Struct Chem (2017) 28:1513–1523
DOI 10.1007/s11224-017-0928-9

ORIGINAL RESEARCH

Accurate prediction of the energetics of weakly bound complexes using the machine learning method kriging

Peter I. Maxwell[1,2] · Paul L. A. Popelier[1,2]

Received: 18 January 2017 / Accepted: 7 February 2017 / Published online: 28 February 2017

Abstract Here, we extend the system energy prediction approach used in the force field FFLUX (Maxwell et al. Theor Chem Acc 135:195, 2016) to complexes bound by weak intermolecular interactions. The investigation features the first application of the approach to bound complex systems, additionally challenged by investigating complexes held together only weakly, through either a predominant dispersion contribution, or through mixed dispersion and hydrogen-bonding. Our approach uses the interacting quantum atoms (IQA) energy partitioning scheme to obtain the intra-atomic, $E_{\text{intra}}^{\text{A}}$, and interatomic, $V_{\text{inter}}^{\text{AA}'}$, energies, which when summed, compose the molecular energy, $E_{\text{IQA}}^{\text{system}}$. The $E_{\text{intra}}^{\text{A}}$ and $V_{\text{inter}}^{\text{AA}'}$ energies are mapped to the positions of the nuclear coordinates through the machine learning method kriging to build atomic energy models. A model's quality is established through its ability to accurately predict the atomic and molecular energies of atoms in an external test set. Mean absolute error percentages (MAE%) of 1.5, 1.5, 1.6, 1.0, 2.6 and 1.7% are obtained in recovering the molecular energy for ammonia…benzene, water…benzene, HCN…benzene, methane…benzene, stacked-benzene (C_{2h}) dimer and T-benzene (C_{2v}) dimer complexes, respectively.

This paper is dedicated to Professor Lou Massa on the occasion of his Festschrift: A Path through Quantum Crystallography.

✉ Paul L. A. Popelier
paul.popelier@manchester.ac.uk

[1] Manchester Institute of Biotechnology (MIB), 131 Princess Street, Manchester M1 7DN, Great Britain

[2] School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, Great Britain

Keywords QTAIM · Quantum chemical topology (QCT) · IQA · Kriging · Machine learning · Force field development · S22 · Complexes · Dispersion

Introduction

Within a protein, one may expect to find several different types of interatomic interactions such as hydrogen bonds, halogen bonds, π-π stacking interactions and ionic bonds. Force fields should be able to cope with these interactions, ideally in a streamlined and conceptually minimal way, rather than by ad hoc modifications or additions to their original standard architecture. The development of the force field FFLUX [1] (formerly called QCTFF [2]) is a sustained and relatively recent effort carried out in this spirit.

At the heart of FFLUX are topological atoms defined by the Quantum Theory of Atoms in Molecules (QTAIM) [3–6]. These atoms emerge naturally [7] (without using parameters) in the electron density of any (quantum chemical) *system*: a single molecule, a cluster of molecules or a piece of solid matter. The topological atoms are space-filling: no overlap and no interatomic gaps. It turns out that topological atoms are also so-called quantum atoms [8], that is, subspaces with a well-defined [9] and unique kinetic energy. This characteristic [10] is important in the design of a force field that stays close to the underlying quantum mechanics. FFLUX is such a force field: it is aware of the *internal* energy of an atom, as well as its various interaction energies, an atom's charge, dipole moment and higher multipole moments. Hence, FFLUX "sees the electrons" unlike the popular classical force fields AMBER or CHARMM. Topological atoms have already been proven to be successful in describing the electrostatic interactions in proteins [11].

FFLUX uses machine learning to predict how a given atom will behave in an atomic environment previously not seen by this atom. More precisely, FFLUX needs to be trained by a sufficient number of relevant geometries such that it can interpolate a property of a given atom of interest between the data learnt. The selected [12] machine learning method is Kriging [13], which has been tested successfully on a variety of systems, including ethanol [14], (peptide-capped) alanine [15], the microhydrated sodium ion [15], *N*-methylacetamide (NMA) and histidine [16], the four aromatic (peptide-capped) amino acids [17], all naturally occurring amino acids [18], helical deca-alanines [19, 20], water clusters [21], cholesterol [22] and carbohydrates [23]. This collective work shows an existing proof-of-concept that kriging models generate sufficiently accurate atomic property models, and they do this *directly from the coordinates of the surrounding atoms.* What all these models have in common is that only *ab initio* wavefunctions are necessary to cover any type of desired interaction. The only requirement is that the input training data consists of system geometries that include examples of the interaction type at hand.

The work presented here follows on from our earlier work [24], where we obtained successful kriging models of atomic multipole moments of seven hydrogen-bonded complexes present in the S22 dataset [25]. The current work concentrates on a different segment of the S22 dataset, now not focusing on hydrogen bonding but on what are sometimes (loosely) called dispersion-dominated complexes. Furthermore, here, we go beyond atomic multipole moments, which cover only long-range electrostatics. The short-range electrostatic interaction can still be treated without using multipole moments. This energy type refers to the situation when the multipole expansion [26, 27] fails to converge. The un-expanded interatomic Coulomb energy can also be successfully kriged as we recently demonstrated [28]. This work also showed that exchange energy and intra-atomic energy could all be kriged with an accuracy of about 1 kJ mol^{-1} or less (for methanol, NMA and peptide-capped glycine). These energy components are defined by the quantum topological method of interacting quantum atoms (IQA) [29].

Here, we obtain the first ever kriging models for the IQA energies of six weakly bound complexes where hydrogen bonding is not the dominant interaction but, instead, dispersion is. The six systems studied all contain benzene: the ammonia…benzene complex, water…benzene, HCN…benzene, methane…benzene, the stacked-benzene (C_{2h}) dimer and the T-benzene (C_{2v}) dimer. For this purpose, we use the density functional M06-2X [30], because it has been shown to mimic the effects of the dispersion interaction. The ammonia…benzene, water…benzene, HCN…benzene and T-benzene (C_{2v}) dimer complexes involve a weak hydrogen bond between the hydrogen atom of the donor non-benzene molecule interacting with the delocalised π-system of the benzene ring. The stacked-benzene (C_{2h}) dimer involves a π-π stacking interaction, and the methane…benzene complex involves a C-H/π bond, common in protein side chains [31].

Methodology

The IQA partitioning

Figure 1 shows the topological atoms as they appear in all six complexes studied. The atoms were generated by the in-house program IRIS, which is based on a finite-element algorithm [32]. QTAIM defines these atoms by allowing a system's electron density to partition itself, using the minimal idea [8] of the gradient path, which is a curve following the direction of steepest ascent. We note again that a system can be single molecule, a cluster of molecules (e.g. a complex consisting of two monomers) or a piece of solid matter. A topological atom consists of all gradient paths terminating at the maximum in the electron density nearest to the nucleus associated with the atom. IQA translates this partitioning idea into the energy domain, augmenting the topological atoms with an atomic energy partitioning scheme. Just like a system can be divided into topological atoms, a system's energy can be divided into a collection of atomic energies. The topological atoms and energy values are allied to one another. Since topological atoms partition a system's space exhaustively, ensuring that every point is attributed to an atom, a system's energy is recovered from the summation over all atomic energies. Note that QTAIM and IQA are both part of an overarching approach called Quantum Chemical Topology (QCT) [33]. The central idea behind QCT is to use the gradient of a quantum mechanical density function to extract chemical information from the wavefunction (or experimental electron density). To date, there are almost a dozen such functions (listed in Box 8.1 of ref. [34]) having been analysed with the QCT context, including ELF [35, 36], for example.

The IQA decomposition of the system energy, used within this work, is now briefly reviewed. The IQA-reconstructed system energy, $E_{\mathrm{IQA}}^{\mathrm{system}}$, is obtained through a summation of atomic energies, $E_{\mathrm{IQA}}^{\mathrm{A}}$, one for each atom A,

$$E_{\mathrm{IQA}}^{\mathrm{system}} = \sum_{\mathrm{A}} E_{\mathrm{IQA}}^{\mathrm{A}} \quad (1)$$

which in turn are a summation of intra-atomic (also known as 'self' energy), $E_{\mathrm{intra}}^{\mathrm{A}}$, and interatomic energies, $V_{\mathrm{inter}}^{\mathrm{AA}'}$:

$$E_{\mathrm{IQA}}^{\mathrm{A}} = E_{\mathrm{intra}}^{\mathrm{A}} + \frac{1}{2} V_{\mathrm{inter}}^{\mathrm{AA}'} \quad (2)$$

where A represents an atom and A' represents the remainder of the system without A present (and hence AA' refers to the interaction between A and A'). Note that VinterAA is halved in order to prevent double counting. This is made possible by attributing only half of the total interaction energy to atom A.

For the purpose of this work, the above decomposition is enough but we point out that both the intra-atomic and interatomic energies can be decomposed further to pursue deeper

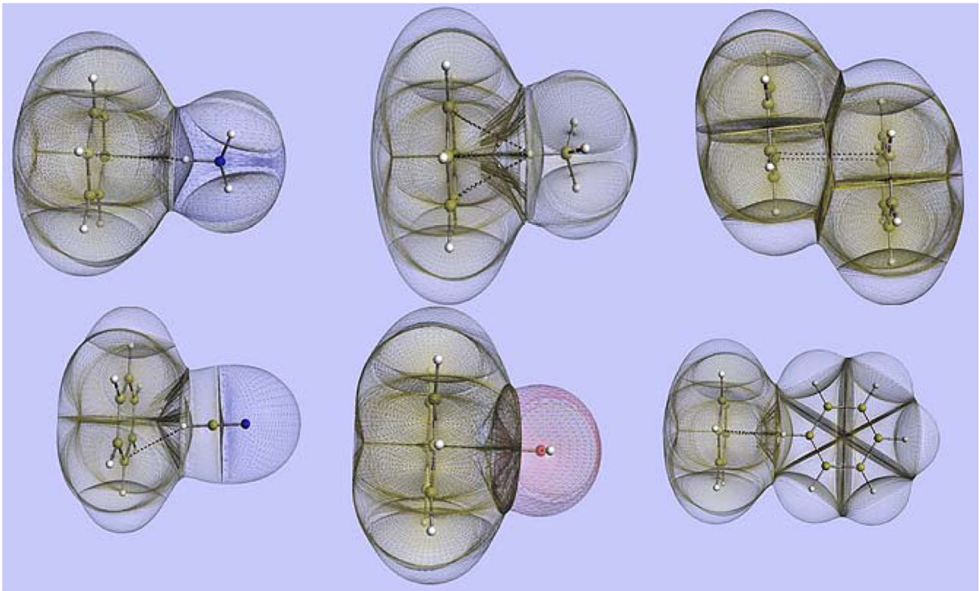

Fig. 1 The six weakly bound complexes studied in this work: ammonia…benzene (*top left*), methane…benzene (*top middle*), stacked-benzene (C_{2h}) dimer (*top right*), HCN…benzene complex (*bottom left*), water…benzene complex (*bottom middle*) and T-benzene (C_{2v}) dimer (*bottom right*). Visualisation [37] of the atomic basins of the topological atoms is made possible by a finite-element algorithm [32]

chemical insight [28]. However, here, we are only interested in testing our building protocol of kriging models to complexes with a more subtle binding nature than the hydrogen-bond dominated complexes studied [24] before. The intra-atomic energy results from the kinetic energy, the electron-electron interaction and the nucleus-electron interaction, confined to electrons within the volume of the topological atom at hand. This energy has recently been shown [38] to be fitted well by an exponential Buckingham-type potential, giving credence to IQA. In summary, in this work, we map two atomic energies (intra-atomic and interatomic) onto the topological atoms, resulting in $2n$ models for a given system, where n represents the number of atoms in the system.

In 2012, Flick et al. [39] analysed the interaction energy contributions in the three S22 subsets (hydrogen-bonded complexes, dispersion-dominated complexes and mixed complexes). In the dispersion and mixed complexes, electrostatics were found not to play the same dominant role they play in hydrogen-bonded complexes. We have chosen to build kriging models with only a single IQA energy representing the interatomic interaction energy for a given atom A, denoted $V_{\mathrm{inter}}^{\mathrm{AA}'}$. This quantity refers to the total interaction energy that atom A experiences as a result of interacting *all* other atoms in the system, A' (except itself). The energy contribution $V_{\mathrm{inter}}^{\mathrm{AA}'}$ incorporates both the Coulombic and non-classical exchange and correlation components. In the previous study [24] on S22 hydrogen-bonded systems, it is the Coulombic component that was expanded using spherical harmonics [40] to give rise to the atomic multipole moments kriged there. The remainder of an atom's energy is collected within the intra-atomic energy, denoted $E_{\mathrm{intra}}^{\mathrm{A}}$. Modelling both energy contributions ($V_{\mathrm{inter}}^{\mathrm{AA}'}$ and $E_{\mathrm{intra}}^{\mathrm{A}}$) for each atom in the system gives us a system model recovering the total energy of the system. Thus, the current treatment of the weakly bound complexes goes beyond the one that was performed before on hydrogen bonded complexes and now offers a complete model of the system's energy. Note that a rigorous, multipolar description of the electrostatic interaction, not used here, is still important for a potential that aims to accurately model the energy profile of larger oligopeptides and proteins, because of long-range electrostatics. However, the six systems investigated here do contain atoms that are far enough from each other that they normally can be represented by multipole moments.

A final note on the IQA partitioning is on its recent inclusion of some density functionals, such as B3LYP and M06-2X. Previously, IQA could only be used in conjunction with computational *ansätze* that generate a well-defined second-order reduced density matrix. A recent publication explains the problem in greater detail [41] and presents a practical solution. An alternative, slightly more recent solution is that [42] of Francisco et al., which is not (yet) implemented in the software (see The GAIA Protocol) we used to generate the IQA contributions. Note that, very recently, IQA can also be used with MP2, MP3 and MP4 wavefunctions, involving the explicit four-dimensional two-particle density matrix, and thereby theoretically recovering the original total energy [43]. The important point is that the system's energy can be recovered (to a practical degree of accuracy) from the atomic IQA energy components with the M06-2X functional used in this study.

Sampling of the molecular complexes

Behind each sampling of system geometries is a generator of geometries. Typically, normal modes are used to distort the geometry of a stationary point on the potential energy surface (i.e. an "equilibrium geometry"). Normal modes ensure a physically (and chemically) informed way of distorting a system's nuclear skeleton. However, in this work, we wanted to enhance the geometric diversity, beyond that of mere distortions around the local energy minima. It is important that a kriging training set also samples geometries of complexes in which the monomers are translated (and rotated) with respect to each other. Figure 2 gives an impression of this enhanced sampling for all six systems (i.e. complexes). In more detail, we used complexes from the extended S22x5 dataset [44] which includes the equilibrium S22 complex geometries as input for normal modes sampling. The resulting dataset includes the S22 systems at four non-equilibrium geometries, where the monomers have been translated along the axis in the direction of the main intermolecular interaction. As a consequence of the non-equilibrium nature of the extra geometries in the S22x5 set, standard normal modes sampling [24], not revised here, was not possible. The first derivative term of the Taylor expansion (used to calculate the vibrational modes) is no longer zero and, thus, must be included in the calculation of the normal modes. Instead, our non-equilibrium normal modes sampling algorithm described in Part B of the Supplementary Material of ref. [45] and implemented in the in-house program EROS [45] were used for the vibrational sampling of the complexes.

We now describe in more detail how the training set was constructed. For each molecular complex, we obtained the five S22x5 geometries (one being the equilibrium S22 complex). Subsequently, each of these five geometries had one molecule in each complex rotated by 90°, 180° and 270°, in turn, in order to give a total of 20 [=(1 + 3) × 5] molecular geometries. The latter are henceforth called *seed geometries*. For HCN…benzene, ammonia…benzene, methane…benzene and T-benzene, the two monomers are almost orientated perpendicular to one another. In these systems, the intermolecular interaction axis is defined as the axis formed by the centre of the benzene monomer and the nearest atom of the second monomer. When a rotation is applied along such intermolecular interaction axis, little monomer displacement occurs (see Fig. 2). However, in the cases of water…benzene and stacked-benzene, the monomers are not perpendicular with respect to each other. Indeed, one monomer is directed towards the second monomer at an acute angle and offset from the centre of the benzene monomer. Here, the intermolecular interaction axis is defined as the axis formed by the two nearest functional groups between the monomers (H-C…H-O in water…benzene and H-C…H-C in stacked-benzene,

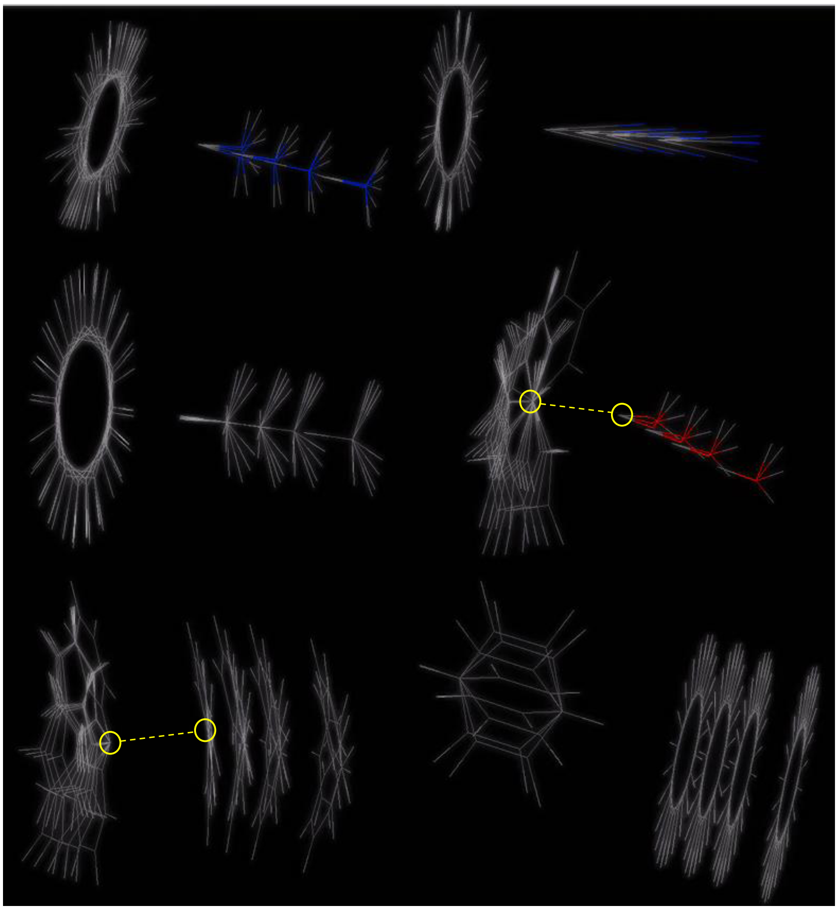

Fig. 2 Wireframe images of 16 sample geometries of the ammonia…benzene complex (*top left*), HCN…benzene (*top right*), methane…benzene (*middle left*), water…benzene (*middle right*), stacked-benzene (C_{2h}) dimer complex (*bottom left*) and T-benzene (C_{2v}) dimer complex (*bottom right*). The intermolecular interaction line (upon which rotation occurs) lies between the centre of the benzene ring, and the nearest atom of the second monomer, except for those where the monomers form an acute angle as a complex, where instead the nearest atoms are used to define the intermolecular interaction line (appended in *yellow*). In the latter systems, the off-centre pivot causes a displacement-like effect in the figure (colour figure online)

denoted in Fig. 2 in yellow). Hence, when a rotation is applied to these systems, the off-centre pivot causes a displacement as illustrated in Fig. 2. All 20 seed geometries were then input as minima to the non-equilibrium normal modes sampling routine within the program EROS. The use of seed geometries from the S22x5 data set provides an additional four non-equilibrium geometries to the geometries found in the S22 set and achieves a greater and more challenging sampling of conformational space. A more challenging sampling gives rise to potentially more useful kriging models as they are able to predict energies for systems with greater flexibility. With the above details in mind, Fig. 2 can now be more thoroughly inspected, showing images of 16 sample geometries for each of the six weakly bound complexes. Note that the 16 samples depicted belong to samples generated around the four S22x5 non-equilibrium seeds. The equilibrium S22 seed was sampled to produce twice as many samples compared to each non-equilibrium seed to ensure a broad sampling in this important region of conformational space.

For each molecular seed, EROS inserts energy into the normal modes in a pseudo-random distribution enabling vibrational distortions of the molecule to be generated. Snapshots can be taken from aforementioned distortions and used as samples in the training set. To ensure that only realistic molecular samples are generated, a bond-stretch and angular-stretch parameter of 1.10 is defined by the user as a threshold. The threshold parameter ensures that the bond and angular stretches are limited to ±10% of the respective values in the seed geometry. Approximately 10% was selected as a chemically reasonable threshold, producing distorted geometries with equivalent bond and angle stretches similar to those obtained through a molecular dynamics simulation at room temperature.

The GAIA protocol

The GAIA protocol is the sequence of computational steps used in FFLUX to build atomic models from scratch. We recently reported [28] the IQA-compatible version of GAIA that is subsequently used in this investigation, which is why only a brief description will be presented here.

The GAIA protocol has five key steps: (1) sampling, (2) *ab initio* calculations, (3) atomic property calculations, (4) kriging model building and (5) validation. Each step is performed in sequence, with the output of the previous step forming the input for the next step. The first four steps involve data being generated, using either in-house software or commercially available software. The final step is a quality check or validation step completed through an analysis of the outputs both by the user and the computer, evaluating the generated models. In short:

1. Sampling – **EROS** (in-house): EROS distorts input seed geometries using the molecular normal modes, creating sample geometries, which collectively describe the molecular conformational space around the seed geometries.
2. *Ab initio* calculations – **GAUSSIAN09** (commercial): GAUSSIAN09 [46] performs single-point energy calculations for each sample, outputting the wavefunctions of all systems.
3. Atomic property calculations – **AIMAll** (commercial): AIMAll (version 14.11.23) [47] uses the system's wavefunction and calculates the intra-atomic and inter-atomic IQA energies (amongst others) for each wavefunction.
4. Model building – **FEREBUS** [48, 49] (in-house): The atomic property data is compiled and 'scrubbed'. Scrubbing removes and discards any sample geometry that has an atomic energy with an integration error [50] (L(Ω)) greater than a given user-defined threshold, which is in our case 0.001 Hartrees. Next, from the remaining samples, a pre-determined amount is set aside as the *test set*, and the remainder, to the nearest hundred, become the *training set*. FEREBUS builds kriging models using the *training set* by mapping the geometrical features to the atomic energies.
5. Validation – kriging models built by FEREBUS are tested, using the *test set* by predicting atomic energies for each test sample, and then comparing them with the known correct values.

Together, the steps outlined above describe the parameterization procedure within FFLUX. In previous literature (e.g. see Appendix of ref. [51]), a different variation of GAIA described the analogous procedure used to build models for atomic multipole moments in place of the atomic IQA energies. Future work will describe a final version which caters for the building and merging of both atomic properties (IQA and multipole electrostatics).

Computational details

The M06-2X functional, used in this work, was developed with the aim of improving the description of intermolecular energies and has been adopted due to its success [52–55]. As a consequence of the widespread use of M06-2X, our group worked with Dr. Keith to have this functional implemented and tested in his program AIMAll. Using the same methodology thoroughly reported in our other research [41], the IQA decomposition can be performed on M06-2X wavefunctions. The other commonly available IQA theory levels (HF and B3LYP) would give poor interaction energies of weakly bound systems without the use of (ad hoc) dispersion corrections [56].

Molecular models were obtained by following the GAIA protocol for each of the six complexes. Five seed geometries for each complex were obtained from the S22x5 datasets optimised [44] at MP2/cc-pVTZ level of theory by Jurecka et al. [25]. One of these seeds is the S22 equilibrium geometry, the remaining non-equilibrium seeds sample the intermolecular distance at translated relative distances of 0.9, 1.2 1.5 and 2.0 to the equilibrium value. The S22 and S22x5 datasets are common benchmarking datasets for non-covalently bonded complexes. Thus, rather than manipulate the geometries by re-optimising at M06-2X level, which would introduce an unnecessary uncertainty into the geometries, the MP2-optimised S22x5 geometries were used as reported by Jurecka et al. [25]. Furthermore, it should be noted that an MP2-IQA approach has recently become computationally possible [43] but is feasible at the moment only for much smaller molecules. For each of the five seeds, one molecule was subjected to rotation by 90°, 180° and 270°, resulting in 20 [=5×(3 + 1)] final seed geometries to be distorted. For each system, 1992 sample geometries were generated from each set of 20 seeds (83 samples per non-equilibrium seed and 166 (=2 × 83) for the equilibrium seed, so 1992 = (16 × 83) + (4 × 166)) using EROS with bond and angle stretch factors of ±10%. All *ab initio* calculations were performed using the GAUSSIAN09 software package at the M06-2X/aug-cc-pVDZ level of theory. The M06-2X [30] functional was chosen for its specific design to correctly provide accurate interaction energies for a range of intermolecular interaction types, in particular van der Waals dimers and the S22 complex set [39]. The aug-cc-pVDZ basis set was chosen for its compromise between speed and accuracy. In keeping with AIMAll's user documentation, each wavefunction file was appended with the 'M062X' keyword to act as a flag to AIMAll, which in turn ensures that the explicit M06-2X IQA algorithm is followed. The IQA calculations were performed by AIMAll (version 14.11.23), using default parameters but with the added request of the IQA energies to be calculated '-*encomp = 3*' (short for *energy components*, and where the value (0 to 4) corresponds to the computation of a given list of IQA energies). The calculated $\Delta E_{\mathrm{IQA}}^{\mathrm{system}}$ energies, across all systems, on average recovered the *ab initio* molecular energies to within approximately 1 kJ mol^{-1}. The kriging models were built with the FEREBUS kriging engine using the following variables: *p* was optimised, *convergence* was set to 200, theta (Θ) was set to a maximum value of 0.1 and the *tolerance* to 10^{-9}. Variable training set sizes between 800 and 1400 examples were used for the six molecular complexes, conditional on the number of samples passing the molecular scrubbing (set to 0.001 Hartrees). The test set consisted of 500 samples, with exception of the two benzene dimers, which used 400 each. The predictions made by FEREBUS were used to construct the so-called *S-curves* (explained in S-curves formulation) for the system's energy predictions, $\Delta E_{\mathrm{IQA}}^{\mathrm{system}}$, and for the intra-atomic energy, $\Delta E_{\mathrm{intra}}^{\mathrm{A}}$, and the interatomic energy predictions, $\Delta V_{\mathrm{inter}}^{\mathrm{AA}'}$.

Results

S-curves formulation

The $E_{\mathrm{intra}}^{\mathrm{A}}$ and $V_{\mathrm{inter}}^{\mathrm{AA}'}$ energies were predicted for 500 test geometries for ammonia…benzene, water…benzene, methane…benzene and HCN…benzene, and 400 test geometries for the T-benzene and stacked-benzene complexes. A smaller test set of 400 samples was required for the benzene dimer complexes due to a greater number of geometries being filtered out with high integration errors in the scrubbing step. The performance of the kriging models, obtained from FEREBUS for the six complexes studied, is displayed using S-curves. Each point in the S-curve is equal to the error for a specific test point, that is, a sample geometry in the test set. The *y*-axis returns the number of test samples represented as a percentile, for example, 500 test points divided by 100%, equates to 0.2% per test point. The *x*-axis plots the absolute energy error between original and predicted values. More precisely, the absolute error for a given system geometry, $\Delta E_{\mathrm{IQA}}^{\mathrm{system}}$, is obtained through a summation of the errors obtained across both atomic $E_{\mathrm{intra}}^{\mathrm{A}}$ and $V_{\mathrm{inter}}^{\mathrm{AA}'}$ energies, and across all atoms, or

$$\Delta E_{\mathrm{IQA}}^{\mathrm{system}} = \left| \sum_{A}^{N_{\mathrm{atoms}}} \left[\left(E_{\mathrm{intra,Act}}^{A} - E_{\mathrm{intra,Pred}}^{A} \right) + \left(\frac{1}{2} V_{\mathrm{inter,Act}}^{AA'} - \frac{1}{2} V_{\mathrm{inter,Pred}}^{AA'} \right) \right] \right| \quad (3)$$

where 'Act' stands for the actual (i.e. original) value and 'Pred' the predicted value.

The mean absolute error (MAE) can be calculated in order to obtain a single error value for a system's model. The MAE is calculated by summing all the $\Delta E_{\mathrm{IQA}}^{\mathrm{system}}$ values and dividing by the number of test set samples:

$$\Delta E_{MAE}^{\mathrm{system}} = \frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} \Delta E_{\mathrm{IQA,i}}^{\mathrm{system}} \quad (4)$$

where N_{test} is the number of samples in the test set, with *i* representing a single test sample.

A final measure, the MAE percentage (MAE %), can also be calculated by dividing $\Delta E_{MAE}^{\mathrm{system}}$ by the size of the energy range sampled by the test set:

$$MAE\% = \frac{\Delta E_{MAE}^{\mathrm{Molec}}}{E_{\mathrm{max}}^{\mathrm{TestSet}} - E_{\mathrm{min}}^{\mathrm{TestSet}}} \quad (5)$$

where 'max' refers to the highest system energy in the test set and 'min' to the lowest. Percentage errors are more *transferable* than MAEs since they free the error from the associated sampled energy range, which is known to influence the error obtained for the model. Thus, the MAE%'s from different molecules are comparable as a transferable performance measure.

The fortuitous cancellation of errors has been described in full in previous work [28], which is why we described it again only briefly here. Using two or more IQA energies to model the system energy results in two or more predicted energies being summed. If a predicted energy is predicted to be less stable than the actual energy, it is called *underestimated*. Accordingly, an energy that is predicted to be more stable is *overestimated*. When an overestimated energy is summed with an underestimated energy, the resulting system energy recovered is more accurate due to a cancellation. In opposition, if two over- or two under-estimated energies are summed, the resulting energy is less accurate through an accumulation of errors. Control of the over- and under-estimation of energies is not possible, but previous research [28] has proven that they often fortuitously cancel.

A final note concerning the formation of S-curves is on the removal of predictions that fall outside the domain of applicability. The domain of applicability is defined as the region of conformational space that can be interpolated by the training points of the kriging model, i.e. the conformational space defined by the training set points. Points that fall outside the training set, and thus outside the domain of applicability, require an extrapolation from the model to make a prediction. Where a point lies far from the domain of applicability, noticeably larger prediction errors are observed. The identification of points outside the domain of applicability can be made by the analysis of the mean signed error (MSE) (or mean signed deviation, MSD). A high MSE or MSD indicates to a user that a particular prediction point is not well trained for in the model and thus is a hallmark of working outside the domain of applicability. Some clear outliers have been removed from the $\Delta E_{\text{intra}}^{\text{A}}$ and $\Delta V_{\text{inter}}^{\text{AA}'}$ S-curves presented in this investigation. However, no outliers are removed from the system energy S-curves, which naturally eliminate those seen in $\Delta E_{\text{intra}}^{\text{A}}$ and $\Delta V_{\text{inter}}^{\text{AA}'}$ through cancellation of errors.

S-curves

Figure 3 shows the system prediction errors for all six systems as S-curves. The ammonia…benzene (blue) and water…benzene (red) complex kriging models perform very similarly and both outperform the models obtained for the remaining four benzene complexes. Of the test points, 90% are accurately predicted within 2.2, 2.3, 4.5, 5.5, 7.7 and 9.8 kJ mol^{-1} for the ammonia…benzene, water…benzene, methane…benzene, stacked-benzene, T-benzene and HCN…benzene complexes, respectively.

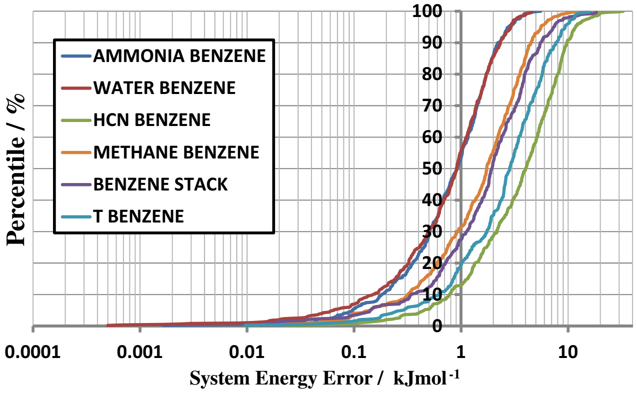

Fig. 3 S-curves displaying the absolute error for a given system geometry ($\Delta E_{\text{IQA}}^{\text{system}}$) defined in Eq. (3) for the six weakly bound complexes: ammonia…benzene (*blue*), water…benzene (*red*), HCN…benzene (*green*), methane…benzene complex (*orange*), stacked-benzene dimer (*purple*) and T-shaped benzene dimer (*turquoise*) (colour figure online)

Table 1 contains the range in the total energy for each weakly bound complex as well as the mean absolute error (MAE) for the predicted molecular energy. Included is the MAE% error, i.e. the MAE as a percentage of the range of said energy. The system energy is predicted within 2.6% for all systems. The values in Table 1 show that as the range in total energy increases, the MAE also increases, but the increase in MAE is slower than that of the range, and therefore the MAE is a smaller percentage of the range. This shows that the FFLUX protocol is capable of handling large ranges in system energies with only a small cost to the accuracy of the kriging predictions.

The kriging performance of the separate $E_{\text{intra}}^{\text{A}}$ and $V_{\text{inter}}^{\text{AA}'}$ energetic terms has also been analysed, where the two terms on the right hand side of Eq. (3) are each plotted as separate S-curves. Thus, each point on the $\Delta E_{\text{intra}}^{\text{A}}$ curve is given by:

$$\Delta E_{\text{intra}}^{\text{A}} = \left| \sum_{A}^{N_{\text{atoms}}} \left(E_{\text{intra,Act}}^{A} - E_{\text{intra,Pred}}^{A} \right) \right| \qquad (6)$$

and each point on the $\Delta V_{\text{inter}}^{\text{AA}'}$ curve given by:

$$\Delta V_{\text{inter}}^{\text{AA}'} = \left| \sum_{A}^{N_{\text{atoms}}} \left(\frac{1}{2} V_{\text{inter,Act}}^{AA'} - \frac{1}{2} V_{\text{inter,Pred}}^{AA'} \right) \right| \qquad (7)$$

The two sets of S-curves are seen in Fig. 4. Both sets of S-curves perform similarly to the total energy S-curve; only the stacked-benzene complex shows a noticeable shift to slightly poorer predictions. However, since this shift to the right (i.e. worse performance) is seen for both the $\Delta E_{\text{intra}}^{\text{A}}$ and $\Delta V_{\text{inter}}^{\text{AA}'}$ energetic terms, we again benefit from a cancellation of errors,

Table 1 Summary of the kriging performance of the weakly bound complexes

System	Ammonia…benzene	Water-benzene	HCN…benzene	Methane…benzene	Stacked-benzene	T-benzene
$E_{\mathrm{IQA}}^{\mathrm{system}}$						
Energy Range	74.58	71.03	301.18	226.45	103.13	210.23
St. Deviation	0.88	0.89	4.17	1.89	2.56	2.89
MAE	1.11	1.10	4.94	2.20	2.64	3.58
MAE % Error	1.49	1.55	1.64	0.97	2.56	1.70
$E_{\mathrm{intra}}^{\mathrm{A}}$						
Energy Range	203.56	233.45	2131.71	316.34	260.11	435.43
St. Deviation	1.12	1.11	4.77	2.40	4.94	2.79
MAE	1.36	1.39	5.61	2.73	5.29	3.69
MAE % Error	0.67	0.59	0.26	0.86	2.03	0.85
$V_{\mathrm{inter}}^{\mathrm{AA}'}$						
Energy Range	218.38	248.96	2259.48	333.17	261.78	384.57
St. Deviation	0.92	0.98	3.67	1.60	5.22	2.37
MAE	1.11	1.23	4.71	1.92	5.04	3.06
MAE % Error	0.51	0.49	0.21	0.58	1.93	0.80

St. Deviation is the standard deviation and MAE is the mean absolute error. All energies are given in kJ mol^{-1}

as in previous work [28], resulting in the overall better prediction of the system energy.

The S-curve MAE values are found in Table 1 alongside the test set energy range sampled for the intra-atomic and interatomic energies. The test set energy ranges for the two separate IQA energy terms are much larger than the test set energy range for the IQA system energy. For example, the ranges in the $E_{\mathrm{intra}}^{\mathrm{A}}$ and $V_{\mathrm{inter}}^{\mathrm{AA}'}$ energies for ammonia…benzene are 203.6 and 218.4 kJ mol^{-1}, respectively, whereas the range in the system energy is only 74.6 kJ mol^{-1}. The lower system energy ranges are a result of cancellation between the energetic components. When two molecules are close to one another, the intra-atomic energy is more positive than when they are at greater separation. A more positive intra-atomic energy is observed because the atoms are deformed [38] when brought close together, resulting in them being less stable. Bringing atoms together to be in closer proximity always gives rise to a positive change in the intra-atomic energy, $E_{\mathrm{intra}}^{\mathrm{A}}$. Conversely, the interatomic energy, $V_{\mathrm{inter}}^{\mathrm{AA}'}$, is more negative to the closer two molecules are because the interatomic, and therefore intermolecular, bonding is stronger. The relationship between IQA's intra-atomic and interatomic energies has been a topic of discussion in previous publications by our group [28, 57, 58]. Table 1 shows that despite the large range in total $E_{\mathrm{intra}}^{\mathrm{A}}$ and $V_{\mathrm{inter}}^{\mathrm{AA}'}$ values, the respective MAEs are relatively similar to the MAE values of the IQA system energy for all complexes, except stacked-benzene. Thus, the MAE% values are often

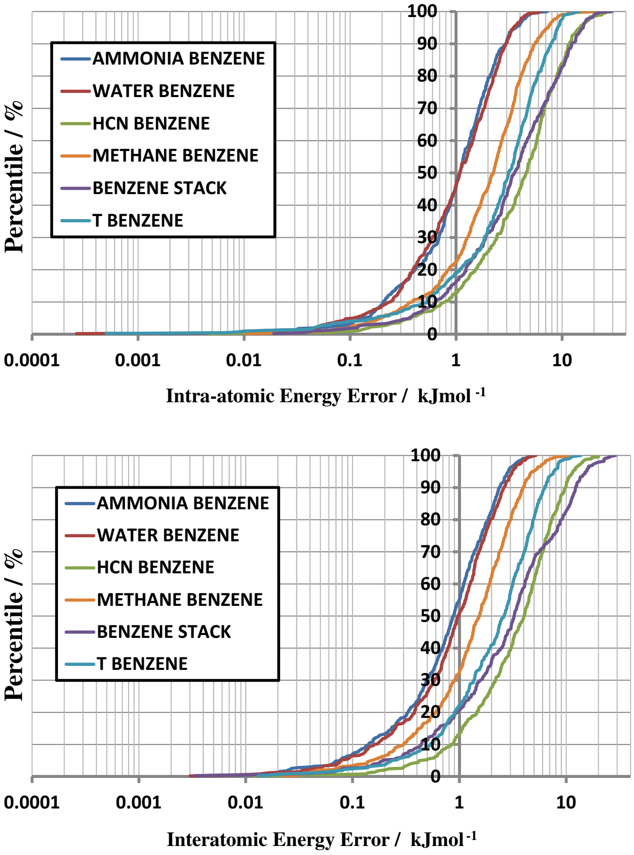

Fig. 4 S-curves displaying the prediction error of the total intra-atomic energy (*top*) and total interatomic energy (*bottom*) for the six weakly bound complexes

much less than 1% of the range in the total intra-atomic and interatomic energies, but slightly higher for the system energy.

From the results, two points must be addressed that arose in the analysis. Firstly, the HCN…benzene complex has an energy sampling range much greater than any of the other complexes, by up to an order of magnitude for the $\Delta E_{\text{intra}}^{\text{A}}$ and $\Delta V_{\text{inter}}^{\text{AA}'}$ energies. Such a large sampled energy range is the reason the S-curve is shifted to higher energy prediction errors. However, obtaining models with a MAE % smaller than 0.26% for energy ranges of ~2200 kJ mol^{-1} is testament to the proficiency of the kriging algorithm and encouraging for the future of FFLUX. The second point to address is the cause of the stacked-benzene ($C_{2\text{h}}$) complex S-curves being shifted for the $\Delta E_{\text{intra}}^{\text{A}}$ and $\Delta V_{\text{inter}}^{\text{AA}'}$ energies. Observing the MSEs of the predictions within the atomic models for the stacked-benzene ($C_{2\text{h}}$) dimer allowed us to identify numerous test points that lay outside the domain of applicability. Those considered very far from the training set region of conformational space (>~10 kJ mol^{-1}) were removed from the plot. However, a number of points within a few kJ mol^{-1} of the training range were still included. The inclusion of such points is one of three possible causes for the shifting of the S-curve, the other two being (1) the PES is undulant for the $\Delta E_{\text{intra}}^{\text{A}}$ and $\Delta V_{\text{inter}}^{\text{AA}'}$ energies, making them independently more difficult to model than the singular $\Delta E_{\text{IQA}}^{\text{system}}$, or (2) the cancellation of errors from the summation of the $E_{\text{intra}}^{\text{A}}$ and $V_{\text{inter}}^{\text{AA}'}$ energy models is particularly high, causing a significantly improved S-curve for the resulting system model.

Conclusions and further work

The results of the investigation demonstrate that the IQA atomic energies can be modelled by kriging as a function of nuclear coordinates to high accuracy for weakly bound intermolecular systems featuring a mixture of intermolecular interactions. As such systems are ubiquitous within chemistry, and the accurate modelling of system energies of bound systems is of great importance in the design of a next-generation force field such as FFLUX, the extension of the modelling approach to incorporate bound complexes was necessary. As the models are built on *ab initio* values for such IQA energies, kriging allows for near-*ab initio* atomic energies to be obtained in a fraction of the time. The models are able to describe bound systems with complex intermolecular interactions, including dispersion and hydrogen bonding, to within 2.6% accuracy for the molecular energy, and within 2.1% for the individual $E_{\text{intra}}^{\text{A}}$ and $V_{\text{inter}}^{\text{AA}'}$ atomic models.

The current work extends the applications that the GAIA protocol can operate on, allowing future progress to be made on larger, more complex chemical systems. For example, knowledge that the hydrogen bond in the water dimer can be kriged to a high accuracy opens the door to working on larger water clusters as well as hydrated molecules. Recent work has been started by others in the group on such systems. Further work will focus on the scaling up of these investigations, along with the creation of *strategic* training sets, designed to reduce the likelihood of errors resulting from a point arising outside of the domain of applicability.

Acknowledgements P.M. thanks the BBSRC for the award of PhD funding and P.L.A.P. acknowledges the EPSRC for funding through the award of an Established Career Fellowship (grant EP/K005472). We thank Timothy J. Hughes for his preliminary work on this topic, as part of his PhD thesis.

Compliance with ethical standards

Ethical statement All ethical guidelines have been adhered.

Conflict of interest There are no conflicts of interest.

References

1. Popelier PLA (2016) Molecular simulation by knowledgeable quantum atoms. Phys Scr 91:033007
2. Popelier PLA (2015) QCTFF: on the construction of a novel protein force field. Int J Quant Chem 115:1005–1011
3. Bader RFW (1990) Atoms in molecules. A quantum theory. Oxford Univ. Press, Oxford
4. Popelier PLA, Smith PJ (2002) Quantum Topological Atoms. In: Hinchliffe A (ed) Chemical Modelling: Applications and Theory, vol 2. Royal Society of Chemistry Specialist Periodical Report, Ch. 8, pp 391–448
5. Popelier PLA (2000) Atoms in molecules. An introduction. Pearson Education, London
6. Matta CF, Boyd RJ (2007) The quantum theory of atoms in molecules. From solid state to DNA and drug design. Wiley-VCH, Weinheim
7. Popelier PLA (2016) Quantum chemical topology. In: Mingos M (ed) The chemical bond - 100 years old and getting stronger. Springer, Switzerland, pp. 71–117
8. Popelier PLA (2016) On quantum chemical topology. In: Chauvin R, Lepetit C, Alikhani E, Silvi B (eds) Challenges and advances in computational chemistry and physics dedicated to “applications of topological methods in molecular chemistry”. Springer, Switzerland, pp. 23–52
9. Bader RFW, Beddall PM (1972) Virial field relationship for molecular charge distributions and the spatial partitioning of molecular properties. J Chem Phys 56:3320–3329

10. Rafat M, Popelier PLA (2007) Atom-atom partitioning of total (super)molecular energy: the hidden terms of classical force fields. J Comput Chem 28:292–301
11. Yuan Y, Mills MJL, Popelier PLA (2014) Multipolar electrostatics for proteins: atom-atom electrostatic energies in Crambin. J Comput Chem 35:343–359
12. Handley CM, Hawe GI, Kell DB, Popelier PLA (2009) Optimal construction of a fast and accurate polarisable water potential based on multipole moments trained by machine learning. Phys Chem Chem Phys 11:6365–6376
13. Cressie N (1993) Statistics for spatial data. Wiley, New York
14. Mills MJL, Popelier PLA (2011) Intramolecular polarisable multipolar electrostatics from the machine learning method kriging. Comput Theor Chem 975:42–51
15. Mills MJL, Popelier PLA (2012) Polarisable multipolar electrostatics from the machine learning method kriging: an application to alanine. Theor Chem Accounts 131:1137–1153
16. Kandathil SM, Fletcher TL, Yuan Y, Knowles J, Popelier PLA (2013) Accuracy and tractability of a kriging model of intramolecular polarizable multipolar electrostatics and its application to histidine. J Comput Chem 34:1850–1861
17. Fletcher TL, Davie SJ, Popelier PLA (2014) Prediction of intramolecular polarization of aromatic amino acids using kriging machine learning. J Chem Theory Comput 10:3708–3719
18. Fletcher TL, Popelier PLA (2016) Multipolar electrostatic energy prediction for all 20 natural amino acids using kriging machine learning. J Chem Theor Comput 12:2742–2751
19. Fletcher TL, Popelier PLA (2015) Transferable kriging machine learning models for the multipolar electrostatics of helical decaalanine. Theor Chem Accounts 134:131–116
20. Fletcher TL, Popelier PLA (2017) Toward amino acid typing for proteins in FFLUX. J Comput Chem 38:336–345
21. Davie SJ, Di Pasquale N, Popelier PLA (2016) Incorporation of local structure into kriging models for the prediction of atomistic properties in the water Decamer. J Comput Chem 37:2409–2422
22. Fletcher TL, Popelier PLA (2016) Polarizable multipolar electrostatics for cholesterol. Chem Phys Lett 659:10–15
23. Cardamone S, Popelier PLA (2015) Prediction of Conformationally dependent atomic multipole moments in carbohydrates. J Comput Chem 36:2361–2373
24. Hughes TJ, Kandathil SM, Popelier PLA (2015) Accurate prediction of polarised high order electrostatic interactions for hydrogen bonded complexes using the machine learning method kriging. Spectrochim Acta A 136:32–41
25. Jurecka P, Sponer J, Cerny J, Hobza P (2006) Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. Phys Chem Chem Phys 8:1985–1993
26. Popelier PLA, Kosov DS (2001) Atom-atom partitioning of intramolecular and intermolecular coulomb energy. J Chem Phys 114: 6539–6547
27. Solano CJF, Pendás AM, Francisco E, Blanco MA, Popelier PLA (2010) Convergence of the multipole expansion for 1,2 coulomb interactions: the modified multipole shifting algorithm. J Chem Phys 132:194110
28. Maxwell P, di Pasquale N, Cardamone S, Popelier PLA (2016) The prediction of topologically partitioned intra-atomic and inter-atomic energies by the machine learning method kriging. Theor Chem Accounts 135:195
29. Blanco MA, Martín Pendás A, Francisco E (2005) Interacting quantum atoms: a correlated energy decomposition scheme based on the quantum theory of atoms in molecules. J Chem Theory Comput 1: 1096–1109
30. Zhao Y, Truhlar D (2008) The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. Theor Chem Accounts 120:215–241
31. Ringer AL, Figgs MS, Sinnokrot MO, Sherrill CD (2006) Aliphatic C-H/ð interactions: methane-benzene, methane-phenol, and methane-indole complexes. J Phys Chem A 110:10822–10828
32. Rafat M, Popelier PLA (2007) Visualisation and integration of quantum topological atoms by spatial discretisation into finite elements. J Comput Chem 28:2602–2617
33. Malcolm NOJ, Popelier PLA (2003) The full topology of the Laplacian of the electron density: scrutinising a physical basis for the VSEPR model. Faraday Discuss 124:353–363
34. Popelier PLA (2014) Chapter 8 The Quantum Theory of Atoms in Molecules. In: Frenking G, Shaik S (eds) The Nature of the Chemical Bond Revisited. Wiley-VCH, Weinheim, pp. 271–308
35. Malcolm N, Gillespie RJ, Popelier PLA (2002) A topological study of homonuclear multiple bonds between the elements of Group 14. Dalton Trans 127(17):3333–3341
36. Becke AD, Edgecombe KE (1990) A simple measure of electron localization in atomic and molecular systems. J Chem Phys 92: 5397–5403
37. Rafat M, Devereux M, Popelier PLA (2005) Rendering of quantum topological atoms and bonds. J Mol Graphics Modell 24:111–120
38. Wilson A, Popelier PLA (2016) Exponential relationships capturing atomistic short-range repulsion from the interacting quantum atoms (IQA) method. J Phys Chem A 120:9647–9659
39. Flick JC, Kosenkov D, Hohenstein EG, Sherrill CD, Slipchenko LV (2012) Accurate prediction of noncovalent interaction energies with the effective fragment potential method: comparison of energy components to symmetry-adapted perturbation theory for the S22 test set. J Chem Theory Comput 8:2835–2843
40. Popelier PLA, Joubert L, Kosov DS (2001) Convergence of the electrostatic interaction based on topological atoms. J Phys Chem A 105:8254–8261
41. Maxwell P, Martin Pendas A, Popelier PLA (2016) Extension of the interacting quantum atoms (IQA) approach to B3LYP level density functional theory. Phys Chem Chem Phys 18:20986–21000
42. Francisco E, Casals-Sainz JL, Rocha-Rinza T, Martin-Pendas A (2016) Partitioning the DFT exchange-correlation energy in line with the interacting quantum atoms approach. Theor Chem Accounts 135:170
43. McDonagh JL, Vincent MA, Popelier PLA (2016) Partitioning dynamic electron correlation energy: viewing Møller-Plesset correlation energies through interacting quantum atom (IQA) energy partitioning Chem. Phys Lett 662:228–234
44. Gráfová L, Pitoňák M, Řezáč J, Hobza P (2010) Comparative study of selected wave function and density functional methods for noncovalent interaction energy calculations using the extended S22 data set. J Chem Theory Comput 6:2365–2376
45. Hughes TJ, Cardamone S, Popelier PLA (2015) Realistic sampling of amino acid geometries for a multipolar polarizable force field. J Comput Chem 36:1844–1857
46. GAUSSIAN09, Revision B.01, Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery JA Jr, Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam JM, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas Ö,

Foresman JB, Ortiz JV, Cioslowski J, Fox DJ (2009) Gaussian, Inc., Wallingford CT, USA, 2009, GAUSSIAN09

47. Keith TA (2016) AIMAll TK Gristmill Software, Overland Park KS, USA, (aim.tkgristmill.com)
48. Di Pasquale N, Davie SJ, Popelier PLA (2016) Optimization algorithms in optimal predictions of atomistic properties by kriging. J Chem Theor Comp 12:1499–1513
49. Di Pasquale N, Bane M, Davie SJ, Popelier PLA (2016) FEREBUS: highly parallelized engine for kriging training. J Comput Chem 37:2606–2616
50. Aicken FM, Popelier PLA (2000) Atomic properties of selected biomolecules. Part 1. The interpretation of atomic integration errors. Can J Chem 78:415–426
51. Fletcher T, Davie SJ, Popelier PLA (2014) Prediction of intramolecular polarization of aromatic amino acids using kriging machine learning. J Chem Theory Comput 10:3708–3719
52. Gu J, Wang J, Leszczynski J (2011) Stacking and H-bonding patterns of dGpdC and dGpdCpdG: performance of the M05-2X and M06-2X Minnesota density functionals for the single strand DNA. Chem Phys Lett 512:108–112
53. Walker M, Harvey AJA, Sen A, Dessent CEH (2013) Performance of M06, M06-2X, and M06-HF density functionals for Conformationally flexible anionic clusters: M06 functionals perform better than B3LYP for a model system with dispersion and ionic hydrogen-bonding interactions. J Phys Chem A 117: 12590–12600
54. Tiwary AS, Datta K, Mukherjee AK (2015) Performance of the M06 family of functionals in predicting the charge transfer transition energies of molecular complexes of TCNE with a series of methylated indoles. Computational and Theoretical Chemistry 1068:123–127
55. Tiwary AS, Mukherjee AK (2014) Performance of the M06 family of functionals in prediction of the charge transfer transition energies of the naphthalene–TCNE and pyrene–TCNE molecular complexes. Chem Phys Lett 610–611:19–22
56. Grimme S, Antony J, Ehrlich S, Krieg H (2010) A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. J Chem Phys 132: 154104–154122
57. Maxwell P, Popelier PLA (2016) Transferable atoms: an intra-atomic perspective through the study of homogeneous oligopeptides. Molec Phys 114:1304–1316
58. Davie SJ, Maxwell PI, Popelier PLA (2016) The long-range convergence of the energetic properties of the water monomer in bulk water at room temperature. Phys Rev. Lett under revision