Calculation of Protein-Ligand Interaction Energies by a Fragmentation Approach Combining High-Level Quantum Chemistry with Classical Many-Body Effects

Söderhjelm, Pär; Aquilante, Francesco; Ryde, Ulf

Published in:
The Journal of Physical Chemistry Part B

DOI:
10.1021/jp810551h

2009

Document Version:
Peer reviewed version (aka post-print)

Link to publication

Citation for published version (APA):
Söderhjelm, P., Aquilante, F., & Ryde, U. (2009). Calculation of Protein-Ligand Interaction Energies by a Fragmentation Approach Combining High-Level Quantum Chemistry with Classical Many-Body Effects. *The Journal of Physical Chemistry Part B*, *113*(32), 11085-11094. https://doi.org/10.1021/jp810551h

Total number of authors:
3

Creative Commons License:
Unspecified

Calculation of protein–ligand interaction energies by a fragmentation approach combining high-level quantum chemistry with classical many-body effects

Pär Söderhjelm*[1], Francesco Aquilante[2], and Ulf Ryde[1]

[1] *Department of Theoretical Chemistry, Lund University,*
Chemical Center, P.O.B. 124, SE-22100 Lund, Sweden

[2] *University of Geneva,*
30 Quai Ernest Ansermet, CH-1211 Geneva 4, Switzerland

* Correspondence to: par.soderhjelm@teokem.lu.se

June 10, 2009

Abstract

We have developed a method to estimate accurate interaction energies between a full protein and a bound ligand. It is based on the recently proposed polarizable multipole interaction with supermolecular pairs (PMISP) method [1], which treats electrostatic interaction by multipoles up to quadrupoles, induction by anisotropic polarizabilities, and non-classical interactions by explicit quantum mechanical (QM) calculations, using a fragmentation approach. For a whole protein, electrostatics and induction is treated the same way, but for the non-classical interactions, a Lennard–Jones term from a standard molecular mechanics (MM) force field (e.g. Amber) is used outside a certain distance from the ligand (4–7 Å). This QM/MM variant of the PMISP method is carefully tested by varying this distance. Several approximations related to the classical interactions are also evaluated. It is found that one can speed up the calculation by using density functional theory to compute multipoles and polarizabilities, but that a proper treatment of polarization is important. As a demonstration of the method, the interaction energies of two ligands bound to avidin are calculated at the MP2/aug-cc-pVTZ level, with an expected relative error of 1–2%.

1 Introduction

Determining potential energies is a common and important task for theoretical methods in chemistry. Depending on the system of interest and the target accuracy, several methods are available, ranging from coarse-grained methods, via all-atom molecular mechanics (MM) force fields, to high-level quantum-mechanical (QM) methods. An example of an application that requires high accuracy is the prediction of the free energy of a ligand binding to a biological receptor (typically a protein). An unbiased prediction of this quantity requires consideration of several intricate effects, such as solvation, conformational changes, and entropic factors. However, no matter how well these effects are treated, the result will not be more reliable than the underlying estimate of the potential-energy surface.

Most studies of protein–ligand interactions with a physical approach employ standard non-polarizable MM force fields [2, 3]. Although these force fields are built on a great amount of experience and are constantly improved, they are obviously missing the polarization part and can thus never include e.g. many-body (non-additive) effects. A more fundamental problem is that by neglecting polarization while still trying to reproduce experimental quantities, one has to depart from the physical picture, and the fitted parameters become just parameters. Clearly, such approaches have a limited transferability and thereby limited accuracy when applied to new systems, although it has been questioned whether this is currently a limiting factor of protein–ligand affinity predictions [4].

Explicit inclusion of polarization has been shown to improve the transferability of force-field parameters [5]. In particular, careful treatment of each physical term can reduce the number of fitted parameters needed or even totally eliminate them [6]. Among the force fields that have been developed along these lines, several have been applied to protein–ligand interactions, including the SIBFA [7, 8, 9, 10], AMOEBA [11, 12], and PFF [13, 14] models. Whereas the parameters for the classical part (e.g. electrostatic and induction energies) of these potentials can normally be derived directly from QM calculations, the non-classical energy contributions (i.e. exchange repulsion, dispersion, and short-range corrections) are more problematic. Although there are expressions based on monomer properties also for these terms [15, 16, 17, 18], it is more common to include some fitting to either experimental data [13], QM interaction energies [19], or decomposed interaction energies from e.g. symmetry-adapted perturbation theory [7, 20].

A different approach to determine the potential energy is to directly exploit quantum-chemical calculations, as these are intrinsically transferable, and also systematically improvable. Protein–ligand systems are in general too large for a full QM treatment. Attempts have been made to use standard hybrid QM/MM methods that treat only the ligand [21, 22] or preferably also the closest residues [23] by QM and the rest of the protein by MM. However, the size of a QM region containing the closest residues is typically 300–800 atoms, whereas most QM methods that treats dispersion interactions reasonably well (at least second-order perturbation theory, MP2, with a sufficiently large basis set) are limited to ∼100 atoms. A more fruitful strategy is to decompose the QM region into smaller subsystems, which are treated more or less independently. Several such *fragmentation methods* have been applied to protein–ligand interactions, including the divide-and-conquer method [24, 25], the fragment molecular orbital (FMO) method [26, 27, 28, 29], the molecular fractionation with conjugate caps (MFCC) method [30, 31, 32] and related approaches [33, 34]. However, all of these studies except Ref. [34] still employ a low level of theory or a small basis set.

Recently, we introduced the *polarizable multipole interaction with supermolecular pairs* (PMISP) method, which is a combination of the polarizable molecular mechanics and fragmentation approaches [1]. It is a general method to estimate the interaction energy between a large molecule and a small molecule at an arbitrary QM level. PMISP differs from other fragmentation methods in that it uses the fragmentation approach only for the non-classical part of the interaction energy, and instead a polarizable multipole approach for the classical part (electrostatics and polarization). Thereby, the method achieves exact (within the QM theory employed) fragment pair potentials, whereas many-body effects are modeled classically. We have previously evaluated the PMISP method by comparing the results with full QM calculations at the same level of theory (MP2/6-31G*) for 250-atom model systems [1]. It was shown to be both faster and more accurate than related fragmentation approaches. Regarded as a force field, PMISP is as far as one can get without explicitly treating the coupling between induction and non-classical terms, i.e. probably as good as a standard polarizable force field can ever become.

In this study, we formulate a QM/MM extension of the PMISP method, approximating the non-classical energy contributions from distant parts of the large molecule by a standard MM force field. This makes the method, which we denote PMISP/MM, suitable for calculating the interaction energy between a full protein and a ligand at a high level of QM theory. We test the approach by computing the interaction energy between the avidin protein (with ∼7800 atoms) and the biotin ligand, as well as a biotin analog, at the MP2/aug-cc-pVTZ level of theory, using a polarizable multipole description of the entire protein. Moreover, we investigate a number of approximations that can be used to speed up the evaluation of the classical part, several of which have general implications for the accuracy of polarizable multipole models. Finally, we evaluate the accuracy of the PMISP/MM approximation by enlarging the PMISP region.

In some sense, the combined use of a polarizable multipole description for long-range interaction and quantum-chemical fragment calculations for short-range interaction is similar to the recent work by Bettens and Lee [34], but in their work no many-body interactions are included, i.e. the polarizable multipole description is only used to approximate the quantum-chemical calculations.

2 Methods

2.1 The PMISP and PMISP/MM methods

We consider the interaction between a large molecule A (typically a protein) and a small molecule B in vacuum. These two molecules will be denoted *monomers*. The geometries of the isolated monomers are kept fixed as the dimer is formed, a common approximation in ligand-binding calculations [35]. In the PMISP method [1], the interaction energy is estimated by the following expression:

$$E_{AB}^{PMISP} = E_{AB}^{ele} + E_{AB}^{ind} + E_{AB}^{nc} \tag{1}$$

where E^{ele} and E^{ind} are the electrostatic and induction interaction energies, respectively, when both monomers are treated using multipoles and polarizabilities, and E^{nc} is the non-classical term, containing mainly dispersion and exchange repulsion but also short-range corrections to the classical terms, e.g. charge penetration. It is estimated by

$$E_{AB}^{nc} = \sum_{i=1}^{n} c_i \left(E_{A_iB}^{sup} - E_{A_iB}^{ele} - E_{A_iB}^{ind} \right) \tag{2}$$

where monomer A has been divided into fragments (A_i) as in the *molecular fractionation with conjugate caps* (MFCC) method [30], and c_i is equal to 1 for a normal (capped) fragment and -1 for a *conjugated cap* (concap) fragment, i.e. the capping groups of two adjacent fragments merged together. $E_{A_iB}^{sup}$ is the counterpoise-corrected supermolecular interaction energy of the A_i–B pair.

A similar formula is used to assemble *properties* (multipoles and polarizabilities) for the whole monomer A from the properties of individual fragments, which are obtained by analysis of the electron density from quantum-chemical calculations (see Section 2.3 below). For a full description of the PMISP method, see Ref. [1]. Note that all energies occuring in this paper are interaction energies.

The physical significance of Eq. 2 is that the non-classical term is assumed to be pairwise additive, so that it can be estimated from supermolecular calculations on smaller subsystems. Thus, all many-body effects are modeled by the classical polarization model. In contrast, in the original MFCC method, the whole interaction energy is assumed to be pairwise additive, so that many-body effects are completely neglected. Other fragmentation methods model many-body effects by *electrostatic embedding* [26, 36].

Let us now consider the situation in which molecule A is so large that only a fraction of the fragments A_i are in close contact with B. For example, this applies to the important case of calculating the interaction energy between a protein and a small ligand. Clearly, the direct use of Eq. 2 is in this case inefficient. For a fragment well separated from B, the corresponding term in Eq. 2 is presumably small and can probably be well approximated by the Van der Waals term of a molecular mechanics force field. However, the use of overlapping fragments necessitates some caution in the treatment of the boundary between the close and far regions. Formally, this can be achieved by using a standard hybrid QM/MM approach,

$$E_{AB}^{PMISP/MM} = E_{MB}^{PMISP} - E_{MB}^{MM} + E_{AB}^{MM} \tag{3}$$

where M is a model of A containing the closest surrounding of B, and E^{MM} is an arbitrary MM potential.

One could in principle employ a standard point-charge/Van der Waals potential in Eq. 3. However, as will be shown below, the energy difference caused by different treatments of the electrostatics and polarization is long-range in nature, and thus one would have to make the model M very large before the PMISP/MM estimate converges. In this work, we instead choose the MM potential to be identical to the PMISP potential except that the non-classical term is replaced by the Van der Waals term from the Amber 1994 force field [37], i.e.

$$E_{XB}^{MM} = E_{XB}^{ele} + E_{XB}^{ind} + E_{XB}^{vdw} \quad X = A, M \tag{4}$$

By inserting Eqs. 4 and 1 into Eq. 3, we note that the total energy can be written as

$$E_{AB}^{PMISP/MM} = E_{AB}^{ele} + E_{AB}^{ind} + E_{MB}^{nc} + E_{AB}^{vdw} - E_{MB}^{vdw} \tag{5}$$

reestablishing that we are simply approximating the non-classical contributions from atoms outside of M. Although the Van der Waals term has formally no meaning on its own, it does account for the long-range dispersion, and its accuracy can easily be tested by extending the model M. The larger one makes the model M, the smaller effect the choice of Van der Waals parameters will have on the energy; this implies that also $E_{AB}^{PMISP/MM}$ (in this limit) is independent of fitted parameters.

2.2 Systems

For testing the method, we used the avidin protein interacting with the seven ligands (biotin analogues) shown in Fig. 1. The studied geometries were the same as in Ref. [1]. They were obtained from snapshots of a molecular dynamics simulation of the protein–ligand complex in explicit water [38], using the Amber 1994 force field [37].

To obtain statistics, 10 snapshots of BTN1 (biotin) were grouped into the *geometry set*, whereas one snapshot of each of the seven ligands were called the *ligand set*. For computational reasons, the calculations involving the full protein were only performed for one snapshot of BTN1, called the *main structure*, and for one snapshot of BTN7. These ligands were chosen because they have different charge (−1 and 0, respectively) and fairly different size. The residue numbering occuring in this study refers to PDB structure 1AVD [39], which was used as starting geometry for the simulations. The full tetramer of the protein was used, but only the interaction energy with one biotin molecule (binding to the subunit labeled B) was calculated, whereas the other three biotin molecules were considered as part of the protein. The total charge of the protein without any ligands is +18.

Several sizes of the model M in Eq. 3 were used. These were all constructed by using a cutoff distance X, so that at least all protein atoms within X Å of the ligand were included. The cutoff was applied in two different ways giving two series of models. In the *chemically cut* models, denoted C_X, only a minimal set of additional atoms were included, viz. those necessary to complete chemically holistic groups such as aromatic rings or peptide bonds. This was accomplished by an automatic cutter that was fed information on which types of bonds are allowed to be cut in a general protein (mainly C–C single bonds). In the other type, the *full residue* models, denoted F_X, all amino acids within X Å of the ligand were fully included and capped with standard caps

To test the use of a standard Van der Waals potential for the long-range part of the non-classical term, we evaluated the effect of enlarging the PMISP region. To this aim, we defined a series of PMISP regions for each of the two ligands BTN1 and BTN7 and calculated the error of the PMISP/MM approximation for each of the regions, using the largest region (F_7) as a reference. The regions were of two types: including all atoms all atoms within X Å of the ligand and a minimal set of additional atoms that are necessary to complete chemically reasonable groups (denoted C_X).

Most calculations were done using the smallest model, C_4, which is the same structure as was used in Ref. [1]. For the main structure, it consists of 216 atoms and is shown schematically in Figure 2. The model is divided into 25 fragments (i.e. scheme *c* of Ref. [1]), out of which 14 are separate molecules (i.e. disconnected already when the model was constructed), 6 are amino-acid-sized fragments (capped with $-COCH_3$ and $-NHCH_3$ groups at the N and C termini, respectively), and 5 are the concap fragments in between (i.e. $CH_3NHCOCH_3$ molecules). The largest model, F_7, contains 45 amino acids, together constituting the following 7 segments of consecutive amino acids: 12–17, 33–44, 47, 68–79, 97–101, 114–120, and 110–111. The total number of atoms in F_7 is 803.

2.3 QM calculations

The multipoles and polarizabilities were obtained by the LoProp method [40] as implemented in MOLCAS [41, 42] (default settings). Expansion centers were placed in the nuclei and in the covalent bond midpoints, and each center contained a multipole expansion up to quadrupoles and an anisotropic polarizability tensor, if not otherwise stated. For the MP2 method, properties were obtained by using the linear-response charge density, which includes all effects from orbital relaxation and therefore gives the same multipole moments as a finite-field perturbative approach [43]. Calculating this density is similar in effort to a gradient evaluation, and thus takes significantly more time than an MP2 energy evaluation. For these calculations, the MOLPRO program [44] was used to generate the density needed by LoProp.

In the PMISP calculations at the MP2/aug-cc-pVTZ level, the properties were computed using the B3LYP/aug-cc-pVTZ method. This method was chosen as an example of a commonly used density functional method; several density functionals were tested and found to give similar results. Because the charge density obtained from density functional theory includes electron correlation, it was assumed to be more similar to the MP2 response density than the one obtained from Hartree–Fock theory. As will be shown below, the estimated error introduced by this approximation is indeed only ∼3 kJ/mol.

The supermolecular calculations were also performed with MOLCAS. To make the large MP2/aug-cc-pVTZ calculations (with up to 2350 basis functions) possible to perform on standard personal computers, we applied the Cholesky decomposition (CD) approximation to the two-electron integrals [45, 46] in combination with the local exchange (LK) algorithm [47]. Based on previous analysis of the accuracy of the CD approximation [47, 48, 49], a decomposition threshold of 10^{-4} was used in all calculations.

2.4 Distance-dependent approximations

In some calculations, we investigated whether time can be saved by using different properties (multipoles and polarizabilities) for the part of the protein closest to the ligand and in the outer part of the protein. The transition between the inner, “good” properties, and the outer, “poor” properties, was characterized by a cutoff distance R and was done residue-wise so that the good properties were used for all residues having at least one atom closer than R to any ligand atom (in the same way as in the construction of the F_X models).

Because the properties are assembled by an MFCC-like approach [1], the sum of charges in each residue is in general not exactly equal to the formal (integer) charge of the residue. These small differences cancel out exactly so that the sum for the whole protein is equal to the formal charge. However, for the distance-dependent transitions between property sets having different charges (i.e. the ff94, ff02, B3LYP, HF, and 6-31G* approximations), the cancellation is lost. To prevent this trivial error from obscuring other errors, the poor charges were in these cases automatically adjusted so that the total charge of each residue was the same as for the good properties. The adjustment was done by distributing the needed charge shift equally over all atoms in the residue. Although such procedure is impractical in a real application (because the purpose of the approximation is to avoid computing the good properties), it gives smoother transitions; the qualitative conclusions are the same without the shift.

Special care is needed when the property sets have different rules for intramolecular polarization, i.e. for the ff02 approximation to the LoProp model. The rule used for the LoProp properties is to exclude polarization between any two centers that have been in the same quantum-chemical calculation [1], whereas the ff02 model excludes interactions between atoms separated by 1 or 2 chemical bonds and scales the 3-bond interactions [50]. For the transition between these two models, a “generous” exclusion protocol was used to ensure that no unphysical polarization occurs. Thus, polarization between two centers treated by different models was omitted if it would have been omitted in either of the two models.

For the transition between the standard LoProp model (good) and a model that neglects polarizability coupling (poor), we applied the following procedure: First the isolated protein was allowed to “pre-polarize”, so that the induced dipoles caused by the field from other parts of the protein were turned into static dipoles. As discussed in Ref. [1], the pre-polarization itself has no effect on the total energies, it only gives a more natural decomposition into electrostatic and induction energy. Another pre-polarization calculation was performed separately, in which the polarizability coupling was neglected (i.e. the field from the static multipoles was directly used to compute the response). The obtained properties from the two calculations were then combined as described above, and in the final energy calculation, polarizability coupling was only included among the good centers (following the standard exclusion rules) and between the good centers and the ligand.

3 Results

3.1 Interaction energy

The interaction energies between biotin and the full avidin protein, as well as the model M, are shown in Table 1 for the main structure. The corresponding results for the BTN7 ligand are also shown. In all these calculations, the smallest model, C_4, was employed. The calculations were performed at the MP2/aug-cc-pVTZ level of theory, but the results with the 6-31G* basis set are also shown for comparison. Note that, apart from the model complex with the smaller basis set (with $E_{ref} = -412$ kJ/mol for the main structure [1]), there is no reference energy available, because the systems are too large for the supermolecular approach. An experimental estimate of the enthalpy of binding would not be useful, because it contains a significant contribution from the surrounding solvent.

Nevertheless, there are several points to note from Table 1. First, there is a significant basis-set dependence of the interaction energy. The result with the 6-31G* basis set is hardly of any use, being 159 kJ/mol less attractive than that with the large basis set (51 kJ/mol for BTN7). A term-wise comparison shows that for the main structure the basis-set dependence originates mainly from the induction (60%) and non-classical terms (36%; predominantly dispersion), whereas the contribution from the electrostatic energy (4%) is minor.

We also see that the increased attraction for the main structure when going from the model to the whole protein comes mainly from the electrostatic energy. Interestingly, there are also significant long-range contributions from the induction and non-classical terms. Note that the difference in the non-classical term between the model complex and the full protein complex, -35 kJ/mol, is taken from the Van der Waals term of the Amber force field, according to Eq. 5. As expected, the contribution from the electrostatic energy is much smaller in magnitude for the neutral ligand; in fact the Van der Waals contribution dominates when extending the system.

We have also compared our estimate of the interaction energy with estimates from two Amber force fields: the non-polarizable 1994 force field (*ff94*) [37] and the polarizable 2002 force field (*ff02*) [50]. These force fields have indentical Van der Waals parameters so for our purposes they differ only in the electrostatic and induction terms. The results are shown in Table 1. The most striking result is that for the main structure the two Amber force fields differ by 90 kJ/mol for the full complex and by 63 kJ/mol for the small model complex. For the model complex, the energy with *ff02* is in closest agreement with the PMISP/MM result (which can be assumed to be close to the true MP2/aug-cc-pVTZ value), whereas for the full complex the situation is reversed. This suggests that interaction energies estimated by standard force fields are far from reliable. For the neutral ligand, the effect of the force field is smaller, 12–20 kJ/mol, but both force fields give results that are far from the PMISP/MM result, with relative differences up to 25%.

Analyzing the individual energy terms, we see that the electrostatic term is significantly smaller with *ff02* than with *ff94*. This is expected, because the charges in *ff94* are deliberately derived using a method (Hartree–Fock with the 6-31G* basis set) that overestimates the absolute charges, with the purpose of compensating for the lack of polarization in the force field. For the main structure, the electrostatic energy with *ff02* (with charges derived using the B3LYP/cc-pVTZ method) is closer to the PMISP/MM electrostatc energy, as could be expected, but for the BTN7 ligand the situation is reversed. For both ligands, the *ff02* force field greatly underestimates the induction energy compared to the PMISP/MM method. It is also noteworthy that the non-classical term is significantly more attractive in the Amber force fields than in the PMISP/MM method, thereby partly correcting for the difference in induction energy.

3.2 Approximations in the classical treatment

The PMISP/MM approximation dramatically reduces the computational cost of evaluating the non-classical term if A is large. However, the use of an accurate multipole and polarizability treatment of the classical terms for the whole system requires a large set of property calculations, which can actually become a bottle-neck in the PMISP/MM calculation. Therefore, we investigate a number of approximations of the classical energy ($E^{ele} + E^{ind}$). One such approximation, necessary to obtain a computationally effective method, was already emplyed in Table 1, viz. the use of B3LYP theory instead of MP2 theory to compute properties.

Throughout this section, the smallest model, C_4 was used as M. Before addressing the protein–ligand interaction, we consider the model–ligand (M–B) interaction in some detail. In the calculation of the PMISP energy, classical energies occur both for the whole interaction (Eq. 1) and for the fragment interactions (Eq. 2). Therefore, we apply each approximation in two different ways: either only for the whole interaction (denoted *single approximation*) or for both the whole interaction and the fragment interactions (denoted *double approximation*). The single approximation tests the accuracy for the total classical interaction energy, whereas the double approximation tests the accuracy for the many-body contribution to the classical interaction energy (because of the perfect cancellation occuring for the two-body contribution [1]). Thus, the error from the double approximation is the one that occurs in the PMISP energy. For all considered approximations, the errors for the main structure and the mean absolute errors (MAE) for the geometry and ligand sets are shown in Table 2. As the reference level for most tests, we use a multipole expansion up to quadrupoles and anisotropic polarizabilities obtained at the MP2/6-31G* level of theory, as in Ref. [1]. However, we also investigate the reduction of basis set size from aug-cc-pVTZ to 6-31G* at the B3LYP level of theory, as well as the neglect of octupoles.

The first two approximations concern the multipole expansion. As can be seen in Table 2, the truncation of the distributed multipole expansion after quadrupoles gives a significant effect with the single approximation, but these are perfectly canceled with the double approximation. This indicates that most of the effect comes from the electrostatic term, where large effects will always be seen when the interaction involves short distances (the energy is probably not converged after octupoles either). On the other hand, if the expansion is truncated already after dipoles, there is a significant effect also with the double approximation, indicating that quadrupoles has a strong influence on the many-body polarization and thus are important for the PMISP method.

The next three approximations are related to the treatment of polarization. First, ignoring the polarization completely gives huge errors in the total energy, but they are rather well cancelled in the double approximation. This type of cancellation is of course the reason why non-polarizable force fields work at all. Still, the ignorance of the polarization gives errors of ∼20 kJ/mol. Next, we consider to approximate each anisotropic polarizability tensor with a scalar isotropic polarizability, defined as the average of the three diagonal elements. Again, the MAEs with the single approximation are significantly larger than with the double approximation. With the double approximation, the error is ∼5 kJ/mol , i.e. similar to the error of the PMISP method itself [1]. Next, we investigate the effect of not including polarizability–polarizability (α–α) coupling, i.e. determining the induced dipoles directly from the electric field from the multipoles without iteration, as in e.g. Ref. [51]. Interestingly, the errors from this approximation are not so dramatic. For example, the error is similar (for the double approximation) to or significantly lower (for the single approximation) than the error caused by using isotropic polarizabilities. This fact is interesting from a computational point of view: The neglect of iterative effects typically reduces the cost of a classical calculation by a factor 10, whereas the commonly used isotropic approximation does not give any significant reduction of the computational time, but apparently has a larger effect on the final result.

The last three approximations in Table 2 concern the possibility of calculating the properties with a cheaper QM method than the one used in the supermolecular calculations. First, we consider calculating the properties by density functional theory, employing the B3LYP functional, instead of MP2, but keeping the same basis set (in this case 6-31G*). The results show that this is a reasonable approximation, affecting the total classical energy less than e.g. the neglect of octupoles. With the double approximation the MAEs are reduced to 1 kJ/mol. Interestingly, Hartree–Fock (HF) properties give similar errors with the double approximation, although the errors using the single approximation are of course significantly higher, because the multipole moments are overestimated at the HF level. Finally, we instead consider the use of a smaller basis set while keeping the same level of theory. This approximation is more problematic than the change of theory, with significant errors also with the double approximation. The reason for this is probably that the polarizabilities are so different with the two basis sets.

Next, we considered the full protein–ligand complex in the main structure. The errors in the total classical energy when applying the same approximations as above are listed in Table 3 (single approximation). All errors have the same sign as the corresponding errors for the model system, but most of them are larger in magnitude, simply because the system is larger. In addition, the use of the classical part of two standard Amber force fields (ff94 and ff02) is included. The Amber energies are treated as an approximation to the polarizable multipole method at the B3LYP/aug-cc-pVTZ level. This is not formally correct, because individual terms are not necessarily comparable between various force fields, but serves as a starting point for the distance-dependent treatment described below.

The results with the double approximation are also reported in Table 3, although not for the Amber approximations because it would involve an equivocal assignment of atomic charges for the fragments. To comply with the computation of the PMISP/MM energy, the approximation in the fragment interactions is only applied within the model M (i.e. in the fragments for which E^{nc} is computed). This means that the two-body interactions are only partially cancelled, and thus the obtained error contains information about the combined effect of the many-body interaction with the closest surrounding and the total interaction with the rest of the protein. For this reason, the errors are significantly larger than for the model system. In fact, only the MP2 → B3LYP approximation (which was the only approximation used in Section 3.1) gives an error below 4 kJ/mol, which can be considered a reasonable limit of the required accuracy in this type of applications (the negligible error for the aug-cc-pVTZ → 6-31G* approximation is only fortuitous, as will be shown below). Evidently, error contributions from the region outside of the model M are also important. Of course, the errors from the double approximation can be reduced by choosing a larger model M, but this would simultaneously mean that the number of supermolecular calculations increases.

An alternative way to reduce the effect of the approximations is to use the "good" set of properties for the ligand and its closest surrounding (up to a cutoff distance) and the "poor" set only for the outer parts of the protein. This is in contrast to the double approximation, where one relies on cancellation by using the poor properties twice for the closest surrounding. Whereas the double approximation approach is specific to PMISP-like methods, the distance-dependent approach is completely general (as it directly affects the total classical energy) and the results should therefore be of interest for any application of polarizable multipole methods.

The results with the distance-dependent approach is also given in Table 3 using two different cutoff distances (4 and 10 Å). The full distance-dependence (up to 20 Å) is shown for the multipole and polarization approximations (i.e. the first five approximations) in Fig. 3 and for the remaining approximations in Fig. 4. Based on these curves, we also give an estimate of the convergence distance in Table 3, i.e. the smallest cutoff distance for which the absolute error is below 4 kJ/mol (and remains so for larger cutoff distances). Obviously, the convergence distance will depend on the studied system and the particular geometry, but it functions as a rough guide to which effects are long-range.

At first sight, it may seem strange that the results with the distance-dependent approximation at 4 Å differ so much from the results with the double approximation, considering that the cutoff for the construction of the employed model, C_4, was also 4 Å. There are four main reasons for the deviation. First, the model C_4 is somewhat smaller than the "good" region in the distance-dependent approach, because the cutoff was applied in different ways (see Sections 2.2 and 2.4). Second, the error in the many-body contribution from the close surroundings is contained in the double approximation but not in the distance-dependent approximation, and this error is significant for some of the approximations (see Table 2). Third, the ligand always has good properties in the distance-dependent approach but poor properties in the double approximation. Fourth, the fragments used in the computation of properties are in general smaller for C_4 than for A, so the properties may be slightly different and thereby not fully cancel in the double approximation.

Looking at each individual approximation, we see that the truncation of the multipole expansion has much smaller effect with the distance-dependent approach than with the double approximation, probably because of the fragmentation discrepancy in the latter. The octupoles can in fact be removed outside a distance of 2 Å and the quadrupoles outside a distance of 5 Å from the ligand without introducing errors larger than 4 kJ/mol. The errors from the neglect of polarization and from using isotropic polarizabilities are more unpredictable and long-range; no convergence is seen before 20 Å. Although the neglect of polarizability coupling has a significant effect in the region close to the ligand, it shows a smooth distance-dependence curve, suggesting that it is safe to neglect this coupling outside a distance of 15 Å. This can significantly decrease the cost of the classical calculations. Interestingly, more detailed tests showed that most of the single-approximation error (29 kJ/mol) comes from the intramolecular polarization of the protein and only a minor part (5 kJ/mol) from the actual induction interaction with the ligand.

Of particular practical interest is the approximation to use B3LYP properties instead of MP2 properties. In accordance with the small errors with the double approximation, the distance-dependent approximation converges already at 3 Å. Thus, one can use B3LYP properties for the major part of the protein. For really accurate results, a better QM method should be used for the closest part of the protein, but obviously one should not regard the MP2 method as the final answer. The results also show that an electron-correlated method is essential for computing properties; the error from using HF properties is more long-range. Note that this is solely due to the better description of the monomers at the correlated level; it has nothing to do with the lack of long-range dispersion interaction common to both methods. Using the small 6-31G* basis set instead of aug-cc-pVTZ also gives a fairly long-range error, but outside of 3 Å, the error appears rather predictable and its magnitude does not exceed 6 kJ/mol.

The approximation to use properties from standard force fields (ff94 and ff02) is also interesting from a computational point of view, because it avoids the computation of properties for most of the protein. Unfortunately, this approach gives significant errors even for large cutoff distances, especially with the (non-polarizable) ff94 model. The polarizable ff02 model appears to give much better results and can probably be applied for the outer region if errors of up to 10 kJ/mol are acceptable. In fact, the long-range distance-dependence of the ff94 approximation is rather similar to the simple neglect of polarization, and analogously the ff02 curve resembles the curve obtained with isotropic polarizabilities. This indicates that, for the outer part of the protein, the treatment of multipoles is less important than the treatment of polarizabilities.

The corresponding results for the neutral BTN7 ligand are given in Figures 5 and 6. As the use of B3LYP properties was found to be such good approximation already for the charged ligand, the B3LYP and HF approximations were omitted and all tests were done using B3LYP properties to save computational time. Because the magnitude of the interaction energy is smaller for the neutral ligand, the results converge faster with distance than for biotin; in fact, all approximations can be applied outside of 9 Å from the ligand, i.e. for ~90% of the residues, without introducing an error larger than 4 kJ/mol. In general, the trends for the individual approximations are the same as for BTN1. The most notable difference is that, for the neutral ligand, the neglect of polarizability coupling has a more long-range effect (8 Å) than the use of isotropic polarizabilities (3 Å), in contrast to what was seen for BTN1. Although both these approximations converge much faster for BTN7 than BTN1, the result suggests that the importance of polarizability coupling may be system-dependent. From a computational point of view, it is interesting to note that, for the neutral ligand, properties can be calculated with a much smaller basis set (6-31G*) outside of 3 Å from the ligand, i.e. for ~98% of the residues

3.3 Enlargement of the PMISP region

To test the use of a standard Van der Waals potential for the long-range part of the non-classical term, we evaluated the effect of enlarging the PMISP region. To this aim, we defined a series of PMISP regions for each of the two ligands BTN1 and BTN7 and calculated the error of the PMISP/MM approximation for each of the regions, using the largest region (F_7) as a reference. The regions were of two types: fully including all amino acids within X Å of the ligand (denoted F_X) or including all atoms all atoms within X Å of the ligand and a minimal set of additional atoms that are necessary to complete chemically reasonable groups (denoted C_X).

From Eq. 5 it follows trivially that the error in the total protein–ligand energy when using a smaller PMISP region M instead of the reference region F_7 is given by

$$\Delta E_{F_7 \to M} = \left(E^{nc}_{MB} - E^{nc}_{F_7 B}\right) - \left(E^{vdw}_{MB} - E^{vdw}_{F_7 B}\right) \tag{6}$$

An error $\Delta E_{F_7 \to M}$ close to zero would indicate that the model M is sufficient.

The results are given in Table 4. For the smallest region C_4, used in the previous sections, the error is -16 kJ/mol for BTN1 and -5 kJ/mol for BTN7. These errors are not alarmingly high, considering that the accuracy of the PMISP method itself is $\sim$10 and 2 kJ/mol, respectively, for these ligands [1]. On the other hand, these results indicate that there are significant differences between the Amber E^{vdw} term and the PMISP E^{nc} term from the part of the model that extends outside C_4 and that the region should be extended to avoid that the error introduced in the PMISP/MM approximation dominates the calculation.

Table 4 gives some indications how this extension should be done in the most effective way. First, there is a dramatic reduction of the error when using full residues instead of chemical cuts without changing the cutoff distance, i.e. when going from C_4 to F_4. However, the number of atoms in M, and thus the computational cost, which is roughly proportional to the number of atoms because of the fragmentation approach, also increases dramatically. In fact, F_4 is significantly larger than both C_5 and C_6 for both ligands.

Nevertheless, even when taking into account the number of atoms included, the error appears to converge more rapidly towards zero if one employs full residues when extending M. However, there is an error cancellation in effect. Note that, because both the E^{nc} and E^{vdw} terms are by definition pairwise additive, the quantity $\Delta E_{F_7 \to M}$ in Eq. 6 can be exactly split into a sum of fragment-wise contributions involving only those fragments that are treated differently in the two regions. Clearly, when M is a full-residue model F_x, all fragments in M are also part of the reference model F_7. Thus, the only contributions to Eq. 6 come from fragments having no atom within X Å from the ligand. In contrast, when M is a chemically cut model C_X, even the closest-lying fragments, e.g. those making hydrogen bonds to the ligand, may have different composition than in the reference model, typically missing the peptide bonds, which, although located further than X Å from the ligand, may influence the interaction energies significantly. If the non-classical term in PMISP was truly atom-wise additive, this would not be an issue, but it is really a rest term which also contains non-additive energy contributions.

Fortunately, this problem is not relevant in a real application. As soon as the model M is large enough that the error in Eq. 6 is comparable to the error of the PMISP approximation itself, there is no way to determine which model is the most "correct" reference, and consequently there is no point in further enlarging M. In conclusion, both the F_4 and C_6 models are appropriate in our test cases, with C_6 having a computational advantage. In fact, we believe that it is even safer to use C_6, because with F_4 one could, in a different system, neglect to include a residue having a polar or charged group as near as 4.1 Å.

To get more detailed information on the origin of the large error when using C_4 for BTN1, we have calculated the residue-wise contributions to Eq. 6. These results are given in Table S1 in the Supporting information. Note that some contributions come from amino acids not at all included in C_4, whereas other contributions come from amino-acids that are truncated in C_4. A summary of these results, where the fragment-wise contributions have been summed within each segment of consecutive residues, is given in Table 5. As can be seen, all E^{vdw} terms (and almost all E^{nc} terms) are negative, indicating that dispersion dominates over repulsion when going outside of C_4. Moreover, the E^{nc} and E^{vdw} terms are significantly different whenever there are short-range interactions, but the $F_7 - C_4$ differences are much more similar.

The two largest fragment contributions to $\Delta E_{F_7 \to C_4}$ come from Ser-73 (-8 kJ/mol) and Asn-118 (-2 kJ/mol). The hydroxyl group of Ser-73 forms a very strong hydrogen bond (H$\cdots$O distance 1.53 Å) to the carboxylate group of biotin. In C_4, the residue is modeled as a methanol molecule, which appears to be a too severe truncation for such a strongly interacting fragment. It is more difficult to explain the large contribution from Asn-118, which forms a normal hydrogen bond with the neutral end of the biotin molecule. Thus, this magnitude of the error contributions must be considered typical for a truncation close to an interacting group.

From this investigation, we conclude that our minimal model C_4 should have included a larger part of the Ser-73 residue because of its exceptionally strong interaction in the main structure. If such cases are considered, the accuracy of the MM approximation appears to be $\sim$10 kJ/mol, i.e. similar to the accuracy of the PMISP approximation itself. A more automatic procedure yielding similar results would be to use C_6 instead. With the huge model F_7, the MM error is completely negligible, because the full MM contribution ($E^{vdw}_{AB} - E^{vdw}_{F_7B}$) is only -6 kJ/mol. Thus, if correcting also for the PMISP error (-7 kJ/mol) found with a smaller basis set [1], our best estimate of the avidin–biotin interaction energy for this particular geometry is $-1419 + 16 - (-7) = -1396$ kJ/mol at the MP2/aug-cc-pVTZ level. The corresponding interaction energy for the avidin–BTN7 complex is $-202.1 + 5.0 - 1.5 = -198.6$ kJ/mol. We have posted the coordinate files as Supporting information to enable a future evaluation of these predictions.

4 Conclusions

We have developed and tested a computationally efficient method (PMISP/MM) to estimate the quantum-mechanical interaction energy between e.g. a protein and a ligand in vacuum. The method combines the recently developed *polarizable multipole interaction with supermolecular pairs* (PMISP) method [1] with a standard molecular mechanics treatment of non-classical energy contributions from distant parts of the protein. The practical limits of the molecular sizes will depend on the applied quantum-chemical method and the computational resources at hand, but for the MP2/aug-cc-pVTZ level used in this study, the ligand may typically contain up to ∼60 atoms, whereas the protein may contain tens of thousands of atoms.

The method was tested by computing the interaction energies of the avidin–biotin and avidin–BTN7 complexes (both with ∼7800 atoms), where BTN7 is a neutral biotin analog. It was found that the basis set dependence of the interaction energy is enormous (up to ∼160 kJ/mol when going from 6-31G* to aug-cc-pVTZ) and that the results with two different Amber force fields differ by up to ∼90 kJ/mol. Moreover, the qualitative description of the interaction varies significantly among these three methods.

We also tested several approximations for the classical part of the interaction energy. It was found that the effect of most of the considered approximations cancels in a pure PMISP calculation as long as they are applied in both the whole-monomer and fragment-wise calculations. When performing PMISP/MM calculations for the whole protein complex, the situation is different. The treatment of polarization (e.g. using anisotropic polarizabilities) is then important and only the approximation of MP2 properties with B3LYP properties can be considered sufficiently accurate (in relation to other errors) to rely on this cancellation. However, several other approximations can be applied for the outer part of the protein without significantly affecting the results. In particular, octupoles and quadrupoles can be neglected outside a distance of 2 and 5 Å from the ligand, respectively, and polarizability coupling outside of 15 Å. A significantly smaller basis set can be used for the computation of properties outside of 15 Å, and also at smaller distances if a small systematic error is acceptable (or if the basis set reduction is less severe). All these effects were found to be substantially smaller for the neutral ligand, and they also become smaller if solvent effects are included [52].

The error of the non-classical MM approximation was tested by varying the region in which fragmented supermolecular calculations were performed. For a region extending 4 Å from the ligand, the error was found to be 16 kJ/mol for biotin (half of which comes from a single interaction, which could easily be identified) and 5 kJ/mol for BTN7. These errors can be reduced by extending the M region and becomes negligible around 6 Å. The error of the PMISP method relative to a supermolecular calculation for these ligands is ∼10 and ∼2 kJ/mol, respectively [1], if we assume that this error is independent of the basis set. We should also include the error from using B3LYP properties, but this error is only ∼3 kJ/mol for biotin (and probably negligible for BTN7), again assuming that the error is independent of basis set. The cumulative error is therefore 10–19 kJ/mol for the (charged) biotin ligand and 2–5 kJ/mol for the (neutral) BTN7 ligand. This corresponds to ∼1–2% of the total interaction energy. The total CPU time is ∼10 days, which is remarkably fast considering that the corresponding QM calculation for the avidin–biotin complex would have 270, 974 basis functions. Moreover, the method is trivially parallellizable, because all quantum-chemical calculations (i.e. both supermolecular calculations and fragment calculations) are completely independent, and the cost of the classical calculations is negligible. The full calculation can be performed automatically with a coordinate file, a cutoff distance for the construction of M, and a choice of basis set as the only input.

Owing to the large basis-set effect, one should of course ask whether the applied level (MP2/aug-cc-pVTZ) is accurate enough or whether we could still be off by hundreds of kJ/mol from the true QM result. Fortunately, this question can be addressed using much smaller molecular systems, because the correlation contributions are fairly additive [1]. One such study [53] showed that the interaction energy with aug-cc-pVTZ is far from the basis set limit, but can be significantly improved by extrapolation. On the other hand, MP2 often overestimates the dispersion energy compared to higher-level theories [54]. It has therefore been suggested to use a smaller basis set (e.g. cc-pVTZ) to compensate for this effect [55], but the situation is complicated by the fact that for hydrogen-bonded complexes, the augmented basis set gives better results [54]. The cc-pVTZ basis set has also been suggested for calculating polarizabilities [13], but in that case, the reason for omitting diffuse functions is to avoid overpolarization due to neglect of coupling between induction and repulsion. Obviously, there is some consistency in these suggestions which could motivate the use of a less diffuse basis set with PMISP. However, in this study we have chosen to use a large and diffuse basis set to demonstrate that the reduction in basis set quality is at least not required for computational reasons. As the PMISP/MM procedure is independent of the level of theory, one can switch to a coupled-cluster treatment with basis-set extrapolation as soon as the progress of computers and quantum-chemical algorithms allows it.

Finally, we note that the principle behind the PMISP/MM method is not limited to vacuum interaction energies, but has the potential to solve many types of problems involving large molecular systems. We currently work on combining the method with the polarizable continuum model for treating solvation effects, and on using the method to investigate the effect from the protein environment on an enzymatic reaction.

Acknowledgments

We thank LUNARC for generously providing computational resources and AstraZeneca and the Swedish Science Research Council for financial support.

Supporting Information Available

Detailed residue-wise results for the enlargement of the PMISP region are found in Table S1. The coordinates of the main structure and the avidin–BTN7 structure (in PDB format) are found in the files `main.pdb` and `btn7.pdb`, respectively. This information is available free of charge via the Internet at http://pubs.acs.org.

References

[1] Söderhjelm, P.; Ryde, U. *J. Phys. Chem. A* **2009**, *113*, 617.

[2] Huang, N.; Kalyanaraman, C.; Bernacki, K.; Jacobson, M. P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5166.

[3] Gilson, M. K.; Zhou, H.-X. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21.

[4] Warshel, A.; Kato, M.; Pisliakov, A. V. *J. Chem. Theory. Comput.* **2007**, *3*, 2034.

[5] Geerke, D. P.; van Gunsteren, W. F. *J. Phys. Chem. B* **2007**, *111*, 6425.

[6] Gordon, M. S.; Freitag, M. A.; Bandyopadhyay, P.; Jensen, J. H.; Kairys, V.; Stevens, W. J. *J. Phys. Chem. A* **2001**, *105*, 293.

[7] Gresh, N.; Piquemal, J.-P.; Krauss, M. *J. Comput. Chem.* **2005**, *26*, 1113.

[8] Antony, J.; Piquemal, J.-P.; Gresh, N. *J. Comput. Chem.* **2005**, *26*, 1131.

[9] Roux, C.; Gresh, N.; Perera, L. E.; Piquemal, J.-P.; Salmon, L. J. *J. Comput. Chem.* **2007**, *28*, 938.

[10] Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J.-P. *J. Chem. Theory. Comput.* **2007**, *3*, 1960.

[11] Ren, P.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23*, 1497.

[12] Jiao, D.; Golubkov, P. A.; Darden, T. A.; Ren, P. *Proc. Natl. Acad.. Sci.* **2008**, *105*, 6290.

[13] Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621.

[14] Maple, J. R.; Cao, Y.; Damm, W.; Halgren, T. A.; Kaminski, G. A.; Zhang, L. Y.; Friesner, R. A. *J. Chem. Theory. Comput.* **2005**, *1*, 694.

[15] Piquemal, J.-P.; Chevreau, H.; Gresh, N. *J. Chem. Theory. Comput.* **2007**, *3*, 824.

[16] Piquemal, J.-P.; Cisneros, G. A.; Reinhardt, P.; Gresh, N.; Darden, T. A. *J. Chem. Phys.* **2006**, *124*, 104101.

[17] Jensen, J.; Gordon, M. S. *Mol. Phys.* **1996**, *89*, 1313.

[18] Adamovic, I.; Gordon, M. S. *Mol. Phys.* **2005**, *103*, 379.

[19] Brdarski, S.; Karlström, G. *J. Phys. Chem. A* **1998**, *102*, 8182.

[20] Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887.

[21] Cho, A. E.; Guallar, V.; Berne, B. J.; Friesner, R. *J. Comput. Chem.* **2005**, *26*, 915.

[22] Gräter, F.; Schwarzl, S. M.; Dejaegere, A.; Fischer, S.; Smith, J. C. *J. Phys. Chem. B* **2005**, *109*, 10474.

[23] Khandelwal, A.; Lukacova, V.; Comez, D.; Kroll, D. M.; Raha, S.; Balaz, S. *J. Med. Chem.* **2005**, *48*, 5437.

[24] Yang, W. *Phys. Rev. Lett.* **1991**, *66*, 1438.

[25] Raha, K.; Merz, K. M. *J. Am. Chem. Soc.* **2004**, *126*, 1020.

[26] Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Letters* **1999**, *313*, 701.

[27] Fukuzawa, K.; Mochizuki, Y.; Tanaka, S.; Kitaura, K.; Nakano, T. *J. Phys. Chem. B* **2006**, *110*, 16102.

[28] Nakanishi, I.; Fedorov, D. G.; Kitaura, K. *Proteins: Struct. Funct. Bioinformatics* **2007**, *68*, 145.

[29] Fedorov, D. G.; Kitaura, K. *J. Phys. Chem. A* **2007**, *111*, 6904.

[30] Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, *119*, 3599.

[31] Zhang, D. W.; Xiang, Y.; Gao, M.; Zhang, J. Z. H. *J. Chem. Phys.* **2004**, *120*, 1145.

[32] Zhang, D. W.; Xiang, Y.; Zhang, J. Z. H. *J. Phys. Chem. B* **2003**, *107*, 12039.

[33] Bettens, R. P. A.; Lee, A. M. *J. Phys. Chem. A* **2006**, *110*, 8777.

[34] Bettens, R. P. A.; Lee, A. M. *Chem. Phys. Letters* **2007**, *449*, 341.

[35] Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. *Acc. Chem. Res.* **2000**, *33*, 889.

[36] Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory. Comput.* **2007**, *3*, 46.

[37] Cornell, W.; Cieplak, P.; Bayly, C.; Gould, I.; Merz, K. M.; Ferguson, D.; Spellmeyer, D.; Fox, T.; Caldwell, J.; Kollman, P. *J. Am. Chem. Soc.* **1995**, *117*, 5179.

[38] Weis, A.; Katebzadeh, K.; Söderhjelm, P.; Nilsson, I.; Ryde, U. *J. Med. Chem.* **2006**, *49*, 6596.

[39] Pugliese, L.; Coda, A.; Malcovati, M.; Bolognesi, M. *J. Mol. Biol.* **1993**, *231*, 698.

[40] Gagliardi, L.; Lindh, R.; Karlström, G. *J. Chem. Phys.* **2004**, *121*, 4494.

[41] Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Computational Materials Science* **2003**, *28*, 222.

[42] *MOLCAS 7, University of Lund, Sweden* **2007**. see http://www.teokem.lu.se/molcas.

[43] Trucks, G. W.; Salter, E. A.; Sosa, C.; Bartlett, R. J. *Chem. Phys. Letters* **1988**, *147*, 359.

[44] Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Rauhut, G.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. *MOLPRO, version 2006.1, a package of ab initio programs* **2006**. see http://www.molpro.net.

[45] Beebe, N. H. F.; Linderberg, J. *Int. J. Quantum Chem.* **1977**, *7*, 683.

[46] Koch, H.; A. Sánchez de Merás; Pedersen, T. B. *J. Chem. Phys.* **2003**, *118*, 9481.

[47] Aquilante, F.; Pedersen, T. B.; Lindh, R. *J. Chem. Phys.* **2007**, *126*, 194106.

[48] Aquilante, F.; Pedersen, T. B. *Chem. Phys. Letters* **2007**, *449*, 354.

[49] Aquilante, F.; Malmqvist, P.-Å.; Pedersen, T. B.; Ghosh, A.; Roos, B. O. *J. Chem. Theory. Comput.* **2008**, *4*, 694.

[50] Cieplak, P.; Caldwell, J.; Kollman, P. *J. Comput. Chem.* **2001**, *22*, 1048.

[51] Straatsma, T.; McCammon, J. *Molecular Simulation* **1990**, *5*, 181.

[52] Söderhjelm, P.; Ryde, U. *J. Comput. Chem.* **2009**, *30*, 750.

[53] Halkier, A.; Klopper, W.; Helgaker, T.; Jorgensen, P.; Taylor, P. R. *J. Chem. Phys.* **1999**, *111*, 9157.

[54] Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.

[55] Riley, K. E.; Hobza, P. *J. Phys. Chem. A* **2007**, *111*, 8257.

Table 1: Interaction energies (in kJ/mol) of the full protein–ligand complex and the model–ligand complex for two structures: BTN1 (main structure) and BTN7, and using two different basis sets. Corresponding results with two Amber force fields are also shown, with the Van der Waals term listed as E^{nc}.

	PMISP/MM (MP2)				**Amber**			
	6-31G*		**aug-cc-pVTZ**		**ff94**		**ff02**	
	Model	**Full**	**Model**	**Full**	**Model**	**Full**	**Model**	**Full**
BTN1:								
E^{ele}	-356.0	-1119.7	-369.8	-1125.6	-565.3	-1299.9	-403.4	-1096.4
E^{ind}	-145.1	-190.4	-222.5	-285.5	0.0	0.0	-98.7	-113.0
E^{nc}	84.2	49.5	26.3	-8.4	-108.5	-143.2	-108.5	-143.2
E_{tot}	-416.9	-1260.6	-566.0	-1419.4	-673.9	-1443.1	-610.6	-1352.6
BTN7:								
E^{ele}	-107.5	-107.1	-113.3	-114.8	-127.5	-118.3	-101.0	-91.1
E^{ind}	-31.7	-34.4	-49.1	-54.4	0.0	0.0	-6.9	-15.2
E^{nc}	4.4	-9.7	-18.7	-32.9	-45.7	-59.8	-45.7	-59.8
E_{tot}	-134.8	-151.3	-181.1	-202.1	-173.2	-178.1	-153.6	-166.2

Table 2: Effect of various approximations on the classical interaction energy between the model M and the ligand. Errors for the main structure (MS) and mean absolute errors for geometry set (Gset) and ligand set (Lset) are given. All values are in kJ/mol.

	Single approx.			**Double approx.**		
	MS	**Gset**	**Lset**	**MS**	**Gset**	**Lset**
No octupoles[a]	11.8	29.8	18.3	0.0	0.9	0.7
No quadrupoles[b]	98.6	123.7	86.9	21.5	19.3	14.2
No polarization[b]	124.0	122.0	75.9	-18.8	22.3	16.2
Isotropic polarizabilities[b]	29.5	26.3	17.7	6.6	4.5	5.1
No α–α coupling[b]	15.0	13.4	10.9	-2.9	4.6	4.5
B3LYP properties[b]	18.1	16.8	10.1	1.7	1.4	1.2
HF properties[b]	-53.3	51.0	34.9	1.4	1.4	1.3
6-31G* properties[c]	91.2	78.8	57.9	-8.8	9.2	6.4

[a] Relative to the MP2/6-31G* (octupoles, anisotropic polarizabilities) result

[b] Relative to the MP2/6-31G* (quadrupoles, anisotropic polarizabilities) result

[c] Relative to the B3LYP/aug-cc-pVTZ (quadrupoles, anisotropic polarizabilities) result

Table 3: Errors (in kJ/mol) of various approximations on the classical interaction energy between the full protein and the ligand for the main structure, when appplying the single, double, and distance-dependent approximations with a cutoff of 4 and 10 Å, respectively. For the distance-dependent approximation, an estimate of the convergence distance R is also given, i.e. the smallest cutoff for which the absolute error is constantly below 4 kJ/mol.

	Single	**Double**	**Distance-dependent**		
			4 Å	**10 Å**	***R*** **(Å)**
No octupoles[a]	18.5	6.7	0.1	0.3	2
No quadrupoles[b]	96.4	19.3	4.5	1.4	5
No polarization[b]	123.5	-19.2	-23.7	-25.5	20
Isotropic polarizabilities[b]	57.3	34.4	-6.3	-12.9	20
No α–α coupling[b]	34.1	16.3	10.7	8.5	15
B3LYP properties[b]	19.1	2.7	2.5	1.6	3
HF properties[b]	-72.7	-17.9	-4.6	-1.5	9
6-31G* properties[c]	100.9	1.1	-3.2	-5.1	12
ff94[c]	111.1	N/A	6.5	-27.4	20
ff02[c]	201.6	N/A	30.4	-3.9	15

[a] Relative to the MP2/6-31G* (octupoles, anisotropic polarizabilities) result

[b] Relative to the MP2/6-31G* (quadrupoles, anisotropic polarizabilities) result

[c] Relative to the B3LYP/aug-cc-pVTZ (quadrupoles, anisotropic polarizabilities) result

Table 4: Error in kJ/mol for the PMISP/MM approximation as a function of the cutoff distance in Å, using the largest PMISP region (F_7) as a reference. Results are given for two ligands (BTN1 and BTN7) and for two different ways of defining the PMISP region: either using full residues or chemical cuts. The total number of atoms in each region M is also specified.

	Full residues			**Chemical cuts**		
Ligand	**Region**	**#Atoms**	**Error**	**Region**	**#Atoms**	**Error**
BTN1	F_7	803	0.0	C_7	587	-1.8
	F_6	698	-1.0	C_6	421	-2.7
	F_5	585	-2.1	C_5	277	-12.5
	F_4	515	-2.5	C_4	216	-16.0
BTN7	F_7	624	0.0	C_7	373	-0.1
	F_6	563	-0.2	C_6	275	-2.2
	F_5	415	-0.3	C_5	196	-3.4
	F_4	358	-0.4	C_4	153	-5.0

Table 5: Contributions from each segment of consecutive amino acids to the change in total interaction energy when enlarging the PMISP region from C_4 to F_7 for the main structure. The quantities are defined in Eq. 6 and energies are in kJ/mol.

Segment	E^{nc}			E^{vdw}			$\Delta E_{C_4 \rightarrow F_7}$
	F_7B	C_4B	$F_7B - C_4B$	F_7B	C_4B	$F_7B - C_4B$	
12-17	5.7	9.8	-4.1	-5.1	-0.6	-4.6	0.5
33-44	14.6	16.8	-2.2	-41.7	-37.1	-4.6	2.4
47	-0.1	0.0	-0.1	-0.4	0.0	-0.4	0.4
68-79	39.0	39.8	-0.9	-35.9	-26.3	-9.6	8.7
97-101	-32.1	-29.0	-3.0	-30.9	-26.9	-4.0	1.0
114-120	-1.5	0.9	-2.3	-9.6	-5.0	-4.6	2.3
110-111	-13.2	-11.9	-1.3	-14.7	-12.6	-2.1	0.8
Total	12.5	26.3	-13.8	-138.4	-108.5	-29.8	16.0

Figure captions

Figure 1: The seven ligands to avidin used in this study: a) BTN1 (biotin), b) BTN2, c) BTN3, d) BTN4, e) BTN5, f) BTN6, g) BTN7. The first three have a molecular charge of -1, whereas the remaining ligands are neutral.

Figure 2: Two-dimensional cartoon of the smallest avidin model (C_4) interacting with biotin. It should give a rough guidance to the location of each fragment; in reality, the fragments surround the ligand completely. For clarity, all hydrogen atoms are omitted. The most prominent hydrogen bonds are marked as dotted lines. The fragments of the model are labeled from A_1 to A_{15}, with A_5 further divided into A_{5a} to A_{5g} (concap fragments are not shown). The avidin residue from which each fragment is derived is shown in brackets. The biotin molecule is labeled B.

Figure 3: The distance dependence of the first five approximations in Table 3 for the main structure. The difference between the approximated and reference classical energies of the full protein–ligand interaction is shown.

Figure 4: The distance dependence of the last five approximations in Table 3 for the main structure. The difference between the approximated and reference classical energies of the full protein–ligand interaction is shown.

Figure 5: The distance dependence of the first five approximations in Table 3 for the BTN7 ligand. The difference between the approximated and reference classical energies of the full protein–ligand interaction is shown.

Figure 6: The distance dependence of the last five approximations in Table 3 for the BTN7 ligand. The difference between the approximated and reference classical energies of the full protein–ligand interaction is shown.

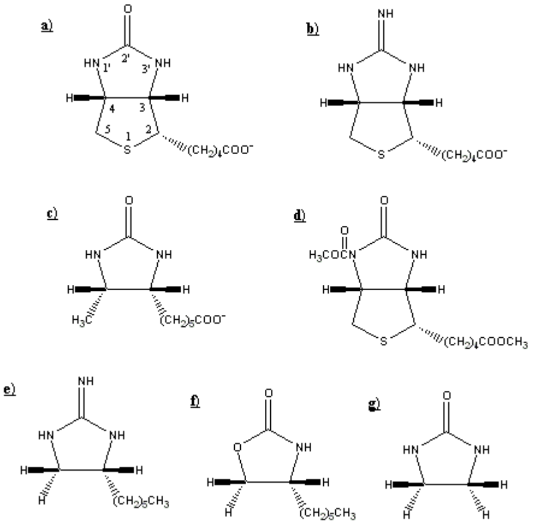

Figure 1: The seven ligands to avidin used in this study: a) BTN1 (biotin), b) BTN2, c) BTN3, d) BTN4, e) BTN5, f) BTN6, g) BTN7. The first three have a molecular charge of -1, whereas the remaining ligands are neutral.

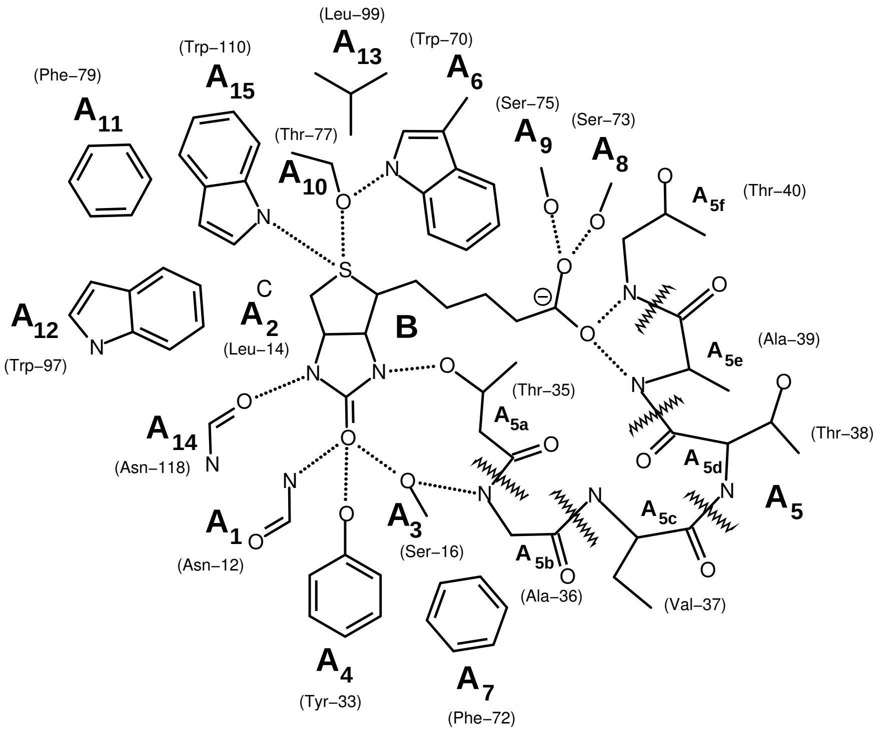

Figure 2: Two-dimensional cartoon of the smallest avidin model (C_4) interacting with biotin. It should give a rough guidance to the location of each fragment; in reality, the fragments surround the ligand completely. For clarity, all hydrogen atoms are omitted. The most prominent hydrogen bonds are marked as dotted lines. The fragments of the model are labeled from A_1 to A_{15}, with A_5 further divided into A_{5a} to A_{5g} (concap fragments are not shown). The avidin residue from which each fragment is derived is shown in brackets. The biotin molecule is labeled B.

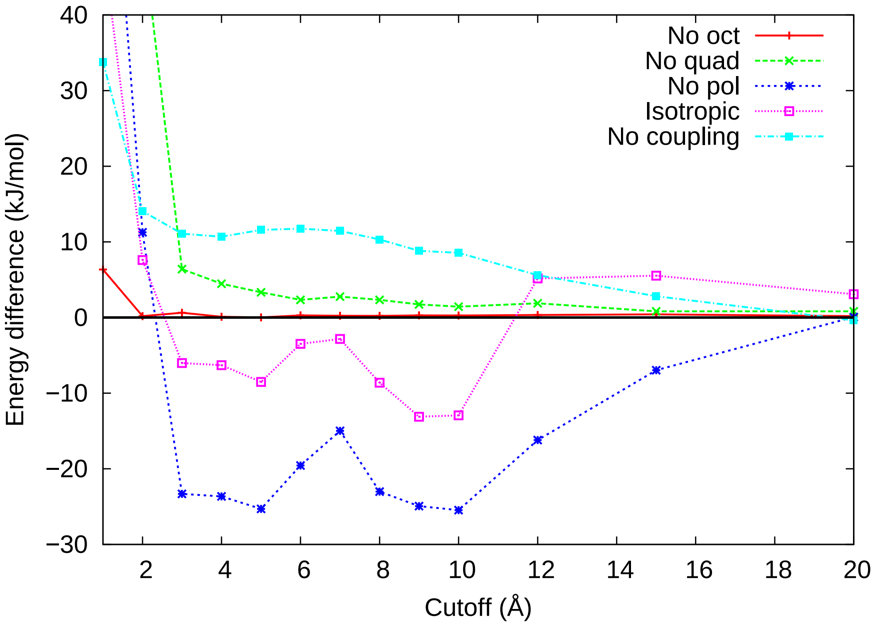

Figure 3: The distance dependence of the first five approximations in Table 3 for the main structure. The difference between the approximated and reference classical energies of the full protein–ligand interaction is shown.

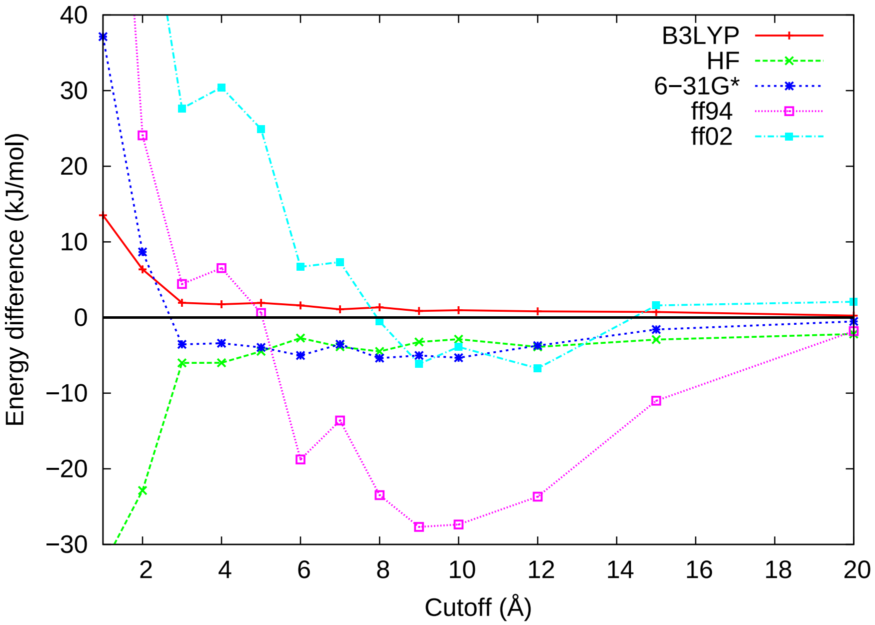

Figure 4: The distance dependence of the last five approximations in Table 3 for the main structure. The difference between the approximated and reference classical energies of the full protein–ligand interaction is shown.

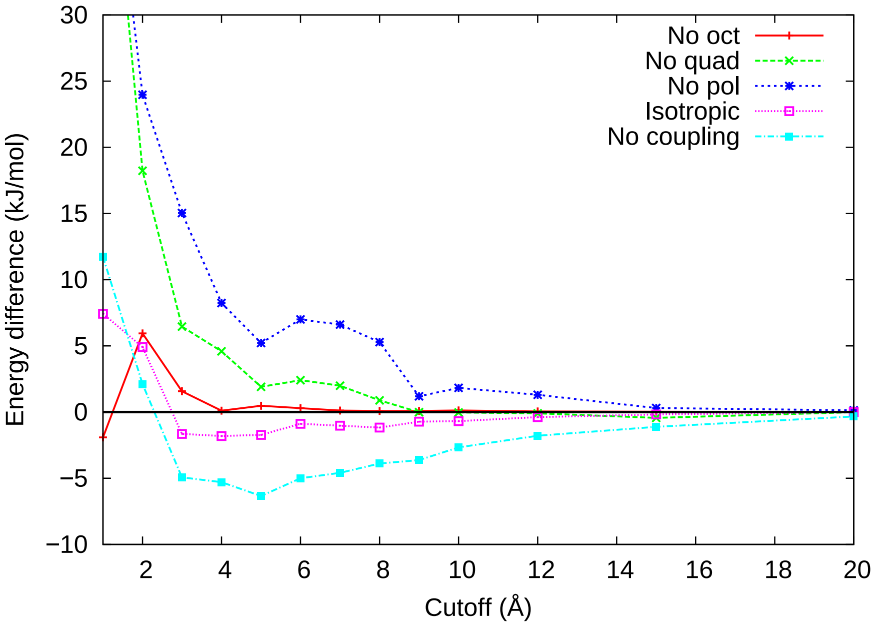

Figure 5: The distance dependence of the first five approximations in Table 3 for the BTN7 ligand. The difference between the approximated and reference classical energies of the full protein–ligand interaction is shown.

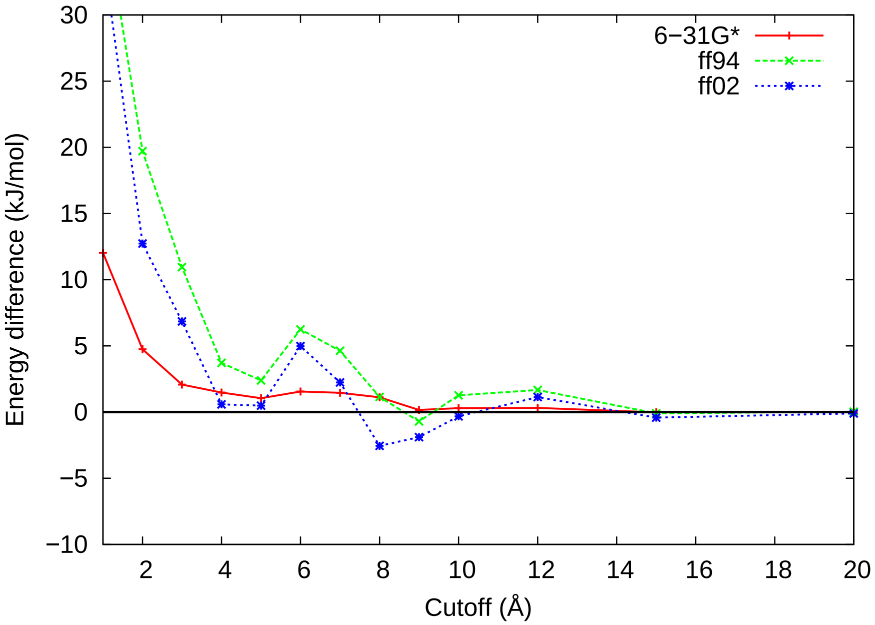

Figure 6: The distance dependence of the last five approximations in Table 3 for the BTN7 ligand. The difference between the approximated and reference classical energies of the full protein–ligand interaction is shown.