

# Elimination of basis set superposition error in linear-scaling density-functional calculations with local orbitals optimised *in situ*

P. D. Haynes<sup>a,\*</sup> C.-K. Skylaris<sup>b</sup> A. A. Mostofi<sup>a,c</sup> M. C. Payne<sup>a</sup>

<sup>a</sup>*Theory of Condensed Matter, Cavendish Laboratory, J. J. Thomson Avenue, Cambridge CB3 0HE, UK*

<sup>b</sup>*Physical & Theoretical Chemistry Laboratory, South Parks Road, Oxford OX1 3QZ, UK*

<sup>c</sup>*Department of Materials Science & Engineering, Massachusetts Institute of Technology, Cambridge MA, USA*

---

## Abstract

Basis set superposition error (BSSE) in density-functional calculations occurs when the extended Kohn-Sham orbitals are expanded in localised basis sets, but is absent when a plane-wave basis is used. Elimination of BSSE is essential for the accurate description of intermolecular forces. Linear-scaling methods are formulated in terms of local orbitals, making plane-waves an inappropriate choice of basis. In this work the BSSE in linear-scaling methods is studied in the context of hydrogen bonds. In particular it is shown that BSSE is eliminated by optimizing the local orbitals *in situ* using a systematic basis set equivalent to a set of plane-waves.

*Key words:* linear-scaling, density-functional theory, basis set superposition error  
*PACS:* 31.50.Bc, 71.15.Ap, 82.30.Rs

---

## 1 Introduction

Density-functional theory (DFT) has had a profound impact on science and technology, well beyond the traditional realms of quantum-mechanical simulations into fields as diverse as geology [1], biochemistry [2] and microelectronics [3]. The broad multidisciplinary appeal of DFT stems from its ability

---

\* Corresponding author. Fax: +44 1223 337356.

*Email address:* pdh1001@cam.ac.uk (P. D. Haynes).

*URL:* <http://www.tcm.phy.cam.ac.uk/~pdh1001/> (P. D. Haynes).

to provide a sufficiently accurate description of electron correlation for most purposes at a computational cost which scales very favorably (with the cube of the system size  $N$ ) compared with correlated wave function methods such as coupled cluster or configuration interaction (which typically exhibit  $N^5$  to  $N^7$  scaling). Nevertheless, even with the most powerful computers, the  $N^3$  scaling of DFT still presents a bottleneck which restricts the size of simulations to a few hundred atoms and thus limits their predictive power and ability to provide insight into complex processes in real materials. In recent years there has therefore been much interest in the development of linear-scaling or order- $N$  methods [4,5] which have the potential to revolutionise the scope and scale of DFT calculations to include entire biological molecules and nanostructures consisting of many thousands of atoms. However in spite of the multiplicity of proposed computational schemes very few successful applications have been reported. Most importantly, there is a severe lack of studies of the accuracy attainable with linear-scaling methods and consequently much uncertainty about their ability to achieve the necessary level of detail to describe correctly the systems they are intended to tackle. This paper compares the accuracy of three approaches to linear-scaling methods for the sensitive and important case of relatively weak hydrogen bonds, which are essential for the structural stability and function of biomolecules such as proteins, DNA and sugars and which are also relevant in nanoscience where many potential applications are expected to operate in an aqueous environment.

Basis set superposition error (BSSE) in the solution of the Schrödinger equation was first addressed for the calculation of accurate intermolecular potentials [6]. The interaction potential (the negated binding energy) of two molecules A and B is defined by

$$\Delta_{AB} = E_{AB} - E_A - E_B \quad (1)$$

where the total energies of the dimer  $E_{AB}$  and of the two isolated molecules  $E_A$  and  $E_B$  are obtained from separate calculations and structural relaxation of the molecules upon formation of the dimer is neglected. Error arises when the parameters or procedures are not consistent between these three calculations. Boys and Bernardi [7] identified several causes of error including the use of different integration grids for the calculation of matrix elements and different basis sets for the expansion of the wave functions. The methods considered in detail here employ identical uniform real-space grids for all calculations, eliminating the former source of error. The latter is known as BSSE and plays a central role in this study.

## 2 Methods

Figure 1 provides a schematic representation of the source of BSSE and the commonly-used counterpoise correction [7] first invoked by Jansen and Ros [6] to treat it. Basis functions (shaded gray) are localised and associated with particular atoms (white disks). The total energy of the dimer is calculated naturally using the basis functions associated with both molecules. For an isolated molecule, the total energy may be obtained using only the basis functions associated with that molecule. However this simple approach restricts the variational freedom of the molecular calculations with respect to that of the dimer, resulting in an overestimate of the binding energy which can be significant [8]. The counterpoise correction uses the basis functions associated with both molecules in calculating the total energies of the isolated molecules, so that identical basis sets are used for all three energies in Eq. 1. Mayer’s chemical Hamiltonian approach [9] provides an alternative *a priori* BSSE-free scheme.

The plane-wave pseudopotential method [10] has established itself as the leading workhorse for DFT calculations. The main advantage of a plane-wave basis in the context of this work is that being homogeneous and unbiased by atomic positions it is free of BSSE. The corresponding drawback is that it is relatively inefficient at treating systems with significant vacuum regions such as isolated molecules.

Linear-scaling methods appropriate for non-metallic systems exploit the “near-sightedness” of quantum many-body systems [11] exhibited in the exponential localisation of the Wannier functions [12] and single-particle density-matrix [13]. In all of these methods, local orbitals play a key part. For example, in the ONETEP method [14] the density-matrix is expressed in separable form [15] as

$$\rho(\vec{r}_1, \vec{r}_2) = \sum_{\alpha\beta} \phi_\alpha(\vec{r}_1) K^{\alpha\beta} \phi_\beta^*(\vec{r}_2) \quad (2)$$

in terms of a density kernel  $K^{\alpha\beta}$  and a set of local orbitals called non-orthogonal generalised Wannier functions (NGWFs)  $\{\phi_\alpha\}$  [16] centred on the atomic positions. Linear scaling is obtained by truncating the kernel and localizing the NGWFs according to variational spatial cut-offs. Most linear-scaling methods fall into two categories: the first equivalent to optimizing the kernel only [17–19] for a fixed (but potentially large) set of local orbitals and the second involving both kernel and NGWF optimisation [16,15,20]. In the first case, the local orbitals are optimised beforehand for isolated atoms of each species and act as the basis set. In the second case the optimisation occurs *in situ* during the self-consistent calculation. Since atom-centred local orbitals are

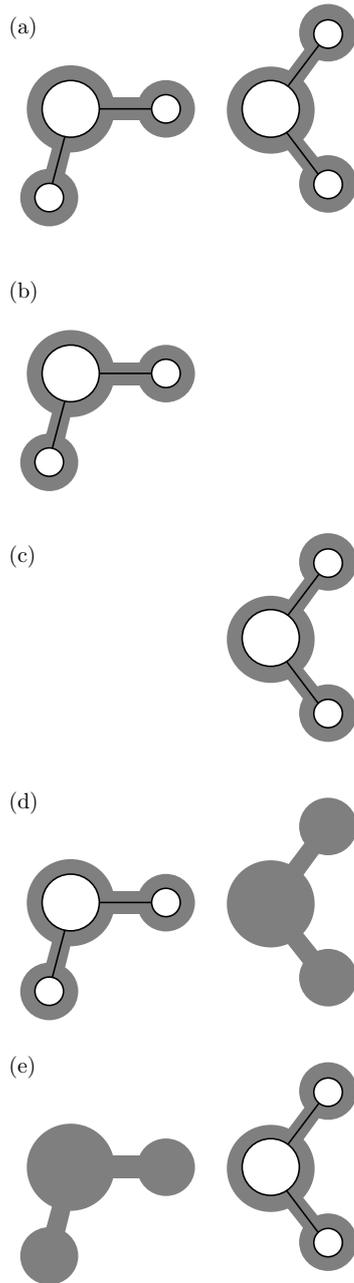


Fig. 1. Basis set superposition error occurs in the calculation of the binding energy of the dimer shown in (a) when the molecular energies are calculated as in (b) and (c) with only the basis functions (in grey) of the atoms (white disks) associated with the individual molecule. The counterpoise correction calculates the molecular energies as in (d) and (e) with the same basis set as for the dimer (a).

involved in both cases, these methods are expected to be susceptible to BSSE. However in the ONETEP method the NGWFs are iteratively improved during the calculation by optimizing an expansion in terms of the psinc basis (defined in [21]) which is equivalent to a set of plane-waves. This work shows that this optimisation eliminates BSSE.

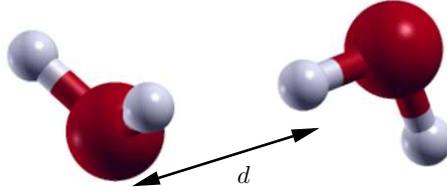


Fig. 2. Water dimer geometry showing the hydrogen bond O-H distance  $d$ .

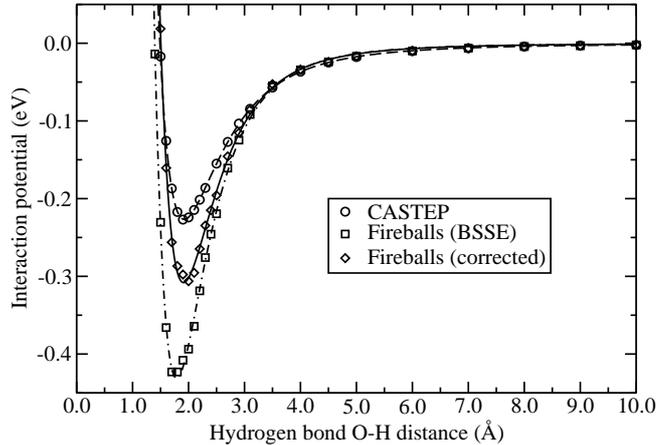


Fig. 3. Interaction potential of a water dimer obtained using a minimal basis set of fireballs

### 3 Results and discussion

The system selected for this study is the water dimer illustrated in Fig. 2, for which there are comprehensive reviews of previous experimental [22] and first-principles computational [23] work. The water dimer is bound by a medium strength hydrogen bond which has been important for the development of the theory of intermolecular forces and is seen as a paradigm of the bonding upon which many biochemical processes so sensitively depend. As a relatively weak bond, it is a challenge for DFT [24,25] and calculations of its binding energy are very sensitive to BSSE [26]. For the purposes of this study comparison is made with results from the BSSE-free plane-wave CASTEP code [27] by varying the hydrogen bond O-H length  $d$  while the orientation and internal geometry of the water molecules remained fixed: structural rearrangements [28] are not considered. As a result, the geometries do not correspond to experimental or computationally optimised structures. The ONETEP and CASTEP calculations used the same norm-conserving pseudopotentials [29] and gradient-corrected exchange-correlation functional [30]. Equivalent energy cut-offs were used for the psinc and plane-wave basis sets.

Figure 3 compares the interaction potential curve from CASTEP with results obtained from ONETEP without *in situ* optimisation of the NGWFs, for which spherically confined pseudoatomic orbitals (fireballs) [31] are used, as in the

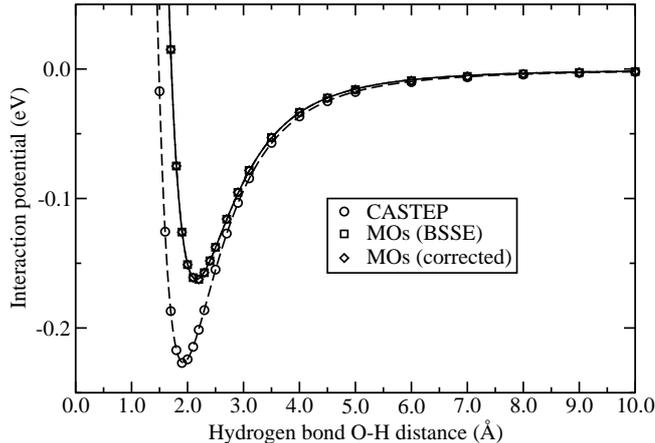


Fig. 4. Interaction potential of a water dimer obtained using a minimal basis set of local orbitals pre-optimised for isolated water molecules (MOs)

SIESTA [17] and DMol<sup>3</sup>[32] codes. This particular choice mimics the use of a minimal LCAO set and involves only kernel optimisation. The uncorrected result overestimates the binding energy by almost 100%, and while the counterpoise correction reduces the error by half, significant disagreement with the CASTEP result remains. This demonstrates the presence of significant BSSE and the need for high-quality basis sets to describe hydrogen bonds. As BSSE results from the reduction in variational freedom for the isolated molecules compared to the dimer, when the dimer basis is used for all three calculations the molecular energies are lowered significantly: by about 65 meV each at  $d=1.8$  Å. While this only amounts to a 0.4% change in the total energy, this correction is significant in calculating the binding energy where significant cancellation occurs in Eq. 1. BSSE is therefore present as expected in linear-scaling methods with fixed local orbitals. While the error may be reduced by employing larger sets of local orbitals than the minimal set used here, the counterpoise correction is always necessary [33].

Since BSSE arises from a relatively poor description of the isolated molecules compared to the dimer, one possible solution which would still retain fixed local orbitals for large-scale calculations would be to use pre-optimised NGWFs corresponding to molecular orbitals (MOs). For the case of the water dimer, this scheme involves first optimizing NGWFs with ONETEP for the isolated water molecules. These NGWFs are then used as the fixed local orbitals for the dimer calculation, in which only the kernel is optimised. In the general case, this approach is essentially equivalent to a third linear-scaling method, Yang’s density-matrix divide-and-conquer scheme [34], in which a large system is split into smaller components for which density-matrices are calculated before being combined to generate an estimate of the total density-matrix of the whole system. Figure 4 compares the CASTEP result with this choice of pre-optimised NGWFs, which lowers the total energy of the isolated water molecules by almost 0.3 eV. BSSE is indeed eliminated since there is nothing

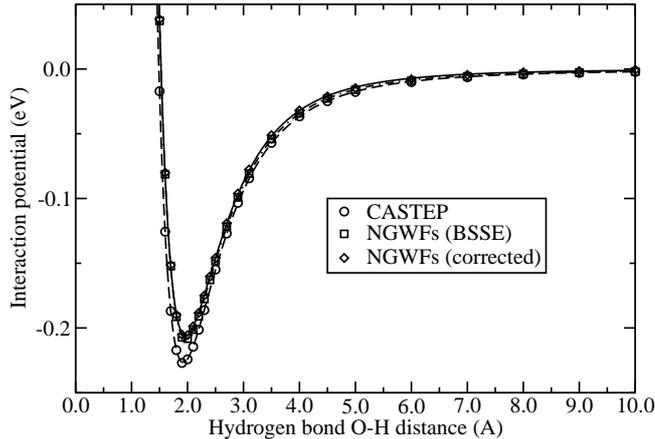


Fig. 5. Interaction potential of a water dimer obtained using a minimal basis set of fully optimised orbitals (NGWFs)

to be gained by adding the MOs of a second molecule to the local orbital set already optimised for the first. Instead of the overestimation of the binding energy in Fig. 3 there is now significant underestimation since the MOs give a poor description of the dimer when there is significant interaction between the molecules at short distances, and this is reflected in the long equilibrium bond distance predicted. These results demonstrate that while this divide-and-conquer approach does avoid BSSE, it does not provide a sufficiently accurate solution for hydrogen-bonded systems, and full optimisation of the local orbitals is required for the system as a whole.

Allowing full optimisation of the NGWFs as well as the kernel for both molecules and dimer results in Fig. 5. The dimer energy is now lowered substantially (by over 0.1 eV for  $d=1.8$  Å) compared to the use of fixed local orbitals. BSSE is again notably absent and furthermore there is extremely close agreement between the CASTEP and ONETEP interaction potential curves. Table 1 gives results within the local density approximation for the equilibrium bond length and binding energy for fully converged calculations from ONETEP, CASTEP and NWChem [35], an all-electron code with an “aug-cc-pVTZ” Gaussian-type orbital basis. The equilibrium bond lengths for this orientation agree to 0.3% and the bond energies to 2%. This demonstrates that *in situ* optimisation of the NGWFs in terms of the systematic psinc basis is indeed sufficient to eliminate BSSE entirely and to describe the hydrogen bond with plane-wave accuracy.

## 4 Conclusions

For subtle interactions like hydrogen bonds important for many potential applications of linear-scaling methods, it has been demonstrated that fixed local

Table 1

Comparison of bond lengths and binding energies for the water dimer. ( $1 \text{ kcal mol}^{-1} = 43.4 \text{ meV}$ )

	Bond length (Å)	Binding energy (meV)
ONETEP (1300 eV)	1.790	182
CASTEP (1300 eV)	1.785	178
NWChem (aug-cc-pVTZ)	1.779	181

orbital schemes suffer from BSSE and divide-and-conquer approaches are insufficiently accurate. In contrast the ONETEP method is BSSE-free and attains the accuracy of traditional plane-wave and near-complete Gaussian basis calculations as a result of the *in situ* optimisation of local orbitals.

P.D.H. and C.-K.S. thank the Royal Society for University Research Fellowships. A.A.M. thanks Christ's College for a Research Fellowship.

## References

- [1] D. Alfè, M. J. Gillan, G. D. Price, *Nature* 401 (1999) 462.
- [2] M. T. Green, J. H. Dawson, H. B. Gray, *Science* 304 (2004) 1653.
- [3] J. Junquera, P. Ghosez, *Nature* 422 (2003) 506.
- [4] G. Galli, *Curr. Opin. Solid St. M.* 1 (1996) 864.
- [5] S. Goedecker, *Rev. Mod. Phys.* 71 (1999) 1085.
- [6] H. B. Jansen, P. Ros, *Chem. Phys. Lett.* 3 (1969) 140.
- [7] S. F. Boys, F. Bernardi, *Mol. Phys.* 19 (1970) 553.
- [8] R. Rurali, N. Lorente, P. Ordejón, *Phys. Rev. Lett.* 95 (2005) 209601.
- [9] I. Mayer, *Int. J. Quantum Chem.* 23 (1983) 341.
- [10] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, J. D. Joannopoulos, *Rev. Mod. Phys.* 64 (1992) 1045.
- [11] W. Kohn, *Phys. Rev. Lett.* 76 (1996) 3168.
- [12] W. Kohn, *Phys. Rev.* 115 (1959) 809.
- [13] S. Ismail-Beigi, T. Arias, *Phys. Rev. Lett.* 82 (1999) 2127.
- [14] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi, M. C. Payne, *J. Chem. Phys.* 122 (2005) 084119.

- [15] E. Hernández, M. J. Gillan, C. M. Goringe, *Phys. Rev. B* 55 (1997) 13485.
- [16] C.-K. Skylaris, A. A. Mostofi, P. D. Haynes, O. Diéguez, M. C. Payne, *Phys. Rev. B* 66 (2002) 035119.
- [17] J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón, D. Sánchez-Portal, *J. Phys.: Condens. Matter* 14 (2002) 2745.
- [18] G. E. Scuseria, *J. Phys. Chem. A* 103 (1999) 4782.
- [19] M. Challacombe, *J. Chem. Phys.* 110 (1999) 2332.
- [20] J. L. Fattebert, J. Bernholc, *Phys. Rev. B* 62 (2000) 1713.
- [21] A. A. Mostofi, P. D. Haynes, C.-K. Skylaris, M. C. Payne, *J. Chem. Phys.* 119 (2003) 8842.
- [22] G. T. Fraser, *Int. Rev. Phys. Chem.* 10 (1991) 189.
- [23] S. Scheiner, *Annu. Rev. Phys. Chem.* 45 (1994) 23.
- [24] X. Xu, I. W. A. Goddard, *J. Phys. Chem. A* 108 (2004) 2305.
- [25] J. Ireta, J. Neugebauer, M. Scheffler, *J. Phys. Chem. A* 108 (2004) 5692.
- [26] J. Garza, J.-Z. Ramírez, R. Vargas, *J. Phys. Chem. A* 109 (2005) 643.
- [27] M. D. Segall, P. J. D. Lindan, M. J. Probert, C. J. Pickard, P. J. Hasnip, S. J. Clark, M. C. Payne, *J. Phys.: Condens. Matter* 14 (2002) 2717.
- [28] T. Taketsugu, D. J. Wales, *Mol. Phys.* 100 (2002) 2793.
- [29] V. Milman, M.-H. Lee, *J. Phys. Chem.* 100 (1996) 6093.
- [30] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* 77 (1996) 3865.
- [31] O. F. Sankey, D. J. Niklewski, *Phys. Rev. B* 40 (1989) 3979.
- [32] B. Delley, *J. Chem. Phys.* 113 (2000) 7756.
- [33] E. Artacho, M. Machado, D. Sánchez-Portal, P. Ordejón, J. M. Soler, *Mol. Phys.* 101 (2003) 1587.
- [34] W. Yang, T.-S. Lee, *J. Chem. Phys.* 103 (1995) 5674.
- [35] H. P. C. C. Group, NWChem, A Computational Chemistry Package for Parallel Computers, Version 4.5, Pacific Northwest National Laboratory, Richland, Washington 99352, USA, 2003.