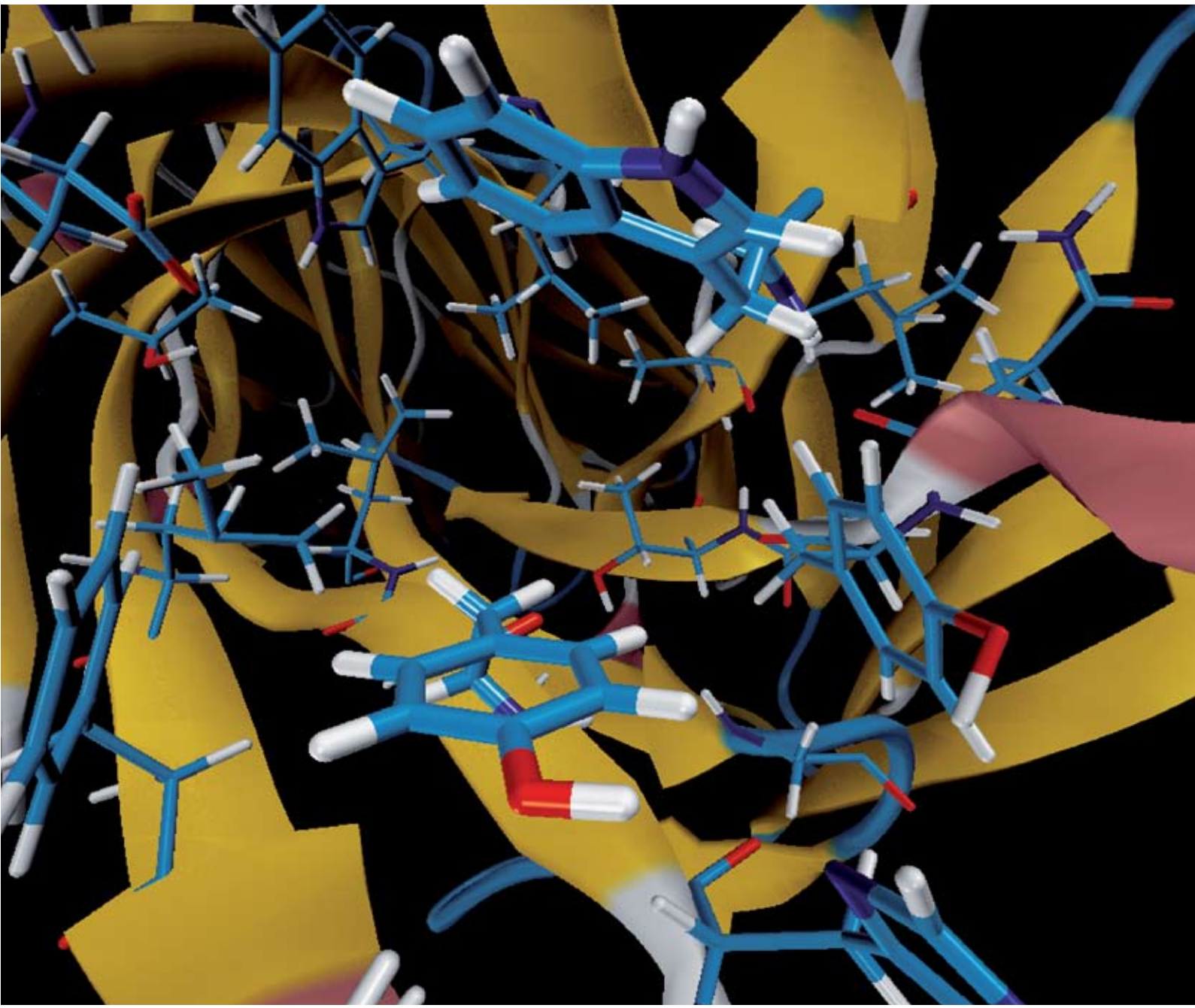
# PCCP

## Physical Chemistry Chemical Physics

www.rsc.org/pccp

Volume 10 | Number 16 | 28 April 2008 | Pages 2125–2272



ISSN 1463-9076

**COVER ARTICLE**
Tuttle and Thiel
OM*x*-D: semiempirical methods with orthogonalization and dispersion corrections. Implementation and biochemical application

**PERSPECTIVE**
Bhushan and Nosonovsky
Phase behavior of capillary bridges: towards nanoscale water phase diagram

1463-9076(2008)10:16;1-1

# OM*x*-D: semiempirical methods with orthogonalization and dispersion corrections. Implementation and biochemical application†

**Tell Tuttle*[a] and Walter Thiel*[b]**

*Received 5th December 2007, Accepted 25th January 2008*
*First published as an Advance Article on the web 25th February 2008*
**DOI: 10.1039/b718795e**

The semiempirical methods of the OM*x* family (orthogonalization models OM1, OM2, and OM3) are known to describe biochemical systems more accurately than standard semiempirical approaches such as AM1. We investigate the benefits of augmenting these methods with an empirical dispersion term (OM*x*-D) taken from recent density functional work, without modifying the standard OM*x* parameters. Significant improvements are achieved for non-covalent interactions, with mean unsigned errors of 1.41 kcal/mol (OM2-D) and 1.31 kcal/mol (OM3-D) for the binding energy of the complexes in the JSCH-2005 data base. This supports the use of these augmented methods in quantum mechanical/molecular mechanical (QM/MM) studies of biomolecules, for example during system preparation and equilibration. As an illustrative application, we present QM and QM/MM calculations on the binding between antibody 34E4 and a hapten, where OM3-D performs better than the methods without dispersion terms (AM1, OM3).



## 1. Introduction

The use of hybrid quantum mechanical/molecular mechanical (QM/MM) methods to study systems of biochemical interest has become increasingly popular in recent years. There has been a steady increase in the number of publications on QM/MM applications to biochemical systems since the mid-1990s, with a particular rise in the publication rate since 2003.[1] Concomitant with the increased use of these methods, as well as more sophisticated analysis techniques, has been the ability to more readily identify the potential sources of error that can arise within a QM/MM approach. Several papers have recently highlighted the manner in which choices made at the level of system preparation can affect the final results (*e.g.*, reaction mechanisms, property calculations, *etc.*) of QM/MM calculations.[2–5]

The preparation of a system for QM/MM studies is time-consuming. One of the challenges often faced during this initial stage concerns the treatment of the "chemically interesting" (QM) region. A reliable MM description is often unavailable as this region may contain ligands (or substrates) that are not included in the standard biochemical force field. This problem is commonly overcome by performing a specific MM parameterization of the ligand; however, this can be quite laborious, and the compatibility with the existing force field is not easily verifiable. Alternatively, one may fix the geometry of the ligand during system preparation and assign "reasonable" charges and van der Waals parameters for the non-covalent interactions with the environment; in this case, the choice of the orientation of the ligand will introduce some bias. Hence, both of these approaches have their disadvantages. Furthermore, the use of either strategy in defining an MM representation for the QM region becomes increasingly difficult if the system is prepared and equilibrated as a transition structure (*e.g.*, for the calculation of enzymatic reaction mechanisms involving charge transfer).[4,6,7] Ideally, one would prefer a consistent treatment at the QM/MM level throughout the entire study.

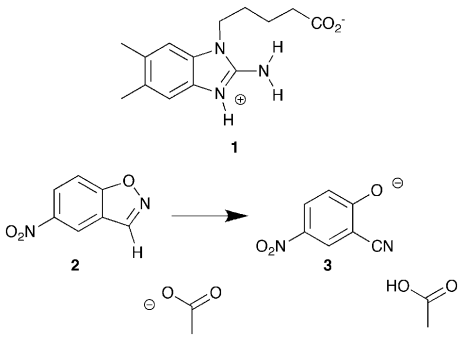
In this context semiempirical QM/MM approaches can be an attractive option. Recently several groups have addressed the development and refinement of semiempirical QM treatments for biomolecular applications.[1,8–20] The present paper examines the latest semiempirical method with orthogonalization corrections (Orthogonalization Model OM3[21]) and its predecessors (OM2[22,23] and OM1[20,24]) and introduces augmented versions (OM*x*-D) with an added empirical dispersion term taken from the recent work by Grimme on density functional theory (DFT).[25]

Empirical dispersion functions have been used before to augment QM methods, *e.g.*, already in 1977 for the Hartree–Fock method[26] and in 2001 for the self-consistent-charge density functional tight binding method (SCC-DFTB).[27] Hillier *et al.* have recently incorporated dispersion terms into the semiempirical AM1[28] and PM3[29] methods, with partial re-parameterization.[13] The relation between this previous work and our present implementation will be discussed in section 3A.

As an initial application, we report the QM/MM equilibration of antibody 34E4, with its bound hapten (**1**, Scheme 1),

[a] *WestCHEM, Department of Pure and Applied Chemistry, University of Strathclyde, Glasgow, UK G1 1XL. E-mail: tell.tuttle@strath.ac.uk; Fax: +44 141 548 4822; Tel: +44 141 548 2290*
[b] *Max-Planck-Institut für Kohlenforschung, D-45470 Mülheim an der Ruhr, Germany. E-mail: thiel@mpi-muelheim.mpg.de; Fax: +49 208 306 2996; Tel: +49 208 306 2150*

† Electronic supplementary information (ESI) available: Details of the statistics and individual energies from the JSCH-2005 database evaluation, extensive material describing the setup and preparation of the 34E4 system, and active atoms in the MD simulations and QM/MM geometry optimizations. Structure of antibody 34E4 in PDB format. See DOI: 10.1039/b718795e



**Scheme 1** The crystal structure of 34E4 contains the bound hapten (**1**). 34E4 catalyzes the conversion of benzisoxazole (**2**) to salicylonitrile (**3**).

using OM3-D and other semiempirical methods as QM components. Antibody 34E4 catalyzes the conversion of benzisoxazoles (**2**) to salicylonitriles (**3**) with high efficiency.[30] The well-organized transition state is expected to be stabilized both by electrostatic interactions and by π-stacking dispersive interactions (Fig. 1).[31] In the current study, we focus on the ability of the semiempirical methods to describe the binding site with the hapten present. The mechanism and factors contributing to the excellent efficiency of this antibody in catalyzing the conversion of benzisoxazoles will be discussed in a separate publication.

## 2. Computational methods

### (A) Parameterization of the dispersion terms

The OM*x* methods include orthogonalization corrections into the one-electron part of the Hamiltonian to account for effects such as Pauli repulsion that arise, at the *ab initio* level, from the transformation of the secular equations from a non-orthogonal to an orthogonal basis. These corrections are incorporated only into the one-center one-electron terms in OM1,[20,24] but also into the two-center one-electron terms in OM2.[22,23] Some of the latter tend to be small and are disregarded in OM3,[21] which is thus a simplified (and somewhat faster) variant of OM2. The results from the OM*x* methods are generally superior to those from established MNDO-type semiempirical methods such as AM1[28] and PM3.[29] OM2 and OM3 results are normally of similar quality, and usually somewhat more accurate than OM1 results.[21,32]

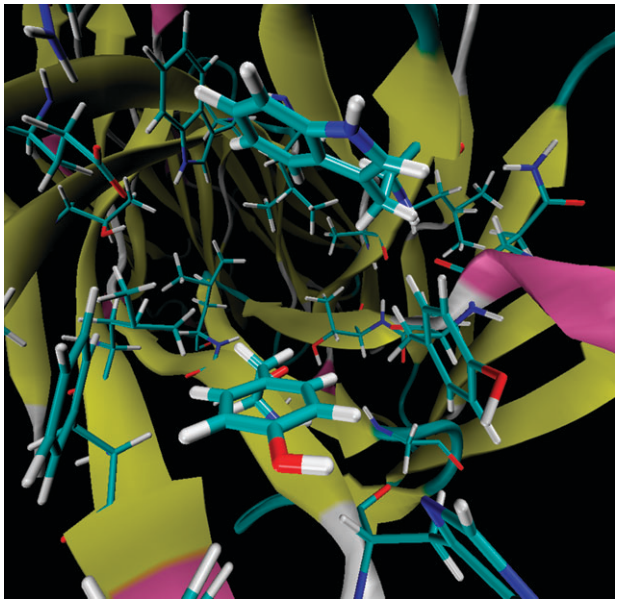


**Fig. 1** Binding site of 34E4 taken from the crystal structure (PDB ID: 1YOL). The residues involved in the expected π-stacking interaction with the hapten (not shown) are highlighted.

As in the case of DFT, dispersion is formally not included in semiempirical molecular orbital methods. Given the improved performance of DFT methods after augmentation with an empirical dispersion function,[25] we decided to test this approach also for the OM*x* methods. Our basic strategy is to keep the OM*x* formalism and parameters unchanged and to explore the effects of adding empirical dispersion terms of the form that had proven to be successful in previous DFT work.[25,33–35] The total energy of the system is given by:

$$E_{\mathrm{tot}} = E_{\mathrm{OM}x} + E_{\mathrm{disp}} \qquad (1)$$

where $E_{\mathrm{OM}x}$ is the energy from a standard OM*x* calculation, and $E_{\mathrm{disp}}$ is the empirical dispersion correction given by:

$$E_{\mathrm{disp}} = -s_6 \sum_{i=1}^{N_{\mathrm{at}}-1} \sum_{j=i+1}^{N_{\mathrm{at}}} \frac{C_6^{ij}}{R_{ij}^6} f_{\mathrm{dmp}}(R_{ij}) \qquad (2)$$

Here, $N_{\mathrm{at}}$ is the number of atoms in the system, $R_{ij}$ is the distance between atoms $i$ and $j$, and $s_6$ is a global scaling factor. $C_6^{ij}$ is the dispersion coefficient for atom pair $ij$ and is calculated from pre-defined atomic coefficients as:

$$C_6^{ij} = 2\frac{C_6^i C_6^j}{C_6^i + C_6^j} \qquad (3)$$

Among the different combination rules that have been investigated,[36] we adopt the simple average of eqn (3).[25] The damping function in eqn (2) is necessary to avoid singularities as $R_{ij} \rightarrow 0$; its form is the same as in previous DFT work:[25]

$$f_{\mathrm{dmp}}(R) = \frac{1}{1 + e^{-\alpha(R/R_0 - 1)}} \qquad (4)$$

where $R_0$ is calculated as the sum of pre-defined atomic radii $R_i$. The $C_6$ and $R_i$ parameters (Table 1) are taken from the work of Grimme.[25] They are not reoptimized as previous investigations have shown them to be sufficiently accurate and transferable.[13,25]

In our semiempirical implementation of the dispersion function, there are thus only two adjustable parameters, *i.e.*, the global scaling factor ($s_6$) and the damping coefficient (α). Within DFT the optimum value of $s_6$ depends on the chosen functional.[25] BP86 and BLYP (which do not account for dispersion) have similar optimum $s_6$ values of 1.3 and 1.4, respectively, whereas the PBE functional (which mimics some

**Table 1** Atomic $C_6$ parameters and radii

| | H | C | N | O | F |
|---|---|---|---|---|---|
| $C_6$ [J nm$^6$ mol$^{-1}$] | 0.16 | 1.65 | 1.11 | 0.70 | 0.57 |
| $R_i$ [pm] | 111 | 161 | 155 | 149 | 143 |

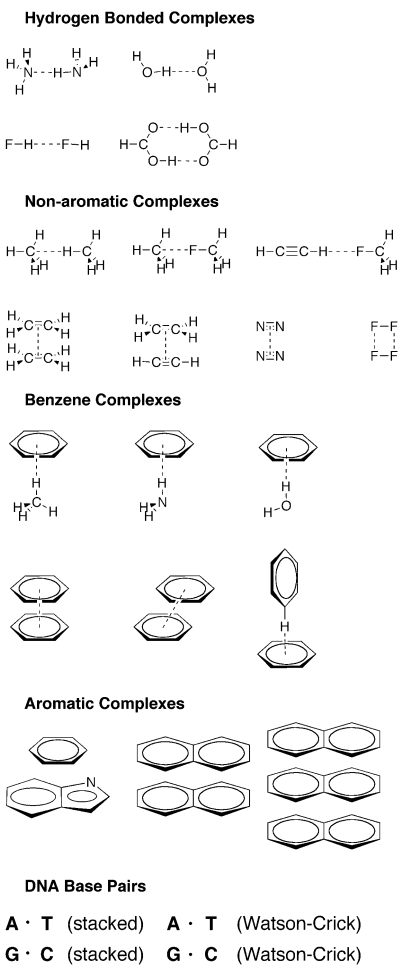


mid-range dispersion) has a smaller scaling factor of 0.7.[25] We have thus scanned the $s_6$ values over the range 0.7–1.6 in increments of 0.1.

In the original DFT-D implementation of the damping function (eqn (4)) the value of the damping coefficient was fixed at $\alpha = 23$.[25,33] In the current work this value was varied over the range 18–27 in increments of 1, to allow for greater flexibility. The complexes that comprise the training set (Scheme 2) feature H-bonding, π-stacking, and a combination of the two interactions. The training set is the same as that used in the original DFT-D parameterization by Grimme.[25]

It should be noted that the chosen empirical dispersion correction is not unique. For example, a recently proposed alternative scheme[37] employs the functional form of eqn (2) and adjusts the damping coefficient ($\alpha$) and a global scaling factor ($s_R$) for the atomic radii, while keeping $s_6 = 1$ for most (but not all) of the density functions and basis sets considered, in conjunction with modified combination rules for the radii and the $C_6$ coefficients. Although these variations[37] as well as re-optimization of selected semiempirical parameters[13] may yield further improvements, we have decided against such fine tuning because our primary goal is to examine the performance of the original DFT-D dispersion correction[25] in the OM*x* framework, with minimal re-parameterization.




**Scheme 2** Training set used in optimizing the dispersion function.

The present validation of the OM*x*-D approaches and the comparison with other semiempirical methods employs both the S22 and JSCH-2005 sets of complexes, which have been assembled to cover the full spectrum of non-covalent interactions commonly found in biochemical systems.[38] The set used in this work is slightly reduced from the complete data base as all (9) complexes involving sulfur were excluded as were the four complexes identified previously as being ambiguously assigned,[39] resulting in a test set of 152 complexes. All semiempirical calculations were performed with a development version of the MNDO99 program.[40]

### (B) QM/MM calculations

The structure of antibody 34E4 with a bound hapten, available from the protein data bank (PDB ID:1YOL),[31] was the starting point of the QM/MM study. Chains L and H of the protein were used to model the system. The program REDUCE,[41,42] which accounts for the generation of H-bond networks and potential steric clashes, was employed to add hydrogen atoms and to adjust the orientation of Asn, Gln and His side chains where necessary. All changes were checked by visual inspection and the complete list of changes made to the structure is provided in the Supporting Information.† The protonation states of ionizable residues were assigned with the empirical p$K_a$ prediction program, PROPKA.[43]

The system was hydrated using the droplet model with a 25 Å sphere of equilibrated TIP3P water, with the origin defined as the geometric centre of **1** (Scheme 1). All waters that overlapped[44] with the protein were deleted. The remaining water molecules were relaxed and subsequently subjected to a molecular dynamics (MD) run for 100 ps at 300 K using Langevin dynamics. A stochastic boundary potential[45] was used to maintain the structure of the water sphere, while all non-water atoms were frozen during both the geometry minimization and the equilibration period. This hydration procedure was performed five times, at which point the number of added water molecules remained approximately constant. All these preparatory calculations were carried out with the CHARMM program,[46–48] using the standard CHARMM force field.[49]

The QM/MM calculations were done with three different semiempirical QM methods—AM1,[28] OM3,[21] and OM3-D. In each case the CHARMM force field was used for the MM part. The modular program package CHEMSHELL[50,51] was used for all QM/MM calculations, where the QM energy and gradients were provided by a development version of MNDO99.[40] CHEMSHELL's internal force field driver using the CHARMM parameter and topology data provided the MM energy and gradients. No electrostatic cut-offs were used, neither between the QM and MM regions nor within the MM region alone. The QM density was electrostatically embedded into the MM environment by including the point charges from the MM atoms into the QM Hamiltonian. The boundary region was treated with the charge-shift scheme,[51,52] which provides a reliable treatment of the coupling between the QM and MM zones.[53] The bonded and van der Waals interactions between the QM and MM regions were handled at the MM level, as described previously.[51]

The QM region contains the hapten (**1**) and the side chains of several residues that compose the binding site: $Trp^{L91}$, $Trp^{H33}$, $Tyr^{H100D}$, and $Glu^{H50}$ resulting in 102 QM atoms. The covalent bonds across the QM/MM boundary are the C–$C^{\alpha}$ bond of each side chain for the residues listed above; in the QM calculations the carbon atom is saturated with an H-link atom. The cut between the QM and MM regions coincides with the CHARMM charge-group definitions such that the QM and MM regions both have integer charge.

In the QM/MM MD simulation an active subset of the total system is allowed to move while the remaining atoms are frozen. The active subset was defined from the initial complex geometry using a distance criterion, whereby any residue that contains an atom within 10 Å of *any* atom of **1** is included. The resulting active region contains 1247 atoms, about one tenth of the total system size 11 393 atoms.

The MD simulations were performed under NVT conditions at $T = 300$ K. During the heating phase (10 ps) the temperature was controlled by a Berendsen thermostat,[54] with a coupling time of 0.1 ps. During the equilibration phase (50 ps) the Nose–Hoover chain thermostat,[55–57] as implemented in the ChemShell dynamics module,[51] was used. All hydrogen atoms were assigned the mass of deuterium and the free water molecules were kept rigid using SHAKE constraints.[58] The time step for both heating and equilibration was 1 fs.

QM/MM geometry optimizations were performed with HDLCOpt[59]—a linear scaling, microiterative algorithm that employs a set of hybrid delocalized coordinates—as implemented in ChemShell. The residues defined for HDLCOpt were taken as the standard CHARMM residues. The QM/MM setup was consistent with that used in the MD equilibration.

## 3. Results and discussion

### (A) Parameterization and performance of OM*x*-D

Optimum $\alpha$ and $s_6$ values were determined for the OM3 method by a grid search covering all possible combinations of parameter values (see section 2A). In each case, the binding energy (BE) was computed through geometry optimization of the complex and its monomers. Comparison with the best available binding energy[25] for each complex in the training set (see section 2A) then gave the root-mean-square deviation (RMSD) for a given combination of parameter values. The five best combinations are presented in Table 2 (a full list is provided in the Supporting Information†). It is obvious that the RMSD values are not very sensitive to the choice of the damping coefficient $\alpha$. For $s_6 = 1.1$, lower damping coefficients ($\alpha$ = 17,16) do not lead to a further reduction of RMSD, and we have therefore selected the values of $s_6 = 1.1$ and $\alpha = 18$ for the dispersion function in OM3-D. We have also run analogous grid searches for OM1 and OM2 and have obtained essentially the same optimum parameters (OM1-D: $s_6 = 1.2$, $\alpha = 18$; OM2-D: $s_6 = 1.0$, $\alpha = 18$). Since the small change in $s_6$ has almost no effect on RMSD (0.031 kcal/mol and 0.008 kcal/mol, respectively; see Supporting Information†) we adopt the OM3-D parameters also for OM1-D and OM2-D, for the sake of consistency.

**Table 2** RMSD of OM3-D binding energy for the variation of $s_6$ and $\alpha$[a]

| $s_6$ | $\alpha$ | RMSD |
|---|---|---|
| 1.1 | 18 | 1.985 |
| 1.1 | 19 | 1.991 |
| 1.1 | 20 | 1.994 |
| 1.1 | 21 | 1.998 |
| 1.1 | 22 | 2.007 |

[a] RMSD in kcal/mol.

The comparison of the OM3-D calculated BEs with the best available BEs for the training set does reveal some problem cases (Table 3). For example, the stability of the naphthalene trimer is underestimated (BE(OM3-D) = −3.4; BE(exp) = −8.7 kcal/mol). Similarly, the ordering of the benzene dimers is not consistent with the best estimates of their BEs. Nonetheless, the dispersion terms clearly stabilize the dispersion-bound complexes in a realistic manner, and the BEs for the H-bonded complexes are also slightly improved. The overall agreement between the results obtained with the OM3-D method and the reference data is reasonable, particularly in light of the inherent simplifications of the OM3 approach.

For further validation, we now turn to the large JSCH-2005 and S22 data bases which provide interaction energies for a wide range of non-covalent complexes. To allow for direct comparisons, our calculations were carried out at the geometries provided in the database, and the energies of the

**Table 3** Comparison of OM3 and OM3-D calculated binding energies for the training set with the reference values. All energies in kcal/mol

| Complex | OM3 | OM3-D | Ref. | Ref. method[a] |
|---|---|---|---|---|
| $NH_3$-dimer | −2.4 | −3.2 | −3.0 | MP4//MP2[b] |
| $H_2O$-dimer | −4.5 | −5.2 | −4.8 | CCSD(T)//MP2[c] |
| HF-dimer | −1.2 | −1.5 | −4.4 | CCSD(T)//MP2[c] |
| HCOOH-dimer | −16.7 | −18.5 | −13.9 | CCSD(T)//MP2[c] |
| $CH_4$-dimer | 0.0 | −0.6 | −0.5 | MP2[c] |
| $CH_3F$–$CH_4$ | 0.0 | −0.4 | −0.7 | MP2[d] |
| $C_2H_2$–$CH_3F$ | 0.0 | −0.2 | −1.7 | MP2[d] |
| $C_2H_4$-dimer | 0.0 | −0.7 | −1.3 | CCSD(T)//MP2[c] |
| $C_2H_2$–$C_2H_4$ | 0.0 | −0.5 | −1.52 | MP2[d] |
| $N_2$-dimer | 0.0 | −0.1 | −0.33 | MP2[d] |
| $F_2$-dimer | 0.0 | −0.3 | −0.27 | MP2[d] |
| $C_6H_6$–$CH_4$ | −0.1 | −1.6 | −1.6 | MP2[d] |
| $C_6H_6$–$NH_3$ | −0.6 | −1.6 | −2.4 | MP2[d] |
| $C_6H_6$–$H_2O$ | −1.0 | −1.9 | −3.9 | MP2[e] |
| $C_6H_6$–S-dimer | 0.0 | −3.8 | −1.8 | CCSD(T)//MP2[f] |
| $C_6H_6$-PD-dimer | 0.0 | −4.2 | −2.8 | CCSD(T)//MP2[f] |
| $C_6H_6$–T-dimer | −0.2 | −2.4 | −2.7 | CCSD(T)//MP2[f] |
| $C_6H_6$-IND | 0.1 | −5.3 | −5.9 | Exp//MP2[g] |
| NAP-dimer | 0.0 | −8.3 | −6.2 | CCSD(T)//MP2[h] |
| NAP-trimer | 2.1 | −3.4 | −8.7 | Exp[i] |
| A:T-stack[j] | — | −9.5 | −11.6 | CCSD(T)//MP2[k] |
| G:C-stack[j] | — | −17.3 | −16.9 | CCSD(T)//MP2[k] |
| A:T-WC-dimer | −15.1 | −18.9 | −15.4 | CCSD(T)//MP2[k] |
| G:C-WC-dimer | −24.9 | −28.9 | −28.8 | CCSD(T)//MP2[k] |

[a] The first entry refers to the binding energy and the second, if different, refers to the geometry optimization. [b] Ref. 60. [c] Ref. 61. [d] Ref. 25. [e] Ref. 62. [f] Ref. 63. [g] Ref. 64. [h] Ref. 65. [i] Ref. 66. [j] The DNA base pair stacking complexes rotated during the optimization into H-bonded complexes, thus the OM3 binding energies are not reported. [k] Ref. 67.

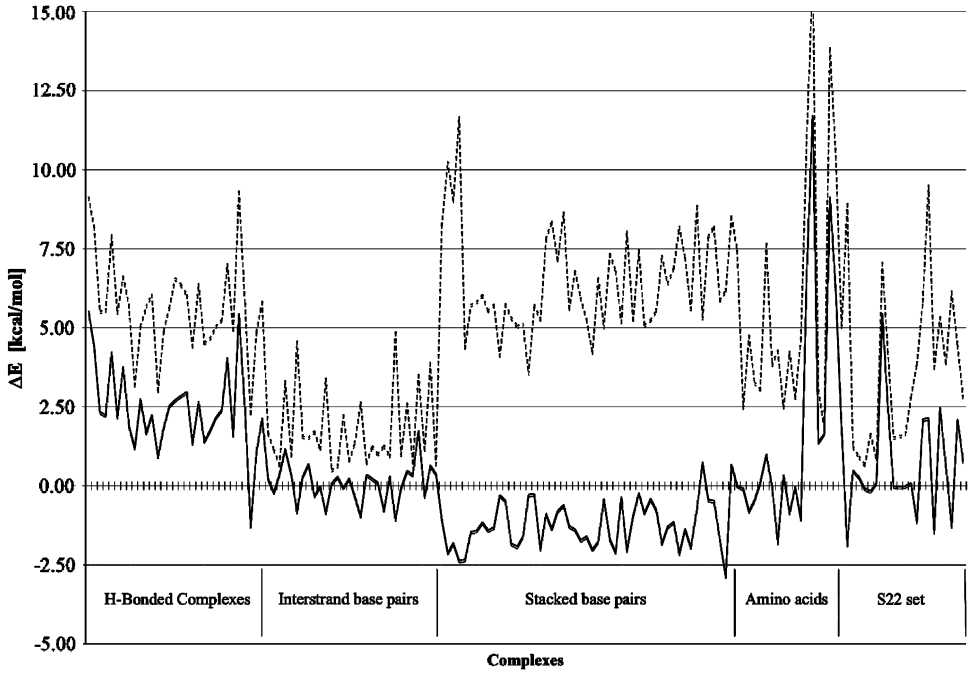



**Fig. 2** Deviation between calculated and reference interaction energies of the JSCH-2005 and S22 sets for OM3 (dotted) and OM3-D (solid).

monomers were computed at their complex geometry. We tested the performance of four semiempirical methods without dispersion terms (AM1, OM1, OM2, OM3) and of the three dispersion-corrected OM*x*-D approaches. As noted above, OM1-D, OM2-D, and OM3-D share the same dispersion function. The resulting interaction energies for all complexes are provided in the Supporting Information.†

It is evident that the inclusion of the dispersion correction into OM3 leads to a general improvement in the computed interaction energies (*cf*. Fig. 2). Not surprisingly, the largest benefits are encountered for the stacked base pairs where the OM3-D deviations are rather small (mean unsigned error of 1.26 kcal/mol) and systematic (interaction energies slightly overestimated). Improvements are also seen for the H-bonded complexes (still not quite stable enough in OM3-D) as well as for the interstrand base pairs and the S22 set. The larger errors observed for several amino acid complexes arise from charged amino acid pairs, which have significantly larger interaction energies. Thus, while the fractional error of these complexes is similar to the rest of the set, the absolute values are much larger. This was also noted in a comparison of the DFT-D methods for this data base.[39] The statistical analysis presented in Table 4 therefore excludes the charged complexes in the amino acid set, as well as the one charged complex in the H-bonded set (statistics including these complexes are available in the Supporting Information).†

The analysis shows that among the methods without dispersion terms, AM1 is significantly less accurate than any of the OM*x* methods. Within the latter family, OM2 and OM3 are superior to OM1, especially for H-bonded complexes (with smaller improvements for the other subsets). Both OM2 and OM3 predict the interaction energies with reasonable accuracy; overall OM2 marginally outperforms OM3.

The inclusion of the dispersion correction improves the accuracy of the OM*x* methods. In the case of OM1-D, the errors for the H-bond complexes remain rather large. For OM2-D and OM3-D, the results are again relatively similar. Given the limitations of semiempirical methods, the MUEs of 1.31 kcal/mol (OM3-D) and 1.41 kcal/mol (OM2-D) are considered acceptable, also in comparison with the MUEs of 0.48–0.75 kcal/mol reported for the investigated DFT-D functionals.[39]

In recent related work of Hillier *et al*., a dispersion correction was incorporated into the AM1[28] and PM3[29] methods in a somewhat different manner.[13] The usual functional form was chosen (as in DFT-D and in our case), but the parameters for the dispersion function were taken from the BLYP-D

**Table 4** Deviations between the semiempirical reference interaction energies for a large data base of non-covalent complexes[a]

| | AM1 | OM1 | OM1-D | OM2 | OM2-D | OM3 | OM3-D |
|---|---|---|---|---|---|---|---|
| *Mean Unsigned Error for the sub-sets* | | | | | | | |
| H-bonded complexes | 14.78 | 10.25 | 6.99 | 5.67 | 2.41 | 5.72 | 2.54 |
| Interstrand base pairs | 1.81 | 1.81 | 0.47 | 1.68 | 0.52 | 1.79 | 0.47 |
| Stacked base pairs | 10.67 | 6.68 | 1.08 | 6.36 | 1.52 | 6.53 | 1.26 |
| Amino acids | 4.52 | 4.46 | 0.47 | 3.27 | 1.17 | 3.96 | 0.60 |
| S22 | 7.02 | 5.32 | 2.52 | 3.28 | 1.15 | 3.81 | 1.23 |
| *Statistics for the full set* | | | | | | | |
| RMSE | 10.35 | 7.01 | 3.72 | 5.16 | 1.80 | 5.38 | 1.72 |
| MSE | 8.67 | 6.04 | 1.43 | 4.55 | −0.06 | 4.77 | 0.17 |
| MUE | 8.67 | 6.04 | 2.35 | 4.55 | 1.41 | 4.77 | 1.31 |
| MAXE-MINE | 23.17 | 14.15 | 13.96 | 9.76 | 8.56 | 11.17 | 8.39 |

[a] Statistics are reported for the modified version of the JSCH-2005 and S22 data bases. Complexes with ambiguous assignments, sulfur-containing complexes, and charged complexes are excluded from the statistical analysis. The resulting set contains 145 complexes. RMSE: Root Mean Square Error; MSE: Mean Signed Error; MUE: Mean Unsigned Error; MAXE-MINE: Difference between largest positive and largest negative errors. All values in kcal/mol.

parameterization ($\alpha$ = 23 and $s_6$ = 1.4),[25] and several of the standard AM1 and PM3 parameters were re-optimized ($U_{ss}$, $\beta_s$, and $\alpha$ for hydrogen; $U_{ss}$, $U_{pp}$, $\beta_s$, $\beta_p$ and $\alpha$ for carbon, nitrogen and oxygen).[13] Given the more extensive re-parameterization (18 modified atomic parameters per method compared with 2 general parameters in our case), it is not surprising that this approach yields slightly better interaction energies across the modified JSCH-2005 set (MUE(AM1-D) = 1.13 kcal/mol, MUE(PM3-D) = 1.26 kcal/mol). Nonetheless, the deviations for OM2-D and OM3-D are in the same ballpark (see above), and these methods are thus expected to perform comparably well for non-covalent complexes, without compromising the superior intramolecular description provided by the OM*x* methods.[21,32]

The geometries of non-covalent complexes, and in particular the intermolecular separations of the fragments, provide another important indicator of the ability to describe intermolecular interactions. Thus, the S22 set of complexes was re-optimized using the dispersion-corrected methods,[68] and the intermolecular distances were compared to those in the reference structures (see Supporting Information† for definitions and detailed results).

Considering the general limitations of semiempirical approaches, all three dispersion-corrected methods perform reasonably well in reproducing the reference geometries, with mean unsigned errors in intermolecular distances of 0.1–0.2 Å (OM1-D < OM3-D < OM2-D, see Table 5). As noted previously, the H-bond lengths predicted by OM2 and OM3 are generally underestimated by *ca.* 0.2 Å.[21,32] This trend is also observed in the H-bond complexes optimized in this work, since the inclusion of the dispersion function neither remedies nor worsens this shortcoming. In an overall assessment covering both interaction energies and intermolecular distances, the OM2-D and OM3-D methods perform similarly well, with slightly lower MUEs of OM3-D (see Tables 4 and 5). We have therefore chosen the OM3-D method for an illustrative QM/MM application which is described in the following.

### (B) Equilibration of 34E4 with the bound hapten

The binding site of 34E4 is stabilized by electrostatic and dispersive interactions (Fig. 3). The hapten is engaged in a π-stacking arrangement involving $Tyr^{H100D}$ and $Trp^{L91}$ (black lines), and is also well anchored into position by two strong ionic H-bonds with $Glu^{H50}$ (pink lines). The three aromatic amino acid side chains ($Tyr^{H100D}$, $Trp^{L91}$, and $Trp^{H33}$) are all stabilized by favorable dispersive interactions with neighbouring aromatic amino acids (black lines). In our QM/MM setup, the interactions of hapten in the π-stacking arrangement are described at the QM level and will thus benefit from the inclusion of dispersion terms in the QM method, while the other stabilizing interactions mentioned above are treated by the standard QM/MM formalism and are thus not affected by the OM3-D dispersion terms.

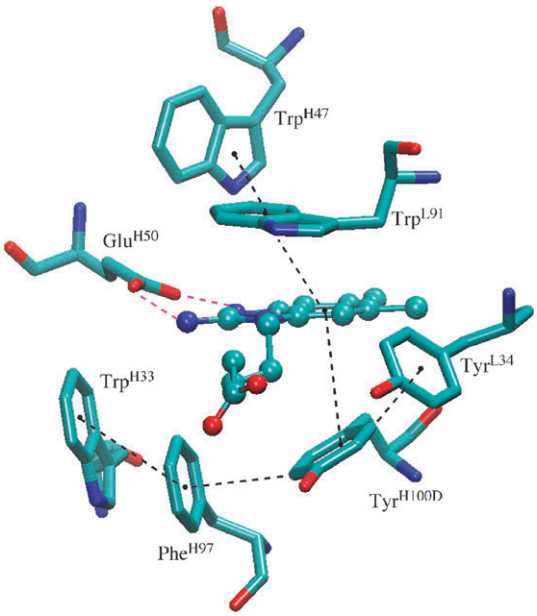



**Fig. 3** The dispersive interactions (black lines) stabilizing the binding site and the ionic H-bonds between $Glu^{H50}$ and the hapten (pink lines).

The results from QM/MM MD equilibration runs are listed in Table 6 for the three semiempirical QM methods that have been applied (AM1, OM3, OM3-D). In each case, the overall structure remains intact, and the RMSD values are rather similar, both for the overall structure and for the binding site. The higher RMSD value for the hapten reflects its higher mobility relative to the rest of the system and is consistent with the lack of covalent bonding to other residues. Comparing the OM3 and OM3-D data for the hapten, the inclusion of the OM3-D dispersion terms makes the hapten less mobile (lower RMSD), presumably due to the attractive dispersive interactions in the π-stack. Overall, however, the inclusion of the OM3-D dispersion term has little geometrical effect, indicating that in a well-structured binding site (such as in 34E4) the

**Table 5** Deviation between calculated and reference data intermolecular distances in the S22 set[a]

| | OM1-D | OM2-D | OM3-D |
|---|---|---|---|
| RMSE | 0.17 | 0.17 | 0.22 |
| MSE | 0.13 | −0.11 | −0.10 |
| MUE | 0.13 | 0.19 | 0.15 |
| MAXE-MINE | 0.38 | 0.68 | 0.73 |

[a] RMSE: Root Mean Square Error; MSE: Mean Signed Error; MUE: Mean Unsigned Error; MAXE-MINE: Difference between largest positive and largest negative errors. All values in Å.

**Table 6** RMSD of the QM/MM MD trajectory relative to the crystal structure[a]

| | AM1 | OM3 | OM3-D |
|---|---|---|---|
| Protein | 0.470 | 0.475 | 0.454 |
| Protein-QM | 0.475 | 0.478 | 0.457 |
| QM | 0.676 | 0.795 | 0.654 |
| QM-hapten | 0.405 | 0.475 | 0.454 |
| Hapten | 1.097 | 1.321 | 1.039 |

[a] RMSD of the active atoms in Å. The first row indicates the QM method used in the QM/MM MD equilibration. Protein – amino acid residues. Protein-QM – all MM amino acid atoms. QM – all QM atoms. QM-hapten – all QM amino acids atoms. Hapten – the hapten molecule.

**Table 7** QM and QM/MM interaction energies of the hapten with 34E4[a]

| QM-Method | $\Delta E_{i\text{-QM/MM}}$ | $\Delta E_{i\text{-QM}}$ | $\Delta E_{i\text{-QM-D}}$ |
|---|---|---|---|
| AM1 | −215.1 | −65.5 | 0 |
| OM3 | −236.5 | −80.2 | 0 |
| OM3-D | −287.7 | −118.1 | −26.0 |

[a] All energies in kcal/mol. $\Delta E_{i\text{-QM/MM}}$ is the interaction energy between the hapten and the full environment at the QM/MM level of theory. $\Delta E_{i\text{-QM}}$ is the full interaction energy between the hapten and the remainder of the QM region computed at the QM level of theory. $\Delta E_{i\text{-QM-D}}$ is the dispersion component included in $\Delta E_{i\text{-QM}}$.

surrounding environment is sufficiently ordered to maintain the binding site structure even in the absence of the QM–QM dispersive interactions.

This notion is quantified by the magnitude of the dispersion contribution to the interaction energy between the hapten and the antibody. As a typical example, the final snapshot of each QM/MM MD trajectory was investigated. The QM interaction energy (*i.e.*, the interaction energy between the hapten and the QM residues, $\Delta E_{i\text{-QM}}$) amounts to 30–40% of the total QM/MM interaction energy ($\Delta E_{i\text{-QM/MM}}$) between the hapten and the environment (Table 7). For the AM1 and OM3 methods, the QM interaction energy does not contain any empirical dispersion stabilization, while in the case of OM3-D, the dispersion term accounts for 22% of the QM interaction energy (*i.e.*, *ca.* 9% of $\Delta E_{i\text{-QM/MM}}$).

Clearly, the interaction of the environment with the hapten is strong enough to stabilize the binding site in each case. However, this does not imply that AM1 or OM3 offer a description as accurate as OM3-D. The intrinsic stability of the binding site can be examined by optimization of the QM region, in the absence of the environment. For this purpose, the final structure of the trajectory for each QM/MM MD run was optimized at the corresponding QM level in the gas phase.

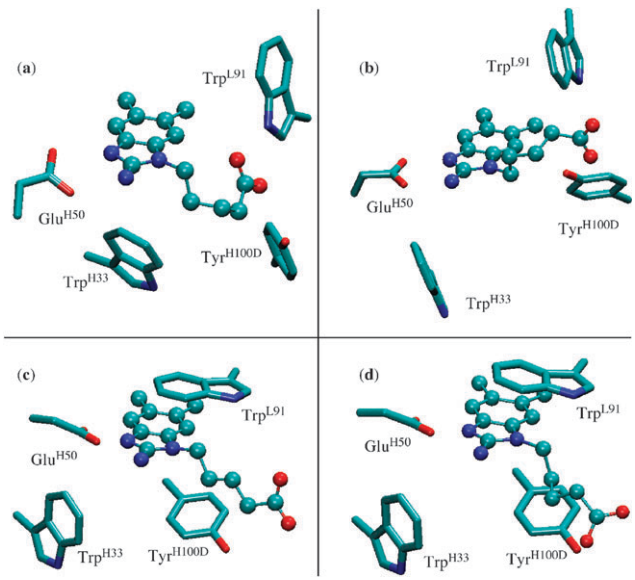



**Fig. 4** QM regions optimized without the environment: (a) QM = AM1, (b) QM = OM3, (c) QM = OM3-D; (d) binding site from crystal structure.

In the case of AM1 and OM3, the lack of dispersion forces results in a rearrangement of the binding site residues (Fig. 4). However, when the empirical dispersion is included in the OM3-D optimization the overall structure of the binding site is retained (*cf.* Fig. 4c and 4d).

## 4. Conclusions

The OM*x*-D methods investigated in this work provide an improved description of non-covalent biochemical interactions, compared to other semiempirical NDDO-based methods without dispersion terms. OM2-D and OM3-D are of similar accuracy. They reproduce the interaction energies from the JSCH-2005 data base with mean unsigned errors of 1.41 and 1.31 kcal/mol, respectively. All OM*x*-D methods provide reasonable geometries for the complexes from the S22 set; the inclusion of the dispersion terms has little effect on H-bond distances. The overall performance of the OM*x*-D methods in biochemical systems is encouraging and supports their use in biochemical QM and QM/MM investigations.

As an initial QM/MM application, the OM3-D method was employed for setting up and equilibrating the structure of the antibody 34E4 with a bound hapten. At the QM/MM level, the structural effect of including QM–QM dispersion is relatively minor because the pre-organized environment essentially determines the geometry of the binding site. The OM3-D dispersion terms contribute 26 kcal/mol to the binding of the hapten. In gas-phase optimizations of the QM region without the environment, the binding site remains intact only when including the OM3-D dispersion terms. The use of dispersion-corrected QM methods thus appears advantageous for a consistent semiempirical QM/MM modeling of biochemical systems.

## Acknowledgements

TT thanks the Royal Society of Edinburgh for support through the RSE-Scottish Executive Personal Research Fellowship.

## References

1 H. M. Senn and W. Thiel, *Top. Curr. Chem.*, 2007, **268**, 173–290.
2 A. Altun, S. Shaik and W. Thiel, *J. Comput. Chem.*, 2006, **27**, 1324–1337.
3 K. E. Ranaghan, L. Ridder, B. Szefczyk, W. A. Sokalski, J. C. Hermann and A. J. Mulholland, *Mol. Phys.*, 2003, **101**, 2695–2714.
4 T. Tuttle and W. Thiel, *J. Phys. Chem. B*, 2007, **111**, 7665–7674.
5 J. Zurek, A. L. Bowman, W. A. Sokalski and A. J. Mulholland, *Struct. Chem.*, 2004, **15**, 405–414.
6 D. Borgis and J. T. Hynes, *Chem. Phys.*, 1993, **170**, 315–346.
7 D. Riccardi, P. Schaefer, Y. Yang, H. B. Yu, N. Ghosh, X. Prat-Resina, P. Konig, G. H. Li, D. G. Xu, H. Guo, M. Elstner and Q. Cui, *J. Phys. Chem. B*, 2006, **110**, 6458–6469.
8 R. A. Kwiecien, M. Rostkowski, A. Dybala-Defratyka and P. Paneth, *J. Inorg. Biochem.*, 2004, **98**, 1078–1086.
9 M. Elstner, T. Frauenheim and S. Suhai, *J. Mol. Struct. (THEOCHEM)*, 2003, **632**, 29–41.
10 Q. Cui, M. Elstner, E. Kaxiras, T. Frauenheim and M. Karplus, *J. Phys. Chem. B*, 2001, **105**, 569–585.
11 F. J. Luque, N. Reuter, A. Cartier and M. F. Ruiz-Lopez, *J. Phys. Chem. A*, 2000, **104**, 10923–10931.
12 W. Thiel, *Adv. Chem. Phys.*, 1996, **93**, 703–757.

13 J. P. McNamara and I. H. Hillier, *Phys. Chem. Chem. Phys.*, 2007, **9**, 2362–2370.
14 P. Dobes, M. Otyepka, M. Strnad and P. Hobza, *Chem.–Eur. J.*, 2006, **12**, 4297–4304.
15 H. Valdes, D. Reha and P. Hobza, *J. Phys. Chem. B*, 2006, **110**, 6385–6396.
16 B. Wang and K. M. Merz, *J. Chem. Theor. Comput.*, 2006, **2**, 209–215.
17 T. J. Giese, E. C. Sherer, C. J. Cramer and D. M. York, *J. Chem. Theor. Comput.*, 2005, **1**, 1275–1285.
18 G. Tresadern, H. Wang, P. F. Faulder, N. A. Burton and I. H. Hillier, *Mol. Phys.*, 2003, **101**, 2775–2784.
19 C. Alhambra, M. L. Sanchez, J. Corchado, J. L. Gao and D. G. Truhlar, *Chem. Phys. Lett.*, 2001, **347**, 512–518.
20 M. Kolb and W. Thiel, *J. Comput. Chem.*, 1993, **14**, 775–789.
21 M. Scholten, PhD Thesis, Universität Düsseldorf, 2003.
22 W. Weber, PhD Thesis, Universität Zurich, 1996.
23 W. Weber and W. Thiel, *Theor. Chem. Acc.*, 2000, **103**, 495–506.
24 M. Kolb, PhD Thesis, Universität Wuppertal, 1991.
25 S. Grimme, *J. Comput. Chem.*, 2004, **25**, 1463–1473.
26 R. Ahlrichs, R. Penco and G. Scoles, *Chem. Phys.*, 1977, **19**, 119–130.
27 M. Elstner, P. Hobza, T. Frauenheim, S. Suhai and E. Kaxiras, *J. Chem. Phys.*, 2001, **114**, 5149–5155.
28 M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902–3909.
29 J. J. P. Stewart, *J. Comput. Chem.*, 1989, **10**, 209–220.
30 S. N. Thorn, R. G. Daniels, M. T. M. Auditor and D. Hilvert, *Nature*, 1995, **373**, 228–230.
31 E. W. Debler, S. Ito, F. P. Seebeck, A. Heine, D. Hilvert and I. A. Wilson, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 4984–4989.
32 N. Otte, M. Scholten and W. Thiel, *J. Phys. Chem. A*, 2007, **111**, 5751–5755.
33 Q. Wu and W. T. Yang, *J. Chem. Phys.*, 2002, **116**, 515–524.
34 X. Wu, M. C. Vargas, S. Nayak, V. Lotrich and G. Scoles, *J. Chem. Phys.*, 2001, **115**, 8748–8757.
35 U. Zimmerli, M. Parrinello and P. Koumoutsakos, *J. Chem. Phys.*, 2004, **120**, 2693–2699.
36 S. Grimme, *J. Comput. Chem.*, 2006, **27**, 1787–1799.
37 P. Jurecka, J. Cerny, P. Hobza and D. R. Salahub, *J. Comput. Chem.*, 2007, **28**, 555–569.
38 P. Jurecka, J. Sponer, J. Cerny and P. Hobza, *Phys. Chem. Chem. Phys.*, 2006, **8**, 1985–1993.
39 J. Antony and S. Grimme, *Phys. Chem. Chem. Phys.*, 2006, **8**, 5287–5293.
40 W. Thiel, *MNDO99*, version 6.1; Max-Planck-Institut für Kohlenforschung, Mülheim an der Ruhr, Germany, 2004.
41 J. M. Word, *Reduce*, 2.21; Durham, NC, 2003.
42 J. M. Word, S. C. Lovell, J. S. Richardson and D. C. Richardson, *J. Mol. Biol.*, 1999, **285**, 1735–1747.
43 H. Li, A. D. Robertson and J. H. Jensen, *Proteins: Struct. Func. Bioinf.*, 2005, **61**, 704–721.
44 Overlapping is defined by the distance between the TIP3P oxygen atom and any non-TIP3P heavy atom: R(O–X). If R(O–X) < 2.8 Å the TIP3P water is deleted.
45 C. L. Brooks and M. Karplus, *J. Chem. Phys.*, 1983, **79**, 6312–6325.
46 *CHARMM*, version c31b1; Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, 2004.
47 B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *J. Comput. Chem.*, 1983, **4**, 187–217.
48 A. D. MacKerell, B. R. Brooks, C. L. Brooks, III, L. Nilsson, B. Roux, Y. Won and M. Karplus, in *Encyclopedia of Computational Chemistry*, ed. P. v. R. Schleyer, Wiley, Chichester, edn, 1998, vol. 1, pp. 271–277.
49 A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus, *J. Phys. Chem. B*, 1998, **102**, 3586–3616.
50 *ChemShell*, version 3.0a3; CCLRC Daresbury Laboratory, Cheshire, UK, 2004.
51 P. Sherwood, A. H. de Vries, M. F. Guest, G. Schreckenbach, C. R. A. Catlow, S. A. French, A. A. Sokol, S. T. Bromley, W. Thiel, A. J. Turner, S. Billeter, F. Terstegen, S. Thiel, J. Kendrick, S. C. Rogers, J. Casci, M. Watson, F. King, E. Karlsen, M. Sjøvoll, A. Fahmi, A. Schäfer and C. Lennartz, *J. Mol. Struct. (THEOCHEM)*, 2003, **632**, 1–28.
52 A. H. de Vries, P. Sherwood, S. J. Collins, A. M. Rigby, M. Rigutto and G. J. Kramer, *J. Phys. Chem. B*, 1999, **103**, 6133–6141.
53 H. Lin and D. G. Truhlar, *Theor. Chem. Acc.*, 2007, **117**, 185–199.
54 H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. Dinola and J. R. Haak, *J. Chem. Phys.*, 1984, **81**, 3684–3690.
55 W. G. Hoover, *Phys. Rev. A*, 1985, **31**, 1695–1697.
56 S. Nose, *J. Chem. Phys.*, 1984, **81**, 511–519.
57 S. Nose, *Mol. Phys.*, 1984, **52**, 255–268.
58 J. P. Ryckaert, G. Ciccotti and H. J. C. Berendsen, *J. Comput. Phys.*, 1977, **23**, 327–341.
59 S. R. Billeter, A. J. Turner and W. Thiel, *Phys. Chem. Chem. Phys.*, 2000, **2**, 2177–2186.
60 J. S. Lee and S. Y. Park, *J. Chem. Phys.*, 2000, **112**, 230–237.
61 S. Tsuzuki and H. P. Luthi, *J. Chem. Phys.*, 2001, **114**, 3949–3957.
62 D. Feller, *J. Phys. Chem. A*, 1999, **103**, 7558–7561.
63 M. O. Sinnokrot, E. F. Valeev and C. D. Sherrill, *J. Am. Chem. Soc.*, 2002, **124**, 10887–10893.
64 J. Braun, H. J. Neusser and P. Hobza, *J. Phys. Chem. A*, 2003, **107**, 3918–3924.
65 S. Tsuzuki, K. Honda, T. Uchimaru and M. Mikami, *J. Chem. Phys.*, 2004, **120**, 647–659.
66 P. Benharash, M. J. Gleason and P. M. Felker, *J. Phys. Chem. A*, 1999, **103**, 1442–1446.
67 P. Jurecka and P. Hobza, *J. Am. Chem. Soc.*, 2003, **125**, 15608–15613.
68 For the semiempirical methods without dispersion terms, optimization of dispersion bound complexes results in re-arrangement to form H-bonded complexes where possible. Comparison of the intermolecular parameters in these cases is therefore meaningless.