PROCEEDINGS OF THE ROYAL SOCIETY A

Proc. R. Soc. A (2009) **465**, 669–683
doi:10.1098/rspa.2008.0398
Published online 9 December 2008

Including dispersion interactions in the ONETEP program for linear-scaling density functional theory calculations

By Quintin Hill and Chris-Kriton Skylaris*

School of Chemistry, University of Southampton, Southampton SO17 1BJ, UK

While density functional theory (DFT) allows accurate quantum mechanical simulations from first principles in molecules and solids, commonly used exchange-correlation density functionals provide a very incomplete description of dispersion interactions. One way to include such interactions is to augment the DFT energy expression by damped London energy expressions. Several variants of this have been developed for this task, which we discuss and compare in this paper. We have implemented these schemes in the ONETEP program, which is capable of DFT calculations with computational cost that increases linearly with the number of atoms. We have optimized all the parameters involved in our implementation of the dispersion correction, with the aim of simulating biomolecular systems. Our tests show that in cases where dispersion interactions are important this approach produces binding energies and molecular structures of a quality comparable with high-level wavefunction-based approaches.

Keywords: dispersion interactions; linear-scaling density functional theory (DFT); ONETEP; biomolecular simulations

1. Introduction

Calculations of properties and processes in materials from 'first principles' quantum mechanical approaches are widely used today in fields as diverse as materials science, biochemistry and engineering (Marzari 2006). These calculations have become established because of their ability to provide an accurate description at the atomic scale, where quantum rather than classical mechanics apply. One of the most widely used first principles methods is density functional theory (DFT; Hohenberg & Kohn 1964) as formulated by Kohn & Sham (1965).

While these calculations have found applications in diverse areas, it would be desirable in many cases to also use them to study nanoscale objects, such as semiconductor nanostructures that have potential applications in electronic devices or entire biological macromolecules. For example, in the case of biological macromolecules such calculations could be used to provide very accurate binding energies as DFT includes, by construction, the electronic changes that take place, such as charge transfer and polarization, which are omitted in the

* Author for correspondence (cks@soton.ac.uk).

Received 6 October 2008
Accepted 11 November 2008

commonly used atomistic force field approaches. However, conventional first-principles methods have a computational cost that scales with the third (or greater) power of the number of atoms in the calculation. Owing to this, unfavourable scaling computations with these methods on more than a few hundred atoms are, in general, not feasible, even on modern supercomputers. To overcome this limitation, several novel reformulations of DFT have been developed, which aim to achieve linear-scaling computational cost (Goedecker 1999). The implementation of these reformulations into robust computational approaches has proved very difficult as challenging mathematical, algorithmic and theoretical problems needed to be overcome (Bowler *et al.* 2008). Nevertheless, intense research efforts spanning more than a decade have allowed satisfactory progress to be made in addressing the challenges involved with linear-scaling DFT and have led to the development of new codes for such calculations by several groups worldwide (Yang 1991; Ordejón *et al.* 2002; Skylaris *et al.* 2005; Bowler *et al.* 2006; Anglada *et al.* 2008).

Linear-scaling DFT approaches greatly extend the length scales that we can access with DFT calculations, from hundreds to many thousands of atoms, and are constructed with the intention of producing as closely as possible the same results as one would obtain with conventional approaches if it were possible to use them on such length scales. Consequently, the well-documented inability of common DFT functionals to describe dispersion interactions correctly remains. Dispersion (or London 1930) forces between atoms arise owing to the instantaneous dipoles, brought about by the fluctuations in the positions of the electrons. These induce dipoles in a nearby atom or molecule which then, in turn, interact with the dipole on the original atom or molecule. The energy of these attractive interactions can be approximated by the London formula, which, for a pair of atoms, has the following form:

$$E(r_{ij}) = -\frac{C_{ij}}{r_{ij}^6} \tag{1.1}$$

and

$$C_{ij} = \frac{3}{2}\alpha_i'\alpha_j'\frac{I_i I_j}{I_i + I_j}, \tag{1.2}$$

where I_j is the ionization potential of atom j; α_j' is its polarizability volume; and r_{ij} is the distance between the atoms i and j. The potential between the two atoms that interact by dispersion forces can be modelled by the Lennard-Jones formula,

$$E(r_{ij}) = \frac{C_{12,ij}}{r_{ij}^{12}} - \frac{C_{6,ij}}{r_{ij}^6}. \tag{1.3}$$

The attractive r^{-6} term comes from the London (1930) formula, while the r^{-12} term represents a repulsive potential. This is required because at a closer range the electrons on each atom repel each other. The form of the repulsive term, however, is mainly a computational convenience as, for example, an exponential form e^{-r/r_0} is a more accurate approximation, but also more costly (Lennard-Jones 1931). Dispersion forces are very weak (e.g. binding energies for noble gas atom pairs are less than 0.25 kcal mol^{-1}) but can collectively be responsible for determining the geometry of many molecules and solids. Important cases include the stacking

interactions between the π-electron systems, such as between graphene sheets in graphite, and base pairs in DNA. In biomolecular simulations, they are often described as 'hydrophobic' interactions and often play an important role in determining the structure and energetics; therefore, they need to be described as well as other non-covalent interactions (such as ion pairs or hydrogen bonding).

The difficulty of describing dispersion with DFT is not an intrinsic failure of the theory as the exact exchange-correlation energy functional would be able to describe all such interactions correctly. However, its form is unknown and approximations are required (Kristyán & Pulay 1994). Commonly, these approximations are based on the local electron density and its gradient, and therefore give a poor description of interactions occurring outside the area of electronic overlap, which is the case with dispersion interactions (Pérez-Jordá & Becke 1995; Meijer & Sprik 1996; Zimmerli *et al.* 2004). DFT does provide an adequate description of the repulsive interactions at a closer range, where the electron densities overlap. The local density approximation will often appear to show binding (in, for example, a noble gas dimer; Pérez-Jordá & Becke 1995; Elstner *et al.* 2001; Wu & Yang 2002) but this binding is spurious as it results from the exchange part of the functional, whereas dispersion is a dynamical correlation effect (Kristyán & Pulay 1994; Rydberg *et al.* 2003). Gradient corrected functionals will usually show no binding at all, although basis set superposition error may often give rise to the appearance of weak binding (Jurečka *et al.* 2007).

Density functionals capable of explicitly including dispersion interactions are being developed by several groups (Langreth *et al.* 2004; Sato *et al.* 2005*a*,*b*; Zhang & Salahub 2007). While these functionals are promising, they have a considerably higher computational cost than conventional DFT functionals as they include non-local terms and their description of binding due to dispersion is not yet consistently comparable with high-level wavefunction-based methods, such as coupled-cluster approaches. More pragmatic efforts to improve the treatment of dispersion in DFT have instead focused on empirical corrections, such as the inclusion of a damped London term in the total energy expression. Following the pioneering application of such approaches by Böhm & Ahlrichs (1982) in Hartree–Fock calculations, which lack dispersion by definition, such schemes are implemented by summing the attractions between all distinct pairs of atoms,

$$E_{\mathrm{disp}}(r_{ij}) = -\sum_{ij,i>j} f_{\mathrm{damp}}(r_{ij}) \frac{C_{6,ij}}{r_{ij}^6}, \tag{1.4}$$

where $f_{\mathrm{damp}}(r_{ij})$ is a damping function that decays to 0 for small r_{ij} and is 1 at large distances. This damping function is required because electronic structure calculations provide an adequate description of short-range attractions, and therefore the empirical correction becomes superfluous at small distances. If a damping function is not applied to the dispersion term then the total energy will be distorted, because of the resulting significant artificial strengthening of every covalent bond.

In this work, we describe the implementation of approaches for empirical dispersion corrections in the ONETEP linear-scaling DFT code, including the optimization of their parameters for use in biomolecular simulations. In §2*a*, a summary of the ONETEP method is given, and its connections with

the conventional plane wave pseudopotential DFT approach are highlighted. Section 2*b* describes the dispersion correction schemes we have explored, and the procedure we have used for the optimization of the parameters follows in §2*c*. In §3, we present tests and comparisons of the methods on a variety of systems demonstrating that they produce dramatic improvements in binding energies and molecular structures in cases where dispersion interactions are important. Section 4 concludes the paper.

2. Theory

(*a*) *The ONETEP program*

ONETEP (Skylaris *et al.* 2005) is a linear-scaling approach for DFT calculations, which is based on the reformulation of DFT in terms of the one-particle density matrix. In terms of Kohn–Sham orbitals, the density matrix is represented as

$$\rho(\boldsymbol{r},\boldsymbol{r}') = \sum_{n=0}^{\infty} f_n \psi_n(\boldsymbol{r}) \psi_n^*(\boldsymbol{r}'), \tag{2.1}$$

where $\psi_n(\boldsymbol{r})$ is a Kohn–Sham orbital and f_n is its occupancy. An equivalent representation is

$$\rho(\boldsymbol{r},\boldsymbol{r}') = \sum_{\alpha\beta} \phi_\alpha(\boldsymbol{r}) K^{\alpha\beta} \phi_\beta^*(\boldsymbol{r}'), \tag{2.2}$$

where $\{\phi_\alpha(\boldsymbol{r})\}$ are localized functions (Hernández *et al.* 1996) and $K^{\alpha\beta}$, which is called the density kernel, is the representation of f_n in the duals of these functions. Most commonly in linear-scaling approaches the density kernel is optimized while keeping $\{\phi_\alpha(\boldsymbol{r})\}$ fixed in some suitable form (e.g. pseudoatomic orbitals). Linear scaling is achieved by truncating the density kernel, thus exploiting the exponential decay of the density matrix (in non-metallic systems; Kohn 1996). A particular characteristic of ONETEP is that the localized functions $\{\phi_\alpha(\boldsymbol{r})\}$ are also optimized during the calculation, subject to a localization constraint, and are thus known as non-orthogonal generalized Wannier functions (NGWFs; Skylaris *et al.* 2002). The NGWFs are expanded in a basis set of periodic sinc (psinc) functions (Mostofi *et al.* 2003), which are equivalent to a plane wave basis as they are related by a unitary transformation. The fact that the NGWFs are optimized *in situ* allows us to achieve plane wave accuracy with only a minimal number of NGWFs (and hence the smallest possible sparse matrices); furthermore, as our basis set is independent of atomic positions and provides a uniform description of space, ONETEP calculations are not affected by basis set superposition error (Haynes *et al.* 2006). The code is parallelized and allows calculations to be performed on large systems containing thousands of atoms (Skylaris *et al.* 2006, 2008).

(*b*) *Dispersion correction*

The various dispersion correction schemes available differ in the form of the damping function $f_{\mathrm{damp}}(r_{ij})$ that they employ. Two major forms for this function have been widely used. The first form is the one introduced by Mooij *et al.* (1999) and later generalized by Elstner *et al.* (2001),

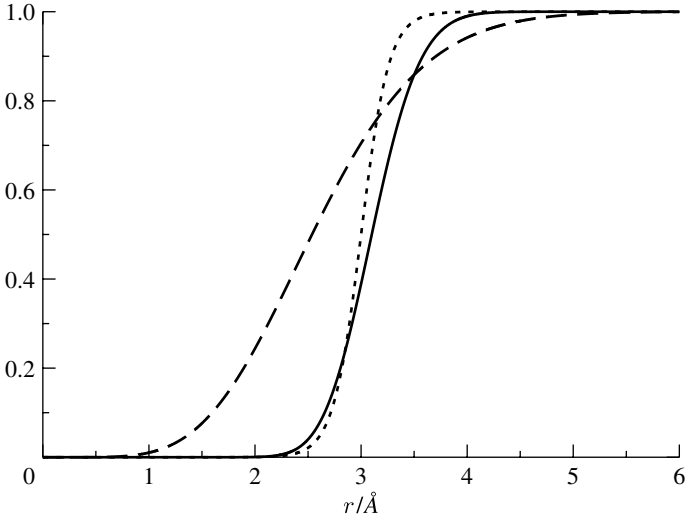

Figure 1. The three damping functions using our optimized parameters for carbon with the PBE exchange-correlation functional. Solid curve, DF1; dashed curve, DF2; dotted curve, DF3.

$$f_{\mathrm{damp}}(r_{ij}) = (1-\exp(-c_{\mathrm{damp}}(r_{ij}/R_{0,ij})^N))^M, \tag{2.3}$$

and the second is a Fermi-like function introduced by Wu & Yang (2002),

$$f_{\mathrm{damp}}(r_{ij}) = \frac{1}{1+\exp(-c_{\mathrm{damp}}(r_{ij}/R_{0,ij}-1))}, \tag{2.4}$$

where c_{damp} is a damping constant and $R_{0,ij}$ is determined by the range of the overlap of atoms i and j (Elstner *et al.* 2001). Elstner *et al.* suggested values of 4 and 7 for M and N, respectively, in equation (2.3), which will be referred to here as damping function 1 (DF1). Mooij *et al.* used $M=2$ and $N=3$ with the damping function in equation (2.3), and this combination will henceforth be referred to as damping function 2 (DF2). Wu and Yang's damping function in equation (2.4) will be labelled in what follows as damping function 3 (DF3). It is possible to obtain heteroatomic $R_{0,ij}$ from the homoatomic values using the following expression (Elstner *et al.* 2001):

$$R_{0,ij} = \frac{R_{0,i}^3 + R_{0,j}^3}{R_{0,i}^2 + R_{0,j}^2}. \tag{2.5}$$

The homoatomic $R_{0,i}$ can be estimated from atomic van der Waals radii (Mooij *et al.* 1999; Wu & Yang 2002).

Figure 1 shows the three damping functions; while they have similar shapes, the range of r for which each damping function has values in the interval [0.01,0.99] varies. DF2 has a notably more gentle decay to zero, while DF3 decays particularly abruptly.

The $C_{6,ij}$ coefficients can be calculated from the homoatomic $C_{6,i}$ coefficients, which can, in turn, be obtained from experimental work or calculated from the atomic polarizabilities α_i using the expression

$$C_{6,i} = \frac{3}{4}\sqrt{N_{\mathrm{eff},i}\alpha_i^3}, \tag{2.6}$$

where $N_{\mathrm{eff},i}$ is the effective number of electrons (Elstner *et al.* 2001). Homoatomic $C_{6,i}$ coefficients can be combined to give heteroatomic $C_{6,ij}$ coefficients using one of the two equivalent forms (Elstner *et al.* 2001; Wu & Yang 2002) of the Slater–Kirkwood combination rule (Slater & Kirkwood 1931),

$$C_{6,ij} = \frac{2C_{6,i}C_{6,j}\alpha_i\alpha_j}{\alpha_i^2 C_{6,j} + \alpha_j^2 C_{6,i}} \tag{2.7}$$

and

$$C_{6,ij} = \frac{2\left(C_{6,i}^2 C_{6,j}^2 N_{\mathrm{eff},i} N_{\mathrm{eff},j}\right)^{1/3}}{\left(N_{\mathrm{eff},i}^2 C_{6,j}\right)^{1/3} + \left(N_{\mathrm{eff},j}^2 C_{6,i}\right)^{1/3}}. \tag{2.8}$$

(*c*) *Optimization of parameters*

Our aim is to use the empirical dispersion correction schemes to improve the description of biomolecular systems in large-scale DFT calculations. We have therefore implemented and tested in ONETEP the schemes described in §2*b*. Our approach also involves the optimization of the parameters involved for each exchange-correlation functional in ONETEP. In order to optimize the parameters, we required a benchmark set of complexes with dispersion interactions where the binding energies are known for high accuracy. Subsets of the JSCH-2005 and S22 sets (Jurečka *et al.* 2006) were chosen for this task. The S22 set is a set of 22 complexes designed to be used as a training set for the inclusion of dispersion corrections and consists of seven hydrogen-bonded complexes, eight complexes with predominant dispersion contribution and seven complexes with significant contributions from both dispersion and hydrogen bonding to the binding. The reference binding energies have been calculated by a combination of MP2 and CCSD(T) methods and extrapolated to the complete basis set limit of CCSD(T). The geometries of the S22 set were obtained by geometry optimizations using MP2 (using a cc-pVTZ basis set and applying a counterpoise correction) for the larger complexes and CCSD(T) (using cc-pVTZ or cc-pVQZ basis sets) for the smaller complexes (Jurečka *et al.* 2006). The JSCH-2005 set provides similarly high-quality binding energies and reference geometries for sets of base pairs and amino acid pairs. The geometries of the complexes in the subset of the JSCH-2005 set that is used in this work were obtained by hydrogen-only geometry optimizations of geometries obtained experimentally. A subset of the stacked base pairs and amino acid pairs from the JSCH-2005 set and the non-hydrogen-bonded complexes from the S22 set were used for our benchmarks. In addition to these, six sulphur-containing complexes from Morgado *et al.* (2007), with binding energies calculated predominantly by MP2, were included so that the parameters for sulphur could also be optimized. The geometries of these complexes were obtained by BLYP-D (with a TZV basis set) optimization (Morgado *et al.* 2007). In total, 60 complexes were chosen for the optimization of 11 parameters. The inclusion of further base pairs from the JSCH-2005 set was deemed undesirable as this could unbalance our chosen training set by giving a bias to base pairs. Also, the hydrogen-bonded complexes in the above sets were omitted, as the empirical dispersion corrections are not designed to describe hydrogen bonding; therefore,

optimizing the parameters in a way that causes them to do so (by imitating the CCSD(T) description of the hydrogen bonds) could compromise their description of dispersion interactions.

Binding energies for the chosen set of complexes were obtained with all of the currently available GGA functionals in ONETEP: PBE (Perdew *et al.* 1996); PW91 (Perdew 1991); revPBE (Zhang & Yang 1998); and RPBE (Hammer *et al.* 1999), as differences of the single-point energies of the bound complex and the two monomers. The geometries were used as provided by the literature and were not modified in these calculations. Subtracting from the reference 'exact' binding energies gave the error in the binding energy (and ideal dispersion energy correction) for each complex. The goal of the optimization was to adjust the parameters in the dispersion formula (1.4) to minimize the difference between the value of the dispersion energy and the error in the binding energy for each complex. The parameters optimized were the $C_{6,i}$ coefficients, the $R_{0,i}$ and the c_{damp} coefficients. Our optimization strategy involved the minimization of the following object function:

$$Err = \sum_{A}^{\text{complexes}} \left[\Delta E_{\mathrm{disp},A} - \left(E^{\mathrm{bind}}_{\mathrm{lit},A} - E^{\mathrm{bind}}_{\mathrm{uncorr},A}\right)\right]^2, \tag{2.9}$$

where the index A runs over the complexes; $\Delta E_{\mathrm{disp},A}$ is the current dispersion energy contribution; $E^{\mathrm{bind}}_{\mathrm{uncorr},A}$ is the pure ONETEP DFT binding energy (without dispersion); and $E^{\mathrm{bind}}_{\mathrm{lit},A}$ is the literature CCSD(T) or MP2 binding energy. For the optimization, the N_{eff} and initial $C_{6,i}$ parameters were taken from Wu & Yang (2002) for carbon, hydrogen, nitrogen and oxygen, and from Halgren (1992) for sulphur. The initial c_{damp} parameters used were: 3.0 for DF1, as used by Elstner *et al.* (2001); 3.54 for DF2, the value Wu & Yang (2002) proposed , rather than Mooij's value of 7.19 (Mooij *et al.* 1999); and 23.0 for DF3 following Wu & Yang (2002). We used $R_{0,i}$ values from Elstner *et al.* (2001). The optimization was considered converged when either of the following two criteria was satisfied.

— The largest change in any parameter from its initial value exceeds 20 per cent. Since the initial parameters are derived from physical quantities, the optimized parameters should not vary considerably in order to preserve transferability and avoid over-optimization to the fitting set.
— An iteration satisfied the following inequality:

$$\frac{\text{maximum percentage change in a parameter in the current step}}{\text{percentage change in } Err \text{ in the current step}} < 0.5, \tag{2.10}$$

which ensures that the parameters were varied only when this led to a significant reduction in the object function.

The c_{damp} parameter was not restricted by the former criterion as it is completely empirical; for example, Mooij's proposed c_{damp} for DF2 is double Wu and Yang's proposed value. The parameters for sulphur were further optimized by starting from the parameters obtained with the entire set (of 60 complexes) and optimizing only the sulphur C_6 coefficient and R_0 with the set of sulphur-containing complexes. In this case, the maximum parameter change was limited to

15 per cent, with the latter convergence criterion the same as above. To eliminate possible effects of the basis set, the single-point energy calculations with ONETEP were performed with a large kinetic energy cut-off of 1200 eV, giving a near-complete psinc basis set. Also, large NGWF radii of $8.0a_0$ were used for all elements (except hydrogen that had NGWF radii of $7.0a_0$).

(*d*) *Atomic forces and geometry optimization*

ONETEP is able to compute atomic forces (as analytic derivatives of the total energy) and use these to perform geometry optimizations. We have included in the forces the contribution from the dispersion interactions so that their effect on determining molecular structure can be taken into account during geometry optimizations. As dispersion interactions dominate only in very weakly bound complexes, a very accurate calculation of all the forces is required. This is possible as the NGWFs are essentially expressed in a plane wave basis, and therefore the 'egg box' effect (Tafipolsky & Schmid 2006) of energy variation with respect to the real-space grid, which is typically observed in real-space techniques, is negligible in ONETEP.

3. Results and discussion

The 60 complexes we used for the fitting of the parameters are presented in table 1, in which the dispersion-including binding energies as obtained with ONETEP are given using the optimized parameters for the three damping functions (DF1, DF2 and DF3) and the PBE (Hammer *et al.* 1999) functional. Table 1 also contains the ONETEP binding energies that are obtained when no dispersion contribution is included. The binding energies are compared with the accurate *ab initio* benchmark binding energies for these complexes, which are subsets of the JSCH-2005, S22 and Morgado *et al.* sets of complexes (Jurečka *et al.* 2006; Morgado *et al.* 2007). We can observe from table 1 that the inclusion of the dispersion contribution dramatically improves the binding energies, in most cases leading to an agreement with the literature results that is better than 1 kcal mol^{-1}. The optimization of the parameters has been necessary to obtain this good agreement as, for example, for DF1 with the PBE functional the value of *Err* (defined in equation (2.9)) was reduced by 78 per cent in the initial parameter optimization, and in the subsequent sulphur parameter optimization of the value of *Err* (for the subset of sulphur complexes) was further reduced by 32 per cent. After optimization, DF1 (with PBE) produced binding energies with the lowest root mean square (r.m.s.) difference from the literature values of 0.813 kcal mol^{-1}. DF3 had a r.m.s. difference only slightly higher, 0.820 kcal mol^{-1}; however, DF2 was notably worse with a r.m.s. of 0.926 kcal mol^{-1}, and a very similar trend was observed for the standard deviations. So, for the PBE functional, DF1 is expected to be the most accurate and consistent.

We argued that a key concern of our approach was to retain the transferability of the parameters. To check if this goal has been achieved, we performed validation calculations on a set of complexes that were not included in the fitting set. These complexes are presented in table 2. They are grouped into four categories: interstrand base pairs; stacked base pairs; hydrogen-bonded base pairs; and other hydrogen-bonded complexes (from the remainder of the S22 set). These systems

Table 1. Binding energies (in kcal mol^{-1}) for the complexes used in the fitting of the parameters. (The ONETEP results are given with and without dispersion interactions with the optimized parameters for the PBE functional and are compared with the 'exact' values from the literature.)

complex	uncorrected	DF1	DF2	DF3	literature (Jurečka *et al.* (2006) and Morgado *et al.* (2007))
2CH_3SH(C_1) a3	−1.61	−2.49	−3.36	−2.40	−2.68
2CH_3SH(C_1) a5	−1.76	−2.79	−3.66	−2.68	−2.50
2CH_3SH(C_i) a4	−1.77	−2.68	−3.15	−2.59	−2.00
AA0-3.24 A–As	2.41	−6.08	−6.02	−6.22	−6.25
AA0-3.24 T–Ts	2.81	−4.32	−4.96	−4.19	−3.86
AA20-3.05 AAs2005	3.03	−5.85	−5.86	−5.99	−6.06
AA20-3.05 TTs2005	−0.68	−2.25	−2.47	−2.21	−4.18
A···C S	1.77	−6.48	−6.28	−6.44	−6.70
adenine–thymine stack	−1.12	−11.40	−12.23	−11.43	−12.23
AG08-3.19 A–Gs	−0.09	−7.47	−7.32	−7.45	−7.58
AG08-3.19 T–Cs	−0.64	−6.31	−6.45	−6.19	−6.07
A···G S	2.35	−6.30	−6.59	−6.34	−6.50
AT10-3.26 A–Ts	0.88	−6.91	−7.00	−6.84	−6.64
A···T S	1.02	−8.42	−8.37	−8.34	−8.10
benzene–ammonia (C_s)	−0.70	−2.38	−2.69	−2.36	−2.35
benzene dimer (C_{2h})	2.02	−3.34	−3.08	−3.41	−2.73
benzene dimer (C_{2v})	−0.05	−2.66	−3.20	−2.59	−2.74
benzene DMS (C_{2v}) a8	−0.15	−3.37	−3.28	−3.39	−3.00
benzene DMS (C_{2v}) a9	−0.66	−1.12	−1.28	−1.10	−1.21
benzene H_2S(C_{2v}) a7	−0.74	−2.43	−2.99	−2.30	−2.74
benzene HCN (C_s)	−3.02	−4.77	−5.55	−4.69	−4.46
benzene–methane (C_3)	0.08	−1.67	−1.87	−1.67	−1.50
benzene–water (C_s)	−1.82	−3.35	−3.83	−3.27	−3.28
CG0-3.19 G–Cs	−1.79	−6.83	−7.09	−6.68	−7.88
C···G S	−2.97	−10.58	−10.53	−10.60	−12.40
ethene dimer (D_{2d})	−0.39	−1.96	−2.30	−1.89	−1.51
ethene–ethine (C_{2v})	−1.32	−2.04	−2.20	−2.01	−1.53
F30-F49	−0.16	−3.20	−3.34	−3.21	−3.30
F30-K46	−1.07	−3.66	−3.82	−3.62	−3.10
F30-L33	−0.40	−5.47	−6.44	−5.25	−5.00
F30-Y13	−1.05	−4.87	−5.05	−4.79	−3.90
F30-Y4	0.85	−6.05	−6.02	−6.03	−7.00
F49 C39	0.28	−2.12	−2.94	−2.03	−2.10
F49 C6	0.70	−5.01	−5.33	−4.90	−5.00
F49-K46	−1.34	−4.95	−5.83	−4.85	−4.80
F49-PB V5-C6	−2.26	−7.93	−8.61	−7.81	−8.20
F49-PB Y4-V5	−0.36	−3.29	−3.38	−3.23	−2.80
F49-V5	−0.85	−6.64	−7.91	−6.47	−6.70
F49-Y37	−0.15	−2.41	−2.47	−2.37	−2.50
F49-Y4	1.41	−3.65	−4.34	−3.56	−3.10

(*Continued.*)

Table 1. (*Continued.*)

complex	uncorrected	DF1	DF2	DF3	literature (Jurečka *et al.* (2006) and Morgado *et al.* (2007))
GA10-3.15 A–Gs	0.45	−9.43	−9.46	−9.52	−9.14
GA10-3.15 T–Cs	0.88	−5.29	−5.26	−5.24	−4.69
GC0-3.25 G–Cs	−2.04	−10.84	−10.81	−10.92	−10.80
G···C S	−2.78	−10.24	−10.43	−10.29	−8.10
G···C S1	0.25	−7.03	−7.05	−7.02	−7.70
G···C S2	−3.46	−7.86	−7.77	−7.81	−11.60
GG0-3.36 CCs036	−3.94	−4.71	−4.70	−4.70	−3.54
GG0-3.36 GGs036	3.41	−2.01	−1.90	−2.01	−1.62
GT10-3.15 A–Cs	2.13	−5.54	−5.40	−5.64	−5.44
GT10-3.15 T–Gs	3.44	−5.27	−5.58	−5.16	−4.96
indole–benzene stack (C_1)	2.46	−5.34	−5.03	−5.50	−5.22
indole–benzene t-shaped (C_1)	−2.14	−5.65	−6.70	−5.59	−5.73
methane dimer (D_{3d})	−0.08	−0.96	−0.91	−0.94	−0.53
phenol dimer (C_1)	−4.33	−7.10	−8.10	−6.97	−7.05
pyrazine dimer (C_s)	0.82	−4.56	−4.48	−4.71	−4.42
TA08-3.16 A–Ts	5.56	−4.44	−5.37	−4.55	−6.07
TG03.19 A–Cs	1.61	−4.40	−4.64	−4.34	−4.96
TG03.19 T–Gs	0.01	−5.22	−5.50	−5.06	−5.67
T···G S	1.82	−6.73	−6.66	−6.64	−6.20
uracil dimer stack (C_2)	−2.66	−9.33	−10.24	−9.12	−10.12
root mean square error	5.915	0.813	0.926	0.820	0

were chosen as they represent a wide range of typical biomolecular environments and also because accurate binding energies are available for these structures in the literature (Jurečka *et al.* 2006). For the interstrand base pairs and the stacked base pairs the dispersion interaction is the dominant interaction; our results show the same dramatic improvement in the binding energies as in the complexes of table 1. Furthermore, the level of improvement in the binding energies of the stacked and the interstrand base pairs is similar even though only stacked base pairs were included in the fitting set, indicating the generality of the empirical dispersion correction. For the hydrogen-bonded base pairs and other hydrogen-bonded complexes, where the binding is mainly due to hydrogen bonds, the inclusion of the empirical dispersion contribution is not as successful. In a few cases, such as the water dimer, for example, the uncorrected ONETEP binding energy is already too large, and the dispersion correction leads to further overbinding. DF2 gave a significant overbinding for every hydrogen-bonded complex (r.m.s. difference 5.777 kcal mol^{-1}), so this function is less applicable to systems with significant hydrogen bonding, which is the norm for many biological molecules. DF3 performed better than DF1 for all but two of the hydrogen-bonded complexes; the r.m.s. differences were 1.225 and 1.367 kcal mol^{-1}, respectively. For the non-hydrogen-bonded complexes, all the damping functions produced binding energies

Table 2. Binding energies in (kcal mol^{-1}) for complexes that were not included in the fitting of parameters. (Values obtained with just the LDA and PBE exchange-correlation functionals are given, as well as values calculated with PBE plus dispersion with DF1, DF2 and DF3.)

complex	LDA	PBE	DF1	DF2	DF3	literature
interstrand base pairs						
AA20 3.05 ATis2005	−2.58	−1.23	−2.41	−2.68	−2.36	−2.34
GA10 3.15 A Cis	0.44	1.13	−0.08	−0.11	−0.06	−0.31
GA10 3.15 T Gis	1.02	1.20	0.56	0.56	0.56	0.58
GG0 3.36 CGis036	−3.21	−2.46	−3.94	−3.94	−3.91	−3.68
TG0319 T Cis	−1.17	−0.95	−1.39	−1.39	−1.39	−1.15
stacked base pairs						
AAst	−6.44	0.27	−8.38	−8.17	−8.52	−8.58
CCst	−8.56	−2.65	−9.61	−9.53	−9.69	−10.02
GGst	−10.39	−2.94	−12.39	−12.30	−12.44	−12.67
UUst	−7.67	−2.31	−8.60	−8.47	−8.61	−7.46
hydrogen-bonded base pairs						
2tU 2tU	−17.40	−10.87	−12.58	−14.16	−12.40	−12.60
6tG C WC pl	−38.08	−28.66	−31.30	−33.40	−31.12	−29.50
A 4tU WC	−18.74	−12.46	−14.80	−16.07	−14.63	−13.20
adenine–thymine	−15.65	−15.65	−18.21	−19.94	−18.01	−16.37
G 2tU	−21.85	−14.82	−16.65	−18.42	−16.47	−16.60
G 4tU	−23.82	−16.37	−18.46	−20.23	−18.26	−17.80
uracil dimer hb (C_{2h})	−27.09	−20.11	−21.91	−23.60	−21.75	−20.65
other hydrogen-bonded complexes						
2-pyridoxine 2-aminopyridine (C_1)	−23.70	−16.86	−19.20	−20.93	−19.02	−16.71
ammonia dimer (C_{2h})	−5.07	−2.93	−3.35	−4.06	−3.31	−3.17
formamide dimer (C_{2h})	−22.48	−16.01	−17.23	−18.80	−17.08	−15.96
formic acid (C_{2h})	−27.28	−19.54	−20.39	−22.40	−20.18	−18.61
water dimer (C_s)	−7.83	−5.20	−5.40	−6.10	−5.31	−5.02
root mean square error	4.494	3.588	1.073	2.048	0.968	0

of similar accuracy; the r.m.s. differences for DF1, DF2 and DF3 were 0.444, 0.452 and 0.430 kcal mol^{-1}, respectively. LDA binding energies have been included for comparison. As expected, these energies are too large for the hydrogen-bonded complexes (Elstner *et al.* 2001). For the non-hydrogen-bonded complexes, LDA (r.m.s. 1.202 kcal mol^{-1}) produced more accurate binding energies than PBE (r.m.s. 5.366 kcal mol^{-1}); however, they are still inferior to the corrected PBE energies.

We have also investigated the effect of the dispersion contribution on the atomic forces by examining the molecular structures obtained during geometry optimization. We have performed full (unconstrained) geometry optimizations on four systems: a benzene dimer; a methane dimer; a methane–benzene complex; and an indole–benzene complex. All the calculations were performed with DF1 and the PBE functional and a rather tight maximum absolute force convergence threshold of 0.001 E_h/a_0 was used as the forces due to dispersion are obviously very weak.

For the case of the benzene dimer, the optimization with dispersion contributions resulted in the equilibrium structure shown in figure 2, where the two benzene molecules are in a conformation with their planes parallel, at a

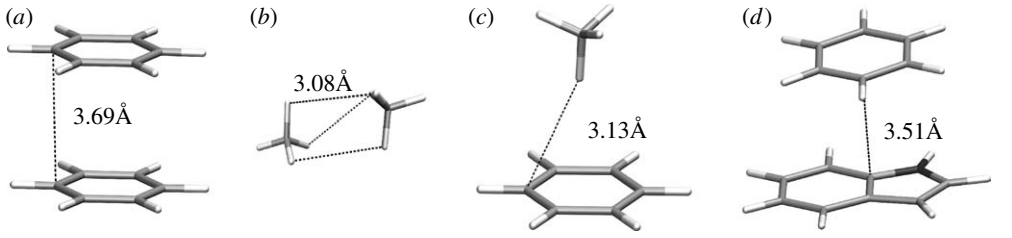

Figure 2. Optimized structures with ONETEP of (*a*) benzene sandwich dimer, (*b*) methane dimer, (*c*) methane–benzene and (*d*) indole–benzene.

separation of 3.7 Å. This is in close agreement with the value of 3.9 Å that has been obtained with CCSD(T) calculations with a near-complete basis set by Sinnokrot & Sherrill (2004). When we do not include dispersion interactions, the pair of benzene molecules experiences only the repulsive potential of the PBE functional and no binding is observed, but the geometry optimization is simply completed when their separation is 5.1 Å, as at this distance the forces are smaller than the set threshold.

For the methane dimer, when dispersion interactions are included ONETEP was able to reproduce the geometry obtained by MP2 calculations using a large Gaussian basis set (cc_pVTZ; Dunning 1989). The final structure obtained with ONETEP is shown in figure 2. The structure has the correct symmetry, and the distance between the hydrogen atoms of the two molecules (3.08 Å) is in agreement with the MP2 value (3.07 Å). When the empirical dispersion contributions are omitted from the ONETEP geometry optimization, the methane molecules end up much further from each other and their orientation is very different from that obtained using the MP2 approach.

In the case of the benzene–methane complex, the benzene–methane distance we obtain after optimization when using our empirical dispersion contribution is 3.15 Å (figure 2), which is in close agreement with the value of 2.98 Å from an accurate MP2 geometry (Jurečka *et al.* 2006). Omitting the dispersion contribution results in a distance of 3.62 Å. For the indole–benzene complex, the indole–benzene distance increased from 3.40 Å in an MP2-optimized geometry (Jurečka *et al.* 2006) to 3.52 Å when optimized with our empirical dispersion and 3.92 Å when optimized without it. Clearly, in all cases, the inclusion of the empirical dispersion contribution has significantly improved the geometries obtained.

4. Conclusions

We have presented an implementation of the empirical dispersion contributions for the ONETEP code, including optimization of parameters for use in biological simulations. Optimization of the parameters significantly improved the obtained binding energies for weakly bound complexes. Further validation calculations, which compare with the literature results from explicitly correlated wavefunction methods, show that the inclusion of dispersion interactions provides an adequate description of the binding energies of weakly bound complexes. We found that the damping functions DF1 and DF3 produced the best binding energies, with DF3 being superior for hydrogen-bonded complexes. Inferior corrected binding energies

were obtained when DF2 was used, especially for hydrogen-bonded complexes. The dispersion contributions have also been included in the calculation of the forces and allow ONETEP geometry optimizations to produce accurate structures. While the emphasis of this work has been on biomolecular simulations, the code is completely general and can in future be extended to other classes of molecules and materials if suitable parameters are provided.

Q.H. would like to thank the EPSRC for research studentship funding. C.-K.S. would like to thank the Royal Society for a University Research Fellowship.

References

Anglada, E. *et al.* 2008 The SIESTA method; developments and applicability. *J. Phys. Condens. Matter* **20**, 064 208. (doi:10.1088/0953-8984/20/6/064208)

Böhm, H.-J. & Ahlrichs, R. 1982 A study of short-range repulsions. *J. Chem. Phys.* **77**, 2028–2034. (doi:10.1063/1.444057)

Bowler, D. R., Choudhury, R., Gillan, M. J. & Miyazaki, T. 2006 Recent progress with large-scale *ab initio* calculations: the CONQUEST code. *Phys. Stat. Sol. B* **243**, 989–1000. (doi:10.1002/pssb.200541386)

Bowler, D. R., Fattebert, J.-L., Gillan, M. J., Haynes, P. D. & Skylaris, C.-K. 2008 Introductory remarks: linear scaling methods. *J. Phys. Condens. Matter* **20**, 290 301. (doi:10.1088/0953-8984/20/29/290301)

Dunning Jr, T. H. 1989 Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **90**, 1007–1023. (doi:10.1063/1.456153)

Elstner, M., Hobza, P., Frauenheim, T., Suhai, S. & Kaxiras, E. 2001 Hydrogen bonding and stacking interactions of nucleic acid base pairs: a density-functional-theory based treatment. *J. Chem. Phys.* **114**, 5149–5155. (doi:10.1063/1.1329889)

Goedecker, S. 1999 Linear scaling electronic structure methods. *Rev. Mod. Phys.* **71**, 1085–1123. (doi:10.1103/RevModPhys.71.1085)

Halgren, T. A. 1992 The representation of van der Waals (vdw) interactions in molecular mechanics force fields: potential form, combination rules, and vdw parameters. *J. Am. Chem. Soc.* **114**, 7827–7843. (doi:10.1021/ja00046a032)

Hammer, B., Hansen, L. B. & Nørskov, J. K. 1999 Improved adsorption energetics within density-functional theory using revised Perdew–Burke–Ernzerhof functionals. *Phys. Rev. B* **59**, 7413–7421. (doi:10.1103/PhysRevB.59.7413)

Haynes, P. D., Skylaris, C.-K., Mostofi, A. A. & Payne, M. C. 2006 Elimination of the basis set superposition error in linear-scaling density-functional calculations with local orbitals optimized *in situ*. *Chem. Phys. Lett.* **422**, 345–349. (doi:10.1016/j.cplett.2006.02.086)

Hernández, E., Gillan, M. J. & Goringe, C. M. 1996 Linear-scaling density-functional-theory technique: the density-matrix approach. *Phys. Rev. B* **53**, 7147–7157. (doi:10.1103/PhysRevB.53.7147)

Hohenberg, P. & Kohn, W. 1964 Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871. (doi:10.1103/PhysRev.136.B864)

Jurečka, P., Černý, J., Hobza, P. & Šponer, J. 2006 Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Phys. Chem. Chem. Phys.* **8**, 1985–1993. (doi:10.1039/b600027d)

Jurečka, P., Černý, J., Hobza, P. & Salahub, D. R. 2007 Density functional theory augmented with an empirical dispersion term. Interaction energies and geometries of 80 noncovalent complexes compared with *ab initio* quantum mechanics calculations. *J. Comput. Chem.* **28**, 555–569. (doi:10.1002/jcc.20570)

Kohn, W. 1996 Density functional and density matrix method scaling linearly with the number of atoms. *Phys. Rev. Lett.* **76**, 3168–3171. (doi:10.1103/PhysRevLett.76.3168)

Kohn, W. & Sham, L. J. 1965 Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138. (doi:10.1103/PhysRev.140.A1133)

Kristyán, S. & Pulay, P. 1994 Can (semi)local density functional theory account for the London dispersion forces? *Chem. Phys. Lett.* **229**, 175–180. (doi:10.1016/0009-2614(94)01027-7)

Langreth, D. C., Dion, M., Rydberg, H., Schröder, E., Hyldgaard, P. & Lundqvist, B. I. 2004 Van der Waals density functional theory with applications. *Int. J. Quant. Chem.* **101**, 599–610. (doi:10.1002/qua.20315)

Lennard-Jones, J. E. 1931 Cohesion. *Proc. Phys. Soc.* **43**, 461–482. (doi:10.1088/0959-5309/43/5/301)

London, F. 1930 Zur Theorie und Systematik der Molekularkräfte. *Z. Phys.* **63**, 245–279. (doi:10.1007/BF01421741)

Marzari, N. 2006 Realistic modeling of nanostructures using density functional theory. *MRS Bull.* **31**, 681.

Meijer, E. J. & Sprik, M. 1996 A density-functional study of the intermolecular interactions of benzene. *J. Chem. Phys.* **105**, 8684–8689. (doi:10.1063/1.472649)

Mooij, W. T. M., van Duijneveldt, F. B., van Duijneveldt-van de Rijdt, J. G. C. M. & van Eijck, P. 1999 Transferable *ab initio* intermolecular potentials. 1. Derivation from methanol dimer and trimer calculations. *J. Phys. Chem. A* **103**, 9872–9882. (doi:10.1021/jp991641n)

Morgado, C. A., McNamara, J. P., Hillier, I. H., Burton, N. A. & Vincent, M. A. 2007 Density functional and semiempirical molecular orbital methods including dispersion corrections for the accurate description of noncovalent interactions involving sulfur-containing molecules. *J. Chem. Theory Comput.* **3**, 1656. (doi:10.1021/ct700072a)

Mostofi, A. A., Haynes, P. D., Skylaris, C.-K. & Payne, M. C. 2003 Preconditioned iterative minimization for linear-scaling electronic structure calculations. *J. Chem. Phys.* **119**, 8842–8848. (doi:10.1063/1.1613633)

Ordejón, P., Soler, J. M. & Sánchez-Portal, D. 2002 The SIESTA method for *ab initio* order-N materials simulation. *J. Phys. Condens. Matter* **14**, 2745–2779. (doi:10.1088/0953-8984/14/11/302)

Perdew, J. P. 1991 *Electronic structure of solids '91*. Berlin, Germany: Akademie Verlag.

Perdew, J. P., Burke, K. & Ernzerhof, M. 1996 Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868. (doi:10.1103/PhysRevLett.77.3865)

Pérez-Jordá, J. M. & Becke, A. D. 1995 A density-functional study of van der Waals forces: rare gas diatomics. *Chem. Phys. Lett.* **233**, 134–137. (doi:10.1016/0009-2614(94)01402-H)

Rydberg, H., Jacobson, N., Hyldgaard, P., Simak, S. I., Lundqvist, B. I. & Langreth, D. C. 2003 Hard numbers on soft matter. *Surf. Sci.* **532–535**, 606–610. (doi:10.1016/S0039-6028(03)00109-2)

Sato, T., Tsuneda, T. & Hirao, K. 2005*a* Van der Waals interactions studied by density functional theory. *Mol. Phys.* **103**, 1151–1164. (doi:10.1080/00268970412331333474)

Sato, T., Tsuneda, T. & Hirao, K. 2005*b* A density-functional study on π-aromatic interaction: benzene dimer and naphthalene dimer. *J. Chem. Phys.* **123**, 104 307. (doi:10.1063/1.2011396)

Sinnokrot, M. O. & Sherrill, C. D. 2004 Highly accurate coupled cluster potential energy curves for the benzene dimer: sandwich, T-shaped, and parallel-displaced configurations. *J. Phys. Chem. A* **108**, 10 200–10 207. (doi:10.1021/jp0469517)

Skylaris, C.-K., Mostofi, A. A., Haynes, P. D., Diéguez, O. & Payne, M. C. 2002 Nonorthogonal generalized Wannier function pseudopotential plane–wave method. *Phys. Rev. B* **66**, 035 119. (doi:10.1103/PhysRevB.66.035119)

Skylaris, C.-K., Haynes, P. D., Mostofi, A. A. & Payne, M. C. 2005 Introducing ONETEP: linear-scaling density functional simulations on parallel computers. *J. Chem. Phys.* **122**, 084 119. (doi:10.1063/1.1839852)

Skylaris, C.-K., Haynes, P. D., Mostofi, A. A. & Payne, M. C. 2006 Implementation of linear scaling plane wave density functional theory on parallel computers. *Phys. Stat Sol. B* **243**, 973–988. (doi:10.1002/pssb.200541328)

Skylaris, C.-K., Haynes, P. D., Mostofi, A. A. & Payne, M. C. 2008 Recent progress in linear-scaling density functional calculations with plane waves and pseudopotentials: the ONETEP code. *J. Phys. Condens. Matter* **20**, 064 209. (doi:10.1088/0953-8984/20/6/064209)

Slater, J. C. & Kirkwood, J. G. 1931 The van der Waals forces in gases. *Phys. Rev.* **37**, 682–697. (doi:10.1103/PhysRev.37.682)

Tafipolsky, M. & Schmid, R. 2006 A general and efficient pseudopotential Fourier filtering scheme for real space methods using mask functions. *J. Chem. Phys.* **124**, 174 102. (doi:10.1063/1.2193514)

Wu, Q. & Yang, W. 2002 Empirical correction to density functional theory for van der Waals interactions. *J. Chem. Phys.* **116**, 515–524. (doi:10.1063/1.1424928)

Yang, W. 1991 Direct calculation of electron density in density-functional theory. *Phys. Rev. Lett.* **66**, 1438–1441. (doi:10.1103/PhysRevLett.66.1438)

Zhang, Y. & Salahub, D. R. 2007 A reparametrization of a meta-GGA exchange-correlation functional with improved descriptions of van der Waals interactions. *Chem. Phys. Lett.* **436**, 394–399. (doi:10.1016/j.cplett.2007.01.074)

Zhang, Y. & Yang, W. 1998 Comment on generalized gradient approximation made simple. *Phys. Rev. Lett.* **80**, 890. (doi:10.1103/PhysRevLett.80.890)

Zimmerli, U., Parrinello, M. & Koumoutsakos, P. 2004 Dispersion corrections to density functionals for water aromatic interactions. *J. Chem. Phys.* **120**, 2693–2699. (doi:10.1063/1.1637034)