text{H}}(\mathbf{r}) + V_{\text{xc}}(\mathbf{r})\right)\psi_i(\mathbf{r}) = \epsilon_i\psi_i(\mathbf{r}), \tag{1}$$

where we have labelled eigenstates by a band index i but not by not by a k-point, as models of biological systems are unlikely to display short-ranged periodicity, and most calculations will be performed at the Γ point only, if a periodic model is used. Here $V_{\text{ext}}(\mathbf{r})$ is the external potential of the nuclei, $V_{\text{H}}(\mathbf{r})$ is the Hartree potential of the electron density, and $V_{\text{xc}}(\mathbf{r})$ is the exchange-correlation potential. Spin labels have been omitted. Atomic units will be used henceforth, in which $\hbar = m = e = 1$. Depending on the basis set used, many calculations on biological systems rely on the replacement of an all-electron representation of the potential of the nuclei by a representation based on pseudopotentials, or the projector augmented wave (PAW) method [27], in which case V_{ext} becomes a nonlocal potential. The term in brackets is the Kohn-Sham Hamiltonian, and it is the computational effort of evaluating and solving this Hamiltonian, in a self-consistent manner, that dominates the computational effort of a DFT calculation. It is therefore of considerable importance to consider appropriate choices of methodology and basis set in which to express the solutions, and the scaling of the computational effort associated with each.

Since we have established a fairly high lower bound for the number of atoms required in realistic models of biological molecules, it becomes necessary to look beyond the traditional eigenstate-based approach, towards linear-scaling approaches, if the entire system is to be simulated within one calculation. Given that such approaches have only become available in recent decades, the early history of the field of electronic structure methods applied to full-sized biomolecules was dominated by methods based on splitting larger model systems into multiple subsystems [28, 29]. Separate calculations can then be run on these subsystems, after which results must be recombined, with some form of self-consistency loop to account for mutual interactions and equilibration and to ensure a single electron chemical potential for the system. In particular, accurate methods require that the calculation of each fragment be performed in the Coulomb field of other fragments, but that the field changes as the density evolves. One of the early methods to accomplish this self-consistent cycle is attributed to Morokuma [30] in 1971. Many of the methods in current use depend on the work of Kitaura [31], which developed the so-called Fragment Molecular Orbital approach. The field was summarised in a review by Gordon et al. [32] in 2012, so we will not address the technical details further in this review. This subsystem strategy can work very well if practised with care, with notable successes describing protein-ligand binding and drug design [33]. However, it is inevitable that as the system size grows it becomes increasingly complex to choose fragments in a way that cuts through as few covalent bonds as possible. There are also inevitable approximations associated with issues of continuity and treatment of kinetic energies at boundaries between regions. The rest of this review will therefore focus on methodologies where all of the QM system is treated concurrently in a single calculation, removing the need for human judgement in breaking the model down into subsystems.

1.2.1. Choice of Basis Set. One of the most important demarcations between different DFT methods is the type of basis set used. There is an apparently fundamental distinction between methods based on an equal treatment of all space (via a plane-wave basis or by real-space grids), and those based on the knowledge that the single-electron wavefunctions of condensed matter systems strongly resemble superpositions of atomic-like solutions (i.e. methods based on local orbitals). As we shall see, this distinction is in fact not so clear, with many of the most successful methods able to harness beneficial aspects of both approaches.

Plane waves have many advantages for total energy calculations: they offer an equal treatment of all space, and systematic convergence using a single parameter with respect to which the total energy is variational. They also offer advantages for calculation of forces, in that there is no need for Pulay terms to ensure accurate forces [34]. They necessitate the use of pseudopotentials or Projector Augmented Wave methods [27], and while in the past this was often a source of error, in recent years there is an emerging consensus that carefully-produced pseudopotentials can be precisely as accurate as all-electron methods in describing quantities associated with valence states. The "Delta" project, a collaborative effort between the developers of major DFT codes to validate their datasets and methodologies, recently showed [35] that pseudopotentials associated with a wide range of codes agree very well with each other, with pairwise differences that are comparable to those between different high-precision experiments.

However, large basis sets become increasingly problematic for large biological systems. One can estimate the scaling of the computational effort with respect to two main factors: the number of basis functions M and the number of occupied eigenstates N. In the case of plane-wave calculations, many parts of the calculation, including those that dominate for small systems, such as construction of the density, scale relatively benignly at first (e.g. as $O(M \ln M) \times O(N)$) due to the advantages of the fast Fourier transform [36]. However, at larger sizes, eventually the requirement of mutual orthogonalisation of all the eigenstates becomes the dominant source of computational effort. This step is associated with $O(N^2) \times O(M)$ scaling, which becomes infeasible for large systems. Furthermore, the representation of large simulation cells with significant vacuum (or implicit solvent) regions is problematic due to the large memory requirements of storing N bands with M coefficients if both N and M are very large. Leading DFT codes that make use of this plane-wave formalism include Quantum Espresso [37], ABINIT [38], CASTEP [39], and VASP [40]: there are many examples of applications of these tools to problems in the field of biochemistry, but few which meet the size criteria we have identified above for inclusion in this review.

Closely related to plane-wave methods are approaches based on their real-space equivalent, namely the use of straightforward representation of eigenstates on a grid, a topic reviewed by Beck [41] in 2008. The grids utilised can be either uniform, or multi-resolution, such that computational effort can be focused selectively on regions where wavefunctions are rapidly-varying. To ensure that the combination of finely-spaced grids with large simulation cells does not result in a rapid explosion of the computational effort required, various approaches have been used to improve the efficiency of grid-based methods. GPAW [42, 43] utilises the projector augmented wave formalism to ensure that relatively coarse grids can be used with high-accuracy, whereas RMGDFT [44, 45, 46] and others use a multigrid approach. The approach used in PARSEC is to bypass the use of an explicit basis and represent wavefunctions explicitly on a real space grid [47, 48]. Liu, Yarne and Tuckerman [49] investigated the use of a so-called Discrete Variable Representation, using continuous functions that satisfy the properties of position eigenfunctions on an appropriate grid. The resulting approach has been demonstrated in *ab initio* MD simulations [50]. RMGDFT, in particular, has been applied to various biological systems, including prion folding [51, 52], but this has usually been as part of a mixed scheme involving embedding a relatively small full-DFT region within an orbital-free DFT treatment of a surrounding environment. There are also grid-based codes with strengths in particular domains such as modelling excited states and spectroscopy, including OCTOPUS [53].

As a result of the poor eventual scaling of plane-wave and grid-based approaches for large systems, methods based on atom-centred local orbitals have traditionally been a much more common choice for biological applications where the full system is represented quantum mechanically. In this approach, a basis set is generated based on the solutions of the Kohn-Sham DFT problem for an isolated atom. Extra flexibility of this basis set is then generated by ‘splitting’ and ‘polarisation’ [54] of the orbitals to allow variational freedom to describe the redistribution of density that occurs as part of bond formation. Resulting basis sets are expressed as the product of a spherical harmonic and a radial function, expressed either via a parameterisation in terms of simple functions such as Gaussians, or numerically on a radial grid, as a so-called numerical atomic orbital (NAO) basis:

$$\phi_\alpha(\mathbf{r}) = \varphi_{n_\alpha l_\alpha}(r_I) Y_{l_\alpha m_\alpha}(\hat{\mathbf{r}}_I) \ , \tag{2}$$

where for an ion labelled by I at $\mathbf{R}_I$, we define $\mathbf{r}_I = \mathbf{r} - \mathbf{R}_I$, and n_α, l_α and m_α are the principle, orbital and angular quantum numbers of a local orbital labelled by α. The major advantages of such a basis set are that it is relatively small compared to plane-waves or grids in terms of the total number of functions required to achieve a given accuracy, and that each basis function is localised, in that it is only non-zero within a certain radius of the atom on which it is centred.

The use of truncated local orbitals means that for large systems, each local orbital only overlaps with an approximately constant number of nearby atoms, a number that does not grow with the size of the system. Constructing the Hamiltonian matrix in this representation requires the evaluation of all non-zero elements of:

$$H_{\alpha\beta} = \langle \phi_\alpha | \hat{H}_{\mathrm{KS}} | \phi_\beta \rangle \ , \tag{3}$$

The effort of each such evaluation does not need to grow with the size of the system, for local orbitals, and thus scales as $O(1)$. Furthermore, the number of such evaluations scales only as $O(N)$, so we can see that, asymptotically, the computational effort of evaluating $H_{\alpha\beta}$ does not need to grow faster than $O(N)$. Similarly, the number of non-zero elements of the overlap matrix grows only linearly with system size:

$$S_{\alpha\beta} = \langle \phi_\alpha | \phi_\beta \rangle$$

However, to solve the Hamiltonian and find the KS eigenstates, one is still required to either fully diagonalise the matrix by solving or to use techniques such as conjugate gradients to find the low-lying eigenstates, both of which are $O(N^3)$ operations.

The 'traditional' mode of operation of a number of local-orbital based codes (some of which also support the linear-scaling and/or adaptive local orbital schemes discussed below) follows this general scheme and utilises NAOs, for example SIESTA [55], FHI-AIMS [56], OpenMX [57], DMOL^3 [58] and Conquest [59, 2]. Due to their advantages in terms of analytical evaluation of matrix elements between local orbitals, Gaussian type orbitals (GTOs) are also a very popular choice for local orbitals in biological calculations, as used in, for example, Gaussian [60], CRYSTAL [61], NWChem [62], FreeON [63] and TeraChem [64]. Finally, there is a class of hybrid methods which combine local orbital and plane-wave basis sets within one calculation. One of the most notable examples of such methodology is the CP2K code [65], which uses a dual basis of atom-centred Gaussian orbitals and plane waves [66], with the former used to represent the wavefunctions and the latter used to represent the electronic density.

Many of these codes have demonstrated good convergence to very close to plane-wave-quality results for specific cases. However, the transferability of basis sets between dissimilar environments must be carefully tested. Furthermore, high-accuracy forces can be rather hard to obtain in local orbital methods due to the need for Pulay terms [34] to incorporate the fact that the basis changes as the atoms move, and the associated possibility of basis set superposition error (BSSE) [67].

In recent years, increasing attention has focussed not just on the total computational effort required to achieve a given precision in large-scale DFT calculations, but also on the degree to which this computational effort can be parallelised. Sustained development of processor architectures is no longer delivering significant improvements in single-core performance, but rather advancements are coming through increasing on-chip parallelism and introducing novel architectures such as Graphical Processing Units (GPUs). The performance characteristics of existing methodologies on these new architectures can differ significantly from previous generations, requiring regular re-evaluation of methodologies to extract best performance from available hardware. DFT methods based on all broad classes identified above have been adapted for GPU-based hardware, including real-space grids [68, 69], plane-waves [70, 71], and GTOs [72, 73, 74].

One relevant distinction in the context of large-scale simulation is between so-called "strong-scaling" and "weak-scaling". "Strong-scaling" refers to the ability to increase the number of parallel cores for a fixed-size simulation and achieve a corresponding decrease in the wall time taken. "Weak-scaling" refers to the ability to scale up the size of the simulation, with a corresponding scaling of the computational resources, while keeping wall time roughly constant. Corsetti investigated the performance of the SIESTA code with respect to both metrics [75], for large boxes of water. The results indicate that traditional DFT approaches are limited not just by the total computational effort, but also by their limited ability to scale to very large computational resources. Similar topics have been investigated by VandeVondele et al. in the context of the CP2K code [76]. There have been recent proposals for DFT methodologies and basis sets whose main virtues are that they can be efficiently parallelised to very large numbers of cores [77].

1.2.2. Linear-Scaling Methods. For biological applications, the traditional DFT approaches described above encounter a scaling-wall of computational effort, prohibiting calculations much beyond around 1000 atoms. At the same time, they also encounter severe limitations in terms of the computational memory required to perform such calculations, which scales at least quadratically with system size. Fortunately, a parallel trend in the DFT methodology literature, dating back as far as the early nineties, has been the development of linear-scaling approaches to density functional theory. A review of the early years of this field was produced by Goedecker in 1999 [78], describing the foundational work of a number of research groups, including those of Yang [79, 80, 81], Kohn [82, 83] and Vanderbilt [84]. A more recent summary was provided by Bowler and Miyazaki [3]. We will not attempt to reproduce the material covered therein, but rather summarise the state-of-the-art in terms of methods appropriate for high-accuracy calculations on biomolecules. For example, while orbital-free DFT is making significant progress in the description of metallic and simple semiconducting systems [85], it is not yet widely applied to biological systems except in the context of improving the description of the region surrounding an embedded system of interest [51] described with a higher-level method. It would appear that for the moment, given the limited accuracy of orbital-free representations of kinetic energy functionals based only on the density, the higher-accuracy representation of the kinetic energy that is possible within Kohn-Sham DFT is required to describe the variety of bonding present within a biological system.

In the previous section, we noted that the use of localised orbitals allows, within KS-DFT based on a semilocal exchange-correlation functional, the construction of the Hamiltonian in linear-scaling computational effort. To create a fully linear-scaling approach, one must also be able to construct and optimise some representation of the sum of the occupied eigenstates in linear-scaling computational effort. Central to many of these schemes is the concept of locality, or "near-sightedness", namely the idea that properties in one region of a system are only weakly influenced by properties spatially far away from this region. This near-sightedness, discussed by Kohn in 1996 [83], is closely associated with the existence of a gap at the Fermi level in the density of states. The existence of such a gap means that, in the Kohn-Sham picture, the system can be well-represented by a set of eigenstates which are either fully-occupied (below the Fermi level) or fully-unoccupied (above the Fermi level). This is true of most components of biological molecules, particularly the more structural components of most proteins, but as we shall see in Section 3, it is not automatically true of model systems which have not been prepared with sufficient attention to considerations of saturating bonds and electrostatic screening.

A second crucial ingredient in linear-scaling methods based on KS-DFT is the use of a density matrix representation. The one-electron density matrix can be shown to completely specify a quantum-mechanical system, within the independent-electron picture used within KS-DFT, such that all observables can be calculated from it. McWeeny [86] provided an early review of the use of the density matrix within electronic structure theory. The aforementioned linear-scaling DFT reviews trace later developments in more detail than is possible here [78, 3]: the key points relevant to the simulation of biological systems are described below.

The density matrix is a representation of the density operator $\hat{\rho}$. For a single-determinant wavefunction, as is used in KS-DFT to describe an insulating system, the density matrix can be expressed in the position representation as:

$$\rho(\mathbf{r},\mathbf{r}') = \sum_i \psi_i^*(\mathbf{r}) f_i \psi_i(\mathbf{r}) \ , \tag{4}$$

that is, as a sum over the orbitals $\psi_i(\mathbf{r})$ multiplied by their occupancy f_i (0 for empty states, 1 for filled states). The expectation value of any operator $\hat{A}$ can be expressed by taking its trace with the density operator, $\mathrm{tr}(\hat{\rho}\hat{A})$. For example, the kinetic and potential energies (for a local potential) can be written as

$$E_{\mathrm{kin}} = -\frac{1}{2}\int \nabla^2_{\mathbf{r}} \rho(\mathbf{r},\mathbf{r}')|_{\mathbf{r}=\mathbf{r}'} \,\mathrm{d}\mathbf{r}'$$

$$E_{\mathrm{pot}} = \int \rho(\mathbf{r},\mathbf{r}) V_{\mathrm{eff}}(\mathbf{r}) \,\mathrm{d}\mathbf{r} \ .$$

As these are the components of the band structure energy, it is possible to express the band structure energy of a system succinctly as $E_{\mathrm{BS}} = \mathrm{tr}(\hat{\rho}\hat{H})$, and the total energy $E_{\mathrm{T}} = E_{\mathrm{BS}} - E_{\mathrm{dc}}$, where E_{dc} is a double-counting term depending only on the electron density. Given that it is thus possible to express the total energy in terms of the density matrix, without use of the eigenstates, it is natural to use it as the fundamental quantity in a total energy calculation.

It is then necessary to minimise the energy with respect to this density matrix, subject to constraints which ensure that the density matrix is a valid representation of the corresponding system of Kohn-Sham single-particle states. The two constraints which must be obeyed are of idempotency, that is to say equivalence to a set of integer-occupied (fully empty or fully filled) states, and normalisation, that is to say, representing the appropriate number of electrons N_e.

The idempotency condition can be expressed as $\hat{\rho}^2 = \hat{\rho}$, or

$$\int \rho(\mathbf{r}, \mathbf{r}'')\rho(\mathbf{r}'', \mathbf{r}') \, \mathrm{d}\mathbf{r}'' = \rho(\mathbf{r}, \mathbf{r}') \; . \tag{5}$$

It is straightforward to verify that this is obeyed for the form given in eq 4. Meanwhile the normalisation condition is $\mathrm{tr}(\hat{\rho}) = N_e$, or

$$\int \rho(\mathbf{r}, \mathbf{r}) \, \mathrm{d}\mathbf{r} = N_e \; . \tag{6}$$

Note that eigenvalues of the density matrix for an insulator will be zero or one: for a spin-degenerate system, therefore, we must multiply by a factor of 2 in equation 6 for doubly-occupied orbitals to obtain the correct number of electrons. Spin labels will be omitted throughout this work for the sake of simplicity.

Extensive effort has been devoted to analysis of the properties of the density matrix, particularly of its decay properties as a function of the separation of the two position operators, $|\mathbf{r} - \mathbf{r}'|$. Via the connection to the range of a corresponding Wannier function description, it can be shown [87] that the density matrix is a highly diagonally-dominant operator, in that asymptotically there is a rapid decay of $\rho(\mathbf{r}, \mathbf{r}')$ as a function of $|\mathbf{r} - \mathbf{r}'|$. This decay can be shown to be exponential for an insulator or a metal at finite temperatures, but only algebraic in the case of a zero temperature metal.

This diagonal dominance means that if represented using localised orbitals similar to those introduced above, the density matrix is expressible as a sparse matrix. Hernandez and Gillan [88] introduced a form which is now widely-used:

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{\alpha\beta} \phi^*_\alpha(\mathbf{r}) K^{\alpha\beta} \phi_\beta(\mathbf{r}') \; , \tag{7}$$

where $K^{\alpha\beta}$ is known as the density kernel. This is a sparse matrix representation of the density matrix, which technically is expressed in terms of the (contravariant) duals $\phi^\alpha(\mathbf{r})$ of the local orbitals $\phi_\alpha(\mathbf{r})$. The dual functions are bi-orthogonal to the direct functions, hence they obey $\int \phi^\alpha(\mathbf{r})\phi_\beta(\mathbf{r}) \, \mathrm{d}\mathbf{r} = \delta^\alpha{}_\beta$, and are in general more delocalised than the direct functions. However, since in most approaches of this type, one never needs to explicitly construct the duals, this is not usually problematic (though care must be taken in general to ensure that all quantities correctly respect the covariant/contravariant nature of localised orbitals).

The use of sparse matrices can thus be seen as the foundation for linear-scaling reformulations of KS-DFT. In terms of the aforementioned sparse matrix forms of the Hamiltonian, overlap, and density kernel, we can express the total energy as:

$$E_\mathrm{T} = \sum_{\alpha\beta} K^{\alpha\beta} H_{\beta\alpha} - E_\mathrm{dc}[\rho(\mathbf{r})] \; , \tag{8}$$

as well as the requirements for idempotency:

$$K^{\alpha\beta} = \sum_{\gamma\delta} K^{\alpha\gamma} S_{\gamma\delta} K^{\delta\beta} \; , \tag{9}$$

and normalisation

$$\sum_{\alpha\beta} K^{\alpha\beta} S_{\beta\alpha} = N_e \; . \tag{10}$$

The problem of finding the total energy is now to self-consistently minimise eq 8 while maintaining the constraints of eqs 9 and 10. Approaches to do this can be divided into three general classes as discussed in more detail in the aforementioned reviews [78, 89, 3]: direct methods, purification methods, and variational methods.

Direct methods, such as the Fermi operator expansion (FOE) [90, 91, 92] and the kernel polynomial method of Voter et al. [93], attempt to find the density kernel corresponding to a given Hamiltonian and electron chemical potential. These must in general be accompanied by self-consistency loops to ensure correct assignment of the chemical potential and self-consistency of the density. Purification-based methods seek to iteratively improve upon an initial guess for the density kernel, without knowledge of the chemical potential. This class includes approaches such as that of Palser and Manolopoulos [94], Niklasson et al. [95], and many more recent variations on this theme, such as Rubensson et al. [96]. Finally, in variational methods, the interacting energy of eq 8 is minimised with an approach that adheres to the constraints as the optimisation proceeds. These approaches include the Li, Nunes and Vanderbilt (LNV) method [84], and methods based on penalty functionals that penalise deviation from idempotency [83, 97].

Within all of these linear-scaling techniques, in the simplest analysis of scaling of computational effort with system size, computational time (for fixed number of CPU cores) would scale as $T(N) = AN + B$, in the limit of large N for constant A, B where A is the prefactor and B is a small constant related to operations that are independent of the size of the system (such as electrostatic calculations in the context of a fixed-size cell). The prefactor A is a hugely important consideration as it determines at what system size there is a crossover to where linear-scaling calculations are more computationally efficient than traditional approaches. To enable large calculations, one needs to keep the size of the local orbital representation as low as possible such that this prefactor is small, and this depends crucially on the choice of basis set. Significant effort has been devoted to finding compact yet accurate combinations of fixed basis sets in the context of linear-scaling approaches such as in OpenMX [57, 98], and SIESTA [55].

1.2.3. In Situ Optimised Local Orbitals. While the previous section made a fundamental distinction between local-orbital methods and approaches such as plane waves that treat all space equally, in fact several linear-scaling methodologies attempt to achieve the best of both worlds. It is possible to harness the flexibility, power and systematic convergence that can be achieved for plane waves and grids, while still obtaining the scaling benefits of local orbitals. One way to achieve this is to use a representation of the density in terms of a set of adaptable local orbitals, themselves expressed in terms of a systematic underlying basis. The local orbitals can be used to represent the density matrix either explicitly or implicitly, but crucially without requiring direct construction of the eigenstates of the Hamiltonian. Codes undergoing active development in this category include BigDFT [99], CONQUEST [59, 2], ONETEP [100, 101] and MGMOL [102, 103]. This type of methodology is particularly attractive for biological applications, because there are in general many different forms of bonding present in a biomolecule, and many different elemental species, yet it is important to treat them all on an equal footing and not bias the results through a choice of basis which may unintentionally be better adapted to describe one type over another.

CONQUEST uses the representation of the density matrix in eq 7, written in terms of support functions that can either be fixed NAOs [104] or adaptive functions expressed using B-spline (blip) functions [59] and optimised *in situ*. The former approach has been used for large-scale demonstrations of modelling biological molecules [105, 106], but the latter approach is not yet widely used for this purpose.

In a similar vein, the BigDFT code uses a basis of Daubechies wavelets of degree 16, on an adaptive real-space mesh with two levels of resolution, to represent its support functions [107, 108]. Despite successful demonstrations to multiple inorganic systems, and benchmarks showing scaling of a calculation for a DNA fragment consisting of 14,300 atoms to over 25,000 parallel cores [99], this combination of approaches is not yet widely used for full-scale biomolecular applications.

The MGMOL code formulates the equations of DFT in terms of general non-orthogonal orbitals instead of eigenfunctions. These are represented on a real-space uniform mesh, using a finite difference discretisation to calculate kinetic energies. The approach used is to directly compute the localised orbitals, truncated beyond a cutoff radius, by minimising an energy functional with localisation constraints that only minimally, and controllably, affect the overall accuracy. This also requires computation of selected elements of the inverse overlap matrix, which is accomplished by inverting principal sub-matrices of the global Gram matrix [109]. The code has been designed to express excellent parallel scaling, with reports of MD calculations of up to 100,000 atoms on 100,000 processors, with a wall-clock time of $O(1)$ minute per molecular dynamics step [110].

ONETEP uses *in situ* optimised local orbitals expressed in a basis of periodic cardinal sine (psinc) functions on a regular real-space grid [100, 101] whose spacing is controlled by a cutoff energy equivalent to a plane-wave cutoff. Multiple options are available for density matrix optimisation: the approach utilises LNV, penalty functional, and purification schemes [111], all employing highly-optimised parallel sparse matrix algebra [112, 113]. The support functions, known as non-orthogonal generalised Wannier functions (NGWFs), are optimised *in situ* and demonstrate controllable variational convergence to plane-wave results [114] with respect to the radius of the functions, and systematic elimination of BSSE [115]. Good computational scaling has been demonstrated on calculations of amyloid fibril systems of 41,907 atoms, to in excess of 30,000 parallel cores [116].

Summary: Linear-scaling methodologies based on systematically-described, adaptive local orbitals are still relatively new and have not yet been widely used for biomolecular simulations outside of their developer communities and research groups collaborating closely with them. However, their success in describing biological systems, to be examined in the following sections, suggests that they will find increasing use as the simulation of full-scale biomolecular systems becomes more widespread. One pressing need is for systematic comparisons between these various approaches, since to date we are not aware of any work that has attempted to examine either the ability of the various schemes to provide equivalent, well-converged energy differences. Another important area of investigation would be to compare the computational effort required to achieve a given level of accuracy for different methods.

2. Required Functionality

2.1. Challenges Involved in Studying Biological Systems

Apart from the question of size, we can identify four further complications that make modelling biological molecules more challenging than modelling the inorganic crystalline systems for which DFT has become so successful. These are: i) the need for very high accuracy, due to the relatively small energy scales associated with the accessible microstates of a system at room temperature (typically overall precision well under 1 kcal/mol would be required for useful total energy comparisons); ii) the need to combine multiple advanced functionalities of electronic structure codes to describe solvation, strong correlations, dispersion, and excited states, possibly all at once; iii) the need to extract useful information beyond just total energies, such as geometry and spectroscopy; and iv), since biological molecules are relatively flexible and function at finite temperature, the need to sample a very large number of geometrical conformations of the biomolecule and solvent. In this section, we examine recent developments in the functionality required to address these challenges.

On the subject of accuracy, it is important to consider whether sufficiently well-converged calculations can be achieved with the linear-scaling methodologies discussed above. In the case of adaptive local orbitals, this has been demonstrated for several approaches. Fox et al. made a careful comparison of binding energies for a system comprising a phenol molecule in a solvation shell of water [14]. Binding energies were calculated firstly with adaptive support functions as in ONETEP (i.e. NGWF optimisation), and with standard GTO basis sets in the NWChem code, with and without counterpoise corrections [67] to address the issue of BSSE. These results show that convergence is achieved relatively rapidly with increasing support function radius in ONETEP, with a convergence to better than 0.05 kcal/mol by $8.0a_0$. By comparison, results for the GTO basis sets without counterpoise corrections showed no clear convergence to a well-defined answer (variations of ±10 kcal/mol) and those with counterpoise required basis sets at the highly-expensive cc-pVTZ level before comparable convergence to $8a_0$ NGWFs was achieved. It has been demonstrated that the use of such accuracy settings is sufficient for high-accuracy forces, e.g. for geometry optimisation and transition state searching [117].

These high-accuracy settings, however, mean that calculations are not as cheap as one might hope for a linear-scaling method. To illustrate the challenge that all the requirements for biomolecular applications introduce over-and-above the normal demands of a DFT calculation, we show in Figure 1 two plots of the total execution time for single-point energy calculations using the ONETEP code on models of amyloid fibrils of a range of sizes, run on 3840 cores of the UK's national supercomputer ARCHER. The size of the simulation cell was kept constant ($80 \times 80 \times 175$ Å) as the size of the amyloid fibril model was varied.

The squares illustrate calculation times at relatively modest accuracy settings, namely support function radii of 7 a_0, expanded in a psinc basis with a 600 eV energy cutoff, with kernel truncation at 20 a_0, and no use of implicit solvent. The PBE generalised gradient approximation was used in all calculations [118]. These settings would be adequate for a preliminary investigation of the energy landscape, and total wall-time for an electronic structure calculation is very clearly linear with system size, and remains well under 1 hour even for a system of nearly 14,000 atoms. By contrast, when adjusting all settings to a level which produces high-accuracy forces (as described in Ref. [117]), namely 8 a_0 support function radii, 800 eV psinc cutoff, kernel truncation at 25 a_0, and an implicit solvent description using a multigrid solver, there are two significant changes to the profile of time as a function of system size. Firstly, there is a constant offset, representing the computational time associated with the multigrid solver. Secondly, the slope is increased by a factor of around five compared to the coarser computational settings. However, recent work to extend the parallelisation via hybrid OpenMP/MPI parallelism means that these calculations can be scaled to many thousands of parallel cores [116], keeping the required wall-time within reasonable bounds.

2.2. Electrostatics and Implicit Solvation

In most functional proteins, one of the purposes of the protein structure is to encapsulate and protect a small number of sites from which the main functionality is derived, for example the active site of an enzyme. As such, the protein structure will have been tuned to optimise the function of this site. One of the ways a protein achieves this is by tuning the site's immediate electrostatic environment, through the charges and dipole moments associated with each residue in the structure. It is therefore of the utmost importance in the simulation of biological molecules to include accurate treatment of long- and short-ranged electrostatics.

In a traditional DFT calculation performed in vacuum, the classical electrostatic part of the local effective potential is given by the sum of the Hartree and nuclear potentials (or the local part of the pseudopotentials). Each of these may be calculated in real or reciprocal space as appropriate to the boundary conditions in use. Most biomolecular calculations are for isolated, non-periodic systems: for charge-neutral systems with no significant dipole moment, it is sometimes adequate to use a supercell approximation as there will be no significant interactions between periodic images. In such cases, the problem of solving the Poisson equation is relatively trivial. However,

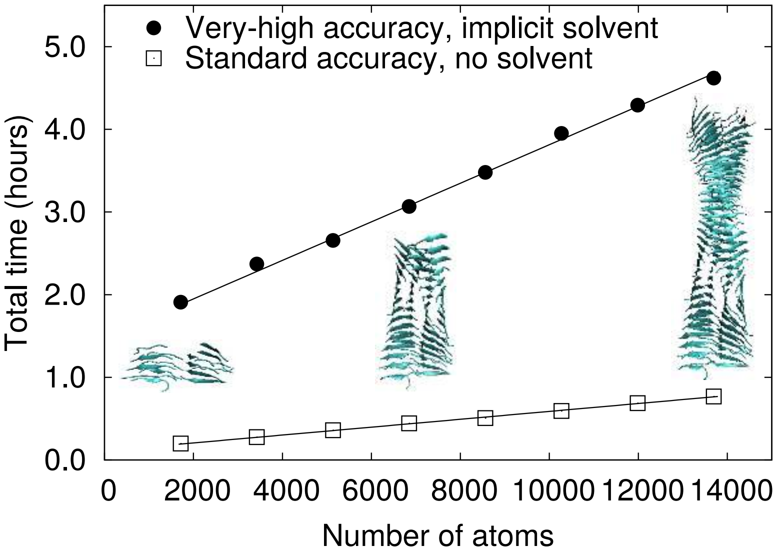

Figure 1. Timings for total energy calculations using ONETEP, with *in situ* optimisation of the NGWF representation, on amyloid fibril structures. Two sets of timings are shown: standard-accuracy total energy calculations (squares) with typical ONETEP parameters (see text for convergence parameters) which have been shown to be of accuracy comparable to a 6-311+G* GTO basis, and for high-accuracy settings (filled circles) equivalent to fully-converged plane-wave accuracy or cc-pVQZ with counterpoise corrections. Trendlines are shown of the form $T(N) = AN + B$ for each case, with the majority of the increase in B for the high-accuracy settings being the result of the time required for applying the multigrid solver for Poisson's equation in a large simulation cell. The structures were obtained with permission from the authors of Ref. [119] and were previously used for similar benchmarks in Ref. [14].

in most real cases it is necessary either to use explicit open boundary conditions or to truncate the Coulomb interaction if periodic boundary conditions are in use. Various "cutoff Coulomb" approaches have been studied and may be efficiently implemented for a range of geometries [120, 121, 122, 123, 124, 125, 126]. Open boundary conditions generally necessitate the use of a multigrid approach to the solution of the Poisson equation. It is important to note that in some cases the solution of the Poisson equation can become the computational bottleneck dominating the overall computational effort, particularly if its solution exhibits poor parallel scaling. Consequently, significant effort has been invested in optimising the solution of this problem [127].

Multigrid approaches also have the significant advantage that they can be combined with the use of a non-uniform dielectric permittivity. This is of great value for the simulation of biological molecules as it allows for an implicit solvent description of the regions beyond the explicitly-simulated QM region. The edges of the molecule are typically exposed to a very different electrostatic environment to that further inside, and without appropriate dielectric screening, incorrect charge distributions will be obtained, leading to problems discussed in more detail in Section 3, which can be significantly alleviated via the use of an implicit solvent model. A very wide range of such approaches has been developed [5], with widely-used implementations including the Polarisable

Continuum Model (PCM) [128], the COSMO model [129], the SMD model [130], and the unified electrostatic and cavitation model of Scherlis et al. [131]. An adaptation of the latter by Dziedzic et al [132] has recently been demonstrated for application to entire proteins [133, 134].

2.3. Dispersion

Van der Waals interactions are generally understood to cover all attractive and repulsive forces between molecules or atoms that do not arise from a covalent bond or directly from electrostatic interactions [135]. While they are relatively weak compared to covalent bonds, they nevertheless account for a significant fraction of the total interaction energy between groups of atoms, particularly, for example, at protein-protein interfaces, or in lipid bilayers, or in non-bonded stacking interactions in DNA [136]. The most significant component in most cases is the so-called London dispersion force, a weak intermolecular force arising from the formation and interaction of instantaneous polarisation multipoles in molecules. Despite their widespread use, traditional semi-local functionals, including the local density approximation and popular generalised gradient approximations, such as PBE, are unable to capture the correct long-ranged tail of the dispersion interaction. A great deal of methodological research in recent years has therefore focused on developing appropriate approaches to correctly model dispersive interactions, and many of these are now approaching extremely high accuracy even for the intermolecular binding of large complexes [137]. These approaches can be categorised into a few main groups [138]: semi-empirical parameterised approaches based on interatomic C_6 coefficients fit to reproduce the results of high-accuracy quantum chemistry calculations [139, 140], schemes that use *ab initio* methods based on the wavefunction or density around each atom to parameterise interaction terms [141, 142, 143], and fully-nonlocal functionals of the density, most notably those based on the van der Waals density functional (vdW-DF) approach [144, 145, 146].

Berland et al. recently reviewed the vdW-DF method [147], covering its applications to biology, notably to DNA [148, 149]. A crucial advance in making applications to large systems feasible was the discovery by Román-Pérez and Soler that the fully-nonlocal (and hence $O(N^2)$-scaling) expressions of the original formulation of vdW-DF could be re-expressed via interpolation of the integration kernel and fast Fourier transforms [146], reducing the scaling to $O(N \log N)$.

There is widespread indication to be found in the works cited in this section that inclusion of vdW effects at some level is crucial for accurate description of both energetics and dynamics in *ab initio* modelling of biological systems. However, there is a shortage of comparative work assessing the relative strengths and weaknesses of these approaches in real-world applications, particularly for large biological systems. It may well be that it is much less important for static calculations based on MM molecular dynamics snapshots or experimental structures, which already implicitly include their effect, since vdW rarely has a significant direct effect on the electronic ground state except via its effect on geometries. We can therefore expect, for example, theoretical spectroscopy to be quite accurate even if the direct effect of vdW interactions are neglected in the DFT calculation.

2.4. Strongly-Correlated Methods and Self-Interaction

Almost half of all enzymes associate with a metal in order to function [150]. Wherever this is the case, it must be assumed that there are specific electronic properties of the metal ion that are required for its function, otherwise organisms would not incorporate elements which are scarce and/or toxic. Often, in the case of partially-occupied d-subshells, this is their ability to exist in multiple oxidation and/or spin states.

For any quantitative computational studies of ligand binding or chemical reactions at metal sites, a quantum mechanical description of the metal centre is clearly required. However, standard DFT treatments of transition metal chemistry is often inadequate due to the strong electron correlation effects associated with the d orbitals. It is well-known that DFT suffers from self-interaction errors [151], which particularly affect the description of transition metal chemistry. There is an extensive literature on attempts to correct these closely-related issues, and many of the resulting approaches have been successfully carried over to large-scale biomolecular simulations.

The class of methods known as DFT+U attempts to use a corrective functional inspired by the Hubbard model to improve deficiencies of the standard DFT description [152, 153, 154]. In practice, DFT+U methods have the effect of penalising non-integer occupations of orbitals of a partially-filled correlated subspace such as the $3d$ electrons of a first-row transition metal. As such they improve upon the DFT description of both self-interaction and strong-correlation effects [155], at the expense of the introduction of an unknown parameter, though more recently methods to calculate this parameter from first principles have been developed [156]. The DFT+U method has been implemented in several of the codes discussed above, notably in ONETEP [157, 158] (applications of this combination of approaches will be discussed in Section 4.3), as well as in SIESTA, CP2K and other codes. An alternative, which in many cases has been shown to have a qualitatively similar effect to DFT+U is to use a hybrid functional incorporating a certain fraction of exact exchange, as this is known to improve the description of self-interaction effects. Popular examples include the B3LYP [159], HSE [160] and PBE0 functionals [161].

However, in cases where strong correlation behaviour is more complex, neither DFT+U nor the use of hybrid functional approaches are able to correct the underlying issues and provide an appropriate description of transition metal chemistry. In such cases, there have been attempts to improve the description of the d orbitals by means of dynamical mean field theory (DMFT), which takes into account both quantum dynamical effects and valence and spin fluctuations, and also explicitly includes the Hund's exchange coupling J [162]. DMFT has recently been coupled with ONETEP to produce a novel linear-scaling DFT+DMFT approach [163, 164], thus enabling the study of the role of quantum many-body effects in large systems [165].

2.5. Spectroscopy via Time-Dependent DFT

Absorption or emission of light, or more generally the properties of electronic excited states, play a vital role in many biomolecular processes. Most notable of these is of course photosynthesis, as well as vision, luminescence, and photo-induced damage to DNA and other biological molecules. As we have discussed in the context of ground state calculations, there are a wide range of possible methodologies that can be used for excited states, but the need to describe large system sizes imposes strong restrictions on what is feasible, in practice leaving only time-dependent forms of DFT (TDDFT) as viable options. Nevertheless, these come in several relevant forms: TDDFT can be performed in a real-time formalism [166], which explicitly propagates time to study dynamical processes, or it can be performed in the linear-response framework [167], whereby the dynamical response of a system to an excitation at a specific frequency is considered, so as to find the energies of specific excitations. Both approaches have found application within the field of modelling small-scale biological molecules such as individual DNA bases and base-pairs [168], structures such as porphyrin rings [169], and chromophores embedded in protein environments such as those found in pigment-protein complexes and fluorescent proteins [170]. Castro et al. have provided a review of theory and applications of TDDFT to biomolecules up to 2009 [169].

As discussed above in the context of ground-state energies, the majority of the field of TDDFT applied to larger biomolecules still utilises fragment-based or QM/MM descriptions of a small chromophore inside a host medium (such as the rest of the protein) described with approximate methods: as before, we will not cover such approaches within this review as they have been widely discussed elsewhere. Recently, however, there has been increasing availability of methodology for large-scale TDDFT calculations. Real space methods including Octopus [53] can be used for calculations of very large systems such as chlorophyll networks in pigment-protein complexes [171], as will be discussed in more detail in Section 4.4. Zuehlsdorff et al. developed an approach to enable linear-response TDDFT within the ONETEP large-scale DFT code [172, 173, 174], and demonstrated application to solvated chromophores and dye molecules embedded in explicit solvent [21]. Tretiak et al. and Challacombe have also developed linear-scaling approaches within the linear-response formalism [175, 176]. O'Rourke and Bowler, meanwhile, have demonstrated a real time, density matrix based approach to TDDFT within the CONQUEST approach [177]. This bears similarity to earlier work by Yam et al. [178].

It should be noted that the use of TDDFT for the prediction of the energies of excited states is not yet as accurate or robust as standard DFT is for ground states: notably, the choice of exchange-correlation functional in TDDFT is generally more complicated than in DFT. The optimal choice often can depend on the character of the excitation of interest, such as whether it is primarily a local excitation or has charge-

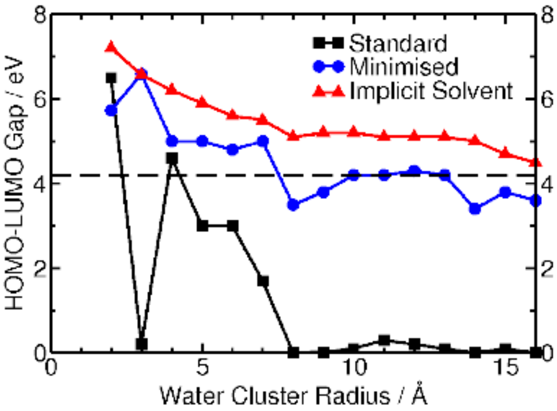

Figure 2. Difference in energy between the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) for a series of spherical water clusters of increasing radius [184]. In the standard simulation, the cluster is cut out of the bulk liquid. In the minimised simulation, the cluster first undergoes structural optimisation. Finally, in the implicit solvent calculation, the QM calculation is performed in a dielectric medium ($\epsilon = 80$). **Reproduced with permission from the Institute of Physics.**

transfer character. This issue carries over to modelling large-scale biomolecules, where charge-transfer states are often expected to play an important role and are generally not well-described by semi-local functionals, requiring a portion of exact exchange, with an associated increase in computational effort.

3. Mind the Gap

Before we begin our review of the applications of DFT in biology, we first address a widely held misconception that pure exchange-correlation functionals (that is, those containing no Hartree-Fock exchange) cannot be applied to large system sizes. If correct, this would significantly hinder the study of biological molecules since computation of the Hartree-Fock energy is expensive due to its inherent non-locality and poor scaling with system and basis set size. The root of these beliefs was a series of observations reporting unphysical vanishing HOMO-LUMO gaps and poor self-consistent field (SCF) convergence in proteins [179, 180, 181, 182] and even water clusters [183]. SCF convergence issues were often attributed to the well-known under-estimation of the HOMO-LUMO gap by pure functionals.

To investigate this issue, we performed a series of large-scale DFT calculations on water clusters and proteins using the ONETEP software [184]. In agreement with previous observations, Figure 2 shows that the HOMO-LUMO gap of a spherical cluster of water molecules extracted from a bulk liquid rapidly falls to zero for system sizes in excess of 200 atoms. However, while the under-estimation of the gap by pure exchange-correlation functionals is undisputed, it seemed unlikely that it should become worse for larger system sizes. Indeed, we were able to show that a 2010-atom bulk periodic supercell of water has a clearly defined band gap of 4.2 eV (Figure 2, dashed line). The observed insulating behaviour of the bulk system thus indicates that there are no difficulties in applying pure functionals to large system sizes, but instead the vanishing HOMO-LUMO gap is caused by the creation of an artificial vacuum-water interface in cluster simulations.

Analysis of the electrostatic properties of a cluster of water molecules reveals a significant dipole moment that increases with system size up to around 75 D in the largest cluster studied [184]. This effect is due to unterminated hydrogen bonds at the surface-water interface and is entirely unphysical as, in reality, the surface water molecules would reorient to counteract the net dipole moment. In the simulation, however, the artificially imposed dipole moment has a significant effect on the electronic properties of the system. The local density of states (LDoS) computed as a function of the distance along the dipole moment vector shows a clear shift in electronic energy levels, with the electric field pushing some states higher in energy and some lower. The HOMO-LUMO gap vanishes when electronic states on opposite surfaces of the cluster coincide.

We therefore hypothesised that there is *no* inherent problem with the use of pure functionals in studying biological systems, but rather that SCF convergence problems are manifested in smaller system sizes than for hybrid functionals due to the formers' inherently smaller HOMO-LUMO gap. Furthermore, we were able to make a number of practical suggestions for simulating cluster models of biological systems. Figure 2 shows that if the water cluster undergoes structural optimisation prior to electronic structure analysis, or if the analysis is run using an implicit solvent model, the expected HOMO-LUMO gap of around 4 eV is recovered. It has also been shown previously that embedding classical point charges outside the electron distribution has a similar effect [182]. In all three cases, the preparation method results in a more physical environment for the water cluster and screens the unrealistic dipole moment observed in vacuum simulations.

We have gone on to show that simulation of the electronic structure of proteins is also possible using the methods described here. In vacuum, all proteins studied, ranging in size from 75 to 1231 atoms, displayed vanishing HOMO-LUMO gaps. However, all of the gaps were recovered when the environment of the protein was modelled using either implicit or explicit solvent [184]. Thus, simulations of very large system sizes are fully feasible with Kohn-Sham DFT as long as the environment is modelled in a physical manner. Furthermore, similar concepts are extremely useful in spectroscopic analysis of cluster models. Time-dependent DFT simulations of chromophores in vacuum-terminated water clusters are problematic due to charge transfer to low-lying excited states at the vacuum-water interface [185, 186]. Zuehlsdorff et al. have shown that the use of an implicit solvent model greatly reduces spurious charge transfer states between the surface waters and the pigment which otherwise arise from the mechanism described above [21].

Summary: We have shown that the commonly observed closure of the HOMO-

LUMO gap in DFT simulations of large-scale biological molecules is a simple electrostatic artefact of the system preparation method, and may be trivially remedied by using an implicit solvent model [184]. We hope that this knowledge, in conjunction with the ever-expanding set of tools that is available for tackling such systems, encourages the community to further explore the wide biochemistry of living organisms.

4. Applications

In this section, we summarise applications in biology where large-scale DFT has made the greatest impact, due to the need for both large system sizes and quantum-mechanical detail in the description of bond-breaking, transition metals, electronic excitations, or intermolecular interactions.

4.1. Biomolecular Structure and Electronic Properties

One of the most fundamental biomolecular properties is that of structure, and how it relates to biological function. The first principles determination of protein structure directly from its amino acid sequence is far beyond the realms of current feasibility, although promising progress is being made in the prediction of secondary structures of increasingly long peptides using *ab initio* molecular dynamics simulations [187]. However, if large-scale DFT can be shown to be able to rapidly and accurately optimise structures, there is significant potential to use these methods to refine experimental structure predictions [188, 189]. Furthermore, significant insight into biomolecular function may be obtained by studying the electronic structure of carefully chosen models. As an example, calculations of the electrostatic potential at the surface of the aldose reductase enzyme using the SIESTA code [55] have been shown to be in good agreement with high-resolution x-ray diffraction data and revealed an interesting electrostatic complementarity between the enzyme's active site and its binding cofactor (a molecule required to activate its function) [190]. In this section, we review the state-of-the-art in large-scale DFT modelling of biomolecular structure and give further examples of functions that have been shown to arise directly from electronic properties.

One of the most studied protein structures is the crambin crystal. The crystal structure is available at extremely high resolution (0.48 Å) [191], and while the protein is quite small (46 amino acids, 642 atoms), it is large enough to exhibit both α-helix and β-sheet secondary structure elements. The structure was first optimised at the HF/4-21G level in a pioneering study in 1998 [192]. Subsequently, its deformation density – defined by subtracting the electron densities of isolated atoms from the total electron density – was computed using the SIESTA large-scale DFT software [55] within the generalised gradient approximation and compared with x-ray diffraction data [193]. The authors found a correlation coefficient of 86% between the theoretical and experimental densities in the peptide bonds of the protein, indicating that DFT gives a reasonable picture of chemical bonding in the protein backbone.

A more complete picture of crambin and its interactions with its environment was recently obtained by modelling the periodic protein crystal structure, which contains two proteins in the repeating unit, along with different levels of lattice water solvation [194]. The crystal was modelled using the B3LYP hybrid functional and a 6-31G(d,p) basis set with Grimme's empirical dispersion correction [139]. Both cell parameters and atomic coordinates were optimised for systems ranging in size from 1284 atoms (no added water molecules) to 1800 atoms (172 solvating water molecules). As expected, the fully hydrated system gives structural parameters in the best agreement with experiment. In this case, computed lattice parameters are around 3% smaller in every direction than experimental values, though this may be mostly attributed to the thermal expansion in the crystal structure which was collected at 290 K. Additional factors which should be noted are the neglect of zero point energy contributions in the simulations and over-binding due to basis set superposition error (BSSE), which may become significant for small basis sets. The root-mean-square deviations of the atomic positions relative to experiment were around 0.4 Å. As expected, α-helix and β-sheet secondary structures were better modelled than the random coil regions of the protein. The authors also computed crystal formation energies of the optimised structures. Of particular note here is that i) even in the fully solvated system, dispersion interactions play a significant role in protein-protein binding (37% of the formation energy); ii) before corrections, around 50% of the cohesive energy is due to spurious BSSE-driven attraction, which indicates that the model is still far from basis set convergence; iii) clathrate-like pentagonal rings were observed around hydrophobic residues indicating the possible importance of ordered water structures at the protein surface.

In order to ascertain the accuracy of *ab initio* methods for protein structure refinement, Kulik et al. studied a much larger benchmark set comprising 58 proteins ranging in size from 70 to 590 atoms [180]. The proteins were optimised using the TeraChem package [64] using both restricted Hartree-Fock (RHF) and the range-corrected ωPBEh density functional with basis sets up to a maximum size of 6-31G. Comparing the root-mean-square deviations between computed C_α positions and experiment, it is perhaps surprising to note that even the most accurate QM methods (around 0.7 Å RMSD) appear to be less accurate than MM force fields (0.6 Å RMSD). However, this may be rationalised by considering that MM force fields are extensively parameterised to produce "healthy" solution phase protein structures. The authors provide convincing evidence that the force field often predicts healthy structures even when the experiment suggests an unusual or disordered structure. In these cases, QM methods are likely to be more reliable. However, there is also room for improvement in the QM methods. Firstly, as discussed above in the case of the crambin crystal, the obtained geometries likely suffer from BSSE anomalies. Indeed, the smallest basis set studied suffers from an unusual deprotonation of amide nitrogen atoms. Secondly, the structural optimisations were performed in the gas phase, and so the obtained structures are likely to differ from the crystalline or solution phase structures.‡ To correct this

‡ The authors also comment on convergence difficulties for DFT methods in these vacuum calculations,

second issue, the authors also performed QM/MM simulations for a subset of 20 proteins embedded in a bath of classical point charges to represent the water medium. Protein solvation now eliminates the spurious amide backbone deprotonation and reduces the C_α RMSD to below 0.4 Å, which is a more encouraging result.

The first *ab initio* molecular dynamics simulation of an entire protein – bovine pancreatic trypsin inhibitor (BPTI) – was also performed using the TeraChem software [64]. Optimisation of the protein and crystal waters at the RHF/6-31G level resulted in a RMSD of 0.3 Å in the positions of the backbone atoms [195]. A molecular dynamics simulation of 8.8 ps was run treating the protein and six water molecules (900 atoms in total) at the RHF/STO-3G level and the rest of the solvent using a classical force field (i.e. a QM/MM simulation). Even with this relatively short equilibration time, the RMSD of the protein backbone showed signs of stabilising to a value of around 1.5 Å relative to the experimental structure. In order to investigate charge transfer at the protein-water interface, 88 geometries were extracted from the dynamics for further analysis at the RHF/6-31G and B3LYP/6-31G levels, now with every atom treated quantum mechanically (2634 atoms). On average, around 2.6 electrons were transferred to the protein from water, thus the net charge of the protein differs significantly from its putative charge of +6 e assigned using standard pH rules. The charge transfer is even larger for B3LYP which may be due to the tendency of DFT to delocalise electrons. In the gas phase, there is a strong intra-protein charge transfer, whereby neutral residues donate around 1 e in total to charged residues. Adding water neutralises the excess of positive charge on the neutral residues and releases polarisation stress in the protein. Thus, as well as demonstrating the feasibility of *ab initio* MD of entire proteins, these simulations also highlight a potentially important role for biological water [195].

It is particularly interesting to study the properties of proteins whose electronic structure is expected to play a specific role in their function. Feliciano et al. study the electronic structure of a small α-helical peptide (known as a pilin) belonging to the bacterium *Geobacter sulfurreducens* [196]. Filaments formed by these peptides are conductive and seem to play a role as the electronic conduits between the cell and extracellular electron acceptors. QM single point calculations were performed on solvated pilin structures (comprising 4580 atoms) using the SIESTA code [55] with the PBE exchange-correlation functional and a DZP basis set. The authors observed a biphasic charge distribution along the length of the helix (positive potential in the mid-region, and negative at the two termini). This in contrast to a polyalanine helix, which displayed a relatively flat charge distribution due to solvent screening of the permanent dipole of the helical backbone. This different electrostatic behaviour is also reflected in the computed HOMO-LUMO gaps which are around 1 eV and 3 eV for the pilin and polyalanine respectively, which is consistent with the filament's role in electron transport. The QM calculations indicate that the closing of the gap is due to specific amino acid motifs, in particular positively charged residues, that localise the HOMO

which we later diagnosed as being due to electrostatic artefacts (Section 3) [184].

in the mid-region of the helix. Interestingly, both the HOMO and LUMO have weight in regions containing aromatic tyrosine residues, which have particularly low oxidation potentials and which may aid hopping of electrons under thermal fluctuations [196].

Similar α-helical structural motifs are found in ion channels, whose function it is to regulate the flow of ions across the cell membrane. Large-scale DFT calculations may be particularly important here, since ion permeation is expected to be strongly dependent on the electrostatic potential, which is difficult to model for confined ions with MM force fields in the absence of electronic polarisation. Todorović et al. use the CONQUEST software [59, 2] to investigate the structural and electronic properties of the gramicidin A ion channel [106]. It is one of the simplest ion channels (552 atoms) and has an unusual structure in which all peptide side chains face outwards towards the membrane. Despite this simple structure, it is selective for monovalent cations and shows no measurable permeability for anions or polyvalent cations [197]. Optimised structures of the channel in vacuum, using the PBE exchange-correlation functional, are in good agreement with experiment (RMSD of 0.20 Å in the positions of the backbone atoms). Interestingly, the substitution of tryptophan (Trp) residues with non-polar phenylalanine has been shown experimentally to reduce the ion permeation rate [198], even though the Trp sidechain is not located along the channel axis. However, large-scale DFT calculations reveal that the electrostatic potential inside the gramicidin A ion channel is identical (to within 1 meV) to that of a hypothetical polyalanine tube [106]. Thus, it seems that the effect of the Trp side chains is a more subtle one that may only be revealed by studying dynamical interactions with the lipid membrane and solvent.

Some of the earliest applications of large-scale DFT in biology studied the question of whether dry DNA could be used as a molecular wire. The fundamental question of whether DNA is an electric conductor or not is difficult to ascertain experimentally as it is difficult to control the molecular environment, for example, the amount of water and counter-ions that are present. To answer this question, a series of large-scale DFT calculations were performed using the SIESTA code [55] on a periodic, dry DNA chain comprising eleven guanine (G) – cytosine (C) base pairs (715 atoms) [199, 200]. Following structural optimisation with the GGA exchange-correlation functional, the authors found a clear band gap of 2.0 eV, though this does not in itself rule out conduction if there is hole donation by defects or counter-ions. The topmost valence band of this pristine DNA model has a very low bandwidth of 40 meV and lies on the guanine nucleobases, while the LUMO lies on the cytosines. To model random sequence (or λ-) DNA, one G–C pair was swapped to C–G. The HOMO of the swapped guanine falls into the lower valence bands, while the LUMO of the cytosine rises in energy. The sequence disorder localises states close to the Fermi level on just a few base pairs and hence acts to decrease DC conduction. The prediction of low conduction in λ-DNA is in good qualitative agreement with experimental measurements of the resistivity of 15 μm DNA chains absorbed on mica (lower limit of 10^6 Ωcm) [199]. These early computational investigations did not rule out conduction via hopping mechanisms. Indeed a later study [201] (albeit on smaller model systems) computed an activation energy for polaron (an electron/hole in an empty/filled band coupled to a lattice distortion) hopping in DNA of 0.15 eV in good agreement with experiment (0.12 eV [202]). Meanwhile, subsequent studies of wet DNA tetramers found that water is able to dope DNA if it is able to enter the structure at structural defect sites [203].

Summary: The studies reviewed in this section have played an important role in demonstrating the feasibility of determining not only the electronic, but also the structural, properties of biological macromolecules. In this respect, an average optimisation step for the largest studied system in the aforementioned crambin study (1800 atoms) required around 23 minutes on 1920 cores of a Cray Cascade XC40 supercomputer [194]. The same authors also report timings of 3 hours for a 4575 atom model of γ-chymotrypsin on the same machine [194]. Meanwhile the authors of the BPTI study showed that single point calculations for a 2634 atom system may be performed in under 3 hours on a single desktop machine with eight GPUs [195]. We caution, however, that care must be taken to accurately represent the surrounding environment (Section 3) and further investigation of the convergence of structural properties with basis set size are needed to avoid spurious artefacts and establish large-scale DFT as the benchmark of choice for these systems.

4.2. Enzymes

Due to their extraordinary ability to catalyse biochemical reactions with high specificity and under mild, physiological conditions, enzymes have been the subject of intense computational scrutiny since the pioneering works of Warshel and Levitt in the 1970s [204]. Computational studies have a particularly important role to play in this field, because the transition state of the chemical reaction has a very short lifetime and, hence, is difficult to characterise experimentally. As well as providing information on the structure of the transition state, computation also has the potential to validate proposed reaction mechanisms, balance various contributions to the catalytic effect and discern the roles of active site residues [205, 206]. Armed with this knowledge, researchers have been able to interpret and predict the effects of mutagenesis [206], design transition state analogues as potent enzyme inhibitors [207], predict the factors that control drug metabolism *in vivo* [208], and even design new enzymes [209].

Computational enzymology is a natural fit with QM/MM methods. In general, bond-breaking is confined to a well-defined active site and is described by semi-empirical methods or DFT or, in recent years, high-level electronic structure methods such as coupled-cluster [210]. Nevertheless, as we described in Section 1.1, there is a growing consensus that QM regions comprising many hundreds of atoms are required to converge enzymatic properties such as activation energy barrier heights. This view is reflected in the growing number of applications that employ large-scale DFT in the elucidation of enzymatic mechanisms [18, 17, 16, 211], and these methods may further contribute towards the important future goal of making simulation methods accessible to non-specialists by removing the complexity of the QM/MM interface. In what follows, we

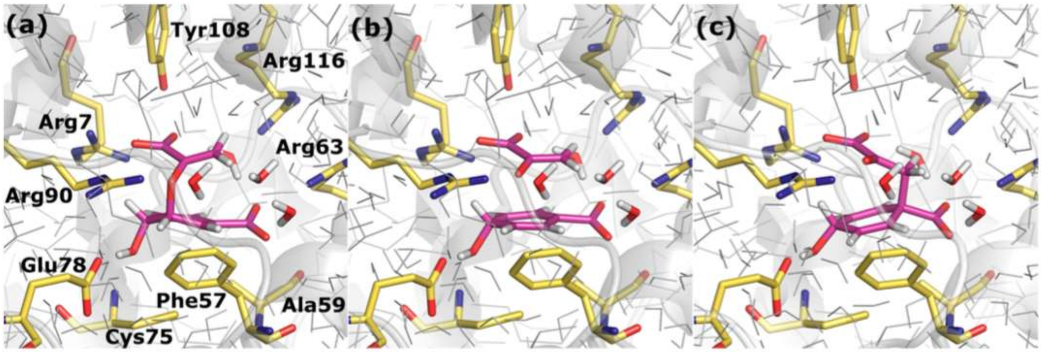

Figure 3. Optimised structures of (a) chorismate, (b) the transition state and (c) prephenate in the chorismate mutase enzyme using large-scale DFT calculations [212]. **Reproduced with permission from the American Chemical Society.**

focus on two recent examples of the application of large-scale DFT to computational enzymology.

4.2.1. Chorismate Mutase. Over the past few decades, a number of enzymes, such as lysozyme and citrate synthase, have become benchmarks against which new functionalities or more accurate methodologies have been assessed [205]. Chorismate mutase is another such example. The enzyme catalyses the Claisen rearrangement of chorismate to prephenate, which maintains the balance of aromatic amino acids in the cells of fungi, bacteria and higher plants [213, 214]. Chorismate mutase is of theoretical interest due to a number of appealing factors. Firstly, the substrate does not covalently bind to the active site [215, 216, 217], hence the system may be readily separated into the QM region (the substrate) and the MM region (the enzyme). Secondly, the reaction proceeds with a similar mechanism in aqueous solution, hence the factors controlling the enzymatic rate enhancement may be readily inferred by contrasting the energetic contributors to the activation energy barrier with those in water. And finally, experimental measurements of both the free energy and enthalpy of activation in both enzyme and water are available for comparison. The activation free energy is much lower in chorismate mutase ($\Delta^{\ddagger}G = 15.4$ kcal/mol, $\Delta^{\ddagger}H = 12.7$ kcal/mol) [216] than for the uncatalysed reaction in water ($\Delta^{\ddagger}G = 24.5$ kcal/mol, $\Delta^{\ddagger}H = 20.7$ kcal/mol) [218], which corresponds to a catalytic rate enhancement of 10^6. In addition, calorimetric investigations have measured an enthalpy of reaction of -13.2 kcal/mol in water [217].

We therefore chose the chorismate to prephenate reaction in chorismate mutase as the ideal system on which to benchmark the use of large-scale DFT for computational enzymology applications [212]. Initial reactant state (RS) and product state (PS) geometries in the enzyme and in water were generated using semi-empirical QM/MM molecular dynamics, followed by structural optimisation at the B3LYP/MM level [219]. Spherical clusters centred on the substrate molecule were extracted and re-optimised using the ONETEP large-scale DFT code [101] with the PBE exchange-correlation functional augmented by empirical dispersion corrections [140]. Transition state (TS) optimisation was performed using a linear and quadratic synchronous transit pathway algorithm (LST/QST) with conjugate gradient optimisation [220].

An important consideration when planning large-scale DFT calculations with structural optimisation is not only the scaling of the QM calculation with system size, but also the rapid increase in the number of evaluations of the potential energy surface that are required to find the minimum energy structure [221]. We have found that for enzymology applications optimising an inner 'mobile' region within a fixed outer region improves the tractability of the problem whilst retaining the scaffold and electrostatic environment of the surrounding protein. However, it is important to show that the computed properties of the system are converged with respect to the size of these optimisation regions. For the chorismate to prephenate reaction in the enzyme, we investigated system sizes ranging from 900 to 1900 atoms and mobile regions from 100 to 200 atoms. The maximum change in the computed activation energy barrier and reaction energy was just 0.3 kcal/mol, indicating convergence of the energetics of the system with respect to the size of the computational model.

Figure 3 shows the optimised RS, TS and PS geometries in chorismate mutase. It is important to note that no information regarding the structure of the transition state or any reaction coordinate was fed into the calculations; the transition-state searching algorithm used only the RS and PS structures as input. After averaging over five reaction pathways to account in an approximate manner for temperature effects, we found that the chorismate mutase enzyme reduces the barrier height by 10.5 kcal/mol relative to water, which is in good agreement with experiment (8.0 kcal/mol [218, 216]). The small discrepancy is likely dominated by the use of the PBE functional to describe bond-breaking. Future studies will examine the stability of the results with respect to changes to the exchange-correlation functional.

Finally, as well as reproducing experimental data, it is important to also use large-scale DFT calculations to extract insight into the catalytic mechanism. To this end, we have performed natural bond orbital analysis [222, 223, 224] to estimate the contribution of each active site residue to enzymatic rate enhancement [212]. It was found that significant orbital overlap between the substrate and a number of charged active site residues act to stabilise the transition state (relative to the reactant and product states). This picture is supported by a previous study, which demonstrated the importance of electrostatic interactions in the active site to TS stabilisation in chorismate mutase [225].

4.2.2. Acetylcholinesterase. In most enzymatic applications employing large QM regions, the focus is inevitably on computing enthalpies of reaction. The reason for this is, of course, that computing *free energies* of reaction involve orders of magnitude longer computational times to sample configurational space. Encouragingly, Fattebert et al. have made some progress in the computation of free energies from first principles by using molecular dynamics simulations in their study of the acylation reaction of acetylcholine (Ach) catalysed by acetylcholinesterase [226]. The active site of the enzyme is modelled as a 612 atom QM subsystem (of which 212 atoms are mobile) in a precomputed external potential field to represent the protein environment [227].

The first stage of the chemical reaction involves the approach of a serine residue on the protein towards the Ach substrate, and the transfer of a hydrogen from serine to a nearby histidine residue. The loss of the hydrogen atom increases the nucleophilicity of serine, and results in the formation of a bond with the substrate. A reaction coordinate was defined as the distance between atoms on the serine and substrate, and first principles molecular dynamics simulations were run at 300 K for at least 4 ps using a constrained reaction coordinate. The forces required to enforce these constraints were recorded and integrated over the reaction coordinate to obtain the free energy pathway. The free energy barrier for the first stage was computed to be 6.0 kcal/mol, and a charged stable intermediate was identified [226].

The second stage of the reaction involves the subsequent transfer of the hydrogen atom from the first stage onto the Ach substrate. The free energy barrier was computed as before, using the hydrogen position as the new reaction coordinate. The authors found that the second stage is the rate-limiting step in the reaction and computed an overall barrier of 8.5 kcal/mol, in good agreement with the experimental value of 11.8 kcal/mol [228]. The remaining discrepancy between theory and experiment is attributed to errors in the use of the PBE exchange-correlation functional and possible incomplete sampling. Nevertheless, the reported run times (100 s per MD step on 363 CPUs [226]) indicate that large-scale DFT calculations are becoming fast enough that accurate and converged free energy calculations are on the horizon.

Summary: The discussed applications are important in demonstrating the feasibility of performing calculations that are converged with respect to system size and largely independent of the choice of computational parameters. This will allow us, in future, to more extensively study the effects on the reaction mechanism of factors such as protein conformational changes and fluctuations, nuclear tunnelling, and the exchange-correlation functional.

4.3. Metalloproteins

Computational modelling of metalloproteins is a challenging endeavour, often requiring both a sophisticated treatment of strong electron correlation effects at the metal centre and accurate treatment of long-ranged interactions between charged species and the surrounding protein. A common moiety in protein biochemistry and the centre of our initial investigations is the haem molecule, which reversibly binds small molecule ligands and plays a crucial role in storing and transporting oxygen (O_2) in vertebrates (Figure 4). The myoglobin protein contains a single haem molecule, bound via its central iron ion (Fe(II)) to a histidine residue on the protein (H93). The protein environment has the effect of reducing haem's natural preference for binding carbon monoxide (CO) over O_2 – the binding free energy of CO, relative to O_2, is reduced from 5.9 kcal/mol in non-polar solvent to 1.9 kcal/mol in the protein [230]. The influence of the protein

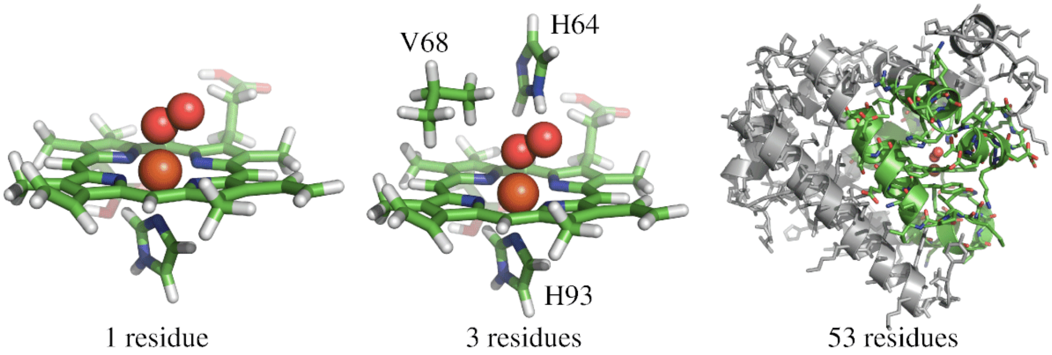

Figure 4. Computational models of haem [229]. Iron and O_2 are shown as orange and red spheres. (right) The myoglobin protein, with the 53 residues modelled in our large-scale DFT simulations shown in green. **Reproduced with permission from the American Chemical Society.**

is traditionally understood to be mediated by i) favourable electrostatic interactions between the residue H64 and the negatively charged O_2 molecule (charge transfer from Fe to CO is negligible) and ii) unfavourable steric interactions between the protein binding site and the linear bound conformation of the CO molecule (the Fe–C–O bond angle is close to 180°, whereas the Fe–O–O angle is closer to 120° and is a better fit to the protein's binding cavity) [231]. Point i) requires an accurate description of both the electronic structure of the Fe–O_2 bond and long-ranged electrostatic interactions, while for point ii), the amount of strain energy stored in the protein depends critically on the accuracy of the description of intramolecular interactions in the protein.

To begin to study the roles of different ligand discrimination effects in myoglobin, we built realistic 1007 atom models of the myoglobin protein in complex with CO and O_2 (Figure 4) and structurally optimised them at the QM level using the PBE exchange-correlation functional [229]. The structures of the models following optimisation were in excellent agreement with experiment [232], with RMS deviations of around 0.1 Å between computed and experimental heavy atom positions in the haem and ligand molecules. Natural population analysis revealed that approximately 0.5 e are transferred from Fe to the O_2 molecule within the DFT approach [224]. By comparing the relative binding energy of the two molecules to haem in vacuum and protein environments, we computed that the protein discriminates in favour of O_2 by 3.7 kcal/mol, in good agreement with the experimental result (4.0 kcal/mol [230]). Furthermore, by decomposing the total protein contribution to binding into intermolecular and intramolecular contributions, we demonstrated that ligand discrimination is dominated by interactions between the charged O_2 molecule and the protein and that steric effects are negligible. Similar methods were used in a study of the CO_2 fixation energy in the ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) magnesium-based enzyme where it was shown that long-ranged interactions between CO_2 and the protein are important determinants of binding [233].

A major problem with the DFT results is a strong energetic imbalance in myoglobin that favours CO binding over O_2 to too large an extent. This imbalance results from the self-interaction errors in DFT, which particularly affect the description of transition metal chemistry. To correct for this in an efficient manner, we used the DFT+U approach as implemented in the ONETEP software [158] to recompute the relative binding energies of the two small molecules to myoglobin [229]. Although the over-estimation of CO binding was reduced, we could not find a reasonable value of the Hubbard U parameter that gave satisfactory results for all experimental energetic observables.

In order to address this issue, we set out to improve the description of the Fe $3d$ orbitals by incorporating quantum many-body effects into our 1007 atom models of myoglobin by means of the novel linear-scaling DFT+DMFT methodology described in Section 2.4. We first validated the model by comparing the optical absorption spectra of ligated myoglobin with experiment [165]. The characteristic infrared absorption band of oxygenated myoglobin and the double-peaked porphyrin Q band are both recovered using the DFT+DMFT approach with a reasonable value of the Hund's exchange coupling ($J = 0.7$ eV). Interestingly, the net charge on the O_2 molecule is increased from half an electron in the DFT picture to close to one electron in DFT+DMFT (-1.1 e using natural population analysis [165] and later confirmed to be -1.0 e using density derived electrostatic and chemical population analysis [234]). Metal-to-ligand charge transfer occurs via π-bonding between Fe $3d$ and O_2 π^* orbitals. Our calculations predict a much stronger π-bonding interaction than has previously been observed, which is supported by the agreement between the computed $d\pi$ hole character in our large-scale models (19 %) [165] and Fe L-edge X-ray absorption spectroscopy measurements (15±5 %) [235]. Furthermore, the classical picture of the ground-state wave function existing in a single spin configuration appears to be a poor approximation in haem. When bound to CO, myoglobin has a dominant singlet character, whilst oxygenated myoglobin displays significant ground state entanglement with higher spin contributions. Finally, we investigated to what extent the metal-to-ligand charge transfer and the multireference high spin character of the ground state influence ligand discrimination in myoglobin [165]. Remarkably, when all of these effects are taken into account, we obtain an excellent agreement with the experimental relative binding (free) energy (1.9 kcal/mol [230]) for $J = 0.7$ eV.

Summary: Both a sophisticated treatment of strong correlation effects and large systems sizes are essential for quantitative measures of ligand binding energetics in transition metal complexes.

4.4. Photosynthesis

The initial stages of photosynthesis employ optically active molecules (pigments) in light-harvesting protein complexes to capture photons through the formation of molecular electronic excited states (excitons) [236]. These excitons are transported through

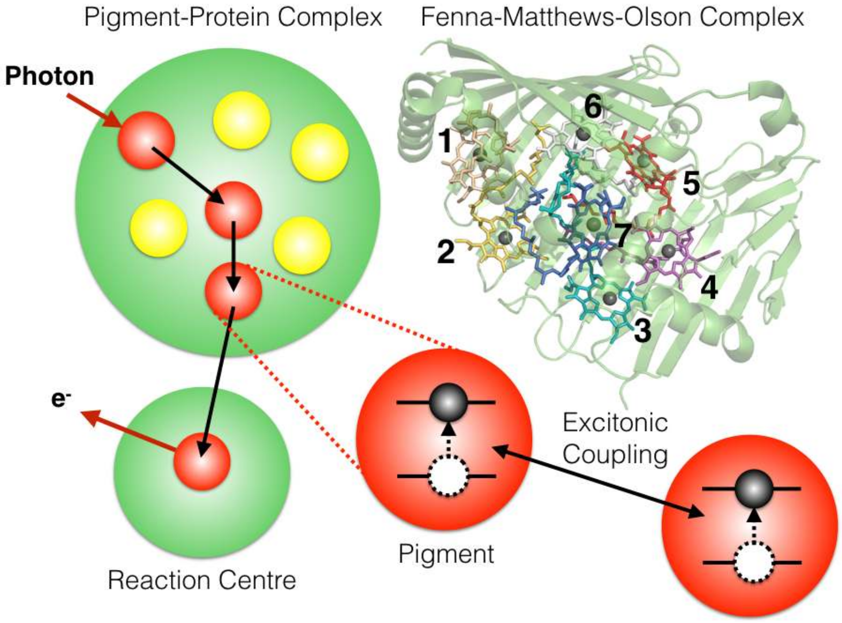

Figure 5. Cartoon of the early stages of photosynthesis. Electron excitations (excitons) are created in optically active pigments via photon absorption. Excitons are moved through light-harvesting protein complexes, via Coulombic coupling between pigments, and are ultimately transported to the reaction centre where they are used to release electrons. (Inset) One monomer of the Fenna-Matthews-Olson pigment-protein complex. Seven chlorophyll pigments are shown in stick representation.

the pigment-protein complexes to the photosynthetic reaction centre, where they are ultimately converted into stored chemical energy (Figure 5). Much of our theoretical understanding of how these processes occur in nature is derived from the Fenna-Matthews-Olson (FMO) light-harvesting complex of green sulphur bacteria. Since these bacteria are often found in low-light environments, their photon-to-electron conversion efficiency is extremely high, and their underlying design principles are therefore of great interest in the field of artificial solar energy transduction [237].

A recent x-ray crystal structure of the FMO complex reveals a trimeric quaternary protein structure, which sequesters seven optically-active pigments within a clam-like architecture (Figure 5) [238]. Interest in this relatively simple pigment-protein complex has been driven by observations of room temperature quantum dynamics in an ensemble of FMO proteins [239]. Understanding how efficient energy transport may occur in the presence of significant external noise from a computational viewpoint requires a multiscale approach incorporating electronic, optical and dissipative interactions. The key parameters in these models are i) the optical transition energies of the pigments in their local environment (site energies), ii) inter-pigment couplings of optical transitions (excitonic couplings) and iii) the spectral density modelling dynamic interactions between the pigments and their environment [240, 241]. Efforts have been made to extract these parameters directly from the crystal structure using a wide variety of computational methods [242, 243, 244, 245, 246]. However, the results appear to be very sensitive to approximations made in the construction of the computational models. For example, in the majority of these approaches it was necessary to treat all, or most, of the system with approximate MM point charges. Given the importance of long-ranged electrostatics and electronic polarisation to the excited state properties of the pigments, the computation of parameters describing exciton transport in the FMO complex is seemingly an excellent candidate for large-scale DFT simulation.

In 2013, we performed large-scale DFT simulations of the FMO pigment-protein complex with the aim of deriving all of the site energies and excitonic couplings directly from first principles [19]. Using the ONETEP software [101], we performed seven DFT calculations, each centred on one of the pigment sites and containing all protein, water and pigment atoms lying within a 15 Å sphere (1600–2200 atoms). The method of Ratcliff et al. was used to optimise both occupied valence and low-energy unoccupied conduction states [247, 248] and Fermi's golden rule was employed to obtain the local optical transition energies and densities. The resulting *ab initio* site energies were in good agreement with those that have been fit to reproduce experimental optical spectra of the FMO complex (the RMSD between the theoretical and fit data is 47 cm^{-1}) [244]. The identification of pigment 3 (Figure 5) as the lowest energy pigment is consistent with its likely role as the exit through which excitons leave the FMO complex [249]. Further calculations identified the permanent dipoles of two α-helices as playing important roles in red shifting the site energies of pigments 3 and 4. Hence, accurate descriptions of the long-ranged electrostatic interactions between the pigments and their environment is vital for accurate spectral determination.

We have also computed the excitonic coupling strengths as the Coulombic interaction between the optical transition densities of the pigments [19]. The correlation between the coupling parameters and those obtained using a point charge model is extremely good [250]. The first principles site energies and excitonic couplings were used to compute linear optical absorption, circular dichroism and linear dichroism spectra. Comparison with experiment [251, 252] revealed good agreement for the low frequency regime but significant disagreement for the higher frequency components of the spectrum. Nevertheless, this is an extremely encouraging first test of the use of large-scale DFT in an extremely complex pigment-protein environment.

One of the limitations in the above study was the use of Fermi's golden rule to compute optical transition energies. Improvements in accuracy are expected when employing a time-dependent Hamiltonian, and towards this goal, researchers have started to apply TDDFT calculations to study natural photosynthesis. Indeed, a recent study performed real-space time-propagation TDDFT calculations of the electronic absorption spectrum of the chlorophyll network of the light-harvesting antenna complex from green plants (LHC-II) [171]. The trimeric pigment-protein complex comprises over 17000 atoms and contains 14 chlorophyll molecules per monomer. Arguing that the overall effect of the protein environment is to introduce a constant red shift in the site energies of the pigments, the authors reduce the scale of the electronic structure calculation by removing the majority of the protein. This still leaves a chlorophyll network comprising more than 6000 atoms. TDDFT calculations were performed using the Octopus code [53] with an electron density propagation time of 40 fs, which allowed a spectral peak resolution of around 0.1 eV. The computed optical absorption spectrum was in good agreement with experiment, particularly in the lower energy chlorophyll Q-band. Furthermore, by decomposing the spectrum into the contributions of individual pigments, it was shown that the stromal (towards the outside of the cell membrane) and lumenal (inner membrane) sides of LHC-II absorb at different frequencies, thus suggesting a mechanism for energy flow across the antenna complex.

Summary: In recent years, some steps towards a fully first principles determination of the optical spectra of complex biological molecules have been made. It will be interesting to investigate the use of time-dependent DFT in the computation of FMO optical properties [173] and we are extending our study to include dynamical effects on the site energies and excitonic couplings [253]. It is exciting to contemplate the role that large-scale DFT calculations may play in the future study of much larger pigment-protein complexes [254] and more generally in the emerging field of quantum biology [255].

4.5. Medicinal Chemistry

Medicinal chemistry is the highly inter-disciplinary study of the design and synthesis of pharmaceutical agents, or drugs, that bind to a biomolecular target and modulate its activity with therapeutic benefit. Quantum mechanical calculations have the potential to contribute throughout the drug discovery process, from the design of catalysts to aid molecular synthesis [256], to predicting sites on new drug candidates that might be susceptible to metabolism [257], to determining the crystal packing of solid pharmaceuticals [258, 259]. Here, we focus our discussion on the problem of determining the strength of the binding between the therapeutic target (usually a protein or DNA) and small molecules or peptides. The binding affinity is at the core of drug discovery – drugs that bind strongly to their target may be taken in smaller doses and are less likely to cause harmful side effects by promiscuously binding at other sites. Moreover, computational determination of intermolecular interactions must be extremely accurate to make an impact in the drug discovery process. A 1 kcal/mol error in binding free energy corresponds to an order of magnitude change in the drug activity.

Despite the tight restrictions on the level of accuracy required, classical MM force field based approaches to computer-aided drug design are widespread and remarkably successful [260, 261, 262, 263]. Nevertheless, there is always room for improvement, and fixed atom-centred point charge models, of course, fail to correctly model effects such as explicit polarisation and anisotropic electron density [264]. In theory, quantum mechanical modelling provides a natural means to improve the accuracy

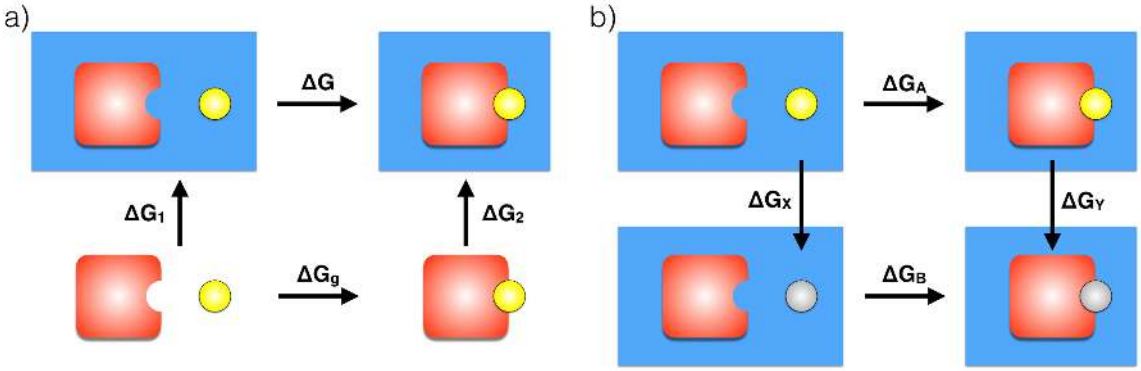

Figure 6. (a) Free energy cycle employed in Refs. [267, 268, 269, 270]. The total binding free energy (ΔG) is given by the sum of the gas phase binding free energy (ΔG_g) and the relative solvation free energies of the receptor (red), ligand (yellow) and complex ($\Delta G_{aq} = \Delta G_2 - \Delta G_1$). (b) Free energy cycle for the computation of relative binding free energies. Molecule A (yellow) is alchemically transformed into molecule B (grey), both in the receptor and in water, and the relative binding free energy is given by $\Delta G_X - \Delta G_Y$.

of the computation of intermolecular interactions. While the large-scale techniques discussed in this review certainly allow realistic models of protein–ligand complexes to be simulated, an obstacle to the widespread use of QM modelling is the sampling requirements. Drug molecules, and their targets, are relatively flexible and, at room temperature, a number of different binding modes can contribute to the binding free energy. We will show, in what follows, some examples of intermittent hydrogen bonds in protein–ligand and protein–protein interactions and the errors that can be introduced if this flexibility is neglected (for example, by determining binding from the crystal structure alone). Furthermore, it has previously been shown that in a range of host–guest systems the free energy of binding is very poorly correlated with the enthalpy of binding [265]. Hence, methods are required that accurately compute the free energy of binding at room temperature. A further important consideration, is the time to solution as pre-clinical drug discovery programmes can progress at a high pace. While computational approaches can certainly compete with experimental approaches on cost (estimates place laboratory costs in excess of $500 000 [266]), a run time on the order of 24 hours is probably required to be competitive.

4.5.1. Quantum mechanical / Poisson Boltzmann Approaches. One of the earliest applications in a biological context of the local orbital methods described in Section 1.2.3 was performed by Heady et al. to compute the free energy of binding of five inhibitors to the cyclin-dependent kinase CDK2 [267]. The motivation for studying the target CDK2 is that it is over-expressed in cancer cells and may contribute to their unregulated growth [271]. However, there are around 500 protein kinases encoded in the human genome each with similar catalytic domains [272], hence the design of cancer therapeutics is hindered by the difficulty of designing inhibitors that bind specifically to CDK2 without unwanted side effects. To determine the relative free energy of binding of the five inhibitors, the authors make use of the free energy cycle shown in Figure 6(a). The total free energy of binding is written in terms of the gas (g) and solution (aq) phase contributions, and the former is further decomposed into enthalpic and entropic terms:

$$\Delta G = \Delta G_g + \Delta G_{aq} = \Delta E_g - T\Delta S_g + \Delta G_{aq} \tag{11}$$

The gas phase binding energy was accurately computed using large-scale DFT calculations. At the time, no model was available to compute the QM solvation free energies for these system sizes, and so they used the classical Generalised Born/Surface Area (GBSA) model with classical point charges. Since only relative free energies of binding were required:

$$\Delta\Delta G = \Delta\Delta E_g - T\Delta\Delta S_g + \Delta\Delta G_{aq}, \tag{12}$$

relative differences in the dispersion and entropic contributions to binding were neglected:

$$\Delta\Delta G = \Delta\Delta E_{DFT} + \Delta\Delta G_{aq} \tag{13}$$

The convergence of the DFT binding energies of the inhibitors to the complex with respect to system size was first ascertained. In agreement with subsequent publications, which are reviewed in Section 1.1, all residues within 7 Å of the inhibitor must be included before convergence is reached. Importantly, it was confirmed that binding energies are insensitive to the choice of O(N) DFT code (ONETEP [101] and SIESTA [55] were used), and that these codes also agreed with the plane-wave DFT code CASTEP [39] for smaller system sizes.

The first point to emphasise is that the application of eq 13 to optimised geometries from the x-ray crystal structures resulted in very poor agreement with experiment, with some errors in relative free energies in excess of 6 kcal/mol. This reinforces the need to explicitly consider finite temperature dynamics when computing free energies of binding. Dynamical simulations using a MM force field revealed two important features of the binding: i) in many cases, binding involves intermittent hydrogen bonds between the inhibitors and CDK2, and ii) water molecules are able to diffuse into the binding pocket and mediate protein-ligand interactions. In light of these findings, the gas phase binding energies were recomputed for a selection of snapshots that best captured the hydrogen bonding networks in the presence of explicit water molecules. The interaction strength was re-weighted by the fraction of time that a particular network was present in the MM simulations. With these corrections to account for finite temperature dynamics and solvation effects, the computed relative free energies of binding are in excellent agreement with experiment – the maximum error is just 1.2 kcal/mol. Considering that the considered inhibitors represent five very different structures, the agreement with experiment is extremely encouraging. The computational methods used are generally

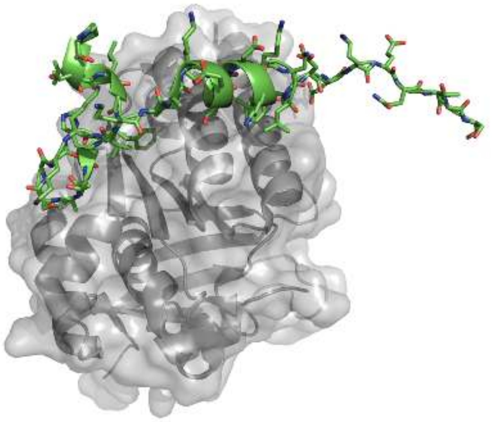

Figure 7. The protein–protein interface between RAD51 (grey) and the BRC4 repeat [276].

applicable to any biomolecular complex – indeed, similar applications have studied the binding of a peptide linked with protein misfolding diseases to a cyclic peptide [273] and a series of carbon nanostructures [274], and also the binding energies of a series of ligands to the FK506 binding protein (FKBP) [275]. However, in order to be used more widely, the methods would need to become less reliant on manual selection of representative structures and this was the focus of subsequent investigations.

Building on the work by Heady et al. [267], large-scale DFT calculations were employed to estimate binding free energies at a protein–protein interaction hotspot [268, 269]. Due to the large contact surface area between the proteins, complexes comprising around 2800 atoms were necessary to converge the energetic results [268]. The system chosen was the interface between the eukaryotic recombinase RAD51 and peptides derived from the protein BRCA2 (known as BRC repeats). BRCA2 is responsible for binding to and delivering RAD51 to sites of DNA damage, where RAD51 mediates the error-free repair of double-stranded DNA breaks [277, 278]. Mutations in BRCA2 have been linked to a predisposition to breast cancer and hence an accurate picture of the interactions at this protein–protein interface is therapeutically important. The interaction between RAD51 and BRCA2 is known to be mediated by eight BRC repeats, each containing subtle sequence variations, however only one crystal structure (between RAD51 and BRC4) has been solved (Figure 7) [276]. A palette of computational tools was therefore used to explore the interface between RAD51 and the BRC repeats (and their cancer-associated mutations) to provide insight into control of RAD51 by human BRCA2.

Molecular dynamics simulations of each of the BRC repeats in complex with RAD51 were run in solution. In close analogy with the classical molecular mechanics Poisson-Boltzmann/Surface Area (MM-PBSA) method [279], the free energy of binding is now computed by averaging the various components of the energy over the snapshots that are extracted from the MD simulation:

$$\Delta\Delta G = \Delta\langle\Delta E_g\rangle + \Delta\langle\Delta G_{aq}\rangle - T\Delta\langle\Delta S_g\rangle, \tag{14}$$

where $\langle\cdots\rangle$ represents an ensemble average. Hence, dynamical effects such as intermittent hydrogen bonds are naturally accounted for. In addition, the gas phase DFT binding energy is augmented by empirical dispersion interactions ($\Delta E_g = \Delta E_{DFT} + \Delta E_{disp}$) [140], and entropic contributions to binding were estimated using a classical normal modes analysis. In agreement with a fluorescence polarisation assay, which measures the ability of the BRC peptides to act as soluble inhibitors of the RAD51-BRC4 interaction, classical MM-PBSA predicted that BRC repeats 1, 2 and 4 bind with relatively high affinity while BRC repeats 3, 5, 6, 7 and 8 are more weakly bound. The calculations were repeated for repeats 1, 4 and 6 using large-scale DFT and eq 14, and the results were in accord with the classical force field approach. Subsequent natural bond orbital analysis of the binding interface between RAD51 and BRC4 identified a possible stabilisation mechanism arising from the delocalisation of electrons along the protein backbone from lone pairs to antibonding orbitals [224] (the so-called $n \to \pi^*$ interaction [280]). This is a good example of a case where relatively little structural information is available and a range of computational techniques is needed to understand both the dynamic and electronic structure effects that determine binding.

Finally, eq 14 was used to analyse the binding of eight small aromatic ligands to the engineered L99A/M102Q double mutant of T4 lysozyme [270]. The relative simplicity of this system makes it attractive for testing new computational methods [281, 282]. Following the methods described above, QM calculations were performed using the entire protein (more than 2600 atoms). Importantly, by using the implicit solvent model implemented in the ONETEP software [132, 133], the authors were able to compute both the gas phase and solvation contributions to the binding free energy using DFT. It was shown that the quantum mechanical estimates of the binding free energy were in markedly better agreement with experiment than the classical counterparts – the RMS error falls from 4.0 kcal/mol (MM) to 2.7 kcal/mol (QM), and 1-phenylsemicarbazide was correctly identified as a non-binder.

Overall, the studies discussed in this section were important in establishing the feasibility of performing routine large-scale DFT calculations on biomolecular assemblies. However, it is safe to conclude that inherent errors in eq 14 will prevent the determination of binding free energies at the 1 kcal/mol accuracy that is required to impact drug discovery programmes. Limitations include the approximate calculation of entropic contributions to binding, the use of implicit solvent (rather than explicitly modelling water molecules), the use of an approximate force field for sampling conformational space, and the assumption that the ligand samples the same conformational space in solution as it does in the bound complex. Hence, in the last three years, the focus has shifted to the use of large-scale DFT in formally exact free energy methods and this is discussed in the following two sections.

4.5.2. Towards Rigorous Quantum Free Energies of Binding. The use of formally rigorous free energy perturbation theory or thermodynamic integration methods for the computation of relative free energies of binding of small molecules to proteins according to Figure 6(b) is well-established [260]. In these schemes, molecule A is alchemically transformed into molecule B both in the protein and in water using a classical MM force field, yielding the free energies changes ΔG_X and ΔG_Y. The desired relative free energy of binding is then computed as:

$$\Delta\Delta G_{MM} = \Delta G_B - \Delta G_A = \Delta G_X - \Delta G_Y. \tag{15}$$

More recently, a number of schemes have been suggested that use an approximate, cheap potential to estimate the relative binding free energy ($\Delta\Delta G_{MM}$), and then compute a correction by evaluating the energies of selected snapshots with a more accurate, expensive Hamiltonian [283, 284]. This procedure computes the free energy correction via the evaluation of the Zwanzig equation [285]:

$$\Delta G_{MM\to QM} = -k_B T \ \ln\langle e^{-(E^{QM}-E^{MM})/k_B T}\rangle_{MM} \tag{16}$$

where $\langle\cdots\rangle_{MM}$ represents an ensemble average over structures obtained from the MM simulations, and E^{QM} and E^{MM} are the total energies of the system using the QM and MM Hamiltonians respectively. In the limit of infinite sampling, evaluation of eq 16 would recover the precise QM free energy change. However, in practice, the procedure is severely hampered by limitations in the phase space overlap of the QM and MM ensembles. To get around this, the QM and MM total energies in eq 16 are often replaced by the corresponding interaction energies (ΔE^{QM} and ΔE^{MM}) between the molecule under study and its environment [286, 134]. In effect, this makes the reasonable assumption that the MM description of intramolecular free energy changes are adequate.

As a surrogate for the full description of protein-ligand binding, Fox et al. used the methods described above to compute the QM hydration free energies of seven small organic molecules [134]. Hydration free energies are an important component of free energy cycles used in drug design and are often used for testing MM force field accuracy due to the relatively low sampling requirements [287]. The MM system comprised the small molecule, described by the generalised Amber force field (GAFF) [288] in a bath of 1545 water molecules, described by the TIP3P force field [289]. QM binding energies were computed using the ONETEP software [101] with the PBE exchange-correlation functional augmented by empirical dispersion corrections [140]. The QM system comprised the small molecule plus 200 water molecules (corresponding to a 9 Å solvation shell). The remaining water molecules were treated as classical embedding charges [14]. Using thermodynamic integration with the MM force field, the experimental hydration free energies were reproduced with a RMS error of 0.9 kcal/mol. Using the QM free energy correction tended to make no difference when the MM result was good, whilst improving outliers that deviate by more than $\sim k_B T$ from experiment. An exception to this trend is thiophenol, though there is evidence for poor convergence of the Zwanzig equation for this case, even for 180 sampled conformations. Thus, to

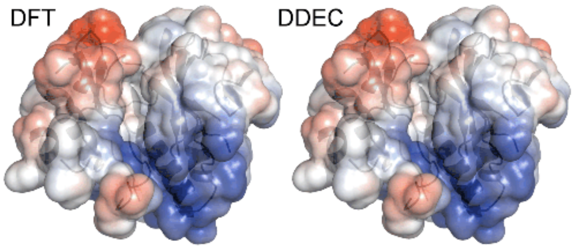

Figure 8. Electrostatic potentials on the solvent-accessible surface of the p38 kinase protein computed using large-scale DFT (left) and DDEC atom-centred point charges (right). The point charge electrostatic potential is virtually indistinguishable from the QM result.

improve the reliability of this method, a key goal is to improve the phase space overlap between the MM and QM ensembles and, hence, the convergence of eq 16.

In order to address these convergence issues in the computation of MM to QM correction free energies, Sampson et al. propose a "stepping stone" approach for the computation of quantum hydration free energies [290]. First, a hybrid Monte Carlo simulation generates a statistically-rigorous QM/MM ensemble of structures from the underlying MM ensemble [291]. Once this ensemble has been generated, a single-step free energy correction may be computed using the approach of Fox et al. to transform the QM/MM hydration free energies to the full QM result [134]. The idea behind the stepping stone approach is that, in the QM/MM ensemble, the polarisation of the ligand by the surrounding solvent is accounted for and, hence, it should provide a much closer representation of the target QM ensemble. In this way, the Zwanzig equation should converge more quickly than for a direct MM to QM correction. Using this method, the quantum free energies of hydration of five organic molecules were computed. Using a classical force field, the RMS error in the hydration free energies is 0.9 kcal/mol. Computing the QM/MM correction from the hybrid Monte Carlo simulation and adding it to the MM energies reduces the error to 0.7 kcal/mol. Finally, performing the QM/MM to full QM perturbation, the authors observe very good convergence of the free energy correction and the RMS error reduces still further to 0.5 kcal/mol. The low error in the full QM hydration free energies is extremely encouraging and it will be interesting to see if the same accuracy is obtained for larger benchmark sets and in the computation of protein–ligand binding free energies.

4.5.3. Classical Force Field Parameterisation. A possible disadvantage of the methods that have been discussed up to now is the computational cost associated with the multiple large-scale electronic structure evaluations that are required to obtain converged free energies. Classical molecular mechanics force fields are, of course, orders of magnitude less expensive and are much more widely used in the field of computer-aided drug design. However, MM force fields typically treat intermolecular interactions using a limited library of empirical parameters. Fitting of these parameters is extremely time-consuming and, because the parameters are developed for small molecules, they do not account for polarisation in larger molecules. That is, they are developed to be transferable rather than to be specific to the system under study.

For small molecule force field design, it is commonplace to fit atomic charges to the QM electrostatic potential (ESP) of the molecule of interest, thus naturally accounting for polarisation effects. However, ESP analysis cannot be applied to large systems with buried atoms, such as proteins. To address this problem in recent years, we have started to investigate the use of atoms-in-molecule (AIM) electron density partitioning in the design of a new class of environment-specific biological force fields [292, 234, 293]. AIM methods partition the total QM electron density into atomic basins. There is no unique method to perform this analysis, but we favour the density derived electrostatic and chemical (DDEC) partitioning method developed by Manz and Sholl [294, 295, 296, 297]. Amongst its many advantages, the method has been shown to yield atomic charges that are chemically reasonable and, vitally for force field design, it yields a rapidly convergent multipole expansion of the underlying QM electrostatic potential. We have implemented the DDEC methodology in the ONETEP linear-scaling DFT code [101], which allows us to perform AIM analysis of systems comprising many thousands of atoms [292, 234]. We have derived environment-specific atomic charges for a number of small proteins, including ubiquitin and lysozyme, and shown that there is a good correlation between the DDEC charges and standard transferable force field charges. However, the protein-specific charges show a greater spread, since they are able to respond to their environment and, as demonstrated in Figure 8, reproduce extremely well the QM electrostatic potential at the surfaces of the proteins (RMS errors of 1.3–1.5 kcal/mol compared to errors of 5–7 kcal/mol for standard force fields) [292, 293]. We have also computed DDEC charges for nine NMR conformers of bovine pancreatic trypsin inhibitor (BPTI). The charges show very good stability with respect to small positional fluctuations but vary more in flexible regions of the protein. Thus, they are suitable for flexible force field design whilst retaining the attractive feature of being able to respond to large conformational or environmental changes [292].

Importantly, it has been shown that environmental polarisation affects not only atomic charges, but also the strength of intermolecular van der Waals (vdW) interactions [141, 142, 298]. Thus, in theory both charge and vdW parameters used in MM force fields should be able to adapt to their environment. It has recently been shown that *all* components of the nonbonded force field may be derived directly from AIM electron density partitioning [293]. This scheme ensures compatibility between the derived charges and vdW parameters and has a very small number of fitting parameters (only five for a protein, compared to many tens or hundreds in standard force fields), thus substantially simplifying the force field parameterisation process. Furthermore, the method is applicable to arbitrarily large system sizes and, thus, a full protein-specific nonbonded force field was derived for a substantial portion of the L99A mutant of T4 lysozyme (1646 atoms) [293]. It was shown that the approximate treatment of vdW interactions by standard force fields using a limited parameter library is a rather crude representation of reality. Furthermore, the relative free energy of binding of two small molecules (benzofuran and indole) was computed using the cycle in Figure 6(b) and the new protein-specific force field (−0.37 kcal/mol), and found to be in excellent agreement with experiment (−0.57 kcal/mol) [299].

Summary: If the full accuracy of large-scale DFT is to be utilised in the calculation of binding affinities of biomolecular complexes then it should be used as part of a rigorous free energy protocol. The most promising schemes are i) the computation of corrections to cheaper Hamiltonians to evaluate rigorous quantum binding free energies [290] and ii) the use of large-scale DFT to compute environment-specific parameters for less expensive model Hamiltonians [293]. Whilst extremely encouraging, many more validation studies will be required before we are able to assess the accuracy and speed of these methods for truely predictive computer-aided drug design applications.

5. Outlook and Conclusions

We have identified four requirements for general-purpose biomolecular modelling with DFT: i) the ability to treat very large system sizes; ii) high-accuracy methods; iii) the prediction of properties going beyond the ground state electron density, such as excited states and reaction pathways; and iv) conformational sampling. One of our main motivations for writing this review is a growing consensus that very large quantum mechanical regions are required to converge many observables in biological simulations. This is best exemplified by detailed studies examining the convergence of activation energy barriers in enzymes [16] and optical absorption spectra of pigment-protein complexes [20]. Interestingly, the size regime at which these properties start to converge (around 500+ atoms) is similar to the crossover point at which linear-scaling DFT methods start to become more computationally efficient than traditional approaches. We have highlighted several strategies by which DFT calculations are able to access this size regime with a feasible computational effort, including the use of a localised orbital basis set and a density matrix representation. State-of-the-art DFT codes also show excellent parallel scaling, and wall times for the simulation of large-scale biological systems on the order of one hour or even less are commonplace.

As for any computational methodology, the balance between accuracy and expense is a crucial one. If large-scale DFT calculations are to become the benchmark of choice for biomolecular simulations (in preference to, for example, MM force field approaches) then it is important to make judicious choices of both the basis set size and the exchange-correlation functional. For computational feasibility, many of the structural optimisations reported in this review use numerical or Gaussian type atomic orbitals and functionals based on the generalised gradient approximation. It is not clear at this stage how much effect basis set superposition errors have on such calculations and it will be interesting, in future, to compare with larger basis sets and *in situ* optimised local orbitals, which alleviate this effect. Similarly, exploration of different exchange-correlation functionals are often beyond the scope of these initial exploratory studies, and it will be interesting to further examine the effects of, for example, hybrid functionals in the description of bond-breaking/forming reactions and range-separated functionals in spectroscopic calculations. However, accuracy considerations extend beyond just the computational parameters, to the details of the system preparation. Just as important in determining the electronic structure in large-scale cluster calculations is a physically-reasonable treatment of the system's environment. We have shown that the use of an implicit solvent model is vitally important in ensuring the presence of a HOMO-LUMO gap (and convergence of the self-consistent field equations) in cluster models of biological molecules [184]. In this respect, an often overlooked factor in large-scale DFT software development is the requirement to code advanced functionalities that go beyond total energy calculations and enable advanced treatments of electrostatics, spectroscopy, dispersion, strongly correlated electronic effects, chemical analysis and so forth. Alongside these developments, it is important to consider the accessibility of the designed software, so that calculations of the type described in this review can be put in the hands of the biologist, rather than being limited to trained electronic structure experts.

Within the limits defined by our criteria (namely that the QM region comprises more than around 500 atoms and is treated concurrently in a single calculation), we have given an overview of the state-of-the-art in large-scale biological DFT calculations. It is extremely encouraging that the computation of the electronic properties, structural optimisation, and even molecular dynamics, of entire proteins is fully feasible with today's software and computing resources [194, 195, 196]. Possible future lines of enquiry are numerous and include, for example, the study of unusual non-covalent bonding interactions in biological molecules that are not accounted for by MM force fields [280], further investigation of the exploitation of entangled spin states in metalloproteins [164, 165], study of long-range electronic transport in microbial nanowires [300], investigations into the effects of mutagenesis on enzymatic reaction pathways [206] and the high-accuracy refinement of experimental structural data [189]. It is worth noting that many of the applications in this review were chosen so as to minimise conformational sampling requirements, and this is still a major obstacle in many systems. In this respect, it is encouraging that rigorous converged free energy calculations are being demonstrated in the fields of computational enzymology [226] and protein–ligand binding [290, 293]. Many of these methods are reaching advanced stages and often all that remains is to bring down the cost of the simulations and demonstrate accuracy on larger data sets.

Given the successes of some of the paradigmatic examples described here, it is interesting to consider whether large-scale DFT is ready for truly first principles *predictive* modelling. Many of the challenges that lie ahead will require combined use of many (or all) of the described functionalities. For example, cytochrome P450s are of significant pharmaceutical interest, because of their role in drug metabolism [208]. A full investigation of their mode of action would incorporate large-scale modelling of the protein, alongside strongly correlated effects to describe their haem centres and transition state searching to elucidate their mode of action on chemical substrates. Meanwhile, the manganese-based water-oxidising cluster of photosystem II is a remarkable photocatalyst that harnesses light energy to split water [301] – potentially key to the elucidation of its mechanism of action are accurate computational studies of its spectroscopic and catalytic properties. We find that to date it has been rare for all of these functionalities to have been combined in a given study, but that ongoing developments hold the promise to make this a reality.

Acknowledgements

DJC is supported by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme. We are grateful to Chris-Kriton Skylaris, Peter Haynes, Jacek Dziedzic, Tim Zuehlsdorff, Greg Lever, Edward Linscott and Alice Allen for helpful discussions. This work used the ARCHER UK National Supercomputing Service (http://www.archer.ac.uk).

References

[1] R. O. Jones. Density functional theory: Its origins, rise to prominence, and future. *Rev. Mod. Phys.*, 87:897–923, 2015.

[2] D. R. Bowler, T. Miyazaki, and M. J. Gillan. Recent progress in linear scaling *ab initio* electronic structure techniques. *J. Phys.: Condens. Matter*, 14:2781–2798, 2002.

[3] D. R. Bowler and T. Miyazaki. O(N) methods in electronic structure calculations. *Rep. Prog. Phys.*, 75:036503, 2012.

[4] H. M. Senn and W. Thiel. QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed.*, 48:1198–1229, 2009.

[5] J. Tomasi, B. Mennucci, and R. Cammi. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.*, 105:2999–3094, 2005.

[6] I. Duchemin, D. Jacquemin, and X. Blase. Combining the GW formalism with the polarizable continuum model: A state-specific non-equilibrium approach. *J. Chem. Phys.*, 144:164106, 2016.

[7] E. Boulanger and W. Thiel. Toward QM/MM simulation of enzymatic reactions with the Drude oscillator polarizable force field. *J. Chem. Theory Comput.*, 10:1795–1809, 2014.

[8] S. Caprasecca, S. Jurinovich, L. Viani, C. Curutchet, and B. Mennucci. Geometry optimization in polarizable QM/MM models: The induced dipole formulation. *J. Chem. Theory Comput.*, 10:1588–1598, 2014.

[9] J. M. H. Olsen, C. Steinmann, K. Ruud, and J. Kongsted. Polarizable density embedding: A new QM/QM/MM-based computational strategy. *J. Phys. Chem. A*, 119:5344–5355, 2015.

[10] L. Hu, P. Söderhjelm, and U. Ryde. On the convergence of QM/MM energies. *J. Chem. Theory Comput.*, 7:761–777, 2011.

[11] K. E. Shaw, C. J. Woods, and A. J. Mulholland. Compatibility of quantum chemical methods and empirical (MM) water models in quantum mechanics/ molecular mechanics liquid water simulations. *J. Phys. Chem. Lett.*, 1:219–223, 2010.

[12] I. Solt, P. Kulhànek, I. Simon, S. Winfield, M. C. Payne, G. Csànyi, and M. Fuxreiter. Evaluating boundary dependent errors in QM/MM simulations. *J. Phys. Chem. B*, 113:5728–5735, 2009.

[13] L. Bondesson, E. Rudberg, Y. Luo, and P. Salek. A linear scaling study of solvent-solute interaction energy of drug molecules in aqua solution. *J. Phys. Chem. B*, 111:10320–10328, 2007.

[14] S. J. Fox, C. Pittock, T. Fox, C. Tautermann, N. Malcolm, and C.-K. Skylaris. Electrostatic embedding in large-scale first principles quantum mechanical calculations on biomolecules. *J. Chem. Phys.*, 135:224107, 2011.

[15] C. V. Sumowski, B. B. T. Schmitt, S. Schweizer, and C. Ochsenfeld. Quantum-chemical and combined quantum-chemical/molecular-mechanical studies on the stabilization of a twin arginine pair in adenovirus ad11. *Angew. Chem. Int. Ed.*, 49:9951–9955, 2010.

[16] K. Sadeghian, D. Flaig, I. D. Blank, S. Schneider, R. Strasser, D. Stathis, M. Winnacker, T. Carell, and C. Ochsenfeld. Ribose-protonated DNA base excision repair: A combined theoretical and experimental study. *Angew. Chem. Int. Ed.*, 53:10044–10048, 2014.

[17] R.-Z. Liao and W. Thiel. Convergence in the QM-only and QM/MM modeling of enzymatic reactions: A case study for acetylene hydratase. *J. Comput. Chem.*, 34:2389–2397, 2013.

[18] H. J. Kulik, J. Zhang, J. P. Klinman, and T. J. Martinez. How large should the QM region be in QM/MM calculations? The case of catechol-O-methyltransferase. 2015. arXiv:1505.05730.

[19] D. J. Cole, A. W. Chin, N. D. M. Hine, P. D. Haynes, and M. C. Payne. Toward *ab initio* optical spectroscopy of the Fenna-Matthews-Olson complex. *J. Phys. Chem. Lett.*, 4:4206–4212, 2013.

[20] C. M. Isborn, A. W. Götz, M. A. Clark, R. C. Walker, and T. J. Martínez. Electronic absorption spectra from MM and *ab initio* QM/MM molecular dynamics: Environmental effects on the absorption spectrum of photoactive yellow protein. *J. Chem. Theory Comput.*, 8:5092–5106, 2012.

[21] T. J. Zuehlsdorff, P. D. Haynes, F. Hanke, M. C. Payne, and N. D. M. Hine. Solvent effects on electronic excitations of an organic chromophore. *J. Chem. Theory Comput.*, 12:1853–1861, 2016.

[22] D. Flaig, M. Beer, and C. Ochsenfeld. Convergence of electronic structure with the size of the QM region: Example of QM/MM NMR shieldings. *J. Chem. Theory Comput.*, 8:2260–2271, 2012.

[23] E. R. Johnson and G. A. DiLabio. Convergence of calculated nuclear magnetic resonance chemical shifts in a protein with respect to quantum mechanical model size. *J. Mol. Struct.: Theochem*, 898:56–61, 2009.

[24] M. Retegan, F. Neese, and D. A. Pantazis. Convergence of QM/MM and cluster models for the spectroscopic properties of the oxygen-evolving complex in photosystem II. *J. Chem. Theory Comput.*, 9:3832–3842, 2013.

[25] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864, 1964.

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

[27] P. E Blöchl. Projector augmented-wave method. *Phys. Rev. B*, 50:17953, 1994.

[28] W. Yang and T.-S. Lee. A density-matrix divide-and-conquer approach for electronic structure calculations of large molecules. *J. Chem. Phys.*, 103:5674–5678, 1995.

[29] V. Gogonea and K. M. Merz, Jr. Fully quantum mechanical description of proteins in solution. Combining linear scaling quantum mechanical methodologies with the Poisson-Boltzmann equation. *J. Phys. Chem. A*, 103:5171–5188, 1999.

[30] K. Morokuma. Molecular orbital studies of hydrogen bonds. III. C=O··· H–O hydrogen bond in $H_2CO\cdots H_2O$ and $H_2CO\cdots 2H_2O$. *J. Chem. Phys.*, 55:1236–1244, 1971.

[31] K. Kitaura, E. Ikeo, T. Asada, T. Nakano, and M. Uebayasi. Fragment molecular orbital method: An approximate computational method for large molecules. *Chem. Phys. Lett.*, 313:701–706, 1999.

[32] M. S. Gordon, D. G. Fedorov, S. R. Pruitt, and L. V. Slipchenko. Fragmentation methods: A route to accurate calculations on large systems. *Chem. Rev.*, 112:632–672, 2012.

[33] D. G. Fedorov, T. Nagata, and K. Kitaura. Exploring chemistry with the fragment molecular orbital method. *Phys. Chem. Chem. Phys.*, 14:7562–7577, 2012.

[34] P. Pulay. *Ab initio* calculation of force constants and equilibrium geometries in polyatomic molecules. *Mol. Phys.*, 17:197–204, 1969.

[35] K. Lejaeghere, G. Bihlmayer, T. Björkman, P. Blaha, S. Blügel, V. Blum, D. Caliste, I. E. Castelli, Stewart J. Clark, A. Dal Corso, S. de Gironcoli, T. Deutsch, J. K. Dewhurst, C. Di Marco, I .and Draxl, M. Dułak, O. Eriksson, J. A. Flores-Livas, K. F. Garrity, L. Genovese, P. Giannozzi, M. Giantomassi, S. Goedecker, X. Gonze, O. Grånäs, E. K. U. Gross, A. Gulans, F. Gygi, D. R. Hamann, P. J. Hasnip, N. A. W. Holzwarth, D. Iuşan, Dominik B. Jochym, F. Jollet, D. Jones, G. Kresse, K. Koepernik, E. Küçükbenli, Y. O. Kvashnin, I. L. M. Locht, S. Lubeck, M. Marsman, N. Marzari, U. Nitzsche, L. Nordström, T. Ozaki, L. Paulatto, C. J. Pickard, W. Poelmans, M. I. J. Probert, M. Refson, K .and Richter, G.-M. Rignanese, S. Saha, M. Scheffler, M. Schlipf, K. Schwarz, S. Sharma, F. Tavazza, P. Thunström, A. Tkatchenko, M. Torrent, D. Vanderbilt, M. J. van Setten, V. Van Speybroeck, J. M. Wills, J. R. Yates, G.-X. Zhang, and S. Cottenier. Reproducibility in density functional theory calculations of solids. *Science*, 351:aad3000, 2016.

[36] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comp.*, 19:297–301, 1965.

[37] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. Dal Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys.: Cond. Matt.*, 21:395502, 2009.

[38] X. Gonze, B. Amadon, P. M. Anglade, J. M. Beuken, F. Bottin, P. Boulanger, F. Bruneval, D. Caliste, R. Caracas, M. Côté, T. Deutsch, L. Genovese, Ph. Ghosez, M. Giantomassi, S. Goedecker, D. R. Hamann, P. Hermet, F. Jollet, G. Jomard, S. Leroux, M. Mancini, S. Mazevet, M. J. T. Oliveira, G. Onida, Y. Pouillon, T. Rangel, G. M. Rignanese, D. Sangalli, R. Shaltaf, M. Torrent, M. J. Verstraete, G. Zerah, and J. W. Zwanziger. ABINIT: First-principles approach to material and nanosystem properties. *Comp. Phys. Commun.*, 180:2582–2615, 2009.

[39] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. J. Probert, K. Refson, and M. C. Payne. First principles methods using CASTEP. *Z. Kristallogr.*, 220:567–570, 2005.

[40] G. Kresse and J. Hafner. *Ab initio* molecular dynamics for liquid metals. *Phys. Rev. B*, 47:558–561, 1993.

[41] T. L. Beck. *Real-Space and Multigrid Methods in Computational Chemistry*, pages 223–285. John Wiley & Sons, Inc., 2009.

[42] J. J. Mortensen, L. B. Hansen, and K. W. Jacobsen. Real-space grid implementation of the projector augmented wave method. *Phys. Rev. B*, 71:035109, 2005.

[43] J. Enkovaara, C. Rostgaard, J. J. Mortensen, J. Chen, M. Dulak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. A. Hansen, H. H. Kristoffersen, M. Kuisma, A. H. Larsen, L. Lehtovaara, M. Ljungberg, O. Lopez-Acevedo, P. G. Moses, J Ojanen, T. Olsen, V. Petzold, N. A. Romero, J. Stausholm-Mller, M. Strange, G. A. Tritsaris, M. Vanin, M. Walter, B Hammer, H. Häkkinen, G. K. H. Madsen, R. M. Nieminen, J. K. Nrskov, M. Puska, T. T. Rantala, J. Schitz, K. S. Thygesen, and K. W. Jacobsen. Electronic structure calculations with GPAW: a real-space implementation of the projector augmented-wave method. *J. Phys.: Condens. Matter*, 22:253202, 2010.

[44] E. L. Briggs, D. J. Sullivan, and J. Bernholc. Real-space multigrid-based approach to large-scale electronic structure calculations. *Phys. Rev. B*, 54:14362–14375, 1996.

[45] M. Hodak, S. Wang, W. Lu, and J. Bernholc. Implementation of ultrasoft pseudopotentials in large-scale grid-based electronic structure calculations. *Phys. Rev. B*, 76:085108, 2007.

[46] M. Hodak, W. Lu, and J. Bernholc. Hybrid *ab initio* Kohn-Sham density functional theory/frozen-density orbital-free density functional theory simulation method suitable for biological systems. *J. Chem. Phys.*, 128:014101, 2008.

[47] J. R. Chelikowsky, N. Troullier, and Y. Saad. Finite-difference-pseudopotential method: Electronic structure calculations without a basis. *Phys. Rev. Lett.*, 72:1240–1243, 1994.

[48] Y. Zhou, Y. Saad, M. L. Tiago, and J. R. Chelikowsky. Self-consistent-field calculations using Chebyshev-filtered subspace iteration. *J. Comp. Phys.*, 219:172–184, 2006.

[49] Y. Liu, D. A. Yarne, and M. E. Tuckerman. *Ab initio* molecular dynamics calculations with simple, localized, orthonormal real-space basis sets. *Phys. Rev. B*, 68:125110, 2003.

[50] H.-S. Lee and M. E. Tuckerman. *Ab Initio* molecular dynamics with discrete variable representation basis sets: Techniques and application to liquid water. *J. Phys. Chem. A*, 110:5549–5560, 2006.

[51] M. Hodak, R. Chisnell, W. Lu, and J. Bernholc. Functional implications of multistage copper binding to the prion protein. *Proc. Natl. Acad. Sci. U.S.A.*, 106:11576–11581, 2009.

[52] M. Hodak and J. Bernholc. Insights into prion protein function from atomistic simulations. *Prion*, 4:13–19, 2010.

[53] M. A. L. Marques, A. Castro, G. F. Bertsch, and A. Rubio. octopus: a first-principles tool for excited electron-ion dynamics. *Comput. Phys. Commun.*, 151:60–78, 2003.

[54] O. F. Sankey and D. J. Niklewski. *Ab initio* multicenter tight-binding model for molecular-dynamics simulations and other applications in covalent systems. *Phys. Rev. B*, 40:3979–3995, 1989.

[55] J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal. The SIESTA method for *ab initio* order-N materials simulation. *J. Phys.: Condens. Matter*, 14:2745–2779, 2002.

[56] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler. Ab initio molecular simulations with numeric atom-centered orbitals. *Comp. Phys. Commun.*, 180:2175–2196, 2009.

[57] T. Ozaki. Variationally optimized atomic orbitals for large-scale electronic structures. *Phys. Rev. B*, 67:155108, 2003.

[58] B. Delley. From molecules to solids with the DMol3 approach. *J. Chem. Phys.*, 113:7756–7764, 2000.

[59] E. Hernández, M. J. Gillan, and C. M. Goringe. Basis functions for linear-scaling first-principles calculations. *Phys. Rev. B*, 55:13485–13493, 1997.

[60] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian09 Revision E.01. Gaussian Inc. Wallingford CT 2009.

[61] R. Dovesi, R. Orlando, A. Erba, C. M. Zicovich-Wilson, B. Civalleri, S. Casassa, L. Maschio, M. Ferrabone, M. De La Pierre, P. D'Arco, Y. Noël, M. Causà, M. Rérat, and B. Kirtman. CRYSTAL14: A program for the ab initio investigation of crystalline solids. *International Journal of Quantum Chemistry*, 114:1287–1317, 2014.

[62] M. Valiev, E. J. Bylaska, N. Govind, K. Kowalski, T. P. Straatsma, H. J. J. Van Dam, D. Wang, J. Nieplocha, E. Apra, T. L. Windus, and W. A. de Jong. NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comp. Phys. Commun.*, 181:1477–1489, 2010.

[63] N. Bock, M. Challacombe, C. K. Gan, G. Henkelman, K. Nemeth, A. M. N. Niklasson, A. Odell,

E. Schwegler, C. J. Tymczak, and V. Weber. FreeON, 2014. Los Alamos National Laboratory (LA-CC 01-2; LA-CC-04-086), Copyright University of California.

[64] I. S. Ufimtsev and T. J. Martinez. Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics. *J. Chem. Theory Comput.*, 5:2619–2628, 2009.

[65] J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing, and J. Hutter. Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach. *Comp. Phys. Commun.*, 167:103–128, 2005.

[66] G. Lippert, J. Hutter, and M. Parrinello. The Gaussian and augmented-plane-wave density functional method for ab initio molecular dynamics simulations. *Theor. Chem. Acc.*, 103:124–140, 1999.

[67] S. F. Boys and F. Bernardi. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Mol. Phys.*, 19:553–566, 1970.

[68] X. Andrade and A. Aspuru-Guzik. Real-Space Density Functional Theory on Graphical Processing Units: Computational Approach and Comparison to Gaussian Basis Set Methods. *J. Chem. Theory Comput.*, 9:4360–4373, 2013.

[69] S. Hakala, V. Havu, J. Enkovaara, and R. Nieminen. Parallel Electronic Structure Calculations Using Multiple Graphics Processing Units (GPUs). In Pekka Manninen and Per Öster, editors, *Applied Parallel and Scientific Computing*, number 7782 in Lecture Notes in Computer Science, pages 63–76. Springer Berlin Heidelberg, 2012. DOI: 10.1007/978-3-642-36803-5_4.

[70] L. Wang, Y. Wu, W. Jia, W. Gao, X. Chi, and L. W. Wang. Large scale plane wave pseudopotential density functional theory calculations on GPU clusters. In *2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, pages 1–10, November 2011.

[71] F. Spiga and I. Girotto. phiGEMM: A CPU-GPU Library for Porting Quantum ESPRESSO on Hybrid Systems. In *2012 20th Euromicro International Conference on Parallel, Distributed and Network-based Processing*, pages 368–375, February 2012.

[72] N. Luehr, A. G. B. Jin, and T. J. Martínez. Ab Initio Interactive Molecular Dynamics on Graphical Processing Units (GPUs). *J. Chem. Theory Comput.*, 11:4536–4544, 2015.

[73] C. M. Isborn, N. Luehr, I. S. Ufimtsev, and T. J. Martínez. Excited-State Electronic Structure with Configuration Interaction Singles and Tamm–Dancoff Time-Dependent Density Functional Theory on Graphical Processing Units. *J. Chem. Theory Comput.*, 7:1814–1823, 2011.

[74] F. Liu, N. Luehr, H. J. Kulik, and T. J. Martínez. Quantum Chemistry for Solvated Molecules on Graphical Processing Units Using Polarizable Continuum Models. *J. Chem. Theory Comput.*, 11:3131–3144, 2015.

[75] F. Corsetti. Performance Analysis of Electronic Structure Codes on HPC Systems: A Case Study of SIESTA. *PLOS ONE*, 9:e95390, 2014.

[76] J. VandeVondele, U. Borštnik, and J. Hutter. Linear Scaling Self-Consistent Field Calculations with Millions of Atoms in the Condensed Phase. *J. Chem. Theory Comput.*, 8:3565–3573, 2012.

[77] W. Hu, L. Lin, and C. Yang. DGDFT: A massively parallel method for large scale density functional theory calculations. *J. Chem. Phys.*, 143:124110, 2015.

[78] S. Goedecker. Linear scaling electronic structure methods. *Rev. Mod. Phys.*, 71:1085–1123, 1999.

[79] W. Yang. Direct calculation of electron density in density-functional theory. *Phys. Rev. Lett.*, 66:1438–1441, 1991.

[80] W. Yang and T.-S. Lee. A density-matrix divide-and-conquer approach for electronic structure calculations of large molecules. *J. Chem. Phys.*, 103:5674–5678, 1995.

[81] D. M. York, T.-S. Lee, and W. Yang. Quantum Mechanical Treatment of Biological Macromolecules in Solution Using Linear-Scaling Electronic Structure Methods. *Phys. Rev. Lett.*, 80:5011–5014, 1998.

[82] W. Kohn. Density functional/Wannier function theory for systems of very many atoms. *Chem. Phys. Lett.*, 208:167–172, 1993.

[83] W. Kohn. Density Functional and Density Matrix Method Scaling Linearly with the Number of Atoms. *Phys. Rev. Lett.*, 76:3168–3171, 1996.

[84] X.-P. Li, R. W. Nunes, and D. Vanderbilt. Density-matrix electronic-structure method with linear system-size scaling. *Phys. Rev. B*, 47:10891–10894, 1993.

[85] M. Chen, J. Xia, C. Huang, J. M. Dieterich, L. Hung, I. Shin, and E. A. Carter. Introducing PROFESS 3.0: An advanced program for orbital-free density functional theory molecular dynamics simulations. *Comput. Phys. Commun.*, 190:228–230, 2015.

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

[87] C. Brouder, G. Panati, M. Calandra, C. Mourougane, and N. Marzari. Exponential Localization of Wannier Functions in Insulators. *Phys. Rev. Lett.*, 98:046402, 2007.

[88] E. Hernandez and M. J. Gillan. Self-consistent first-principles technique with linear scaling. *Phys. Rev. B*, 51:10157–10160, 1995.

[89] H. Shang, H. Xiang, Z. Li, and J. Yang. Linear scaling electronic structure calculations with numerical atomic basis set. *Int. Rev. Phys. Chem.*, 29:665–691, 2010.

[90] S. Goedecker and L. Colombo. Efficient Linear Scaling Algorithm for Tight-Binding Molecular Dynamics. *Phys. Rev. Lett.*, 73:122–125, 1994.

[91] S. Goedecker and M. Teter. Tight-binding electronic-structure calculations and tight-binding molecular dynamics with localized orbitals. *Phys. Rev. B*, 51:9455–9464, 1995.

[92] S. Goedecker. Low Complexity Algorithms for Electronic Structure Calculations. *J. Comput. Phys.*, 118:261–268, 1995.

[93] A. F. Voter, J. D. Kress, and R. N. Silver. Linear-scaling tight binding from a truncated-moment approach. *Phys. Rev. B*, 53:12733–12741, 1996.

[94] A. H. R. Palser and D. E. Manolopoulos. Canonical purification of the density matrix in electronic-structure theory. *Phys. Rev. B*, 58:12704–12711, 1998.

[95] A. M. N. Niklasson. Expansion algorithm for the density matrix. *Phys. Rev. B*, 66:155115, 2002.

[96] E. H. Rubensson, E. Rudberg, and P. Salek. Density matrix purification with rigorous error control. *J. Chem. Phys.*, 128:074106, 2008.

[97] P. D. Haynes and M. C. Payne. Corrected penalty-functional method for linear-scaling calculations within density-functional theory. *Phys. Rev. B*, 59:12173–12176, 1999.

[98] T. Ozaki and H. Kino. Numerical atomic basis orbitals from H to Kr. *Phys. Rev. B*, 69:195113, 2004.

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

[100] C. K. Skylaris, A. A. Mostofi, P. D. Haynes, O. Dieguez, and M. C. Payne. The non-orthogonal generalised wannier function pseudopotential plane-wave method. *Phys. Rev. B*, 66:035119, 2002.

[101] C. K. Skylaris, P. D. Haynes, A. A. Mostofi, and M. C. Payne. Introducing ONETEP: Linear-scaling density functional simulations on parallel computers. *J. Chem. Phys.*, 122:084119, 2005.

[102] J.-L. Fattebert and F. Gygi. Linear-scaling first-principles molecular dynamics with plane-waves accuracy. *Phys. Rev. B*, 73(11):115124, March 2006.

[103] J.-L. Fattebert. Adaptive localization regions for O(N) density functional theory calculations. *Journal of Physics: Condensed Matter*, 20(29):294210, 2008.

[104] A. S. Torralba, M. Todorović, V. Brázdová, R. Choudhury, T. Miyazaki, M. J. Gillan, and D. R. Bowler. Pseudo-atomic orbitals as basis sets for the O(N) DFT code CONQUEST. *Journal of Physics: Condensed Matter*, 20(29):294206, 2008.

[105] T. Otsuka, T. Miyazaki, T. Ohno, D. R. Bowler, and M. J. Gillan. Accuracy of order- N density-functional theory calculations on DNA systems using CONQUEST. *Journal of Physics: Condensed Matter*, 20(29):294201, 2008.

[106] M. Todorović, D. R. Bowler, M. J. Gillan, and T. Miyazaki. Density-functional theory study of gramicidin A ion channel geometry and electronic properties. *J. R. Soc. Interface*, 10:20130547, 2013.

[107] L. Genovese, A. Neelov, S. Goedecker, T. Deutsch, S. A. Ghasemi, A. Willand, D. Caliste, O. Zilberberg, M. Rayson, A. Bergman, and R. Schneider. Daubechies wavelets as a basis set for density functional pseudopotential calculations. *J. Chem. Phys.*, 129(1):014109, July 2008.

[108] S. Mohr, L. E. Ratcliff, P. Boulanger, L. Genovese, D. Caliste, T. Deutsch, and S. Goedecker. Daubechies wavelets for linear scaling density functional theory. *J. Chem. Phys.*, 140(20):204110, May 2014.

[109] D. Osei-Kuffuor and J.-L. Fattebert. Accurate and Scalable $O(N)$ Algorithm for First-Principles Molecular-Dynamics Computations on Large Parallel Computers. *Phys. Rev. Lett.*, 112(4):046401, January 2014.

[110] D. Osei-Kuffuor and J. Fattebert. A Scalable $O(N)$ Algorithm for Large-Scale Parallel First-Principles Molecular Dynamics Simulations. *SIAM Journal on Scientific Computing*, 36(4):C353–C375, January 2014.

[111] P. D. Haynes, C. K. Skylaris, A. A. Mostofi, and M. C. Payne. Density kernel optimisation in the onetep code. *J. Phys. Condens. Matter*, 20:294207, 2008.

[112] N. D. M. Hine, P. D. Haynes, A. A. Mostofi, C.-K. Skylaris, and M. C. Payne. Linear-scaling density-functional theory with tens of thousands of atoms: Expanding the scope and scale of calculations with ONETEP. *Comput. Phys. Commun.*, 180(7):1041–1053, July 2009.

[113] N. D. M. Hine, P. D. Haynes, A. A. Mostofi, and M. C. Payne. Linear-scaling density-functional simulations of charged point defects in Al[sub 2]O[sub 3] using hierarchical sparse matrix algebra. *J. Chem. Phys.*, 133:114111, 2010.

[114] C.-K. Skylaris and P. D. Haynes. Achieving plane wave accuracy in linear-scaling density functional theory applied to periodic systems: A case study on crystalline silicon. *J. Chem. Phys.*, 127:164712, 2007.

[115] P. D. Haynes, C. K. Skylaris, A. A. Mostofi, and M. C. Payne. Elimination of basis set superposition error in linear-scaling density-functional calculations with local orbitals optimised *in situ*. *Chem. Phys. Lett.*, 422:345–349, 2006.

[116] K. A. Wilkinson, N. D. M. Hine, and C.-K. Skylaris. Hybrid MPI-OpenMP Parallelism in the ONETEP Linear-Scaling Electronic Structure Code: Application to the Delamination of Cellulose Nanofibrils. *Journal of Chemical Theory and Computation*, 10(11):4782–4794, November 2014.

[117] N. D. M. Hine, M. Robinson, P. D. Haynes, C.-K. Skylaris, M. C. Payne, and A. A. Mostofi. Accurate ionic forces and geometry optimization in linear-scaling density-functional theory with local orbitals. *Phys. Rev. B*, 83(19):195102, May 2011.

[118] J. P. Perdew, K. Burke, and M. Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865–3868, 1996.

[119] J .T. Berryman, S. E. Radford, and S. A. Harris. Thermodynamic Description of Polymorphism in Q- and N-Rich Peptide Aggregates Revealed by Atomistic Simulation. *Biophysical Journal*, 97(1):1–11, 2009.

[120] G. J. Martyna and M. E. Tuckerman. A reciprocal space based method for treating long range interactions in ab initio and force-field-based calculations in clusters. *J. Chem. Phys.*, 110(6):2810, 1999.

[121] C. A. Rozzi, D. Varsano, E. K. U. Marini, A.and Gross, and A. Rubio. Exact coulomb cutoff technique for supercell calculations. *Phys. Rev. B*, 73(20):205119, May 2006.

[122] L. Genovese, T. Deutsch, A. Neelov, S. Goedecker, and G. Beylkin. Efficient solution of poisson's equation with free boundary conditions. *J. Chem. Phys.*, 125(7):074105, August 2006.

[123] L. Genovese, T. Deutsch, and S. Goedecker. Efficient and accurate three-dimensional poisson solver for surface problems. *J. Chem. Phys.*, 127(5):054704, 2007.

[124] I. Dabo, B. Kozinsky, N. E. Singh-Miller, and N. Marzari. Electrostatics in periodic boundary

conditions and real-space corrections. *Phys. Rev. B*, 77(11):115139, March 2008.

[125] N. D. M. Hine, J. Dziedzic, P. D. Haynes, and C.-K. Skylaris. Electrostatic interactions in finite systems treated with periodic boundary conditions: Application to linear-scaling density functional theory. *J. Chem. Phys.*, 135(20):204103, November 2011.

[126] A. Cerioni, L. Genovese, A. Mirone, and Vicente A. Sole. Efficient and accurate solver of the three-dimensional screened and unscreened Poisson's equation with generic boundary conditions. *J. Chem. Phys.*, 137(13):134108, October 2012.

[127] P. García-Risueño, J. Alberdi-Rodriguez, M. J. T. Oliveira, X. Andrade, M. Pippig, J. Muguerza, A. Arruabarrena, and A. Rubio. A survey of the parallel performance and accuracy of Poisson solvers for electronic structure calculations. *Journal of Computational Chemistry*, 35:427–444, 2014.

[128] G. Scalmani and M. J. Frisch. Continuous surface charge polarizable continuum models of solvation. I. General formalism. *J. Chem. Phys.*, 132(11):114110, March 2010.

[129] A. Klamt and G. Schüürmann. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2*, pages 799–805, 1993.

[130] A. V. Marenich, C. J. Cramer, and D. G. Truhlar. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B*, 113(18):6378–6396, May 2009.

[131] D. A. Scherlis, J.-L. Fattebert, D. Gygi, M. Cococcioni, and N. Marzari. A unified electrostatic and cavitation model for first-principles molecular dynamics in solution. *J. Chem. Phys.*, 124:074103, 2006.

[132] J. Dziedzic, H. H. Helal, C.-K. Skylaris, A. A. Mostofi, and M. C. Payne. Minimal parameter implicit solvent model for ab initio electronic structure calculations. *Europhys. Lett.*, 95:43001, 2011.

[133] J. Dziedzic, S. J. Fox, T. Fox, C. S. Tautermann, and C. K. Skylaris. Large-scale DFT calculations in implicit solvent – a case study on the T4 lysozyme L99A/M102Q protein. *Int. J. Quantum Chem.*, 113:771–785, 2013.

[134] S. J. Fox, C. Pittock, C. S. Tautermann, T. Fox, C. Christ, N. O. J. Malcolm, J. W. Essex, and C.-K. Skylaris. Free energies of binding from large-scale first-principles quantum mechanical calculations: Application to ligand hydration energies. *J. Phys. Chem. B*, 117:9478–9485, 2013.

[135] J. F. Dobson and T. Gould. Calculation of dispersion energies. *J. Phys.: Condens. Matter*, 24:073201, 2012.

[136] R. A. DiStasio, Jr., O. A. von Lilienfeld, and A. Tkatchenko. Collective many-body van der Waals interactions in molecular systems. *Proc. Natl. Acad. Sci. U.S.A.*, 109:14791–14795, 2012.

[137] L. Kronik and A. Tkatchenko. Understanding molecular crystals with dispersion-inclusive density functional theory: Pairwise corrections and beyond. *Acc. Chem. Res.*, 47:3208–3216, 2014.

[138] J. Klimes and A. Michaelides. Perspective: Advances and challenges in treating van der Waals dispersion forces in density functional theory. *J. Chem. Phys.*, 137:120901, 2012.

[139] S. Grimme. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.*, 27:1787–1799, 2006.

[140] Q. Hill and C. K. Skylaris. Including dispersion interactions in the onetep program for linear-scaling density functional theory calculations. *Proc. R. Soc. A*, 465:669, 2009.

[141] A. Tkatchenko and M. Scheffler. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.*, 102:073005, 2009.

[142] A. Tkatchenko, R. A. DiStasio, R. Car, and M. Scheffler. Accurate and efficient method for many-body van der Waals interactions. *Phys. Rev. Lett.*, 108:236402, 2012.

[143] L. Andrinopoulos, N. D. M. Hine, and A. A. Mostofi. Calculating dispersion interactions using maximally localized Wannier functions. *J. Chem. Phys.*, 135:154105, 2011.

[144] M. Dion, H. Rydberg, E. Schröder, D. C Langreth, and B. I Lundqvist. Van der Waals density functional for general geometries. *Phys. Rev. Lett.*, 92:246401, 2004.

[145] T. Van Voorhis and O. A Vydrov. Nonlocal van der Waals density functional made simple. *Phys. Rev. Lett.*, 103:063004, 2009.

[146] G. Román-Pérez and J. M. Soler. Efficient implementation of a van der Waals density functional: Application to double-wall carbon nanotubes. *Phys. Rev. Lett.*, 103:096102, 2009.

[147] K. Berland, V. R. Cooper, K. Lee, E. Schröder, T. Thonhauser, P. Hyldgaard, and B. I. Lundqvist. van der Waals forces in density functional theory: a review of the vdW-DF method. *Reports on Progress in Physics*, 78(6):066501, 2015.

[148] V. R. Cooper, T. Thonhauser, A. Puzder, E. Schröder, B. I. Lundqvist, and D. C. Langreth. Stacking interactions and the twist of DNA. *J. Am. Chem. Soc.*, 130:1304–1308, 2008.

[149] S. Li, V. R. Cooper, T. Thonhauser, B. I. Lundqvist, and D. C. Langreth. Stacking Interactions and DNA Intercalation. *J. Phys. Chem. B*, 113:11166–11172, 2009.

[150] K. J. Waldron, J. C. Rutherford, D. Ford, and N. J. Robinson. Metalloproteins and metal sensing. *Nature*, 460:823–830, 2009.

[151] A. J. Cohen, P. Mori-Sánchez, and W. Yang. Insights into Current Limitations of Density Functional Theory. *Science*, 321:792794, 2008.

[152] V. I. Anisimov, J. Zaanen, and O. K. Andersen. Band theory and Mott insulators: Hubbard U instead of Stoner I. *Phys. Rev. B*, 44:943–954, 1991.

[153] B. Himmetoglu, A. Floris, S. de Gironcoli, and M. Cococcioni. Hubbard-corrected DFT energy functionals: The LDA+U description of correlated systems. *Int. J. Quant. Chem.*, 114:14–49, 2014.

[154] H. J. Kulik. Perspective: Treating electron over-delocalization with the DFT+U method. *J. Chem. Phys.*, 142:240901, 2015.

[155] D. A. Scherlis, M. Cococcioni, P. Sit, and N. Marzari. Simulation of heme using DFT+U: A step toward accurate spin-state energetics. *J. Phys. Chem. B*, 111:73847391, 2007.

[156] M. Cococcioni and S. de Gironcoli. Linear response approach to the calculation of the effective interaction parameters in the LDA+U method. *Phys. Rev. B*, 71:035105, 2005.

[157] D. D. O'Regan, N. D. M. Hine, M. C. Payne, and A. A. Mostofi. Projector self-consistent dft+u using non-orthogonal generalized wannier functions. *Phys. Rev. B*, 82:081102(R), 2010.

[158] D. D. O'Regan, N. D. M. Hine, M. C. Payne, and A. A. Mostofi. Linear-scaling dft+u with full local orbital optimization. *Phys. Rev. B*, 85:085107, 2012.

[159] A. D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.*, 98:5648–5652, 1993.

[160] J. Heyd, G. E. Scuseria, and M. Ernzerhof. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.*, 118:8207–8215, 2003.

[161] C. Adamo and V. Barone. Toward reliable density functional methods without adjustable parameters: The pbe0 model. *J. Chem. Phys.*, 110:6158–6170, 1999.

[162] A. Georges, G. Kotliar, W. Krauth, and M. J. Rozenberg. Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions. *Rev. Mod. Phys.*, 68:13–125, 1996.

[163] C. Weber, D. D. O'Regan, N. D. M. Hine, M. C. Payne, G. Kotliar, and P. B. Littlewood. Vanadium dioxide : A Peierls-Mott insulator stable against disorder. *Phys. Rev. Lett.*, 108:256402, 2012.

[164] C. Weber, D. D. O'Regan, N. D. M. Hine, P. B. Littlewood, G. Kotliar, and M. C. Payne. Importance of many body effects in the kernel of hemoglobin for ligand binding. *Phys. Rev. Lett.*, 110:106402, 2013.

[165] C. Weber, D. J. Cole, D. D. O'Regan, and M. C. Payne. Renormalization of myoglobin-ligand binding energetics by quantum many-body effects. *Proc. Natl. Acad. Sci. U.S.A*, 111:5790–5795, 2014.

[166] K. Yabana and G. F. Bertsch. Time-dependent local-density approximation in real time. *Phys. Rev. B*, 54(7):4484–4487, August 1996.

[167] D. P. Chong. *M. E. Casida, in Recent Advances in Density Functional Methods.* World Scientific,

1995.

[168] D. Varsano, R. Di Felice, M. A. L. Marques, and A. Rubio. A TDDFT study of the excited states of DNA bases and their assemblies. *J. Phys. Chem. B*, 110(14):7129–7138, 2006.

[169] A. Castro, M. A. L. Marques, D. Varsano, F. Sottile, and A. Rubio. The challenge of predicting optical properties of biomolecules: What can we learn from time-dependent density-functional theory? *Comptes Rendus Physique*, 10(6):469–490, July 2009.

[170] X. Lopez, M. A. L. Marques, A. Castro, and A. Rubio. Optical absorption of the blue fluorescent protein: A first-principles study. *Journal of the American Chemical Society*, 127(35):12329–12337, 2005.

[171] J. Jornet-Somoza, J. Alberdi-Rodriguez, B. F. Milne, X. Andrade, M. A. L. Marques, F. Nogueira, M. J. T. Oliveira, J. J. P. Stewart, and A. Rubio. Insights into colour-tuning of chlorophyll optical response in green plants. *Phys. Chem. Chem. Phys.*, 17:26599–26606, 2015.

[172] T. J. Zuehlsdorff, N. D. M. Hine, J. S. Spencer, N. M. Harrison, D. J. Riley, and P. D. Haynes. Linear-scaling time-dependent density-functional theory in the linear response formalism. *J. Chem. Phys.*, 139:064104, 2013.

[173] T. J. Zuehlsdorff, N. D. M. Hine, M. C. Payne, and P. D. Haynes. 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.*, 143:204107, 2015.

[174] J.-H. Li, T. J. Zuehlsdorff, M. C. Payne, and N. D. M. Hine. Identifying and tracing potential energy surfaces of electronic excitations with specific character via their transition origins: application to oxirane. *Phys. Chem. Chem. Phys.*, 17:12065–12079, 2015.

[175] S. Tretiak, C. M. Isborn, A. M. N. Niklasson, and M. Challacombe. Representation independent algorithms for molecular response calculations in time-dependent self-consistent field theories. *J. Chem. Phys.*, 130(5):054111, February 2009.

[176] M. Challacombe. Linear Scaling Solution of the Time-Dependent Self-Consistent-Field Equations. *Computation*, 2(1):1–11, March 2014.

[177] C. O'Rourke and D. R. Bowler. Linear scaling density matrix real time TDDFT: Propagator unitarity and matrix truncation. *J. Chem. Phys.*, 143(10):102801, September 2015.

[178] ChiYung Yam, Satoshi Yokojima, and GuanHua Chen. Linear-scaling time-dependent density-functional theory. *Phys. Rev. B*, 68(15):153105, October 2003.

[179] E. Rudberg, E. H. Rubensson, and P. Salek. Kohn-Sham density functional theory electronic structure calculations with linearly scaling computational time and memory usage. *J. Chem. Theory Comput.*, 7:340–350, 2011.

[180] H. J. Kulik, N. Luehr, I. S. Ufimtsev, and T. J. Martinez. *Ab initio* quantum chemistry for protein structures. *J. Phys. Chem. B*, 116:12501–12509, 2012.

[181] J. Antony and S. Grimme. Fully *ab initio* protein-ligand interaction energies with dispersion corrected density functional theory. *J. Comput. Chem.*, 33:1730–1739, 2012.

[182] E. Rudberg. Difficulties in applying pure Kohn-Sham density functional theory electronic structure methods to protein molecules. *J. Phys.: Cond. Matt.*, 24:072202, 2012.

[183] E. H. Rubensson and E. Rudberg. Bringing about matrix sparsity in linear-scaling electronic structure calculations. *J. Comput. Chem.*, 32:1411–1423, 2011.

[184] G. Lever, D. J. Cole, N. D. M. Hine, P. D. Haynes, and M. C. Payne. Electrostatic considerations affecting the calculated HOMO-LUMO gap in protein molecules. *J. Phys.: Cond. Matt.*, 25:152101, 2013.

[185] C. M. Isborn, N. Luehr, I. S. Ufimtsev, and T. J. Martinez. Excited-state electronic structure with configuration interaction singles and Tamm Dancoff time-dependent density functional theory on graphical processing units. *J. Chem. Theory Comput.*, 7:1814–1823, 2011.

[186] C. M. Isborn, B. D. Mar, B. F. E. Curchod, I. Tavernelli, and T. J. Martinez. The charge transfer problem in density functional theory calculations of aqueously solvated molecules. *J. Phys. Chem. B*, 117:12189–12201, 2013.

[187] M. Rossi, V. Blum, P. Kupser, G. von Helden, F. Bierau, K. Pagel, G. Meijer, and M. Scheffler.

Secondary structure of ac-ala$_n$-lysh$^+$ polyalanine peptides ($n = 5, 10, 15$) in vacuo: Helical or not? *J. Phys. Chem. Lett.*, 1:3465–3470, 2010.

[188] U. Ryde, L. Olsen, and K. Nilsson. Quantum chemical geometry optimizations in proteins using crystallographic raw data. *J. Comput. Chem.*, 23:1058–1070, 2002.

[189] K. M. Merz Jr. Using quantum mechanical approaches to study biological systems. *Acc. Chem. Res.*, 47:2804–2811, 2014.

[190] N. Muzet, B. Guillot, C. Jelsch, E. Howard, and C. Lecomte. Electrostatic complementarity in an aldose reductase complex from ultra-high-resolution crystallography and first-principles calculations. *Proc. Natl. Acad. Sci. USA*, 100:8742–8747, 2003.

[191] A. Schmidt, M. Teeter, E. Weckert, and V. S. Lamzin. Crystal structure of small protein crambin at 0.48Å resolution. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.*, 67:424–428, 2011.

[192] C. Van Alsenoy, C.-H. Yu, A. Peeters, J. M. L. Martin, and L. Schäfer. *Ab Initio* geometry determinations of proteins. 1. Crambin. *J. Phys. Chem. A*, 102:2246–2251, 1998.

[193] M. V. Fernández-Serra, J. Junquera, C. Jelsch, C. Lecomte, and E. Artacho. Electron density in the peptide bonds of crambin. *Solid State Commun.*, 116:395–400, 2000.

[194] M. Delle Piane, M. Corno, R. Orlando, R. Dovesi, and P. Ugliengo. Elucidating the fundamental forces in protein crystal formation: the case of crambin. *Chem. Sci.*, 7:1496–1507, 2015.

[195] I. S. Ufimtsev, N. Luehr, and T. J. Martinez. Charge transfer and polarisation in solvated proteins from *ab initio* molecular dynamics. *J. Phys. Chem. Lett.*, 2:1789–1793, 2011.

[196] G. T. Feliciano, A. J. R. da Silva, G. Reguera, and E. Artacho. Molecular and electronic structure of the peptide subunit of *geobacter sulfurreducens* conductive pili from first principles. *J. Phys. Chem. A*, 116:8023–8030, 2012.

[197] V. B. Myers and D. Haydon. Ion transfer across lipid membranes in the presence of gramicidin A: II. The ion selectivity. *Biochim. Biophys. Acta*, 274:313–322, 1972.

[198] M. D. Becker, D. V. Greathouse, R. E. Koeppe, and O. S. Andersen. Amino acid sequence modulation of gramicidin channel function: effects of tryptophan-to-phenylalanine substitutions on the single-channel conductance and duration. *Biochemistry*, 30:8830–8839, 1991.

[199] P. J. de Pablo, F. Moreno-Herrero, J. Colchero, J. Gómez Herrero, P. Herrero, A. M. Baró, P. Ordejón, J. M. Soler, and E. Artacho. Absence of dc-conductivity in λ-DNA. *Phys. Rev. Lett.*, 85:4992–4995, 2000.

[200] E. Artacho, M. Machado, D. Sánchez-Portal, P. Ordejón, and J. M. Soler. Electrons in dry DNA from density functional calculations. *Mol. Phys.*, 101:1587–1594, 2003.

[201] S. S. Alexandre, E. Artacho, J. M. Soler, and H. Chacham. Small polarons in dry DNA. *Phys. Rev. Lett.*, 91:108105, 2003.

[202] K.-H. Yoo, D. H. Ha, J.-O. Lee, J. W. Park, J. Kim, J. J. Kim, H.-Y. Lee, T. Kawai, and H. Y. Choi. Electrical conduction through poly(dA)-poly(dT) and poly(dG)-poly(dC) DNA molecules. *Phys. Rev. Lett.*, 87:198102, 2001.

[203] A. Hübsch, R. G. Endres, D. L. Cox, and R. R. P. Singh. Optical conductivity of wet DNA. *Phys. Rev. Lett.*, 94:178102, 2005.

[204] A. Warshel and M. Levitt. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, 103:227, 1976.

[205] R. Lonsdale, K. E. Ranaghan, and A. J. Mulholland. Computational enzymology. *Chem. Commun.*, 46:2354–2372, 2010.

[206] M. W. Van der Kamp and A. J. Mulholland. Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. *Biochemistry*, 52:2708–2728, 2013.

[207] V. L. Schramm. Enzymatic transition states, transition-state analogs, dynamics, thermodynamics, and lifetimes. *Annu. Rev. Biochem.*, 80:703–732, 2011.

[208] R. Lonsdale, K. T. Houghton, J. Zurek, C. M. Bathelt, N. Foloppe, M. J. de Groot, J. N. Harvey,

and A. J. Mulholland. Quantum mechanics/molecular mechanics modeling of regioselectivity of drug metabolism in cytochrome P450 2C9. *J. Am. Chem. Soc.*, 135:8001–8015, 2013.

[209] D. Baker. An exciting but challenging road ahead for computational enzyme design. *Protein Sci.*, 19:1817–1819, 2010.

[210] F. Claeyssens, J. N. Harvey, F. R. Manby, R. A. Mata, A. J. Mulholland, K. E. Ranaghan, M. Schütz, S. Thiel, W. Thiel, and H. J. Werner. High-accuracy computation of reaction barriers in enzymes. *Angew. Chem. Int. Ed.*, 45:6856–6859, 2006.

[211] E. D. Hedegård, J. Kongsted, and U. Ryde. Multiscale modeling of the active site of [Fe] hydrogenase: The H_2 binding site in open and closed protein conformations. *Angew. Chem. Int. Ed.*, 54:6246–6250, 2015.

[212] G. Lever, D. J. Cole, R. Lonsdale, K. E. Ranaghan, D. J. Wales, A. J. Mulholland, C.-K. Skylaris, and M. C. Payne. Large-scale density functional theory transition state searching in enzymes. *J. Phys. Chem. Lett.*, 5:3614–3619, 2014.

[213] P. Kast, C. Grisostomi, I. A. Chen, S. Li, U. Krengel, Y. Xue, and D. Hilvert. A strategically positioned cation is crucial for efficient catalysis by chorismate mutase. *J. Biol. Chem.*, 275:36832–36838, 2000.

[214] R. Qamra, P. Prakash, B. Aruna, S. E. Hasnain, and S. C. Mande. The 2.15Å crystal structure of mycobacterium tuberculosis chorismate mutase reveals an unexpected gene duplication and suggests a role in host-pathogen interactions. *Biochemistry*, 45:6997–7005, 2006.

[215] W. J. Guildford, S. D. Copley, and J. R. Knowles. The mechanism of the chorismate mutase reaction. *J. Am. Chem. Soc.*, 109:5013–5019, 1987.

[216] P. Kast, M. Asif-Ullah, and D. Hilvert. Is chorismate mutase a prototypic entropy trap? – activation parameters for the bacillus subtilis enzyme. *Tetrahedron Lett.*, 37:2691–2694, 1996.

[217] P. Kast, Y. B. Tewari, O. Wiest, D. Hilvert, K. N. Houk, and R. N. Goldberg. Thermodynamics of the conversion of chorismate to prephenate: Experimental results and theoretical predictions. *J. Phys. Chem. B*, 101:10976–10982, 1997.

[218] P. R. Andrews, G. D. Smith, and I. G. Young. Transition-state stabilization and enzymic catalysis. kinetic and molecular orbital studies of the rearrangement of chorismate to prephenate. *Biochemistry*, 12:3492–3498, 1973.

[219] F. Claeyssens, K. E. Ranaghan, N. Lawan, S. J. Macrae, F. R. Manby, J. N. Harvey, and A. J. Mulholland. Analysis of chorismate mutase catalysis by QM/MM modelling of enzyme-catalysed and uncatalysed reactions. *Org. Biomol. Chem.*, 9:1578–1590, 2011.

[220] N. Govind, M. Petersen, G. Fitzgerald, D. King-Smith, and J. Andzelm. A generalized synchronous transit method for transition state location. *Comp. Mater. Sci.*, 28:250–258, 2003.

[221] S. Goedecker, F. Lançon, and T. Deutsch. Linear scaling relaxation of the atomic positions in nanostructures. *Phys. Rev. B*, 64:161102, 2001.

[222] E. D. Glendening, J. K. Badenhoop, A. E. Reed, J. E. Carpenter, J. A. Bohmann, C. M. Morales, and F. Weinhold. *NBO 5.* Theoretical Chemistry Institute, University of Wisconsin, Madison, 2001.

[223] A. E. Reed, L. A. Curtiss, and F. Weinhold. Intermolecular interactions from a natural bond orbital, donor-acceptor viewpoint. *Chem. Rev.*, 88:899–926, 1988.

[224] L. P. Lee, D. J. Cole, M. C. Payne, and C.-K. Skylaris. Natural bond orbital analysis in the onetep code: Applications to large protein systems. *J. Comput. Chem.*, 34:429–444, 2013.

[225] B. Szefczyk, A. J. Mulholland, K. E. Ranaghan, and W. Andrzej Sokalski. Differential transition-state stabilization in enzyme catalysis: Quantum chemical analysis of interactions in the chorismate mutase reaction and prediction of the optimal catalytic field. *J. Am. Chem. Soc.*, 126:16148–16159, 2004.

[226] J.-L. Fattebert, E.Y. Lau, B. J. Bennion, P. Huang, and F. C. Lightstone. Large-scale first-principles molecular dynamics simulations with electrostatic embedding: Application to acetylcholinesterase catalysis. *J. Chem. Theory Comput.*, 11:5688–5695, 2015.

[227] J.-L. Fattebert, R. J. Law, B. Bennion, E. Y. Lau, E. Schwegler, and F. C. Lightstone. Quantitative assessment of electrostatic embedding in density functional theory calculations of biomolecular systems. *J. Chem. Theory Comput.*, 5:2257–2264, 2009.

[228] H. C. Froede and I. B. Wilson. Direct determination of acetyl-enzyme intermediate in the acetylcholinesterase-catalyzed hydrolysis of acetylcholine and acetylthiocholine. *J. Biol. Chem.*, 259:11010–11013, 1984.

[229] D. J. Cole, D. D. O'Regan, and M. C. Payne. Ligand discrimination in myoglobin from linear-scaling DFT+U. *J. Phys. Chem. Lett.*, 3:1448–1452, 2012.

[230] J. S. Olson and G. N. Phillips Jr. Myoglobin discriminates between O_2, NO, and CO by electrostatic interactions with the bound ligand. *J. Biol. Inorg. Chem.*, 2:544–552, 1997.

[231] T. G. Spiro and P. M. Kozlowski. Is the CO adduct of myoglobin bent, and does it matter? *Acc. Chem. Res.*, 34:137, 2001.

[232] J. Vojtěchovský, K. Chu, J. Berendzen, R. M. Sweet, and I. Schlichting. Crystal structures of myoglobin-ligand complexes at near-atomic resolution. *Biophys. J.*, 77:2153, 1999.

[233] M. M. El-Hendawy, N. J. English, and D. A. Mooney. Comparative studies for evaluation of CO_2 fixation in the cavity of the rubisco enzyme using QM, QM/MM and linear-scaling DFT methods. *J. Mol. Model.*, 19:2329–2334, 2013.

[234] L. P. Lee, N. Gabaldon Limas, D. J. Cole, M. C. Payne, C.-K. Skylaris, and T. A. Manz. Expanding the scope of density derived electrostatic and chemical charge partitioning to thousands of atoms. *J. Chem. Theory Comput.*, 10:5377–5390, 2014.

[235] S. A. Wilson, T. Kroll, R. A. Decreau, R. K. Hocking, M. Lundberg, B. Hedman, K. O. Hodgson, and E. I. Solomon. Iron L-edge X-ray absorption spectroscopy of oxy-picket fence porphyrin: Experimental insight into FeO_2 bonding. *J. Am. Chem. Soc.*, 135:1124–1136, 2013.

[236] A. W. Chin, J. Prior, R. Rosenbach, F. Caycedo-Soler, S. F. Huelga, and M. B. Plenio. The role of non-equilibrium vibrational structures in electronic coherence and recoherence in pigment-protein complexes. *Nature Phys.*, 9(2):113–118, 2013.

[237] G. D. Scholes, G. R. Fleming, A. Olaya-Castro, and R. van Grondelle. Lessons from nature about solar light harvesting. *Nature Chem.*, 3:763–774, 2011.

[238] D. E. Tronrud, J. Wen, L. Gay, and R. E. Blankenship. The structural basis for the difference in absorbance spectra for the fmo antenna protein from various green sulfur bacteria. *Photosynth. Res.*, 100:79, 2009.

[239] G. S. Engel, T. R. Calhoun, E. L. Read, T.-K. Ahn, T. Mančal, Y.-C. Cheng, R. E. Blankenship, and G. R. Fleming. Evidence for wavelike energy transfer through quantum coherence in photosynthetic systems. *Nature*, 446:782–786, 2007.

[240] T. Renger and F. Müh. Understanding photosynthetic light-harvesting: a bottom up theoretical approach. *Phys. Chem. Chem. Phys.*, 15:3348–3371, 2013.

[241] C. Curutchet and B. Mennucci. Quantum chemical studies of light harvesting. *Chem. Rev.*, 2016. In press.

[242] F. Müh, M. E. Madjet, J. Adolphs, A. Abdurahman, B. Rabenstein, H. Ishikita, E. W. Knapp, and T. Renger. α-helices direct excitation energy flow in the fenna-matthews-olson protein. *Proc. Natl. Acad. Sci. USA*, 104:16862–16867, 2007.

[243] S. Shim, P. Rebentrost, S. Valleau, and A. Aspuru-Guzik. Atomistic study of the long-lived quantum coherences in the fenna-matthews-olson complex. *Biophys. J.*, 102:649–660, 2012.

[244] J. Adolphs, F. Müh, M. E. Madjet, and T. Renger. Calculation of pigment transition energies in the fmo protein. *Photosynth. Res.*, 95:197, 2008.

[245] C. Olbrich, T. L. C. Jansen, J. Liebers, M. Aghtar, J. Strumpfer, K. Schulten, J. Knoester, and U. Kleinekathofer. From atomistic modeling to excitation transfer and two-dimensional spectra of the fmo light-harvesting complex. *J. Phys. Chem. B*, 115:8609, 2011.

[246] C. König and J. Neugebauer. Protein effects on the optical spectrum of the fenna-matthews-olson complex from fully quantum chemical calculations. *J. Chem. Theory Comput.*, 9:1808–1820, 2013.

[247] L. E. Ratcliff, N. D. M. Hine, and P. D. Haynes. Calculating optical absorption spectra for large systems using linear-scaling density-functional theory. *Phys. Rev. B*, 84:165131, 2011.

[248] L. E. Ratcliff and P. D. Haynes. *Ab initio* calculations of the optical absorption spectra of C_{60}-conjugated polymer hybrids. *Phys. Chem. Chem. Phys.*, 15:13024, 2013.

[249] M. T. W. Milder, B. Brüggemann, R. van Grondelle, and J. L. Herek. Revisiting the optical properties of the fmo protein. *Photosynth. Res.*, 104:257–274, 2010.

[250] J. Adolphs and T. Renger. How proteins trigger excitation energy transfer in the fmo complex of green sulfur bacteria. *Biophys. J.*, 91:2778, 2006.

[251] M. Wendling, M. A. Przyjalgowski, D. Gülen, S. I. E. Vulto, T. J. Aartsma, R. van Grondelle, and H. van Amerongen. The quantitative relationship between structure and polarized spectroscopy in the fmo complex of prosthecochloris aestuarii: refining experiments and simulations. *Photosynth. Res..*, 71:99–123, 2002.

[252] D. Hayes and G. S. Engel. Extracting the excitonic hamiltonian of the fenna-matthews-olson complex using three-dimensional third-order electronic spectroscopy. *Biophys. J.*, 100:2043–2052, 2011.

[253] A. S. Fokas, D. J. Cole, and A. W. Chin. Constrained geometric dynamics of the Fenna-Matthews-Olson complex: the role of correlated motion in reducing uncertainty in excitation energy transfer. *Photosynth. Res.*, 122:275–292, 2014.

[254] J. Adolphs, F. Müh, M. E. Madjet, M. Schmidt am Busch, and T. Renger. Structure-based calculations of optical spectra of photosystem i suggest an asymmetric light-harvesting process. *J. Am. Chem. Soc.*, 132:3331–3343, 2010.

[255] N. Lambert, Y.-N. Chen, Y.-C. Cheng, C.-M. Li, G.-Y. Chen, and F. Nori. Quantum biology. *Nature Phys.*, 9:10–18, 2012.

[256] A. D. Gammack-Yamaguta, S. Datta, K. E. Jackson, L. Stegbauer, R. S. Paton, and D. J. Dixon. Enantioselective desymmetrization of prochiral cyclohexanones via organocatalytic intramolecular michael additions to α,β-unsaturated esters. *Ang. Chem. Int. Ed.*, 127:4981–4985, 2015.

[257] J. Kirchmair, A. H. Göller, D. Lang, J. Kunze, B. Testa, I. D. Wilson, R. C. Glen, and G. Schneider. Predicting drug metabolism: experiment and/or computation? *Nat. Rev. Drug Discov.*, 14:387–404, 2015.

[258] S. Datta and D. J. W. Grant. Crystal structure of drugs: Advances in determination, prediction and engineering. *Nat. Rev. Drug Discov.*, 3:42–57, 2004.

[259] A. M. Reilly and A. Tkatchenko. Seamless and accurate modeling of organic molecular materials. *J. Phys. Chem. Lett.*, 4:1028–1033, 2013.

[260] W. L. Jorgensen. Efficient drug lead discovery and optimization. *Acc. Chem. Res.*, 42:724–733, 2009.

[261] J. D. Chodera, D. L Mobley, M. R. Shirts, R. W. Dixon, K. Branson, and V. S. Pande. Alchemical free energy methods for drug discovery: progress and challenges. *Curr. Opin. Struct. Biol.*, 21:150–160, 2011.

[262] L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M. Dahlgren, J. Greenwood, D. L. Romero, C. Masse, J. L. Knight, T. Steinbrecher, T. Beuming, W. Damm, E. Harder, W. Sherman, M. Brewer, R. Wester, M. Murcko, L. Frye, R. Farid, T. Lin, D. L. Mobley, W. L. Jorgensen, B. J. Berne, R. A. Friesner, and R. Abel. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.*, 137:2695–2703, 2015.

[263] D. J. Cole, J. Tirado-Rives, and W. L. Jorgensen. Molecular dynamics and Monte Carlo simulations for protein-ligand binding and inhibitor design. *Biochim. Biophys. Acta, Gen. Subj.*, 1850:966–971, 2015.

[264] C. Kramer, A. Spinn, and K. R. Liedl. Charge anisotropy: Where atomic multipoles matter most. *J. Chem. Theory Comput.*, 10:4488–4496, 2014.

[265] J. D. Chodera and D. L. Mobley. Entropy-enthalpy compensation: Role and ramifications in

biomolecular ligand recognition and design. *Annu. Rev. Biophys.*, 42:121–142, 2013.

[266] W. L. Jorgensen. Challenges for academic drug discovery. *Angew. Chem. Int. Ed.*, 51:2–7, 2012.

[267] L. Heady, M. Fernandez-Serra, R. L. Mancera, S. Joyce, A. R. Venkitaraman, E. Artacho, C. K. Skylaris, L. Colombi Ciacchi, and M. C. Payne. Novel structural features of CDK inhibition revealed by an *ab initio* computational method combined with dynamic simulations. *J. Med. Chem.*, 49:5141–5153, 2006.

[268] D. J. Cole, C. K. Skylaris, E. Rajendra, A. R. Venkitaraman, and M. C. Payne. Protein-protein interactions from linear-scaling first-principles quantum-mechanical calculations. *EPL*, 91:37004, 2010.

[269] D. J. Cole, E. Rajendra, M. Roberts-Thomson, B. Hardwick, G. J. McKenzie, M. C. Payne, A. R. Venkitaraman, and C. K. Skylaris. Interrogation of the protein-protein interactions between human brca2 brc repeats and rad51 reveals atomistic determinants of affinity. *PLoS Comp. Bio.*, 7:e1002096, 2011.

[270] S. J. Fox, J. Dziedzic, T. Fox, C. S. Tautermann, and C.-K. Skylaris. Density functional theory calculations on entire proteins for free energies of binding: Application to a model polar binding site. *Proteins*, 82:3335–3346, 2014.

[271] K. Vermeulen, D. R. Van Bockstaele, and Z. N. Berneman. The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer. *Cell Prolif.*, 36:131–149, 2003.

[272] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298:1912–1934, 2002.

[273] N. Todorova, L. Yeung, A. Hung, and I. Yarovsky. "janus" cyclic peptides: A new approach to amyloid fibril inhibition? *PLoS ONE*, 8:e57437, 2013.

[274] N. Todorova, A. J. Makarucha, N. D. M. Hine, A. A. Mostofi, and I. Yarovsky. Dimensionality of carbon nanomaterials determines the binding and dynamics of amyloidogenic peptides: Multiscale theoretical simulations. *PLoS Comput. Biol.*, 9:e1003360, 2013.

[275] T. Otsuka, N. Okimoto, M. Taiji, D. R. Bowler, and T. Miyazaki. Structural relaxation and binding energy calculations of FK506 binding protein complexes using the large-scale DFT code CONQUEST. *J. Phys.: Conf. Ser.*, 454:012057, 2013.

[276] L. Pellegrini, D. S. Yu, T. Lo, S. Anand, M. Lee, T. L. Blundell, and A. R. Venkitaraman. Insights into DNA recombination from the structure of a RAD51-BRCA2 complex. *Nature*, 420:287, 2002.

[277] S. C. West. Molecular views of recombination proteins and their control. *Nat. Rev. Mol. Cell Biol.*, 4:435–445, 2003.

[278] A. R. Venkitaraman. Linking the cellular functions of BRCA genes to cancer pathogenesis and treatment. *Annu. Rev. Pathol.*, 4:461, 2009.

[279] J. Srinivasan, T. E. Cheatham III, P. Cieplak, P. A. Kollman, and D. A. Case. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. *J. Am. Chem. Soc.*, 120:9401–9409, 1998.

[280] G. J. Bartlett, A. Choudhary, R. T. Raines, and D. N. Woolfson. $n \to \pi^*$ interactions in proteins. *Nature Chem. Biol.*, 6:615–620, 2010.

[281] B. Q. Wei, W. A. Baase, L. H. Weaver, B. H. Matthews, and B. K. Shoichet. A model binding site for testing scoring functions in molecular docking. *J. Mol. Biol.*, 322:339–355, 2002.

[282] D. L. Mobley, A. P. Graves, J. D. Chodera, A. C. McReynolds, and B. K. Shoichet. Predicting absolute ligand binding free energies to a simple model site. *J. Mol. Biol.*, 371:1118–1134, 2007.

[283] R. P. Muller and A. Warshel. *Ab Initio* calculations of free energy barriers for chemical reactions in solution. *J. Phys. Chem.*, 99:17516–17524, 1995.

[284] M. Štrajbl, G. Hong, and A. Warshel. *Ab Initio* QM/MM simulation with proper sampling: "First principle" calculations of the free energy of the autodissociation of water in aqueous solution. *J. Phys. Chem. B*, 106:13333–13343, 2002.

[285] R. W. Zwanzig. High temperature equation of state by a perturbation method. I. nonpolar gases.

J. Chem. Phys., 22:1420–1426, 1954.

[286] F. R. Beierlein, J. Michel, and J. W. Essex. A simple QM/MM approach for capturing polarization effects in protein ligand binding free energy calculations. *J. Phys. Chem. B*, 115:4911–4926, 2011.

[287] J. Z. Vilseck, J. Tirado-Rives, and W. L. Jorgensen. Evaluation of cm5 charges for condensed-phase modeling. *J. Chem. Theory Comput.*, 10:2802–2812, 2014.

[288] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25:1157–1174, 2004.

[289] W. L. Jorgensen, J. Chandrasekhar, and J. D. Madura. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 1983.

[290] C. Sampson, T. Fox, C. S. Tautermann, C. Woods, and C.-K. Skylaris. A "stepping stone" approach for obtaining quantum free energies of hydration. *J. Phys. Chem. B*, 119:7030–7040, 2015.

[291] C. J. Woods, F. R. Manby, and A. J. Mulholland. An efficient method for the calculation of quantum mechanics/molecular mechanics free energies. *J. Chem. Phys.*, 128:014109, 2008.

[292] L. P. Lee, D. J. Cole, C.-K. Skylaris, W. L. Jorgensen, and M. C. Payne. Polarized protein-specific charges from atoms-in-molecule electron density partitioning. *J. Chem. Theory Comput.*, 9:2981–2991, 2013.

[293] D. J. Cole, J. Z. Vilseck, J. Tirado-Rives, M. C. Payne, and W. L. Jorgensen. Biomolecular force field parameterization via atoms-in-molecule electron density partitioning. *J. Chem. Theory Comput.*, 12:2312–2323, 2016.

[294] T. A. Manz and D. S. Sholl. Chemically meaningful atomic charges that reproduce the electrostatic potential in periodic and nonperiodic materials. *J. Chem. Theory Comput.*, 6:2455–2468, 2010.

[295] T. A. Manz and D. S. Sholl. Improved atoms-in-molecule charge partitioning functional for simultaneously reproducing the electrostatic potential and chemical states in periodic and nonperiodic materials. *J. Chem. Theory Comput.*, 8:2844–2867, 2012.

[296] T. A. Manz and N. Gabaldon Limas. Introducing DDEC6 atomic population analysis: Part 1. Charge partitioning theory and methodology. *RSC Adv.*, 6:47771–47801, 2016.

[297] N. Gabaldon Limas and T. A. Manz. Introducing DDEC6 atomic population analysis: Part 2. Computed results for a wide range of periodic and nonperiodic materials. *RSC Adv.*, 6:45727–45747, 2016.

[298] V. V. Gobre and A. Tkatchenko. Scaling laws for van der Waals interactions in nanostructured materials. *Nat. Commun.*, 4:2341, 2013.

[299] A. Morton, W. A. Baase, and B. W. Matthews. Energetic origins of specificity of ligand binding in an interior nonpolar cavity of T4 lysozyme. *Biochemistry*, 34:8564–8575, 1995.

[300] N. S. Malvankar and D. R. Lovley. Microbial nanowires: A new paradigm for biological electron transfer and bioelectronics. *Chem. Sus. Chem.*, 5:1039–1046, 2012.

[301] D. J. Vinyard, G. M. Ananyev, and G. C. Dismukes. Photosystem II: The reaction center of oxygenic photosynthesis. *Annu. Rev. Biochem.*, 82:577–606, 2013.