Ab initio atom–atom potentials using CamCASP: Theory and application to many-body models for the pyridine dimer.

Alston J. Misquitta[1] and Anthony J. Stone[2]

[1]*School of Physics and Astronomy, Queen Mary, University of London, London E1 4NS, U.K.*[*]
[2]*University Chemical Laboratory, Lensfield Road, Cambridge, CB2 1EW, U.K.*

(Dated: November 7, 2016)

Creating accurate, analytic atom–atom potentials for small organic molecules from first principles can be a time-consuming and computationally intensive task, particularly if we also require them to include explicit polarization terms, which are essential in many systems. We describe how the CamCASP suite of programs can be used to generate such potentials using some of the most accurate electronic structure methods currently applicable. We derive the long-range terms from monomer properties, and determine the short-range anisotropy parameters by a novel and robust method based on the iterated stockholder atom approach. Using these techniques we develop distributed multipole models for the electrostatic, polarization and dispersion interactions in the pyridine dimer, and develop a series of many-body potentials for the pyridine system. Even the simplest of these potentials exhibits r.m.s. errors of only about 0.6kJ mol^{-1} for the low-energy pyridine dimers, significantly surpassing the best empirical potentials. Our best model is shown to support eight stable minima, four of which have not been reported in the literature before. Further, the functional form can be made systematically more elaborate so as to improve the accuracy without a significant increase in the human-time spent in their generation. We investigate the effects of anisotropy, rank of multipoles, and choice of polarizability and dispersion models.

PACS numbers:

I. INTRODUCTION

Electronic structure methods for the interaction energy have come a long way since the mid-nineties, when the water dimer represented the largest system for which accurate, *ab initio* intermolecular interaction energies could be calculated. We can now calculate interaction energies for small organic molecules like pyridine and benzene in hours on a single processor [1–3], and medium sized molecules like cyclotrimethylene trinitramine (RDX) [4], base pairs [5], and tetramers of amino acids [6]. Part of the reason for this is the increase in our computational resources, but more important are the new developments in electronic structure methods. For the field of intermolecular interactions, the development of symmetry-adapted perturbation theory based on density-functional theory, or SAPT(DFT), has done much to improve both the accuracy and the range of applicability of theoretical methods. [1, 2, 7–13]

However, such calculations cannot be used on the fly in most molecular simulations, as the computational cost is too high, and we need to represent the interaction energy by an analytic potential. Such potentials are commonly expressed in terms of the many-body expansion, where the interaction energy of a cluster of interacting molecules is partitioned into two-body contributions plus corrections arising from triplets, quartets and larger clusters of molecules. That is,

$$V_{ABC\ldots} = \sum_{X<Y} V_{XY} + \sum_{X<Y<Z} \Delta V_{XYZ} + \cdots , \qquad (1)$$

where V_{XY} is the interaction energy between molecules X and Y in the absence of all other molecules, but in the geometry found in the complete system, while ΔV_{XYZ} is the three-body correction, defined as

$$\Delta V_{XYZ} = V_{XYZ} - V_{XY} - V_{XZ} - V_{YZ}$$

and V_{XYZ} is the energy of the XYZ cluster in the absence of all other molecules, but in the geometry found in the complete system. Four-body, five-body and other many-body corrections are defined in a similar manner.

The success of this expansion depends on its rapid convergence. In any molecular system with distinct interacting units, the two-body terms will dominate, but the many-body terms can contribute as much as 30% of the interaction energy for clusters of polar molecules [14–16], and can be essential for getting the structure and properties correct. For example, three and four-body effects have been shown to be responsible for the tetrahedral structure of liquid water [17]. The many-body polarization energy has also been shown to be an important discriminator in the relative lattice energies of molecular crystals when the structures differ considerably in their hydrogen-bonding motifs [18].

A three-body implementation of SAPT(DFT) does exist [19], but the computational cost makes on-the-fly methods even more impractical, and although three-body non-additive interactions make up the bulk of the many-body non-additivity in systems like water, non-additive effects beyond this level cannot be neglected [17]. If the constituent bodies in a cluster are small enough, it would be possible to use an electronic structure method like SAPT(DFT) or CCSD(T) (coupled-cluster singles and doubles with non-iterated triples) for the two and three-body terms in the many-body expansion, and an appropriate approximation for the other terms. But more generally this approach would make formidable computational

[*]Electronic address: a.j.misquitta@qmul.ac.uk

demands, and it is necessary to use analytic intermolecular potentials in many applications.

Analytic intermolecular potentials have been in use for many decades. (See ref. 20 for a review.) In the past, most have been 'pair potentials', including only two-body terms. In any molecular system with distinct interacting units, the two-body terms will dominate, but the many-body terms can be essential for getting the structure and properties correct. The effects of many-body terms have often been included in an approximate 'average' manner through adjustment of the empirical parameters. This is done in empirical potentials for water, which typically feature an enhanced dipole moment to mimic the increased average dipole of the water molecule in the condensed phase. While such pair potentials are still widely used, it is increasingly recognised that it is necessary to take account of the many-body effects explicitly, particularly to account for the effects of electrostatic polarization [18, 21, 22], but also to account for many-body dispersion effects [23–25], and, as we shall see, to account for intermolecular charge delocalisation, or charge transfer (CT).

Potentials with this level of complexity, accuracy and detail cannot be obtained empirically. Instead we must turn to theoretical methods. *Ab initio*-derived potentials are by no means new, and indeed there are a number of accurate potentials in the published literature (see for example refs. 26–29). These potentials have typically been obtained for small dimers, but recently examples involving medium sized systems have become available [4, 30–32]. There are a few common ideas used in the creation of these and other *ab initio* potentials. The first is that they are all based on a distributed model; that is, the interaction energy between molecules is represented as the sum of contributions between pairs of atoms. Secondly, most are not polarizable, so many-body polarization terms are missing (though polarization may be included at the two-body level). Thirdly, in all cases, long-range parameters have been derived from the unperturbed molecules, which can dramatically simplify the number of free parameters in the fit. Finally, the short-range parameters are usually then fitted to a set of *ab initio* interaction energies calculated using a suitable electronic structure method.

The above procedure works reasonably well, but it has a number of deficiencies. First and foremost is the usual lack of many-body polarization effects. Second, there is much uncertainty associated with fitting the short-range exponential terms in a system of medium sized molecules. These uncertainties are largely related to sampling: we are usually not sure that we have enough data to define the terms in the potential. This is particularly troublesome for the larger systems, which not only have a larger number of free parameters to fit, but which also incur considerable computational expense to calculate the *ab initio* interaction energies needed for the fit. Additionally, the short-range terms are usually exponential in form, and it is very difficult to fit a sum of exponentials while also requiring that the fit parameters remain physically sensible and transferable. Some of these difficulties can be partially alleviated by iterating the process and adding additional data at important configurations [30], but on the whole this approach is unsatisfactory and tedious, and an alternative is needed.

The alternative we describe in this paper is to compute directly most of the potential parameters, including those associated with the short-range part of the potential, and keep the fitting to a minimum. In many ways this is not a new strategy; indeed, a similar technique has been implemented by Schmidt and co-workers [33–35], who have used many of the techniques we will describe in this paper to develop a family of transferable potentials with a strong physical basis. However, so far these have been isotropic potentials of moderate accuracy, with a strong focus on ease of creation and transferability. As we will demonstrate here, we bring a new level of fidelity, accuracy and reliability to the procedure, using the many tools we have developed in recent years and have implemented in the CamCASP [36] program. We begin this paper with a description of the overall strategy, then describe some of the algorithms we have implemented in the CamCASP suite of programs to implement the strategy. In particular, partitioning the electron density using the iterated stockholder atom procedure is very effective in overcoming the difficulties in fitting the short-range potential. We shall apply these methods to the pyridine dimer and discuss the resulting potentials.

II. THE PROBLEM AND DEFINITIONS

The goal is to find an analytic potential $V_{\rm int}$ that accurately models the two-body SAPT(DFT) interaction energy

$$E_{\rm int}^{(1-\infty)} = E_{\rm elst}^{(1)} + E_{\rm exch}^{(1)} + E_{\rm IND}^{(2)} + E_{\rm DISP}^{(2)} + \delta_{\rm int}^{\rm HF}. \qquad (2)$$

(We will use E throughout to denote the computed energy terms and V to denote their analytic representations.) Here $E_{\rm elst}^{(1)}$ and $E_{\rm exch}^{(1)}$ are the first-order electrostatic and exchange-repulsion energies, $E_{\rm IND}^{(2)} = E_{\rm ind,pol}^{(2)} + E_{\rm ind,exch}^{(2)}$ is the total second-order induction energy, $E_{\rm DISP}^{(2)} = E_{\rm disp,pol}^{(2)} + E_{\rm disp,exch}^{(2)}$ is the total dispersion energy [37], and $\delta_{\rm int}^{\rm HF}$ is the estimate of effects of third and higher order, primarily induction [38, 39]. The broad strategy we have adopted to determine $V_{\rm int}$ has been described in some detail in a review article [40]. While many of the details have changed, the essence of the method remains as described there, so only a high-level description will be provided here.

First of all, we represent the potential $V_{\rm int}$ as

$$V_{\rm int} = \sum_{a\in A}\sum_{b\in B} V_{\rm int}[ab](r_{ab}, \Omega_{ab}), \qquad (3)$$

where, a and b label sites (usually taken to be atomic sites) in the interacting molecules A and B, r_{ab} is the inter-site separation, Ω_{ab} is a suitable set of angular coordinates that describes the relative orientation of the local axis systems on these sites (see ch. 12 in ref. 20), and $V_{\rm int}[ab]$ is the site–site potential defined as

$$V_{\rm int}[ab] = V_{\rm sr}[ab] + V_{\rm elst}[ab] + V_{\rm disp}[ab] + V_{\rm pol}[ab]. \qquad (4)$$

The terms in $V_{\rm int}[ab]$ model the corresponding terms in $E_{\rm int}^{(1-\infty)}$. $V_{\rm sr}[ab]$ is the short-range term, which mainly describes the exchange–repulsion energy, but also includes some

other short-range effects, discussed in §VI:

$$V_{\rm sr}[ab] = G\exp\left[-\alpha_{ab}(\Omega_{ab})(r_{ab}-\rho_{ab}(\Omega_{ab}))\right], \qquad (5)$$

where $\rho_{ab}(\Omega_{ab})$ is the shape function for this pair of sites, which depends on their relative orientation Ω_{ab}, and α_{ab} is the hardness parameter which may also be a function of orientation. G is a constant energy which we will take to be 10^{-3} hartree. $V_{\rm elst}[ab]$ is the expanded electrostatic energy:

$$V_{\rm elst}[ab] = V_{\rm elst}[ab](r_{ab},\Omega_{ab},Q^a_t,Q^b_u,\beta^{ab}_{\rm elst}); \qquad (6)$$

Q^a_t is the multipole moment of rank t for site a, where, using the compact notation of ref. 20, $t = 00, 10, 11c, 11s, \cdots$, and $\beta^{ab}_{\rm elst}$ is a damping parameter. The dispersion energy $V_{\rm disp}[ab]$ depends on the anisotropic dispersion coefficients $C^{ab}_n(\Omega_{ab})$ for the pair of sites, and on a damping function f_n that we will take to be the Tang–Toennies [41] incomplete gamma functions of order $n+1$:

$$V_{\rm disp}[ab] = -\sum_{n=6}^{12} f_n\left(\beta^{ab}_{\rm disp}r_{ab}\right)C^{ab}_n(\Omega_{ab})r^{-n}_{ab} \qquad (7)$$

The final term $V_{\rm pol}[ab]$ is the polarization energy, which is the long-range part of the induction energy [42]. $V_{\rm pol}[ab]$ depends on the multipole moments and the polarizabilities α^a_{tu}, which are indexed by pairs of multipole components tu (for details see refs.20, 43):

$$V_{\rm pol}[ab] = V_{\rm pol}[ab](Q^a_t,Q^b_u,\alpha^a_{tu},\alpha^b_{tu},\beta^{ab}_{\rm pol}). \qquad (8)$$

There are a few points to note about the particular form of the potential $V_{\rm pol}[ab]$. Although formally written in the form of a two-body potential, many-body polarization effects are included through the classical polarization expansion [20]. Also, we will normally define the multipole moments and polarizabilities to include *intra*molecular many-body effects implicitly, that is, we use the multipoles and polarizabilities of atoms-in-a-molecule, localized appropriately. To this form of the potential we could add a three-body dispersion model, but this is not addressed in this paper.

III. STRATEGY

There are many parameters in such a potential and our goal is to *compute* as many of these parameters as possible, and keep the fitting of the remainder to a minimum. Additionally, we will adopt a hierarchical approach to the fitting process that helps to guarantee confidence in the parameter values. There are three main parts to the process, and these involve the following:

- *Long-range terms*: The electrostatic, polarization and dispersion interaction energy components possess expansions in powers of $1/R$, where R is the centre-of-mass separation (for small systems) or, more generally, the inter-site distance in a distributed expansion. Multipole moments are functions of the unperturbed molecular densities and may be derived using a variety of methods, the most common being the distributed multipole analysis (DMA) technique [44, 45]. But, using a basis-space algorithm of the iterated stockholder atom (ISA) procedure[46] termed the BS-ISA algorithm [47], we have demonstrated that the ISA/BS-ISA-based distribution yields a more rapidly convergent multipole expansion with properties that make it ideal for modelling. The distributed polarizabilities and dispersion coefficients are obtained using the Williams–Stone–Misquitta (WSM) technique [43, 48–50]. With this approach we may consider the long range parameters in the potential $V_{\rm int}$ as fixed, though, we may optionally tune them if appropriate.

- *Damping*: All three multipole expansions need to be damped at short range, when overlap effects become appreciable and the $1/R$ terms start to exhibit mathematical divergences. Damping will not be applied to the electrostatic expansion as it is not usually needed, but it can be applied if necessary [51]. It is crucial to damp the polarization and dispersion expansions as the mathematical divergence of these expansions is usually manifest at accessible separations, and must be controlled if sensible expansions are needed. For the dispersion expansion we use a single damping coefficient based on the vertical ionization potentials I_A and I_B (measured in atomic units) of the interacting molecules [50]:

$$\beta^{ab}_{\rm disp} \equiv \beta^{AB}_{\rm disp} = \sqrt{2I_A} + \sqrt{2I_B}. \qquad (9)$$

  This single-parameter damping is almost certainly not ideal, and we should rather use damping parameters that depend on the atomic types, and optionally, on their relative orientation. We will propose such a more elaborate, but still non-empirical model in a forthcoming paper [52].

  The damping of the polarization expansion is less straightforward and will be discussed in detail below.

- *Short-range energies*: If the damped multipole (DM) expanded energies are removed from the interaction energy $E^{(1-\infty)}_{\rm int}$, we obtain the remainder which is the short-range energy:

$$\begin{aligned} E^{(1-\infty)}_{\rm sr} &= E^{(1)}_{\rm exch} + (E^{(1)}_{\rm elst} - V^{(1)}_{\rm elst}[\rm DM]) \\ &\quad + (E^{(2)}_{\rm IND} + \delta^{\rm HF}_{\rm int} - V^{(2-\infty)}_{\rm pol}[\rm DM]) \\ &\quad + (E^{(2)}_{\rm DISP} - V^{(2)}_{\rm disp}[\rm DM]) \\ &= E^{(1)}_{\rm sr} + E^{(2-\infty)}_{\rm sr}. \end{aligned} \qquad (10)$$

  Here we have partitioned the short-range energy into a first-order contribution $E^{(1)}_{\rm sr}$ which will be dominant, and the second- to infinite-order contribution $E^{(2-\infty)}_{\rm sr}$ which will be primarily the infinite-order charge-transfer energy. In the above expression, $V^{(1)}_{\rm elst}[\rm DM]$ and $V^{(2)}_{\rm disp}[\rm DM]$ are the multipole expanded forms of the electrostatic and dispersion energies, and $V^{(2-\infty)}_{\rm pol}[\rm DM]$ is the infinite-order (iterated) multipole-expanded polarization energy. In principle, the various contributions

to $E_{\rm sr}^{(1-\infty)}$ are not expected to depend on dimer geometry in the same way and they should be modelled separately. However, we have previously showed that the dominant contributions to $E_{\rm sr}^{(1)}$ —the first-order exchange and penetration energies—are proportional to each other,[47] and here we will show that the charge-transfer contribution is also nearly proportional, so we shall model all parts of $E_{\rm sr}^{(1-\infty)}$ together as a single sum of exponential terms:

$$V_{\rm sr} = \sum_{a\in A}\sum_{b\in B} V_{\rm sr}[ab] \tag{11}$$

where each $V_{\rm sr}[ab]$ has the form of eq. (5).

- *Sampling dimer configuration space*: In order to ensure a balanced fit, it is important to ensure that we sample the six dimensional dimer configuration space adequately. For such a high dimensional space the sampling needs to be (quasi) random, and in earlier work [31, 32, 40] we have described how this can be done using a quasi random Sobol sequence and Shoemake's algorithm [53] (see the supplementary information for a brief description of this algorithm). This algorithm has been implemented in the CamCASP program and ensures that we cover orientation space randomly, but uniformly. Unless otherwise indicated, dimer configurations will be obtained using this algorithm.

- *Fitting the short-range terms: first-order energies*: A direct fit to the terms in $V_{\rm sr}$ usually leads to unphysical parameters and therefore should be avoided. Additionally, it is difficult to sample the high-dimensional configuration space densely enough to define the shape anisotropy of the interacting sites. This is particularly true for the larger molecular systems, for which the computational cost of calculating the second to infinite order SAPT(DFT) interaction energies can be appreciable, thus precluding the possibility of adequate sampling. One possibility in this case is to reduce the complexity of $V_{\rm sr}$ by, for example, keeping only isotropic terms in the expansions for the hardness parameter and the shape functions, but this may not be appropriate when high accuracies are needed.

  In previous work [31] we addressed this problem using the density-overlap model [54–56] to partition the first-order short-range energies, $E_{\rm sr}^{(1)}$, into contributions from pairs of atoms. This partitioning allows us to fit the terms for each pair of sites *ab* and obtain a first guess at $V_{\rm sr}^{(1)}[ab]$, while avoiding fitting the sum of exponential terms directly. In §VI B we provide more detail on how this is done, and show how the parameters in eq. (11) can be determined with a high degree of confidence if we use a density partitioning method based on the ISA method. As we shall see, this procedure effectively eliminates the basis-set limitations seen in our earlier attempts. Moreover, this step uses the first-order energies only, and these energies are not only computationally inexpensive, but may be calculated using a monomer basis set, so a dense coverage of configuration space may be used to determine good initial guesses for the parameters in $V_{\rm sr}^{(1)}$. In this manner, atomic shape functions may be determined easily and reliably.

- *Constrained relaxation*: At various stages in the fitting process we will relax a fit with constraints applied. The idea here is to obtain a good guess for the parameters in the fit in a manner that ensures that they are well-defined. Subsequently, these parameters may be relaxed while pinning them to the predetermined values. Consider a fitting function $g(p_0, p_1, \cdots, p_n)$, where p_i are the free parameters in the fit. If our initial guess for these are p_i^0, then in a constrained relaxation we would optimize the function

$$G(p_0, p_1, \cdots, p_n) = g(p_0, p_1, \cdots, p_n) + \sum_{i=0}^{n} c_i (p_i - p_i^0)^2, \tag{12}$$

  where c_i are suitable constraint strength parameters that should be associated with our confidence in the initial parameter guesses p_i^0. In a Bayesian sense, the p_i^0 are our prior values and the c_i will be related to the prior distribution. As data is included, the parameters p_i may deviate from their initial values. In this manner, a fit may be performed with very little data and we ensure that no parameter attains an unphysical value.

- *Relaxing $V_{\rm sr}^{(1)}$ to $E_{\rm sr}^{(1)}$*: Having obtained the first guess for $V_{\rm sr}^{(1)}$, we may now perform a constrained relaxation of the parameters in $V_{\rm sr}^{(1)}$ to fit $E_{\rm sr}^{(1)}$ better. Symmetry constraints to the shape-function parameters may also be imposed at this stage.

- *Relaxing $V_{\rm sr}^{(1)}$ to include higher-order energies*: The parameters in $V_{\rm sr}^{(1)}$ may now be further relaxed to account for the higher order short-range energies, $E_{\rm sr}^{(2-\infty)}$, thereby obtaining the full short-range potential $V_{\rm sr}$. The higher-order short-range energies will normally be evaluated on a much sparser set of points, so the constraints used in this relaxation step usually need to be fairly tight, and the anisotropy terms should probably be kept fixed at this stage unless enough data can be made available.

- *Overall relaxation and iterations*: The relaxation steps may be repeated as additional data is added. This is a common strategy, but here we do the relaxation with fairly tight constraints. Additional dimer energies are best calculated at special configurations on the potential energy surface. These would include stable minima and regions of configuration space near minima. A suitably *converged* fit is one which is stable with respect to the inclusion of additional data.

Some of these steps have already been used to create accurate *ab initio* potentials [31, 32], and indeed, some of these ideas have been used and developed by other research groups

(see for example, Refs. 30, 56, 57). What is unique to this work is the manner in which these steps have been combined with advanced density-partitioning methods, distribution techniques and a hierarchical calculation of intermolecular interaction energies, so as to obtain intermolecular interaction potentials easily and reliably and with high accuracy. We describe most of these steps in detail below.

IV. NUMERICAL DETAILS

The geometry of the pyridine molecule was optimized using the GAUSSIAN03 program[58] using the PBE0 functional [59] and the cc-pVTZ Dunning basis set [60]. The C_{2v} point group symmetry was imposed during the optimization.

A. Comments on the kinds of basis sets

We use several kinds of basis sets to calculate the various data needed for the intermolecular potential of pyridine. The SAPT(DFT) interaction energies require diffuse monomer basis sets augmented with mid-bond basis functions to converge the dispersion energy, and additionally basis functions located on the partner monomer – the so called far-bond functions — to converge the charge-transfer component of the induction energy. The resulting basis is referred to as the MC+ basis type [61]. The $\delta_{\rm int}^{\rm HF}$ term requires a calculation of the supermolecular interaction energy at the Hartree–Fock level, and therefore needs to be calculated using a dimer-centered basis. In both cases the density-fitting needed for the SAPT and SAPT(DFT) energies is done in a dimer-centered auxiliary basis, possibly augmented with a suitable mid-bond set. For high accuracies the Cartesian form of the auxiliary basis is used.

We compute the large set of first-order energies in a monomer-centered basis and subsequently rotate all quantities to the required dimer orientation. However, for accurate first-order interaction energies, the auxiliary basis used in these calculations must still be the dimer-centered type. Additionally, in this case we use the spherical form of the basis functions as the CAMCASP programme is, as yet, unable to rotate objects calculated using Cartesian functions.

Monomer properties are normally calculated in a monomer-centered basis that is taken to be the monomer part of the basis set used for the SAPT(DFT) energies. However this is not optimal as the additional off-atomic basis functions used in the MC+ basis form have the effect of increasing the size of the equivalent monomer-centered basis set. Consequently, it is advantageous to calculate the monomer properties in a larger, more diffuse monomer basis as this would better match the multipole expanded energies with those from the non-expanded SAPT(DFT) calculations.

B. Basis set details

The distributed molecular properties were calculated using asymptotically corrected PBE0 (PBE0/AC) with the d-aug-cc-pVTZ Dunning basis [62]. The density-functional calculation was performed using a modified version of the DALTON 2.0 program [63] with modifications made using a patch provided as part of the SAPT2008 suite of programs [64]. The asymptotic correction was performed using the Fermi–Amaldi long-range form of the exchange potential with the Tozer–Handy splicing function [65] and a vertical ionization potential of 0.3488 a.u., calculated using a Δ-DFT procedure with the PBE0 functional and an aug-cc-pVTZ basis set. The CAMCASP program [36] was used to evaluate the distributed multipole moments using both DMA and ISA algorithms, and distributed static and frequency-dependent polarizabilities and dispersion coefficients using the WSM algorithm [43, 48–50]. For the ISA calculation the auxiliary basis was constructed from the RI-MP2 aug-cc-pVQZ fitting basis [66, 67] with *s*-functions replaced with those from ISA-set2 supplied with the CAMCASP program [47].

Interaction energy calculations using SAPT(DFT) were performed using the CAMCASP program with molecular orbitals and eigenvalues calculated with the DALTON 2.0 program using the PBE0/AC functional described above. Second-order SAPT(DFT) interaction energy calculations were performed using the Sadlej-pVTZ basis in the MC+ format (monomer basis plus mid-bond functions) with a $3s2p1d$ mid-bond set [23] placed on a site determined using a dispersion-weighted algorithm [68]. The DC+ form of the RI-MP2 aug-cc-pVTZ auxiliary basis [66, 67] with Cartesian GTOs was used for density-fitting with a $3s3p3d2f1g$ fitting mid-bond set with exponents *s*: (1.1061,0.5017,0.2342), *p*: (0.94,0.5,0.25), *d*: (0.9,0.6,0.3), *f*: (0.7,0.4), *g*: (0.65). The hybrid ALDA+CHF kernel was used in all SAPT(DFT) calculations. Kernel integrals were calculated initially within the DALTON 2.0 program, but subsequently they were computed internally in CAMCASP with the ALDA part of the kernel constructed from Slater exchange components and PW91 correlation kernel [69]. The $\delta_{\rm int}^{\rm HF}$ correction was evaluated using the DC+ form of the Sadlej pVTZ basis with a corresponding DC+ auxiliary basis set formed from the RI-MP2 aug-cc-pVTZ fitting basis and $3s3p3d2f1g$ fitting mid-bond set.

Additionally, first-order SAPT(DFT) interaction energies used in the initial stage of the fit were calculated using a monomer-centered (MC) Sadlej-pVTZ basis [70] and a dimer-centered (DC) RI-MP2 aug-cc-pVTZ fitting basis [66, 67]. The density-functional calculations on the monomers were performed once using the PBE0/AC functional, and the molecular orbitals were suitably rotated within CAMCASP for subsequent first-order interaction energy calculations. For this purpose, due to current requirements within CAMCASP, the spherical form of the Gaussian-type orbitals (GTOs) was used for the auxiliary basis set.

C. Data sets

In §III we have described how intermolecular interaction potentials may be developed in multiple stages, with more accurate, but less extensive data sets used in each successive

stage. The pyridine dimer potentials we describe below have been developed using three data sets:

- *Dataset(0)*: First-order energies calculated on a set of 3515 pseudo-random dimer geometries obtained using Shoemake's algorithm as described above. This data set was used in the first stage of the fitting process to obtain the initial short-range parameters using the distributed density-overlap model.

- *Dataset(1)*: Infinite-order SAPT(DFT) interaction energies calculated on a set of 500 pseudo-random dimers also obtained using Shoemake's algorithm. This data set was used in refining the dispersion model, in fitting the charge-transfer contribution to the interaction energy, and, in the final stage, to tune the total interaction energy models.

- *Dataset(2)*: Infinite-order SAPT(DFT) interaction energies calculated on a set of 257 dimers obtained as special points (minima) from early versions of the pyridine potential development. These dimers are significantly lower in energy than those from Dataset(1). This data set served two purposes: Firstly, as it contained dimer geometries significantly different from those found in Dataset(1), it provided us with an independent means of assessing the quality of the fits. Secondly, in the final stage, this data set was used to tune the total interaction energy models.

- *Dataset(3)*: Infinite-order SAPT(DFT) interaction energies calculated on a set of 250 pseudo-random dimers in a manner similar to Dataset(1). This set will be used only in the assessment of the models.

V. LONG-RANGE METHODS

One of the fundamental advantages of intermolecular perturbation theories like SAPT and SAPT(DFT) over supermolecular methods is that the energy components from perturbation theory have well-defined multipole expansions [71]. Therefore the long-range form of these energies can be derived from molecular properties such as the multipole moments and static and frequency-dependent density-response functions. This has the advantage that the asymptotic part of the potential energy surface is obtained directly, that is, without fitting. Additionally, the long-range potential parameters are fully consistent with the short-range energies from the perturbation theory.

In the CamCASP suite of programs, we have implemented a number of algorithms for calculating the distributed forms of the long-range expansions of the electrostatic, polarization (induction) and dispersion energies. The algorithms permit a considerable degree of freedom in the model, so models may be more or less complex as the application requires. The long-range terms in the model can be derived directly from monomer properties, but there is a conflict between accuracy and computational efficiency. We will aim to model most of the contributions to the interaction energy separately, using several versions ranging from accurate but computationally expensive to less accurate but cheaper. For example, electrostatic models may be constructed using multipole models from rank 0 (charges only) up to rank 4; or mixed rank models may also be considered, with high ranking multipoles included only on some sites. This allows a considerable degree of flexibility in constructing the total interaction energy model. For this approach to work, we will need to ensure that each part of the model is sufficiently accurate, with accuracy measured in a meaningful manner. Typically, we will expect to reduce r.m.s. errors against some SAPT(DFT) reference to less than 1 kJ mol^{-1}, and preferably less than 0.5 kJ mol^{-1}.

A. Electrostatic models

Distributed multipole analysis is a well established procedure for obtaining accurate electrostatic models from an *ab initio* wavefunction. We use the revised version of the procedure[45] which reduces the dependence of the multipole description on basis set, at the cost of longer computation times. This procedure uses a scheme based on real-space grids for the density contributions arising from the diffuse functions, while for the more compact functions in the basis the original scheme is used. In this work the parameter controlling the switch between compact and diffuse functions is set at 4.0, so the method is denoted DMA4.

Until recently, the DMA approach has been the standard for distributed moments, but recently we have demonstrated [47] that the ISA-based distributed multipole analysis (ISA-DMA) forms a significantly better basis for potential development as it guarantees fast and systematic convergence with respect to the rank of the expansion and a well-defined basis limit to the multipole components, and yields penetration energies (calculated as the difference from the non-expanded $E^{(1)}_{\rm elst}$) more strongly proportional to the first-order exchange energy $E^{(1)}_{\rm exch}$. The last aspect of the ISA-DMA is particularly useful in model building, since the proportionality of the electrostatic penetration energy to the first-order exchange-repulsion energy allows us to combine the two and model their sum with a single function. For the purposes of this paper we will define the electrostatic penetration energy as [47]

$$E^{(1)}_{\rm pen} = E^{(1)}_{\rm elst} - V^{(1)}_{\rm elst}[\mathrm{DM}], \tag{13}$$

where $V^{(1)}_{\rm elst}[\mathrm{DM}]$ denotes the electrostatic energy calculated from the distributed multipole (DM) expansion evaluated at convergence, which we will take to be the model with terms from ranks 0 (charge) to 4 (hexadecapole).

In fig. 7 of ref. 47 we demonstrated this aspect of the ISA-DMA moments: in contrast to the DMA4 moments, the penetration energy derived from the ISA-DMA model at rank 4 is indeed significantly more proportional to $E^{(1)}_{\rm exch}$ for the pyridine dimer. This alone makes the ISA-DMA model more appropriate for this system—or indeed, any other, as this proportionality seems to be generally true. Here we will look at the data presented in ref. 47 differently, to show more clearly

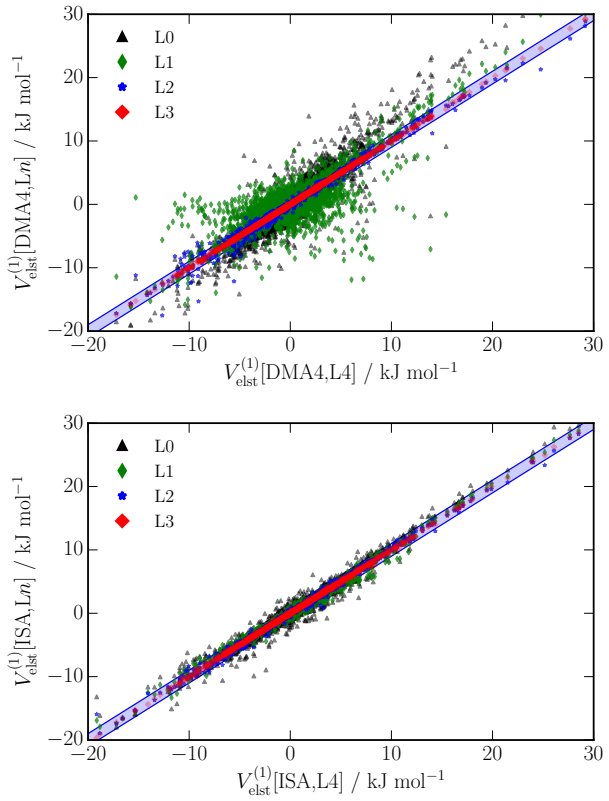

FIG. 1: Scatter plot of model electrostatic energies from the DMA4 and ISA-DMA models at various ranks. The multipole expanded electrostatic energies $V^{(1)}_{\rm elst}$[DM] for rank n models, $n = 0, 1, 2, 3$, (i.e. including multipole moments only up to rank n) are plotted against the energies calculated with the rank 4 model (on the x-axis). No damping has been used. The DMA4 results are in the top panel and the ISA-DMA (BS-ISA, $\zeta = 0.1$) results are presented in the bottom panel. The blue bar represents the ± 1 kJ mol^{-1} error range.

how rapidly the DMA4 and ISA-DMA multipole expansions converge with rank.

For the construction of accurate electrostatic models, it is advisable to include atom charges, dipoles and quadrupoles. The dipoles are needed to describe features such as lone pairs, while quadrupoles are needed to describe π-orbital features. Octopoles and hexadecapoles can improve the description further but the improvement is not generally worth the increased computational cost of the model. However, for many applications, particularly for large molecules, due to program design limitations or more fundamentally, due to computational limitations, only charge models may be permissible. So the question arises: How do the multipole models behave when truncated to lower orders in rank? In Figure 1 we have plotted $V^{(1)}_{\rm elst}$[DM] calculated with each of the two multipole models with truncated rank against the same with all terms to rank 4 (deemed to be converged) included. We clearly see that while the rank 4 terms are not needed in the DMA4 model, any further truncation results in unacceptably large errors and very little correlation is left between the converged results (terms to rank 4) and those with ranks limited to 0 (charges) and 1 (charges and dipoles). In contrast, the ISA-DMA multipoles are distinctly better behaved upon truncation, with a strong correlation between all truncated models and the fully converged energies. This has some advantages: it may be possible to truncate the ISA-based distributed multipole model to much lower rank, perhaps even to rank 0, without the need to re-parametrize the potential. We shall return to this issue below.

We point out here that while the DMA4 multipole model is not directly amenable to rank truncation, there is a way to perform a rank transformation that generally does not result in significant errors. This is done using by optimizing a distributed-multipole description using the Mulfit program of Ferenczy *et al.*[72, 73], in which the effects of higher-rank multipoles on each atom are represented approximately by multipoles of lower ranks on neighbouring atoms. In this way, a model including multipoles up to quadrupole can incorporate some of the effects of higher multipoles. This approach has recently been used effectively to generate simple electrostatic models for a wide range of polycyclic aromatic hydrocarbons occurring in the formation of soot.[32, 74] However the ISA-DMA treatment is consistently better.

B. Polarization and charge-transfer

In this paper we distinguish between the *polarization* energy and the *induction* energy. In SAPT (or SAPT(DFT)), the polarization energy and charge-transfer are combined in the induction energy. We use regularised SAPT [75] to separate these two contributions [42], and by polarization energy we mean that part of the induction energy that is not associated with charge transfer.

The importance of polarizability in the interactions between polar and polarizable molecules is now well recognized [18, 76], as is the inadequacy of the common approximation of polarization effects by the use of enhanced static dipole moments. In CamCASP we use coupled Kohn–Sham perturbation theory to obtain an accurate charge-density susceptibility, $\alpha(\mathbf{r}, \mathbf{r}')$, which describes the change in charge density at $\mathbf{r}$ in response to a change in electrostatic potential at $\mathbf{r}'$. Using a constrained density-fitting-based approach [48], the charge density susceptibility is partitioned between atoms to obtain a distributed-polarizability model α^{ab}_{tu} that gives the change in multipole Q^b_u on atom b in response to a change in the electrostatic potential derivative V^a_t at atom a. Here $u = 00$ for the charge, $10 = z$, $11c = x$ or $11s = y$ for the dipole, 20, $21c$, $21s$, $22c$ or $22s$ for the quadrupole components, and so on; while $t = 00$ for the electrostatic potential, 10, $11c$ or $11s$ for the components of the electrostatic field, 20 etc. for the field gradient, and so on. Note that the electric field components are $E_x = E_{11c} = -V_{11c}$, $E_y = E_{11s} = -V_{11s}$ and $E_z = E_{10} = -V_{10}$.

This is a *non-local* model of polarizability. That is, the electric field at one atom of a molecule can induce a change in the multipole moments on other atoms of the same molecule. This is an impractical and unnecessarily complicated description that seems to be needed only for special cases such as low-

dimensional extended systems [77]. For most finite systems, the moments induced on neighbouring atoms b by a change in electric field on atom a can be represented by multipole expansions on atom a, giving a *local* polarizability description in which the effect of a change in electric field at atom a is described by changes in multipole moments on that atom alone. This is a somewhat over-simplified description of the procedure, and more detailed accounts have been given by Stone & Le Sueur[78], and by Lillestolen & Wheatley[79]. The latter is a more elaborate approach that deals rather better with the convergence issues arising from induced moments on atoms distant from the one on which the perturbation occurs. The local polarizability model is a much more compact and useful description. In particular, the local picture removes charge-flow effects, where a difference in potential between two atoms induces a flow of charge between them. Such flows of charge still occur, but they are described in terms of local dipole polarizabilities. We point out here that the 'self-repulsion plus local orthogonality' (SRLO) distribution method [80] can be used to eliminate the charge-flow terms altogether (for most molecules). This technique, which is a modification of the constrained density-fitting-based distribution method [48] is available in CamCASP but has not been used for the results of this paper. The SRLO polarizabilities are non-local and will typically need localization to be usable by most simulation programs.

The resulting localized polarizability description can be refined by the method of Williams & Stone [81] using the point-to-point responses: the change in potential at each of an array of points around the molecule in response to a point charge at any of the points. An important advantage of this method is that the final, refined polarization model can be chosen to suit the problem—for example a simple isotropic dipole–dipole model, or an elaborate model with anisotropic polarizabilities up to quadrupole–quadrupole or higher. For a given choice of model, the refinement procedure ensures that we obtain the highest accuracy (in an unbiased sense if sufficiently dense grids of point-to-point responses are used) subject to the limitations of the model. The combination of the SAPT(DFT) calculation of local (point-to-point) responses with this refinement procedure is referred to here as the WSM method [43, 49].

The quality of the WSM description can be judged by the accuracy of the interaction energy of a point charge with the molecule. This interaction comprises the classical electrostatic energy of interaction of the point charge with the molecular charge distribution, and the additional term, the polarization energy, that arises from the relaxation of the molecular charge distribution in response to the point charge. These components can be separated using SAPT(DFT). The polarization energy of pyridine in the field of a point charge is mapped in the left-hand picture of Figure 2(a). We construct a grid on the vdW×2 surface of pyridine—that is, the surface made up of spheres of twice the van der Waals radius around each atom—and the polarization energy is calculated for a unit point charge at each point of the grid in turn. The remaining three maps in Figure 2(a) show the error in the polarization energy for three local polarizability descriptions: L1 uses dipole polarizabilities on each atom, L2 includes dipole–quadrupole and quadrupole–quadrupole polarizabilities, and L1,iso uses isotropic dipole polarizabilities on each atom. It is clear that the dipole-polarizability models are rather poor, and that an accurate description needs to include quadrupole polarizabilities.

1. *Polarization damping*

If the polarization interaction between molecules is calculated using distributed multipoles for the electrostatic potential and distributed polarizabilities for the polarization model, the effects of molecular overlap are absent and damping is needed to avoid the so-called polarization catastrophe which results in unphysical energies. In our early work on this issue [43] we advocated damping the classical polarization expansion to best match the total induction energies from SAPT(DFT). Through numerical simulations of the condensed phase and the work of Sebetci and Beran [76] we now know this to be incorrect, as it leads to excessive many-body polarization energies. The polarization damping must instead be determined by requiring that the classical polarization model energies best match the true polarization energies from SAPT(DFT) [42]. As noted above, perturbation theories like SAPT and SAPT(DFT) do not define a true polarization energy, but rather the induction energy, which is the sum of the polarization energy and the charge-transfer energy. Recently one of us described how regularized SAPT(DFT) can be used to split the second-order induction energy into the second-order polarization and charge-transfer components [42] which are defined as follows:

$$\begin{aligned} E^{(2)}_{\mathrm{POL}} &= E^{(2)}_{\mathrm{IND}}(\mathrm{Reg}) \\ E^{(2)}_{\mathrm{CT}} &= E^{(2)}_{\mathrm{IND}} - E^{(2)}_{\mathrm{IND}}(\mathrm{Reg}), \end{aligned} \tag{14}$$

where $E^{(2)}_{\mathrm{IND}}(\mathrm{Reg})$ is the regularized second-order induction energy. This definition leads to a well-defined basis limit for the second-order polarization and charge-transfer energies [42]. We determine the damping needed for the classical polarization expansion by requiring that the non-iterated model energies best match $E^{(2)}_{\mathrm{POL}}$. Once a suitable damping has been found, an estimate for the infinite-order polarization energy $E^{(2-\infty)}_{\mathrm{POL}}$ is obtained by iterating the classical polarization model to convergence.

In principle the above procedure gives us a straightforward way to define the damping: once the form of the damping function is chosen (we use Tang–Toennies damping in this work) all we need to do is determine the damping parameters needed by fitting to $E^{(2)}_{\mathrm{POL}}$ energies calculated for a suitable set of dimer orientations. Since the many-body polarization energy is built up from terms involving pairs of sites, we should expect that the damping parameters depend on the pairs of interacting sites, and potentially on their relative orientations. Indeed, one of us has shown [42] that for small dimers the damping parameters do depend quite strongly on the site types involved. A single-parameter damping model that depends only on the types of interacting molecules may

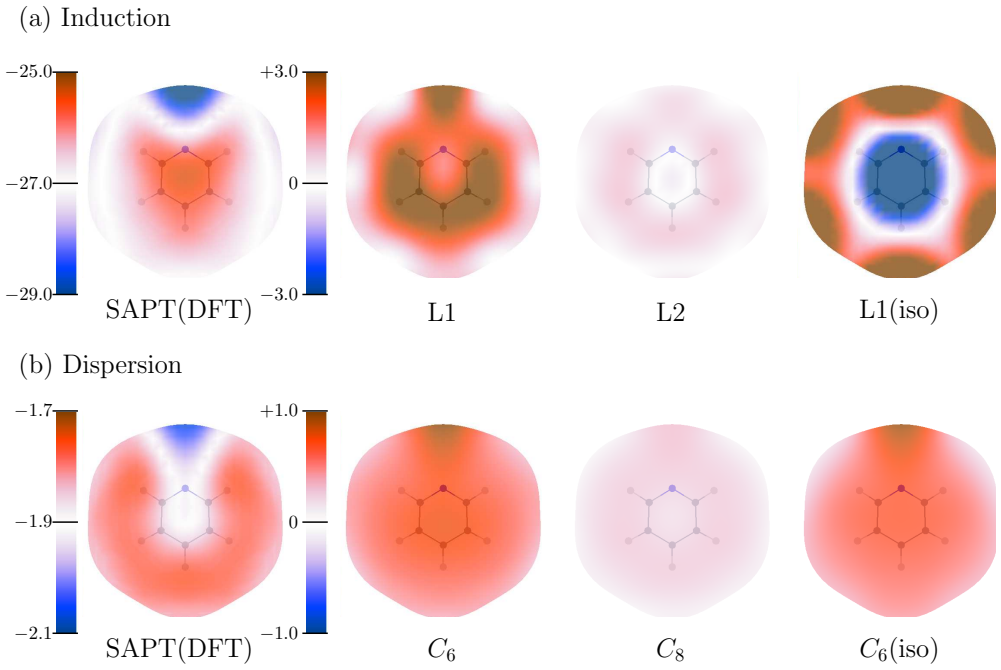

FIG. 2: (a) Polarization and (b) dispersion energy maps and difference maps on the 2×vdW surface of pyridine. Polarization energies have been calculated using a $+1e$ point-charge probe and dispersion energies with a neon atom probe. Energies in kJ mol^{-1}.

be constructed, but such a model is a compromise, and must usually be determined by fitting to data biased towards the important dimer configurations only [42]. The advantage of this approach is that the model is simpler and very few evaluations of $E^{(2)}_{\mathrm{POL}}$ are needed to determine the damping parameter, but the disadvantage is that the model is almost certainly biased towards a few dimer orientations, and additionally, these important orientations need to be known before the final potential is constructed. The last requirement—that we need to have knowledge of the potential—is not as serious as it may seem, as the choice of damping has no effect on the *two-body* interaction potential: this choice affects the many-body polarization energy only. So it is possible to make an informed guess for the damping parameter, determine the parameters of the intermolecular potential, and subsequently re-assess this choice by examining the performance of the polarization model at the important dimer configurations, and, if necessary, alter the model and re-fit.

The initial choice for the damping parameter in pyridine was obtained using two dimer orientations: the doubly hydrogen-bonded C_{2h} dimer, and a T-shaped dimer with the nitrogen of one molecule pointing to the ring of the other. These were chosen so as to sample both H$\cdots$N and N$\cdots$C interactions, though in retrospect the latter proved to be unimportant. In Figure 3 we display the second-order polarization energies calculated using various single-parameter damping models for the C_{2h} structure. Energies for only two of the three polarization models are shown, as the isotropic rank 1 (L1(iso)) model is nearly identical in behaviour to the L1 model. The optimum damping parameter for the L1 model lies between 1.2 and 1.3 a.u., while for the L2 model a stronger damping between 1.0 and 1.1 a.u. is needed. To an extent, the deficiencies of the L1 model are compensated by using a weaker damping.

The single-parameter damping approach has a serious limitation. In Figure 4 we display similar data for the T-shaped dimer orientation with the N of one molecule pointing to the centre of the ring of the other. Here we see that the polarization models need to be considerably more heavily damped with a damping coefficient of 0.9 a.u. for the L1 (and L1(iso)) model and one less than 0.9 a.u. for the L2 model. It is possible that we observe this large variation in the damping because of the strong anisotropy of the molecule, and also because a single damping coefficient is not enough. Perhaps we need to use separate damping parameters for each pair of atoms [42], or even to make the damping parameters orientation-dependent. As a compromise, we have chosen to use the simpler L1 model with a damping coefficient of $\beta_{\mathrm{pol}} = 1.25$ a.u. This model seems capable of describing the polarization in both orientations presented here.

This approach to choosing the damping parameter remains the most problematic part of our approach to potential development. The choice of damping parameters may seem somewhat arbitrary and biased to the choice of dimer configura-

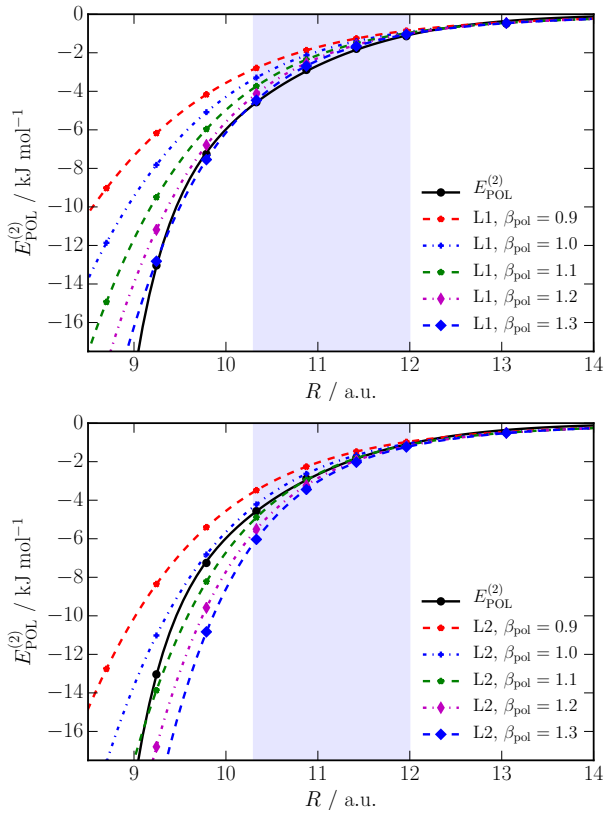

FIG. 3: Second-order polarization energies vs. centre-of-mass separation R for the doubly-hydrogen-bonded pyridine dimer, obtained from regularized SAPT(DFT) and from distributed polarization models. Model polarization energies are reported with local WSM models with a maximum rank of 1 (L1, top) and 2 (L2, bottom). Models are shown with a range of damping parameters using damping functions described in the text. The basin of the minimum along the radial direction is indicated by the light blue shaded region.

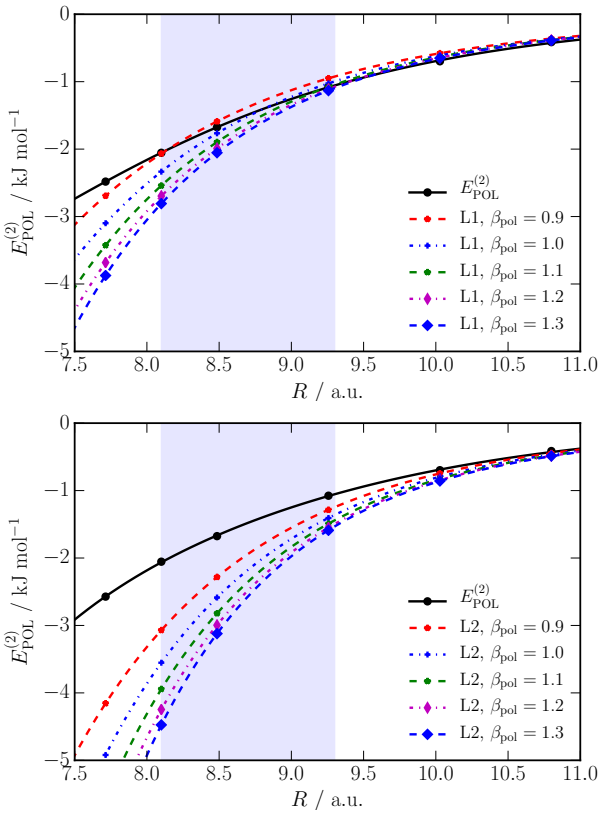

FIG. 4: Second-order polarization energies vs. centre-of-mass separation R for the T-shaped pyridine dimer with N pointing to the centre of the ring. See the caption of Figure 3 for a description.

tions used to determine the damping, but this is probably too pessimistic a view for the following reasons:

- The choice of damping does not affect the two-body interaction energy as the error in the induction energy will be absorbed in the short-range part of the potential. The damping does however alter the many-body polarization energy.

- We should regard this as an iterative process: the damping model will normally be assessed and possibly changed once we have a better understanding of the full PES. Indeed this was done in the present work; we will re-visit this issue in §XII A.

C. Dispersion models

In CamCASP, we normally calculate atom–atom dispersion coefficients using polarizabilities computed at imaginary frequency and localised using the WSM localization scheme. The procedure involves integrals over imaginary frequency[82], and because the imaginary-frequency polarizability is a very well-behaved function of the imaginary frequency the integrals can be carried out accurately and efficiently using Gauss-Legendre quadrature[50]. Since the dispersion coefficients are derived from the WSM polarizability model, it is possible to choose the dispersion model to suit the problem, for example by limiting the polarizabilities to isotropic dipole–dipole, leading to an isotropic C_6R^{-6} model, or by including all polarizabilities up to quadrupole–quadrupole, which yields a model including anisotropic dispersion terms up to R^{-10}. (This latter procedure omits some R^{-10} terms arising from dipole–octopole polarizabilities, but they could be included too if desired.) Within the constraints of the model, the WSM polarizabilities, and hence the WSM dispersion models will be optimized to be the best in an unbiased sense. Within these constraints, intramolecular through-space polarization effects are fully or partially accounted for in the WSM models.

The dispersion energy of pyridine with a neon atom probe placed on the vdW×2 surface of pyridine is mapped in the left-hand picture of Figure 2(b). In the remaining three maps in this figure we show the error in the dispersion energy for three local dispersion models: the C_6 model includes anisotropic C_6 terms on all atoms; the C_8 model additionally includes

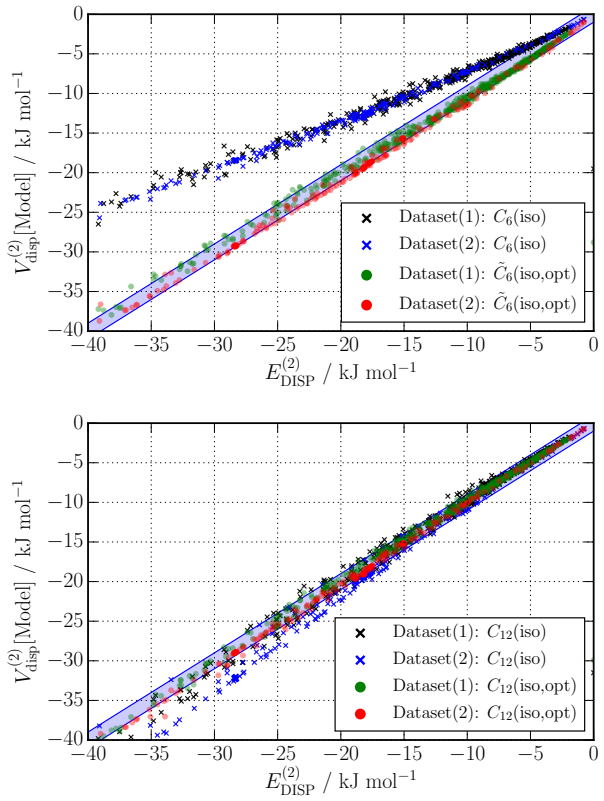

FIG. 5: Scatter plot of pyridine dimer dispersion energies using various models, against reference $E^{(2)}_{\rm DISP}$ energies calculated using SAPT(DFT). See text for details. The bar represents the ±1 kJ mol^{-1} error range.

C_7 and C_8 contributions between the heavy atoms; and the $C_{6,\rm iso}$ model includes only isotropic C_6 terms. The C_{10} and C_{12} models are not shown as they exhibit errors close to zero on the scale shown. It should be clear that to achieve a high accuracy we need to include higher-rank dispersion effects — the dispersion anisotropy is not apparently important in this system, though we may expect it to be important in larger systems. Also, the errors made by both the C_6 models are fairly uniform, and so the lack of higher-order terms in these models may be compensated for by scaling the C_6 coefficients. Indeed, we have demonstrated this in a previous publication [50] and will address this below.

The WSM dispersion models described above need to be suitably damped for them to be applicable in a potential. We have used the Tang–Toennies [83] damping functions and a single damping parameter for all pairs of sites. The damping model needs to account for two effects: First, the SAPT(DFT) dispersion energy, $E^{(2)}_{\rm DISP}$, includes the effects of penetration and exchange, which are absent from the C_nR^{-n} expansion. Secondly, the dispersion expansion suffers from an unphysical mathematical divergence as $R \to 0$. For both reasons the models have to be damped. Damping using the Tang–Toennies functions cancels out the mathematical divergence at small R and, with an appropriate damping parameter, is also able to account for the penetration and exchange effects, albeit approximately. We have opted for the simplest damping model, in which $\beta_{\rm disp}$ depends on the interacting molecules only and is given by eq. (9). With $I_A = I_B = 0.3488$ a.u. we get $\beta_{\rm disp} = 1.67$ a.u.

Figure 5 (bottom) shows the performance of the C_{12}(iso) isotropic dispersion models for the pyridine dimer, As can be seen from the Figure, the above damping works reasonably well for the C_{12}(iso) model with (unweighted) r.m.s. errors of 0.86 kJ mol^{-1} for dispersion energies from Dataset(1) in the energy range −40 to 0 kJ mol^{-1}. However, for Dataset(2) which includes more strongly bound dimers, the model performs less well with an r.m.s. error of 2.30 kJ mol^{-1} in the same energy range. The model dispersion energies are systematically overestimated for the low energy dimers, with errors as large as 4 kJ mol^{-1}. While these errors are just within 'chemical accuracy', they are too large for our purposes. They may stem from the choice of damping function, the damping parameter chosen (in particular, our use of a single, atom-pair independent isotropic damping parameter) and also the WSM dispersion coefficients. To account for some of these deficiencies, while maintaining the isotropy of the model, we have chosen to relax the dispersion coefficients in the C_{12}(iso) model. The relaxation was done using constrained optimisation with harmonic constraints in the form given by eq. (12) used to pin the dispersion coefficients to the values obtained from the WSM procedure. We used tight constraints to prevent the model parameters from taking on unphysical (negative) values. The relaxation was done using only the random dimers from Dataset(1), with the low energy dimers from Dataset(2) used to assess the quality of the relaxation. The relaxed model, C_{12}(iso, opt), is a significant improvement, with r.m.s. errors of 0.41 kJ mol^{-1} on the training set of random dimers and 0.68 kJ mol^{-1} on the test set of low energy dimers.

In a similar manner we have created an isotropic C_6 dispersion model for this system. From Figure 5 (top) we see that the C_6(iso) model systematically underestimates the second-order dispersion energy. This is to be expected, as the higher ranking dispersion contributions are significant for close dimer separations. We have previously argued [43] that rather than use the C_6(iso) model directly, we should instead use a *scaled* model in which all dispersion coefficients are scaled by a constant to match the reference $E^{(2)}_{\rm DISP}$ energies. Here we additionally optimise the scaled model in the manner described above. The resulting model, $\tilde{C}_6$(iso, opt) (here the tilde indicates that this is a scaled model), exhibits an r.m.s. error of 0.68 kJ mol^{-1} on the training set and 1.00 kJ mol^{-1} on the test set. However, such a scaled model will systematically overestimate the long-range contribution to the dispersion energy, and this is a significant drawback: while the scaled C_6(iso) model may be used to model small, gas-phase clusters, it is not suitable for the condensed phase because the scaling causes an excessive van der Waals pressure and the resulting structures are significantly more dense. As one of our aims is to use the resulting potentials in the study of the condensed phase, we cannot use the scaled model. However, we can simplify the C_{12}(iso) model by dropping the R^{-12} terms, which contribute an insignificant amount to the dispersion energy, so we have used

a C_{10}(iso, opt) model in the potentials for pyridine.

VI. SHORT-RANGE ENERGY MODELS

The short-range part of the potential comprises several effects. All of the long-range terms are modified at short range, as mentioned above. The multipole expansion on which the long-range expressions are based converges more slowly or not at all at short distances, and is incorrect when the charge densities overlap, even if it does converge. Damping can be used to correct the dispersion and polarization terms at short range, but in addition there are corrections arising from electron exchange, electrostatic penetration, and charge tunneling, or charge transfer, between the molecules.

The dominant short-range term is the exchange-repulsion: the wavefunction for two overlapping molecules cannot be treated as a simple product of isolated-molecule wavefunctions, but has to be antisymmetrized with respect to electron exchanges between the molecules. This modifies the electron distribution and results in a repulsive energy. It is straightforward to calculate the exchange-repulsion energy *ab initio*, but it has to be fitted by a suitable functional form for use in an analytic potential.

The electrostatic interaction is also modified by the effects of overlap. If a distributed multipole expansion is used, it will still converge at moderate overlap, but it does not converge to the non-expanded energy, $E^{(1)}_{\rm elst}$. The difference between $E^{(1)}_{\rm elst}$ and the converged multipole energy $V^{(1)}_{\rm elst}$[DM] is the electrostatic penetration energy, $E^{(1)}_{\rm pen}$. We have previously shown [47] that $E^{(1)}_{\rm pen}$ is approximately proportional to the first-order exchange energy, so the two terms can, in principle, be modelled together. Alternatively a separate model for $E^{(1)}_{\rm pen}$ can be developed, possibly based on suitable damping functions [51], but we have not explored this possibility.

The contribution to the interaction energy from charge transfer — or, more appropriately, the intermolecular charge delocalisation energy — appears at second and higher orders in the perturbation expansion. Previously one of us has shown that this energy can be interpreted as an energy of stabilization due to electron tunneling [42], so we may expect the charge transfer energy to decay exponentially with separation. In principle, the charge transfer energy should be modelled as a separate exponentially decaying term, but as we shall see, it is approximately proportional to the first-order exchange energy and may therefore be modelled together with $E^{(1)}_{\rm exch}$.

Finally we will use the short-range potential to account for any residual differences between the multipole expansions and the reference SAPT(DFT) energies. The full form of the short-range energy, $E^{(1-\infty)}_{\rm sr}$, is shown in eq. (10) where we have also implicitly defined the first-order short-range energy, $E^{(1)}_{\rm sr}$, and the contributions from second to infinite order, $E^{(2-\infty)}_{\rm sr}$.

A. Fitting the short-range potential

The short-range part of the potential has often been represented by empirical R^{-12} Lennard-Jones atom–atom terms, but for accurate potentials a Born–Mayer (exponential) atom–atom form is usually preferred (eq. (11)), and it is essential in most cases to allow it to be anisotropic, since the non-spherical nature of bonded atoms can have a profound effect on the way that they pack together. Unfortunately, the parameters of the various atom–atom terms are strongly correlated, and this makes the already difficult non-linear fitting problem even more troublesome. A direct fit is generally not possible: it is hard to converge and tends to wander off into unphysical parameter space. Parameters can be forced to stay within reasonable limits, but this introduces an element of arbitrariness in the procedure.

It has however been found empirically that there is a close proportionality between the overlap of the electron densities on two atoms and the exchange–repulsion energy between them. This observation has been used to construct repulsion potentials directly from the density overlap, with varying degrees of success[84]. A better solution, which we adopt here, is to use the density overlap only to guide the parameters in a fitted potential function to a physically meaningful region of parameter space. Once an initial guess to the parameters has been obtained, the fit can be improved using constrained optimisation. Further, we will achieve the final fits to $E^{(1-\infty)}_{\rm sr}$ in stages, first by fitting to only $E^{(1)}_{\rm sr}$, and then by constrained relaxation to include the higher-order contributions from $E^{(2-\infty)}_{\rm sr}$.

B. The density-overlap model

It is useful at this point to review the theoretical basis for the density-overlap model. In the mid-1970's Kita, Noda & Inouye [54], and later, in the early 1980's Kim, Kim & Lee[55] proposed that the intermolecular repulsion energy of rare gas atoms could be modelled as

$$E^{(1)}_{\rm exch}(\mathbf{R}) \approx K(S_\rho(\mathbf{R}))^\gamma, \tag{15}$$

where K and γ are constants and the overlap S_ρ of the two interacting densities ρ^A and ρ^B separated by generalised vector $\mathbf{R}$ is defined as

$$S_\rho(\mathbf{R}) = \int \rho^A(\mathbf{r})\rho^B(\mathbf{r})d\mathbf{r}. \tag{16}$$

Kita *et al.* had $\gamma = 1$ and did not consider the possibility of varying this power, but Kim *et al.* observed that the constant γ was close to, but less than, unity. This model was subsequently used by a number of groups and was successfully applied to study the interactions of polyatomic molecules, and has been investigated [84, 85] together with many other variants. Curiously, to the best of our knowledge, no one seems to have realised the reason for the success of this model, nor why the constant γ is always less than one. Before going on to the numerical details of this model we will discuss both these issues

as we will be led to a better understanding of the model and the exchange-repulsion energies.

First of all we should realise that although the exchange-repulsion and penetration energies are the short-range parts of the interaction energy, these energies result from the overlap of the *density tails* of the interacting densities. That is, we must consider the asymptotic form of the interacting densities for an atomic system [86]:

$$\rho(r) = Cr^{2\beta}e^{-2\alpha r}, \tag{17}$$

where, with I as the vertical ionization energy, and Z the atomic number, we have $\alpha = \sqrt{2I}$ and $\beta = -1 + Q/\alpha$, where for an atom with nuclear charge $+Z$ and electronic charge $-N$, $Q = Z - N + 1$. Both I and r here are in atomic units. In principle, the asymptotic form of the density overlap integral can be obtained by using this density in eq. (16), but the exact integral is not important. Instead we can use the result of Nyeland & Toennies [87] who evaluated eq. (16) using only the exponential term in eq. (17) to get

$$S_\rho(R) = \mathcal{P}(R)e^{-2\alpha R}, \tag{18}$$

where $\mathcal{P}(R)$ is a low-order polynomial in the internuclear separation R. For identical densities $\mathcal{P}(R) = (4/3)\alpha^2R^2 + 2\alpha R + 1$, and for the more general case of different densities, the results of Rosen[88] may be used to obtain a closed-form expression for $\mathcal{P}(R)$ that is now not a low order polynomial, but also includes exponential terms. Since S_ρ is not a pure exponential, Nyeland & Toennies argue that the exchange-repulsion energy should be proportional to $S_\rho(R)/R^2$, but this assumes that the exchange-repulsion itself is a pure exponential, which is not the case.

The asymptotic form of the exchange-repulsion energy has been worked out by Smirnov & Chibisov [89] using the surface-integral approach and later, with a corrected proof, by Andreev [90]. Their result is

$$E^{(1)}_{\text{exch}} = KR^{(7/2\alpha)-1}e^{-2\alpha R} \tag{19}$$

where K is an angular momentum-dependent constant [91]. We observe that:

- The exchange-repulsion energy is not a pure exponential, as is often assumed, but is better represented as an exponential times a function of R. This has been empirically verified by Zemke and Stwalley [92] using spectroscopic data for alkali diatomic molecules. Also, accurate analytic potentials for small van der Waals complexes have tended to use functional forms that include a pre-exponential polynomial term [26–28]. The prefactor function in eq. (19) is not a polynomial, but it is close to linear in R for relevant values of α and R.
- The exchange-repulsion energy has an asymptotic form that is very similar to that of the density overlap, eq. (18), but the prefactor is different. Consequently we should not expect a direct proportionality between the two, and a better form of the density-overlap model might use

$$E^{(1)}_{\text{exch}}(\mathbf{R}) \approx \mathcal{K}(\mathbf{R})S_\rho(\mathbf{R}), \tag{20}$$

  where $\mathcal{K}(\mathbf{R})$ is a low-order polynomial in $\mathbf{R}$.
- The exponents in the asymptotic forms of the density overlap and the exchange–repulsion will be the same only if the wavefunctions used to evaluate them are the same. In general this will not be the case. While the exchange–repulsion could be evaluated with electron correlation effects included, the density-overlap integrals are more typically evaluated using Hartree–Fock densities. Therefore, the α in the exponent of eq. (18) must be replaced by $\alpha_{\text{HF}} = (-2\epsilon_{\text{HOMO}})^{1/2}$, where ϵ_{HOMO} is the energy of the highest occupied molecular orbital from Hartree–Fock theory. In this case, there will be a better agreement between the exchange–repulsion energy and the density overlap if the exponents are made the same by raising the latter by the power $\gamma = (-I/\epsilon_{\text{HOMO}})^{1/2}$ as is done in eq. (15). Now in Hartree–Fock theory $|\epsilon_{\text{HOMO}}| > I$, so γ is always less than unity, and for the helium, neon and argon dimers we obtain values between 0.99 and 0.97 in reasonable agreement with the empirical results of Kim *et al.*.

We will now use these observations to construct models for the short-range energies.

Electron charge densities obtained from density functional theory are exact, in principle. In practice, because of the now well understood self-interaction problem with standard local and semi-local exchange-correlation functionals, they tend to be too diffuse. This can be corrected by applying a suitable asymptotic correction to the exchange-correlation potential [65, 93]. It is now usual to apply this correction in any SAPT(DFT) calculation; without it, even energies that depend on the unperturbed monomer densities, like the electrostatic energy, can be significantly in error. With the asymptotic correction, the asymptotic form of the density given by eq. (17) is enforced, and consequently $\gamma = 1$ in eq. (15).

This has important consequences for multi-atom systems where we use the overlap model to partition $E^{(1)}_{\text{exch}}$ into contributions from pairs of atoms. This idea goes back to the work of Mitchell & Price [85] and begins with a partitioning of the densities into spatially localised contributions that will usually be centered on the atomic locations. If we can write

$$\rho^A(\mathbf{r}) = \sum_a \rho^A_a(\mathbf{r}), \tag{21}$$

where ρ^A_a is the partitioned density centered on (atomic) site a, and likewise for ρ^B, then from eqs. (15) and 16 we get

$$\begin{aligned} E^{(1)}_{\text{exch}}(\mathbf{R}) &\approx \sum_{ab} K \int \rho^A_a(\mathbf{r})\rho^B_b(\mathbf{r})d\mathbf{r} \\ &\approx \sum_{ab} KS^{ab}_\rho(\mathbf{R}), \end{aligned} \tag{22}$$

where S^{ab}_ρ is the site–site density overlap. This expression may be generalised by introducing a site-pair dependence on K as follows:

$$E^{(1)}_{\text{exch}}(\mathbf{R}) \approx \sum_{ab} K_{ab}S^{ab}_\rho(\mathbf{R}) = \sum_{ab} E^{(1)}_{\text{exch}}[ab](\mathbf{R}), \tag{23}$$

where $E^{(1)}_{\text{exch}}[ab]$ is the first-order exchange contribution assigned to site-pair (ab). This is the distributed density overlap model. This is essentially the result obtained by Mitchell & Price but in their case, because of their use of electronic densities from Hartree–Fock theory, they had $\gamma < 1$ and so obtained an expression for the partitioning that is necessarily approximate.

There are a few important issues about the overlap model given in eq. (23):

- The model was originally formulated for the first-order exchange repulsion only, but, as the other short-range energy contributions are also roughly proportional to $E^{(1)}_{\text{exch}}$, we may use the density-overlap model more generally for all of the short-range energy, $E^{(1-\infty)}_{\text{sr}}$. Henceforth we will use the model in this general sense, that is, to model the short-range energy, E_{sr}, however we may choose to define it.

- The model allows us to partition the short-range energy into terms associated with pairs of sites. With this partitioning, we may fit an analytical potential to individual site pairs rather than fit the sum of exponential terms given in eq. (11). The fit to each individual term $V_{\text{sr}}[ab]$ (eq. (5)) is numerically better defined and may be achieved with relative ease.

- This is an approximation: Since the density overlap model cannot exactly model the short-range energy, we have $E_{\text{sr}}(\mathbf{R}) \neq \sum_{a,b} E_{\text{sr}}[ab](\mathbf{R})$. That is, there is a residual error that originates from the original ansatz given in eq. (15).

- Although the residual error is small compared with E_{sr}, it needs to be accounted for to achieve an accurate fit, particularly as the error may be a non-negligible fraction of the total interaction energy, which is generally much smaller in magnitude than E_{sr}. This may be achieved by constrained relaxation of the final short-range potential $V_{\text{sr}} = \sum_{ab} V_{\text{sr}}[ab]$.

VII. ISA-BASED DISTRIBUTED DENSITY OVERLAP

Formally, the distributed density overlap integrals, $S^{ab}_{\rho}(\mathbf{R})$, defined through eqns. (21) and (23), are particularly straightforward to evaluate using the BS-ISA algorithm [47] as this algorithm provides basis-space expansions for the atomic densities $\rho^A_a(\mathbf{r})$. However, basis-set limitations mean that while the BS-ISA algorithm results in fairly well-defined atomic shape-functions, the atomic densities are not well described in the region of the atomic density tails, where the density can even attain small negative values. This not only leads to distributed density overlap integrals that can be negative, but also results in a relatively poor correlation between the first-order exchange energies and the density overlap integrals. This problem may be alleviated using better basis sets for the atomic expansions, but we have not as yet explored this option.

An alternative is to evaluate $S^{ab}_{\rho}(\mathbf{R})$ using the atomic densities defined as

$$\rho^A_a(\mathbf{r}) = \rho^A(\mathbf{r}) \times \frac{\tilde{w}^a(\mathbf{r})}{\sum_{a'} \tilde{w}^{a'}(\mathbf{r})}, \tag{24}$$

where $\tilde{w}^a$ is the tail-corrected shape-function for site a as defined in Ref. 47 as a piece-wise function:

$$\tilde{w}^a(\mathbf{r}) = \begin{cases} w^a(\mathbf{r}) & \text{if } |\mathbf{r}| \leq r^a_0 \\ w^a_{\text{L}}(\mathbf{r}) & \text{otherwise,} \end{cases} \tag{25}$$

where $w^a(\mathbf{r})$ is the atomic shape-function that is the spherical average of atomic density $\rho^A_a(\mathbf{r})$, and the long-range form of the shape-function is defined as $w^a_{\text{L}}(\mathbf{r}) = A_a \exp(-\alpha_a|\mathbf{r} - \mathbf{R}_a|)$, where r^a_0 is a cutoff distance, and the constants in w^a_{L} are defined to enforce continuity and charge-conservation. [47] The shape-functions may be thought of as pro-atomic densities that encode the ionic state of the atom in its molecular environment. This ionic state is not fixed and is instead determined self-consistently through the ISA iterations [46]. While the atomic shape-functions are spherically symmetrical, the atomic densities are not. Now, the distributed density overlap integral is defined as

$$S^{ab}_{\rho}(\mathbf{R}) = \int \left(\rho^A(\mathbf{r}) \frac{\tilde{w}^a(\mathbf{r})}{\sum_{a'} \tilde{w}^{a'}(\mathbf{r})}\right)\left(\rho^B(\mathbf{r}) \frac{\tilde{w}^b(\mathbf{r})}{\sum_{b'} \tilde{w}^{b'}(\mathbf{r})}\right) d\mathbf{r}. \tag{26}$$

Due to the piece-wise nature of $\tilde{w}^a$, this integral must be evaluated numerically using a suitable atom-centered integration grid. Using techniques described by us earlier [47], we evaluate the terms in eq. (26) in $O(N^0)$ computational effort. This is done by defining local neighbourhoods, $\mathcal{N}_a$ and $\mathcal{N}_b$, for sites a and b. These neighbourhoods are based on the dimer configuration, so $\mathcal{N}_a$ may include sites that belong to monomer B, and vice versa for $\mathcal{N}_b$. The neighbourhoods are usually defined using an overlap criterion that naturally takes the basis set used into account with basis sets containing more diffuse functions resulting in larger neighbourhoods. The integration grid, and various terms in the integral S^{ab}_{ρ} are then evaluated using sites in the intersection set $\mathcal{N}_a \cap \mathcal{N}_b$. This intersection set may be null for monomers that are sufficiently far apart. In this manner the density overlap integrals are calculated with linear effort.

VIII. POTENTIALS FOR PYRIDINE

We now apply the methodology presented above to develop a set of many-body potentials for pyridine. In a study such as this is, it is important to use a system that simultaneously presents a challenge and also allows tests to be performed to validate the method sufficiently. We have chosen to use the pyridine dimer as our example as it is small enough to permit accurate interaction energy calculations using SAPT(DFT) on as dense a grid as is needed, but large enough to exhibit a varied and complex potential energy surface (PES) with—as we shall see below—eight distinct minima. Additionally, the pyridine molecule has a sizeable dipole moment and polarizability, so polarization effects are expected to be important,

and, as we shall see, the two-body charge-transfer, or charge-delocalisation [42], energy is also significant. Finally, from the crystallographic studies by Price and co-workers [94] it is known that the crystal energy landscape of this molecule is complex and poses a significant challenge for seemingly accurate empirical potentials. While we will not attempt to use the results of this study in a crystal structure prediction, we intend to perform this test in later work.

IX. SHORT-RANGE FIT

The distributed density-overlap fits were performed using the CamCASP program using the Gaussian/Log weighting scheme [95] in which $w_{\mathrm{GL}}(e) = \exp(-\alpha(\ln(e/E_0))^2)$, where $\alpha = 1/\ln 10$ and $E_0 = 100\ \mathrm{kJ\,mol^{-1}}$. Here the parameter E_0 sets the energy-scale for the fit, and it is usually chosen to be some large multiple of the absolute global minimum dimer energy so as to obtain a reliable fit to the repulsive wall. The fits to individual site–site potentials $V_{\mathrm{sr}}[ab]$ were performed with the Orient program using the same Gaussian/Log weighting scheme.

All relaxation steps were performed using the Orient program using the Boltzmann weighting function

$$w_{\mathrm{Bol}}(e) = \begin{cases} \exp\left((e_{\mathrm{low}} - e)/E_0\right) & \text{for } e > e_{\mathrm{low}} \\ 1.0 & \text{otherwise.} \end{cases} \tag{27}$$

Here e_{low} is typically set to the smallest energy in the data set and the energy-scale for the fit is set by $E_0 = 40\ \mathrm{kJ\,mol^{-1}}$ to increase the weight to lower energies. We used $e_{\mathrm{low}} = 0\ \mathrm{kJ\,mol^{-1}}$ for the relaxation of the repulsive energies, and $-10\ \mathrm{kJ\,mol^{-1}}$ in the final relaxation step involving the total interaction model.

A. Fitting strategy and atomic shape

We set out the fitting strategy for the short-range part of the potential in some detail in §III above. In this multi-stage approach we first fit to $E_{\mathrm{sr}}^{(1)}$ calculated on the dense, pseudo-random set of 3515 dimers in Dataset(0). This is done via the distributed density-overlap model which allows us to partition $E_{\mathrm{sr}}^{(1)}$ into contributions from pairs of sites, and fit the terms in the potential for each atom-pair individually. However, if the atoms are close to spherical, as is the case for the ISA atom densities, the atom-pair shape function $\rho_{ab}(\Omega_{ab})$ that appears in the potential (see eq. (5)) may be written to a good approximation as the sum of shape functions for the interacting atoms (see ch. 12 in ref. 20)

$$\rho_{ab}(\Omega_{ab}) \approx \rho_a(\Omega_a) + \rho_b(\Omega_b). \tag{28}$$

Here Ω_a is a generalised angular coordinate that describes the direction of the vector from site a to site b in the local coordinate system of site a, and likewise for Ω_b, and ρ_a and ρ_b are the atomic shape functions for atoms a and b. The atomic shape functions for all atoms of a given type should be the same.

The shape-function additivity is observed in the first stage of the fitting when the terms in $V_{\mathrm{sr}}^{(1)}[ab]$ are fitted individually via the density-overlap model, but it is not exact, probably in part because of grid sampling variability around the sites. It can however be exactly enforced in the next stage when the short-range parameters are collectively relaxed in a constrained manner to the $E_{\mathrm{sr}}^{(1)}$ energies in Dataset(1). We find it best to perform this relaxation iteratively, with only those parameters associated with a particular subset of sites relaxed at each step. With this approach the constrained relaxation can be performed rapidly, in a computationally efficient manner. At each step, shape-function additivity is imposed by using pinning (prior) values for the parameters from the averaged shape-function parameters from the previous step.

In a similar manner, we may relax the resulting potential parameters to include effects from second and higher orders in the interaction operator. However, there is no reason to expect the shape-function additivity to hold at this stage, as the higher-order short-range effect, which is predominantly the charge-transfer (or charge-delocalisation) energy, depends on the pair of atoms involved in a non-additive manner. In the absence of additivity, the number of independent parameters in the potential would depend quadratically on the number of interacting atoms, but fortunately, as we will demonstrate below, the higher-order correction can be treated as isotropic. That is, the atom-pair shape function now becomes

$$\rho_{ab}(\Omega_{ab}) = \rho_a(\Omega_a) + \rho_b(\Omega_b) - \delta_{ab}, \tag{29}$$

where δ_{ab} is the isotropic higher-order correction.

We will now examine the effectiveness of this strategy in obtaining a series of fits to the short-range potential for the pyridine dimer.

B. Fitting using the distributed density overlap model

In principle, it is straightforward to use the distributed density-overlap model described above. We have used this approach[31, 32], as have others [85, 87, 95, 96], with a reasonable degree of success. The problem lies in the choice of density partitioning method. There is no unique way of decomposing a density into atom-like domains, yet the tacit assumption of the distributed density-overlap model is that the partitioned density ρ_a^A is well-behaved and may be used to extract properties such as size and shape of the atom located on site a. If this were not the case, then the potential parameters extracted from the model would be meaningless, and indeed, a fit to eq. (5) could even be so poor as to be useless. In the past we have used a density-fitting-based scheme to partition the density [31]. This works by expressing the electronic density as a single sum over an auxiliary basis set with functions located on the atomic nuclei, which then naturally suggests a partitioning scheme:

$$\begin{aligned} \rho(\mathbf{r}) &= \sum_k d_k \chi_k(\mathbf{r}) \\ &= \sum_a \sum_{k \in a} d_k \chi_k(\mathbf{r}) = \sum_a \rho_a(\mathbf{r}). \end{aligned} \tag{30}$$

Here the d_k are expansion coefficients and χ_k are Gaussian basis functions from the auxiliary basis. We have previously argued [48] that since the auxiliary basis sets are optimised on free atoms, or homo-diatoms, they may be used in the above manner to partition the molecular density into atom-like parts. This does seem to work, but only if small enough auxiliary basis sets are used, and even then, the resulting atomic domains may be meaningless.

In Figure 6 we present the density-fitting-based (DF-based) atomic isodensity surfaces for the atoms in the pyridine molecule. The total electronic density of pyridine was obtained with the d-aug-cc-pVTZ basis using the PBE0/AC functional. We had to use the relatively less diffuse def-TZVPP basis for the density-fitting as results with any of the more diffuse RIMP2 auxiliary basis sets were so full of artifacts associated with the basis set over-completeness as to lead to completely nonsensical results for the density partitioning. However, even with the relatively small def-TZVPP basis, the DF-based density partitioning results in carbon atoms with rather unusual shapes. If this partitioning method is used to construct a short-range potential using the density-overlap model as described above, we obtain potentials with spurious terms in the atomic anisotropies and overall very poor fit qualities.

In contrast, we can see in Figure 7 that the ISA-based atomic shapes obtained using the algorithm described in §VII are very well-behaved. These have been obtained with the significantly larger aug-cc-pVQZ/ISA-set2 fitting basis and show none of the artifacts seen with the DF-based scheme. Additionally, the ISA-based atoms do not show any significant differences in shape when other basis sets are used, as long as these are large and diffuse enough. This is a significant result: if we wish the atomic shapes to be, in some sense, universal or transferable (properties we will not explore in this paper), we must be able to calculate the atomic shapes with an algorithm that possesses a well-defined basis-set limit. The ISA approach is not the only such method, but for reasons discussed in the Introduction and in ref. 47, it is one of the few partitioning methods that has desirable numerical properties while satisfying physical and chemical expectations.

In Figure 8 we present the ISA-atomic shapes viewed in the molecular plane, along the bond axis, or, in the case of the nitrogen atom, along the N···C3 axis. In order to highlight the atomic anisotropies we have superimposed on the 10^{-3} a.u. isodensity surfaces some contours showing the intersection with spheres centred on the atomic nuclei. These contours clearly illustrate the shape symmetries of each of the atoms. Also included in the figure are the important shape anisotropies for these atoms. These have been calculated by fitting $E_{\mathrm{sr}}^{(1)}$ via the distributed overlap model using a set of local axis frames located on the atomic centres with the x axis pointing along and out of the bond, and the z axis perpendicular to and pointing out of the plane of the molecule. During the relaxation step in this fit we eliminate all terms that are less than a threshold, taken to be 0.01 a.u. The picture that emerges is remarkably simple and convincing:

- *Nitrogen*: The largest anisotropy term for the nitrogen atom in pyridine is the 22c term that is associated with

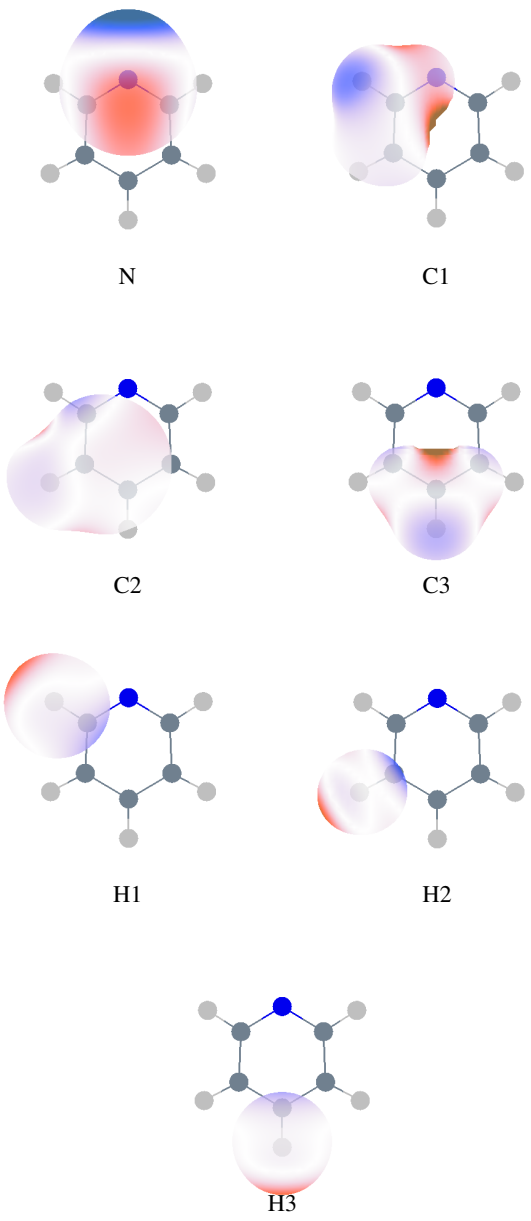

FIG. 6: The 10^{-3} a.u. iso-density surfaces of the density-fitting-based 'atoms' in pyridine. The pyridine density was computed using a d-aug-cc-pVTZ basis, and the density-fitting was performed using the TZVPP auxiliary basis. The colour coding indicates the anisotropic component of the electrostatic potential on the surface arising from ISA-based atomic multipoles located on the nuclei; that is, the atomic charge contributions are not included. The scale used varies from −0.5 V (blue), through 0 V (white), to +0.5 V (red).

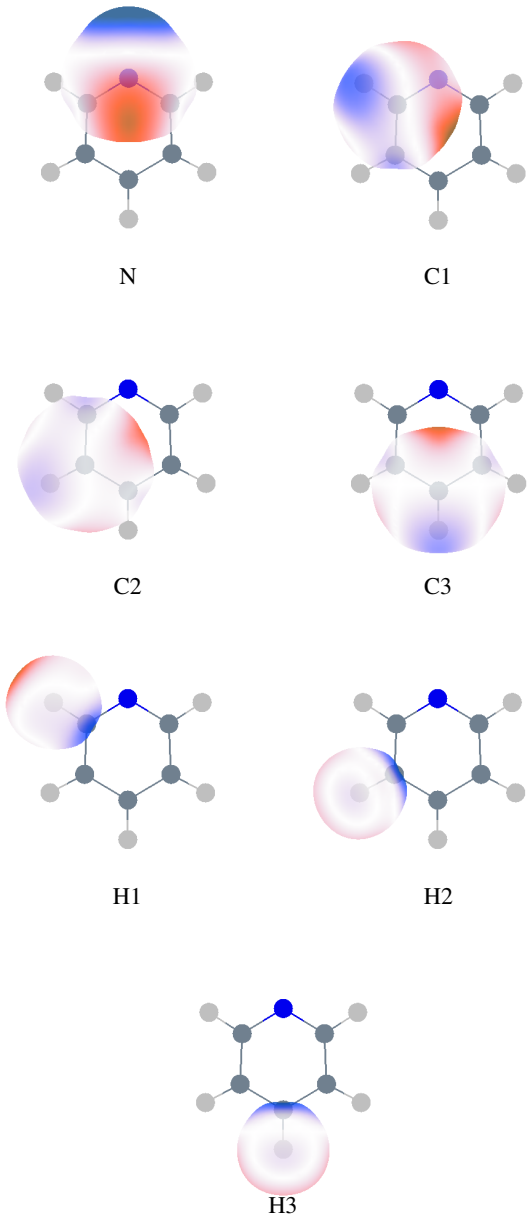

FIG. 7: The 10^{-3} a.u. iso-density surfaces of the ISA-based 'atoms' in pyridine. The pyridine density was computed using a d-aug-cc-pVTZ basis and the ISA calculations were performed using the aug-cc-pVQZ/ISA-set2 auxiliary basis set. Colour coding as described in Figure 6.

the lone pair. Additionally one may include the 11c and 20 terms on the nitrogen atoms, though these are smaller. All other terms are negligible.

- *Carbon*: The 20 term associated with the p_z orbitals is the dominant source of anisotropy on all carbon atoms. Of the other symmetry-allowed terms, the 11c term associated with the C–H bond is relatively strong. The 22c terms are present, but small. Finally, C1 and C2 contain 11s terms due to the proximity of the N atom. These terms describe the in-plane distortion of the C1/C2 densities due to N.

- *Hydrogen*: We have limited all hydrogen atoms to rank 1 terms only. All hydrogen atoms possess a 11c term to describe the electronic distortion along the C–H bond and, both H1 and H2 additionally have 11s terms.

We have developed three models for the short-range terms: srModel(1) contains only isotropic terms, in srModel(2) we have included the 22c anisotropy term on the nitrogen atom, and in srModel(3) we have used all the anisotropy terms shown in Figure 8. In all three models, the hardness parameters α_{ab} in eq. (5) were kept isotropic. The constrained relaxation was performed using eq. (12) with constraint strength parameters c_i chosen to be 0.1 for the isotropic parameters and 1.0 for the anisotropic terms in the $\rho_{ab}(\Omega)$ expansions. This choice was made empirically on the basis that the appropriate parameters were those that when further reduced did not result in any appreciable improvement in the fit quality. The distributed density-overlap fits were performed using the CamCASP program, and the fits to individual site–site potentials $V_{\rm sr}[ab]$ were performed using the Orient program. The weighting schemes used in these fits are described in §IV. The relaxation step was also performed using the Orient program but this time using the Boltzmann weighting function as described in §IV. The scatter plots of these models at various stages in the fitting process are shown in Figure 9. Weighted r.m.s. errors at the final stage are 1.03, 0.90, and 0.61 kJ mol^{-1} for models 1, 2 and 3, respectively. These uncertainties are less than our target of 1 kJ mol^{-1} for all three models, but the performance of srModel(3) is quite remarkable, with errors less than or close to 1 kJ mol^{-1} for energies as large as 100 kJ mol^{-1}.

C. Infinite-order charge transfer (delocalisation) energy

The infinite-order charge-transfer energy is the dominant short-range contribution at second and higher orders in the intermolecular interaction operator. While we can use regularised SAPT(DFT) [42, 75] to determine the second-order charge-transfer energy, the contributions from higher orders cannot, at present, be computed within the SAPT framework. Unfortunately, where charge-transfer is important, these higher-order effects appear to be too large to be ignored, so we need to account for them, if only approximately. As it turns out, the discussion of the infinite-order polarization in §V B readily suggests an approximation. If we argue that the

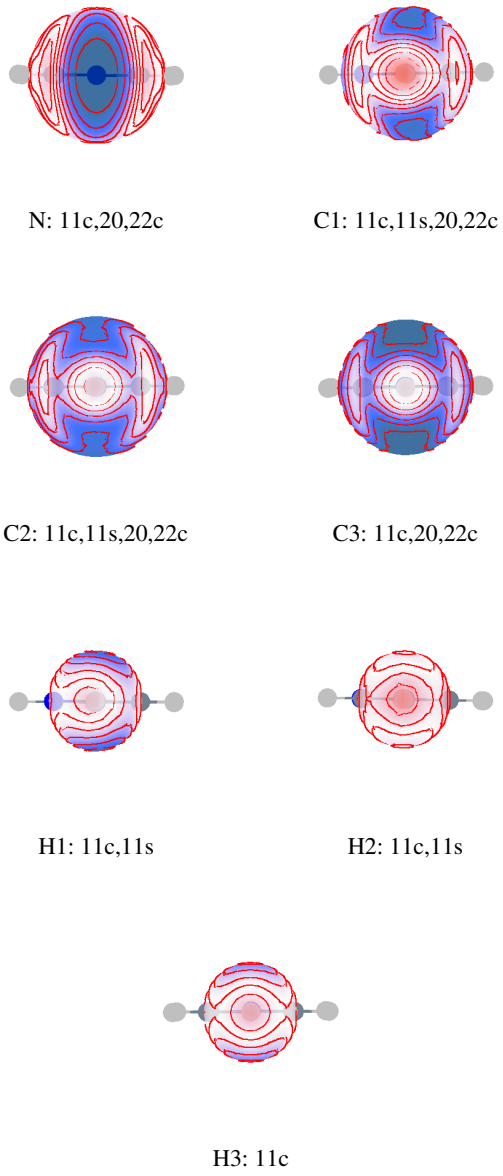

FIG. 8: Along-the-bond views of the ISA-based 'atoms' in pyridine. Here we illustrate the anisotropy of the atom shapes by contours showing the intersections with spherical surfaces centred at the atomic nuclei. The dominant anisotropy terms for each atom are listed for local axis frameworks with the x axis pointing out of the bond (out of the page) and the z axis normal to the plane of the molecule.

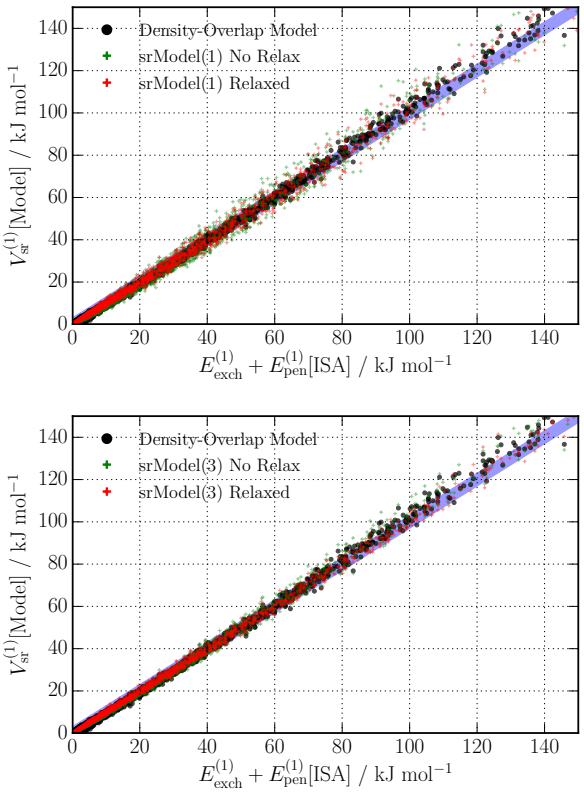

FIG. 9: Performance of two of the short-range models fitted to $E_{sr}^{(1)}$. srModel(1) is fully isotropic and srModel(3) contains the anisotropy terms described in the text and indicated in Figure 8. srModel(2) results are only slightly better than those from srModel(1) and are not shown. The black circles are results directly from the distributed density-overlap model; the green plus signs are data obtained from the model fitted to eq. (5) before relaxation, and the red plus signs are the same after relaxation to $E_{sr}^{(1)}$. The blue bar represents the ±1 kJ mol^{-1} range.

infinite-order induction energy is the sum of just the infinite-order charge-transfer and polarization terms (i.e., assuming that there are no cross terms present), then if we know any two, we can compute the third. Here we approximate the infinite-order induction energy as:

$$E_{\rm IND}^{(2-\infty)} \approx E_{\rm IND}^{(2)} + \delta_{\rm int}^{\rm HF} \tag{31}$$

and define the two-body infinite-order charge-transfer energy to be

$$\begin{aligned} E_{\rm CT}^{(2-\infty)} &= E_{\rm IND}^{(2-\infty)} - E_{\rm POL}^{(2-\infty)} \\ &\approx E_{\rm IND}^{(2)} + \delta_{\rm int}^{\rm HF} - V_{\rm pol}^{(2-\infty)}[\rm DM]. \end{aligned} \tag{32}$$

While this expression is readily implemented, it has a drawback in that the definition depends on the type of polarization model used.

In Figure 10 we have plotted the infinite-order charge-transfer energy calculated using eq. (32) against the first-order short-range energy $E_{sr}^{(1)}$. First of all, at about 20% of $E_{sr}^{(1)}$,

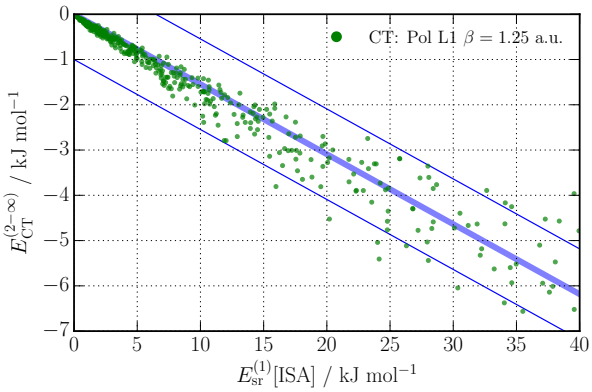

FIG. 10: The infinite-order charge delocalisation (charge-transfer) energy plotted against the first-order short-range energy $E_{\rm sr}^{(1)}$. The thin blue lines represent the ±1 kJ mol^{-1} limits.

$E_{\rm CT}^{(2-\infty)}$ is a significant contribution to the short-range energy and it cannot be ignored. Second, while these two energies are roughly proportional, there is a significant scatter, particularly at the larger charge-transfer energies. Nevertheless, the scatter is rarely more than ±1 kJ mol^{-1}. If we argue that the charge-transfer contribution to the intermolecular interaction energy arises from a tunneling process [42], then it is natural to assume that the tunneling probability will be roughly proportional to the electron density overlap, but further work needs to be done to see whether this holds for other systems.

We may include $E_{\rm CT}^{(2-\infty)}$ into our models for the short-range energy by constrained relaxation of the parameters in the models already obtained for $E_{\rm sr}^{(1)}$, or we may exploit the approximate proportionality of $E_{\rm sr}^{(1)}$ and $E_{\rm CT}^{(2-\infty)}$ and absorb the bulk of the charge-transfer effects by scaling as follows. If we assume a proportionality with constant $k < 0$:

$$E_{\rm CT}^{(2-\infty)} \approx k E_{\rm sr}^{(1)} \approx k V_{\rm sr}^{(1)}, \tag{33}$$

then we can include $E_{\rm CT}^{(2-\infty)}$ into the short-range energy model by scaling it by $(1-k)$ yielding

$$\begin{aligned} V_{\rm sr}^{(1-\infty)} &\approx (1-k) V_{\rm sr}^{(1)} \\ &\approx (1-k) \sum_{a,b} G \exp\left[-\alpha_{ab}(r_{ab} - \rho_{ab}(\Omega_{ab}))\right], \end{aligned} \tag{34}$$

then, re-writing $1 - k = \exp\left[-\alpha_{ab}\delta_{ab}\right]$, where $\delta_{ab} = -\ln(1-k)/\alpha_{ab}$, we get

$$V_{\rm sr}^{(1-\infty)} = \sum_{a,b} G \exp\left[-\alpha_{ab}(r_{ab} - (\rho_{ab}(\Omega_{ab}) - \delta_{ab}))\right]. \tag{35}$$

That is, the isotropic atom-pair radii are reduced by δ_{ab} by the attractive effects of the charge delocalisation process. The atom-pair shape-function $\rho_{ab}(\Omega_{ab})$ remains additive in the sense of eq. (28), but there is an isotropic non-additive correction δ_{ab}, as shown in eq. (29).

For the pyridine dimer we get $k \approx 0.16$ (it varies slightly with the type of polarization model used). Therefore the pair-radius reduction is of the order 0.05 Bohr, which is small but not negligible as it leads to an overall reduction in the intermolecular separation of a few tenths of a Bohr in some dimer orientations. These effects may be expected to be larger in more strongly hydrogen-bonded systems where the charge-delocalisation is stronger.

The above scaling absorbs the bulk of the charge-transfer energy into our short-range energy models. The remainder may be included in a subsequent relaxation step, but we find that this is not necessary as it is usually small, and in any case, this and all other errors against the SAPT(DFT) reference energies will be accounted for in the final relaxation stage that we describe next.

X. TOTAL ENERGY FITS: COMBINING THE TERMS

The analytic fits to the various components of the total interaction energy model may be combined as appropriate, and optionally relaxed, using constraints, to the total SAPT(DFT) interaction energies calculated for a suitable set of dimer geometries. These models have been obtained with a significant amount of data derived directly from the density and transition densities using various partitioning methods. The limited amount of fitting has been largely restricted to the short-range energy model, and even here, our approach ensures that the parameters are well-defined and physically meaningful, with little of the uncertainty usually associated with fits to sums of exponentials. Further, the target residual error for each of the models has been 0.5 to 1 kJ mol^{-1}, and we have largely succeeded in achieving this target. Consequently, as we shall see, these models may be combined without further relaxation to produce reasonably accurate models for the total interaction energy.

In this paper, we have reported the following models:

- *Short-range*: Three models have been obtained. srModel(1) is fully isotropic; srModel(2) contains a 22c anisotropy term on the nitrogen atoms; and srModel(3) contains all the dominant anisotropy terms needed. These short-range energy models include the first-order exchange, the electrostatic penetration, and infinite-order charge-transfer energies.

- *Electrostatic*: A rank 4 ISA-based distributed multipole model.

- *Polarization*: Three distributed polarization models obtained from the WSM procedure. The L1(iso) and L1 models include rank 1 polarizabilities, with the former being isotropic, and the L2 model includes terms to rank 2 on the heavy atoms. All these models are damped. The many-body contributions are obtained through the polarization models. We will consider only the L1 model in this paper.

- *Dispersion*: Two damped isotropic dispersion models have been obtained. The C_6(iso) model contains only (scaled) isotropic C_6 coefficients for all pairs of atoms. And the C_{12}(iso) model consists of isotropic terms to

C_{12} between pairs of heavy atoms, isotropic terms to C_{10} between any hydrogen atom and a heavy atom, and only isotropic C_6 terms between pairs of hydrogen atoms. As the C_{12} terms in the C_{12}(iso) are found to have a minimal effect on the quality of the model, we will instead use the equivalent C_{10}(iso) in the remainder of this work. All models are damped. At present we do not include any three-body dispersion non-additivity.

This gives us 18 possible ways of combining these models into total interaction energy potentials. Of these, we explore three combinations in this paper:

- *Model(1)*: Isotropic short-range model, with rank 4 ISA-DMA, L1 polarizability model, and C_{10}(iso) dispersion model.
- *Model(2)*: Short-range model containing isotropic terms on all atoms and an additional 22c term on the nitrogen atoms, with rank 4 ISA-DMA, L1 polarizability model, and C_{10}(iso) dispersion model.
- *Model(3)*: Anisotropic short-range model, combined with rank 4 ISA-DMA, L1 polarizability model, and C_{10}(iso) dispersion model.

These models differ only in their description of the short-range repulsion.

In Table I we report r.m.s. errors made by these models before relaxation against the SAPT(DFT) interaction energies. The r.m.s. errors are remarkably small at this stage, with models (1) and (3) exhibiting errors less than 1 kJ mol^{-1} for the most energetically important dimers. Surprisingly, Model(2) fares slightly worse than the simpler Model(1) with r.m.s. errors of 1.5 kJ mol^{-1} in this energy range. All models fare reasonably well for the positive energy dimers, with r.m.s. errors between 1.8 to 2.9 kJ mol^{-1}.

The models may be improved by constrained relaxation to SAPT(DFT) total interaction energies. We initially relaxed the models against energies from the random dimers in Dataset(1), but this led to a reduction in the quality of the fits for the test set of low-energy dimers. It appears that while the random dimers are suitable for an unbiased parametrization of the individual components of the model, they are not suitable for relaxing the sum of these components. The principal reason for this seems to be that the random dimer set does not contain low-energy dimers, as can be seen in Figure 11. Consequently, relaxing to this set causes the models to represent these relatively high-energy dimers better at the cost of the more physically important low-energy configurations. Because of this, we have performed the relaxation of the models using both Dataset(1) and Dataset(2).

The constrained relaxation was performed using the ORIENT program with the weighting scheme described in §IV. Constraints were imposed using eq. (12) with tight constraint strength parameters c_i chosen to be 1.0 for the isotropic parameters and the C_8 and C_{10} parameters. The C_6 terms were kept unaltered so as to preserve the long-range dispersion interaction. The anisotropy parameters were not allowed to vary. Rather than relax all parameters simultaneously, the relaxation was performed in stages, with parameters associated with particular sites allowed to vary in each stage. This procedure, though computationally efficient, needed to be iterated to ensure that the relaxation was adequate.

In Table I we also report r.m.s. errors made by the relaxed models. After relaxation, all three models show r.m.s. errors of only 0.5 to 0.6 kJ mol^{-1} for the most strongly bound dimers, and somewhat larger errors for the higher energy dimers. Perhaps unsurprisingly, Model(3) fares best, with r.m.s. errors less than 1 kJ mol^{-1} for all dimers with energies less than or equal to 20 kJ mol^{-1}.

In Figure 11 we display scatter plots of the interaction energies calculated with Model(3) against SAPT(DFT) energies both before and after relaxation. The excellent performance of the unrelaxed Model(3) is evident. At no stage in the development of Model(3) were the total interaction energies from Dataset(1) included; rather we only used the charge-transfer energies in the development of this model. Additionally, none of the low-energy dimers in Dataset(2) were used in any way in the construction of Model(3), yet these energies are accurately predicted by the unrelaxed Model(3), with very few outliers. This model may be improved by relaxing it to the dimer energies in both data sets. This relaxation was performed with the anisotropic terms in the potential frozen and only the isotropic parameters, including the low-ranking dispersion coefficient, allowed to vary with tight anchors imposed (see the SI for additional information). As seen in Figure 11, this relaxed model exhibits an excellent correlation with the SAPT(DFT) reference energies, and has fewer low energy outliers compared with the unrelaxed model. In the remainder of this paper by 'Model(3)' we will refer to this relaxed model.

Similar figures for Model(1) and Model(2) can be found in the SI. As may be expected from the r.m.s. errors reported in Table I, the performance of the unrelaxed Model(1) is excellent given the simplicity of the model, but the unrelaxed Model(2) shows somewhat larger errors for the most strongly bound dimers. However, both of these models improve considerably on relaxation.

The quality of the relaxed potentials can be assessed using Dataset(3) which was not used at any point in the model development process. In Table I we report r.m.s. errors made by the models (before and after relaxation) against Dataset(3). We see that the errors made are largely in accordance with those made against Dataset(1) and Dataset(2), indicating that the models are predictive. In particular, the good performance of the un-relaxed models, particularly for the un-relaxed Model(3), suggests that the algorithm we have described is indeed robust. The comparison with Dataset(3) indicates that the relaxed models show somewhat larger inaccuracies in the repulsive configurations, particularly when compared to the coresponding un-relaxed models. However, these errors are not very large, and may be an acceptable price to pay for the increased accuracy in the low-energy dimer configurations.

In Table I we also report r.m.s. errors for a model functionally identical to Model(3), but created using the DF-AIM approach and with DMA4 multipoles. Apart from these two differences, this model, termed Model(3)-DF-DMA4, has been

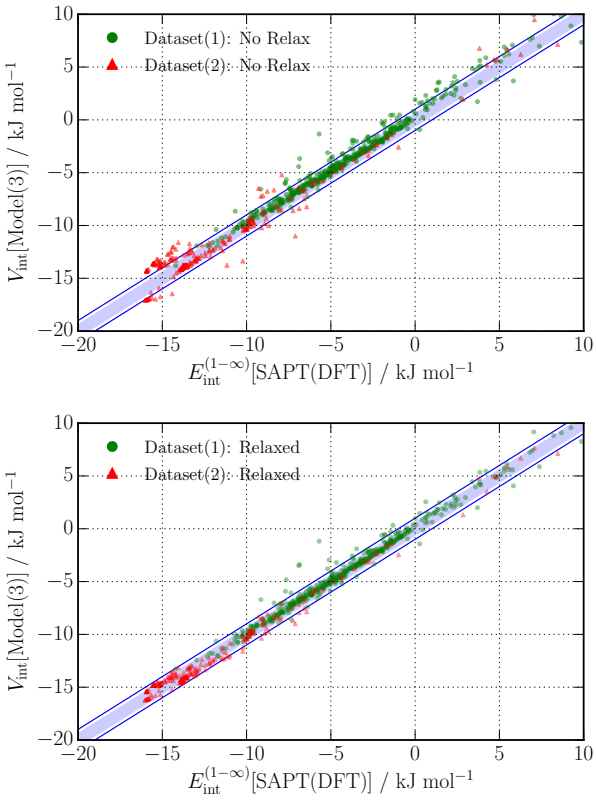

FIG. 11: The total interaction energy models for Model(3). The upper panel shows energies from Model(3) before relaxation to the dimers in Dataset(1) and Dataset(2), and the lower panel shows model energies after relaxation. In both cases these energies are plotted against the total SAPT(DFT) interaction energy $E_{\rm int}^{(1-\infty)}$. The blue bar represents the ± 1 kJ mol^{-1} deviation from SAPT(DFT).

created in an identical manner to the others reported in this paper. This is the kind of model that might have been created using the approach we have described in an earlier paper on atom–atom potentials[40]. We see that across the $-20:20$ kJ mol^{-1} energy range the r.m.s. errors made by this model are twice as large as those from Model(3). This is mainly a consequence of the unphysical AIM atoms that result from the DF-AIM approach that are shown in Figure 6. This approach results in the wrong atomic anisotropies that the fit cannot correct with the limited amount of SAPT(DFT) data in Datasets (1) and (2). This is an inevitable consequence of the Bayes-like approach we have adopted: the role of the first step in the fitting process — the first-order fits through the distributed density overlap model — is to determine prior values for the fitting parameters (see §III). The subsequent relaxation steps merely refine these prior values. However, if the prior values are very poor, as they are with the DF-AIM approach, then we require a considerable amount of data to move them to the correct values. This is not needed with the ISA-based AIM approach, and demonstrates the superiority of this method.

XI. RESULTS

A. Minima

We have used the basin-hopping algorithm (see Ref. 97 for a review) as implemented in the ORIENT program to search for stable dimers on the potential energy surfaces. In contrast to the rather simple PES of the benzene dimer [30, 32] which supports only three minima, we have found eight minima for the pyridine dimer. The minimum-energy structures, which are illustrated in Figure 13, may be classified according to their bonding:

- Hydrogen-bonded: These include Hb1, Hb2 and Hb3. Of these, Hb1 is doubly hydrogen-bonded and has been found in a DFT-D (BLYP+Grimme D1 correction) search [98] and has also been investigated at the CCSD(T)/CBS level of theory [99] to be around -15.5 kJ mol^{-1} (estimated from Figure 5 in Ref. 99). This compares well with our SAPT(DFT) energy of -16.6 kJ mol^{-1}. The Hb2 and Hb3 structures do not appear to have been reported in prior literature.

- Stacked: The S1 and S2 minima are the stacked dimers which are largely dispersion-bound. Both these structures have been found in the DFT-D search, however we see no evidence of the two other stacked structures reported in that study.

- T-shaped: None of these minimum energy dimers are exactly T-shaped, but the T1 and T2 minima are nearly so, and the bT minimum is a very bent-T-shaped structure. The bT structure is similar to one of the T-shaped structures found in the DFT+D search. We do not find the 'T-shaped 1' structure in the DFT+D search by Piacenza and Grimme [98].

The minimum configurations are displayed in Figure 13 and their energies are reported in Table II and Figure 14. For comparison, we have calculated SAPT(DFT) interaction energies for the dimer configurations obtained from the relaxed Model(3) PES. Not all of the models support all the minima. Model(2) does not support the Hb3 minimum, which instead relaxes to the T1 structure on this model PES. The relaxed Model(3)-DF-DMA4 supports only five of the eight minima, and two of those (S2 and T1) differ in structure from the corresponding structures on the ISA-based surfaces: in the S2 structure on this surface the molecules are not parallel, and the T1 is bent. The three missing structures relax to either the Hb2 or the T1 structures. The Hb2 minimum is the global energy minimum on this PES.

For Model(3) we have reported energies for the minima on both the unrelaxed and relaxed models. These largest energy differences in the minima on these two PESs differ by just over 1.1 kJ mol^{-1} (just over 7% of the interaction energy). This is a remarkable result as it indicates that the unrelaxed models can be predictive without the need for fitting to the SAPT(DFT) total interaction energies, in particular, no information about total interaction energies of the stable, low-energy dimers was

TABLE I: R.m.s. errors (kJ mol^{-1}) for the total interaction energy models for the pyridine dimer. Errors are calculated against SAPT(DFT) total interaction energies, and are reported both for the models relaxed to the set of SAPT(DFT) energies in Dataset(1) and Dataset(2), and for the models obtained by combining the different terms in the potential as described in the text. Additionally we report r.m.s. errors agains Dataset(3) which was not used at any stage in the potential development process. The errors for these unrelaxed models are reported in parantheses. Model(3)-DF-DMA denotes a model functionally similar to Model(3) but created using the DF-AIM approach with multipoles from the DMA4 model.

Energy range	Model(1)	Model(2)	Model(3)	Model(3)-DF-DMA4
R.m.s. errors against Dataset(1) and Dataset(2):				
$E \leq -10$	0.59 (1.26)	0.59 (1.22)	0.53 (1.08)	0.97 (1.85)
$-10 < E \leq 0$	0.80 (0.99)	0.72 (0.95)	0.56 (0.70)	1.21 (1.39)
$0 < E \leq 20$	1.69 (2.71)	1.19 (2.58)	0.95 (1.53)	2.17 (3.16)
R.m.s. errors against Dataset(3):				
$E \leq -10$	0.75 (0.66)	0.57 (0.51)	0.33 (0.45)	1.22 (1.53)
$-10 < E \leq 0$	0.65 (0.69)	0.61 (0.65)	0.37 (0.42)	0.87 (0.93)
$0 < E \leq 20$	1.89 (1.58)	1.76 (1.54)	1.66 (1.23)	2.81 (2.59)

used in creating the three unrelaxed models. Further, the similarity of the relaxed and unrelaxed models suggests that the procedure used here appears to be free of artifacts usually introduced by fitting procedures, and is robust to the inclusion of additional data. However this data needs to be biased to low energy dimers, as has been noted above. We will explore this issue in a forthcoming paper [52].

The agreement between the ISA-based models (relaxed and unrelaxed) is made even clearer in Figure 12 where we display PES sections at representative minima. The agreement between SAPT(DFT) and all models — including the unrelaxed Model(3) — for the minima is generally very good, both in the overall shape of the PESs and the location and depth of the radial minimum. Plots for the remainder of the minima can be found in the SI.

In Table III we report the lowest harmonic vibrational frequencies at these minima. These frequencies give us an indication of how different the shapes of the three PESs are at the stable minima configurations. There is generally a good agreement between the minima on all ISA-based models, but the frequencies seem to vary more with the models than the corresponding energies. This may reflect the importance of the anisotropy in determining the shape of the PES. This agreement, though imperfect, is reassuring as it gives us some confidence that the minima we observe are real and not artifacts of the fitting function used. The largest differences are between the ISA-based models and the DF-based Model(3)-DF-DMA4. The lowest vibrational frequencies of the Hb1 and S1 minima are only half as large as the corresponding frequencies for Model(3), indicating that the shape of the PES of Model(3)-DF-DMA4 differs from that of Model(3) in the regions of these minima. This should not be a surprise given the rather significant differences in the AIM shapes from the ISA- and DF-based density partitioning schemes as shown in Figures 6 and 7.

B. Second virial coefficients

The second pressure virial coefficient $B(T)$ represents a necessary, but not sufficient, test of the quality of the two-body PES. As the virial coefficients average over the PES, it is possible to construct an infinity of PESs that yield the correct values of $B(T)$ in a finite temperature range. Nevertheless, it is important that any model PES reproduces the experimental values as a minimum requirement. In Figure 15 we display second virial coefficients calculated for the pyridine dimer. We have calculated $B(T)$ at a range of temperatures using the Orient program. Only the Classical results are presented as the quantum corrections were found to be insignificant over the range of temperatures reported here. We used a stochastic integration sampling algorithm with 102 radial steps and 262,144 dimer orientations in order to integrate $B(T)$ sufficiently accurately. From Figure 15 we see that all three models show good agreement with the experimental data of Andon *et al.* [101] and Cox & Andon [102]. As the models all slightly overestimate $B(T)$ across the temperature range of the figure, they may, on the whole, be somewhat too attractive. We will return to this issue later in this paper.

XII. ANALYSIS & DISCUSSION

A. Polarization damping revisited

In developing the damping model for our polarizability models in §V B we recognised an uncertainty in our choice for damping model. This arose because the damping parameter β_{pol} depends on the choice of dimer configurations used to determine it. Here we re-examine this issue by assessing the damping models against data obtained at the eight minimum energy dimer orientations at various separations. In Figure 16 we compare the second-order polarization energies from the polarization models described in §V B with second-order polarization energies from regularised SAPT(DFT), $E^{(2)}_{\text{POL}}$, It

TABLE II: Interaction energies (kJ mol^{-1}) of the pyridine dimers at the energy minima reported in Figure 13. The SAPT(DFT) reference energies have been calculated at the dimer geometries obtained on the relaxed Model(3) PES. The energies reported for all models are for the stationary points on the model PES, therefore the dimer geometries at which the energies are evaluated will depend on the model and will differ from the geometries used to obtain the SAPT(DFT) reference energies. Where a structure is not supported as a minimum we report in parentheses the structure it relaxes into. Thus Model(2) does not support the Hb3 structure which instead relaxes to the T1 minimum on this PES. Structures on the Model(3)-DF-DMA4 surfaces that are only approximately the same as those on the other surfaces are indicated by an asterisk.

Minimum	SAPT(DFT)	Model(1)	Model(2)	Model(3)		Model(3)-DF-DMA4	
		Relaxed	Relaxed	No Relax	Relaxed	No Relax	Relaxed
Hb1	−16.67	−16.11	−16.00	−17.28	−16.37	−14.38	−15.04
S1	−16.22	−15.64	−15.55	−14.54	−15.61	−13.60*	−15.46
S2	−15.45	−15.38	−15.38	−14.17	−15.35	−12.71*	−14.42*
T1	−14.57	−14.54	−14.73	−14.65	−15.02	−14.63	−14.84*
T2	−14.70	−14.54	−14.69	−14.68	−14.92	(Hb2)	(Hb2)
Hb2	−14.70	−15.03	−14.65	−14.57	−14.76	−15.19	−15.61
bT	−14.01	−14.00	−14.12	−13.97	−14.25	(Hb2)	(Hb2)
Hb3	−13.84	−14.60	(T1)	−13.88	−14.08	−14.00	(T1*)

TABLE III: Lowest harmonic vibrational frequencies for the minima on the relaxed model PESs. For Model(3) we also include data for the unrelaxed version of this model. Model(2) does not support the Hb3 minimum. All frequencies are reported in cm^{-1}.

Minimum	Model(1)	Model(2)	Model(3)		Model(3)-DF-DMA4	
	Relaxed	Relaxed	No Relax	Relaxed	No Relax	Relaxed
Hb1	15.96	12.76	15.79	15.08	7.01	7.70
S1	6.04	3.83	4.79	6.69	1.94	3.42
S2	9.99	10.53	8.98	11.24	9.60	11.81
T1	3.74	4.45	9.57	6.62	5.37	8.22
T2	1.94	3.89	7.38	6.07	—	—
Hb2	12.05	9.94	12.15	12.19	11.92	10.91
bT	5.55	7.43	3.13	6.28	—	—
Hb3	12.07	—	11.36	10.88	6.88	—

should be apparent that while our choices for the damping models are reasonable, with errors typically less than 1 kJ mol^{-1} for the attractive dimers, there is a systematic overdamping, with the polarization energies of some (repulsive energy) dimers underestimated by as much as 2.5 kJ mol^{-1}. This problem can be largely remedied by increasing the value of β_{pol}. In the same figure we also display polarization energies calculated with the anisotropic L2 polarization model with $\beta_{\mathrm{pol}} = 1.0$ a.u. This small increase causes a significant improvement to the match between the model and $E^{(2)}_{\mathrm{POL}}$.

In this manner, we are able to determine a new set of models with the appropriate polarization damping chosen self-consistently. As we emphasised in §V B, the choice of β_{pol} does not affect the quality of the two-body potential. Indeed, Model(3) with this change to the damping is nearly identical in every respect to the original model. The effects will however be manifest in the many-body polarization energies. We are currently investigating this issue.

B. Multipole model rank reduction

Our simplest model, Model(1), contains anisotropic terms only in the ISA-DMA multipole and the polarization models. In §V A we have argued that the ISA-DMA model shows better convergence properties than the usual DMA procedure of Stone [103]. Based on that discussion and the results presented in Figure 1, we may ask whether we can truncate the rank of the ISA-DMA model without incurring a significant loss in accuracy. In Figure 17 we display interaction energy profiles for Model(1) using the ISA-DMA model at various ranks. As before, these calculations have been performed at two representative dimer orientations: Hb1 and S1. At the doubly hydrogen-bonded Hb1 orientation there is no appreciable change on reducing rank to $l = 3$, but any further reduction results in a significant change in the PES with the interaction energy getting systematically smaller (in magnitude). At the dispersion-bound S1 orientation there is almost no change to the model when the rank of the multipole expansion is reduced all the way to $l = 0$ (charges only). This is perhaps to be expected as the electrostatic interaction is relatively insignificant for the S1 (and S2) complexes. What is surprising is that the

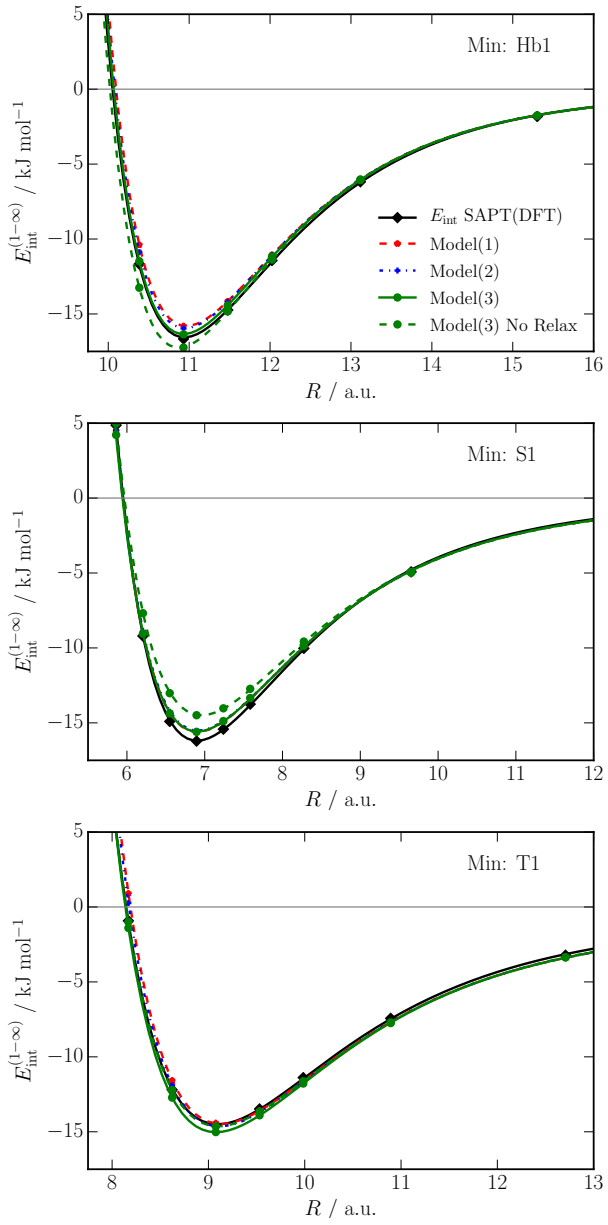

FIG. 12: PES sections at the Hb1, S1 and T1 dimer orientations. Sections at the other minima are provided in the supplementary information.

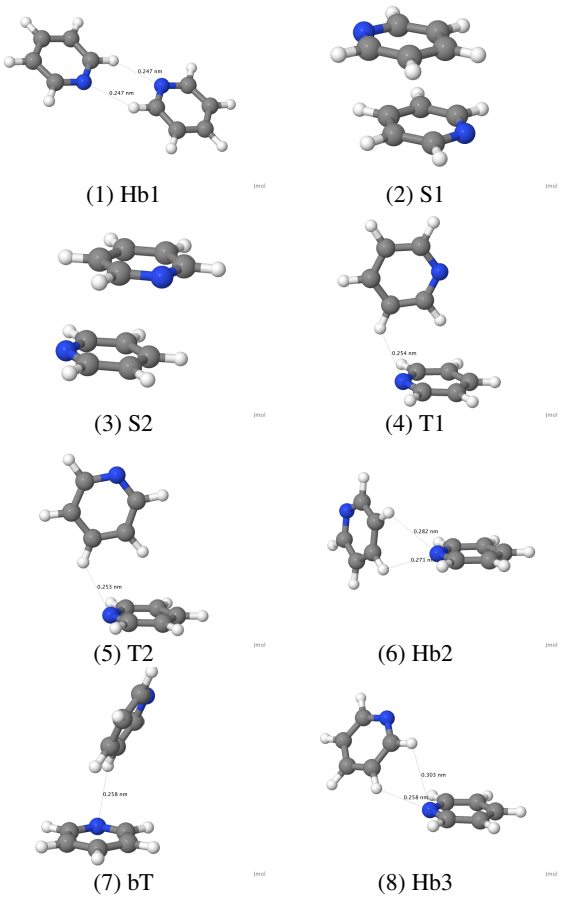

FIG. 13: Structures of pyridine dimers at stable minima on the three PESs. The structures are ordered according to their energies calculated using SAPT(DFT). These images have been produced using the JMOL program [100].

T1 complex also shows a relative insensitivity to the rank of the multipole expansion.

The behaviour of the models at the doubly-hydrogen-bonded Hb1 dimer configuration needs some explanation. The rank of multipoles on the hydrogen atoms do not appear to matter as the model interaction energies do not alter significantly if only rank 0 (charge) terms are included on these atoms. However, the nitrogen and carbon atoms appear to need the octopolar terms to model the electrostatic interaction correctly in this configuration. At least for the nitrogen atom this should not be surprising as the octopolar terms are needed to describe the effects from the lone pairs, but it is surprising that the carbon atoms also require these terms. In any case, it may be possible to improve the quality of the charge-only model by including additional sites around the nitrogen and carbon atoms to account for these terms in much the same way as is done for the oxygen atom in water models. If successful, this would provide us with a route to construct a fully isotropic interaction model for pyridine and other systems. This would be important as, with some exceptions such as the ORIENT and DMACRYS [104] programs, simulation programs cannot nor-

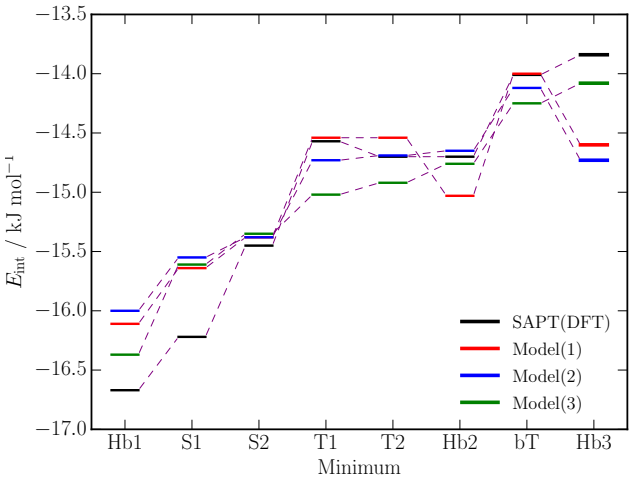

FIG. 14: Visualisation of the data in Table II. The energies of stable dimers on the three PESs are displayed as solid horizontal bars. The dashed lines link the energies levels associated with each of the three models. SAPT(DFT) reference energies have been calculated at the dimer geometries from Model(3). Data for Model(3)-DF-DMA4 are not shown here.

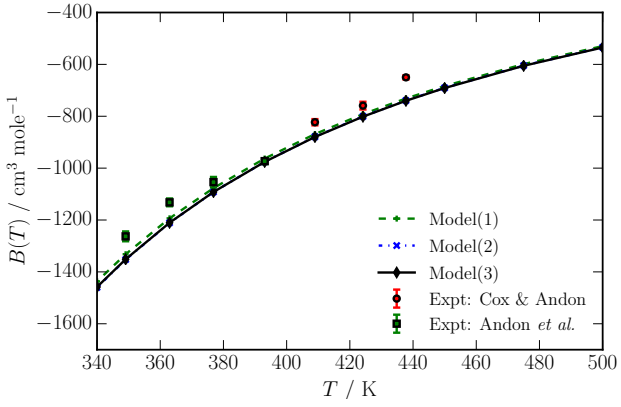

FIG. 15: Classical second virial coefficients for pyridine. The experimental data and error bars are from Andon *et al.* [101] and Cox & Andon [102]. Quantum corrections contribute very little and would not make a visible difference on the scale of this graph.

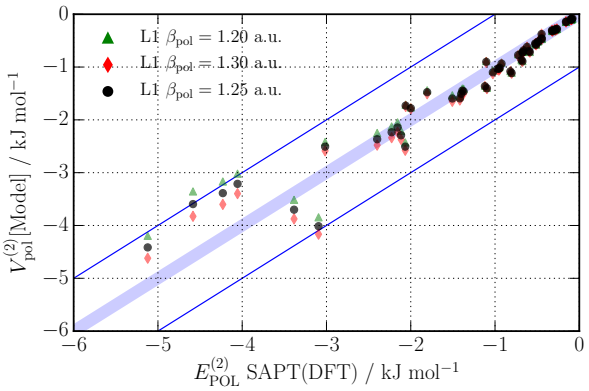

FIG. 16: Second-order polarization energies from regularised SAPT(DFT) compared with the L1 polarization model. The energies have been calculated using minimum energy dimer configurations obtained on the Model(3) PES. Dimers with attractive total interaction energies are indicated with filled symbols, and those with repulsive energies with open symbols. The thin blue lines indicate the ±1 kJ mol^{-1} error limits and the blue bar is present just as a visual aid.

mally use potentials with anisotropic terms, a restriction that significantly limits the usage of the accurate potentials we are able to develop.

XIII. CONCLUSIONS & DIRECTIONS

We have described a robust and relatively easy to implement algorithm for developing accurate intermolecular potentials in which most of the potential parameters are derived from the charge density and density response functions, and the remaining, short-range, parameters are robustly determined by associating these with specific atom-pairs using a basis-space implementation of the iterative stockholder atoms (ISA) algorithm. With this algorithm, accurate, many-body potentials can be derived using a relatively small number of dimer energies calculated using SAPT(DFT). This significantly reduces the computational cost of the approach. Importantly, as all of the long-range and most of the short-range parameters are *derived*, the predictive power of the resulting potentials is significant.

One of the major obstacles to intermolecular potential development has been the derivation of the short-range parameters. We have demonstrated that these can be relatively easily and robustly derived from the non-interacting charge densities using the distributed density-overlap model based the ISA. In this manner, even the atomic anisotropy terms, which are usually poorly defined in a direct fit, are robustly determined with a relatively small amount of computational effort. Using these techniques on the pyridine dimer, we have demonstrated that features such as the density distortions due to the π-bonding on the carbon atoms, and the lone pair on the nitrogen atom in pyridine are well-defined using our approach. Indeed, only terms with a physical origin are present in this approach.

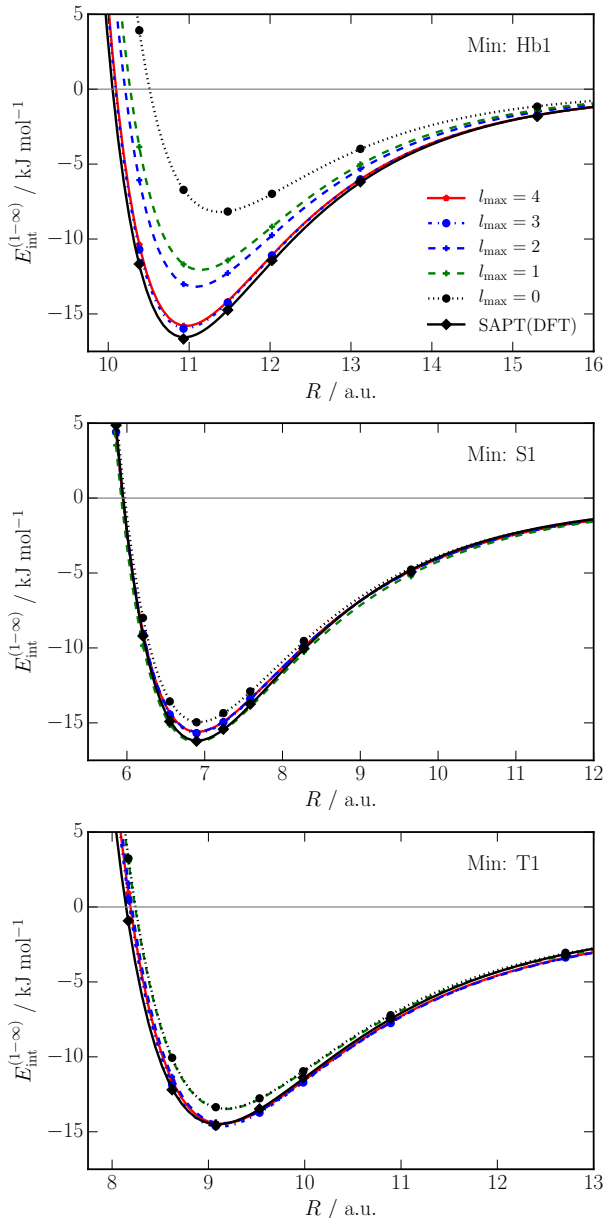

FIG. 17: The effect of rank reduction of the multipole model for Model(1).

The main features of the methodology we describe in the paper are:

- *Efficient use of data*: The potentials are derived using a hierarchy of data sets; the more extensive data sets include only first-order energies and can be very easily calculated, while the second-order energies are included through a significantly smaller data set.

- *Priors*: We use the first and most extensive data set to determine prior values for most of the short-range parameters. These priors may subsequently be modified using the second, smaller data set. These steps may be repeated thus leading to a multi-stage procedure which significantly reduces the amount of data needed to tune the potential.

- *ISA*: The short-range parameters are determined using the ISA method for partitioning the molecular densities into atomic contributions. The BS-ISA algorithm allows this to be performed using extensive basis sets with a well-defined basis set limit. The ISA atoms are as close to spherical as is possible and account for charge movement within the molecule, consequently the resulting short-range repulsion parameters may be expected to be free from basis set artifacts, and be the most isotropic possible. This compares favourably with the density-fitting-based partitioning scheme we have proposed in earlier papers [31, 40] which does not fulfil either of these properties. Indeed the r.m.s. errors made by the ISA-based models are half as much as those from the density-fitting-based models.

- *Long-range models*: The long-range parameters of the potentials are determined using distributed multipoles, polarizabilities and dispersion coefficients. The ISA-DMA multipoles are obtained from the BS-ISA approach and have been demonstrated to exhibit systematic convergence with rank. The WSM distribution scheme has been used to calculate the distributed polarizabilities and dispersion coefficients, the latter of which we have tuned to SAPT(DFT) dispersion energies.

- *Predictive power*: Most of the parameters are derived from or fitted to molecular properties, consequently they are physically meaningful and the resulting potentials exhibit a considerable predictive power.

- *Hierarchy of models*: The methods we have described allow us to determine potentials of various levels of complexity in a meaningful manner. These may be fully isotropic at the atom–atom level or contain as much anisotropy as is needed.

We have used these techniques to develop a set of potentials of varying levels of detail for the pyridine dimer. The simplest of these include only isotropic short-range terms, and the most detailed includes all significant anisotropy terms up to rank two. The predictive power of these potentials is quite significant and all are able to predict SAPT(DFT) interaction

energies for low energy dimers not included in the fit. As a consequence, the potentials are robust to the inclusion of additional data: parameters alter very little on relaxation, and features on the potential energy landscape change only slightly. This robustness is particularly important in the development of multidimensional potentials, as we will generally be unable to sample dimer configuration space adequately, especially for larger monomers.

We have compared our newly derived pyridine potentials to the rather limited set of data available in the literature. Of the eight stable minima found on the Model(3) PES, the double hydrogen-bonded Hb1 dimer has been found in previous DFT+D work by Piacenza and Grimme [98], and the CCSD(T) energy for this structure [99] differs from our SAPT(DFT) interaction energy by only 7%. The two other hydrogen-bonded structures, Hb2 and Hb3, have not been seen before. Both the stacked structures, S1 and S2, have been found earlier [98]. Of the three T-shaped structures, only the bT structure resembles a previously found structure [98], while the T1 and T2 structures appear to be unique to the models developed in this paper. As the DFT+D method cannot be relied on to correctly describe the subtle balance of dispersion, electrostatic, polarization and charge-transfer interactions seen in the eight dimers of pyridine, it is possible that the set of eight minima we have found are a more accurate representation of this system. Further tests are needed at the CCSD(T) level of theory if we are to be sure of this.

In this paper we have provided solutions to some of the most significant issues related to potential development, and, as a consequence, have inevitably exposed other minor issues that need resolving. Some of these are:

- The WSM method for deriving distributed polarization and dispersion models is a good one, but it is based on a less than ideal partitioning method [48] that seems to result in some artifacts in the models and a small, but undesirable basis-set dependence.
- The current damping of the dispersion model based on molecular ionisation potentials only is less than ideal and there is good reason to expect a site–site damping model to perform better.
- More needs to be done to understand the origin of the polarization damping. Like the dispersion damping, here too it is clear that the damping model needs to depend on the pair of interacting sites, but there is evidence [42] that the polarization damping differs strongly from that used for the dispersion. This is probably the least understood issue at present.
- The resulting potentials are for rigid monomers only. However, as the potential parameters are closely associated with the properties of the atoms in the interacting molecules through either the ISA or the DF-based partitioning methods, it is possible that these models may be applicable to flexible monomers. This conjecture needs to be tested.
- One of the most serious limitations of the approach we have described here is that there are very few simulation programs capable of using these potentials. Most simulation programs use the simpler Lennard-Jones models with point-charge electrostatic models. However, distributed multipoles are being increasingly available in simulations codes: both OpenMM [105] and DL_POLY [106] allow the use of distribute multipoles and simple polarization models, but only the Orient [107] and DMACRYS [104] programs currently support the use of the anisotropy terms present in our more complex potentials. We do not doubt that this situation will change as potential development using the methods described in this paper becomes more streamlined and easy to use, and as we accumulate evidence that these more elaborate potentials do result in higher predictive accuracy.

It should be apparent that the ISA — in particular, the BS-ISA algorithm — plays a central role in the methodology we have described. Consequently it should come as no surprise that some of the issues listed above may be resolved using data extracted from the ISA atomic densities. In a forthcoming paper [52] we will describe how the dispersion damping issue may be resolved using the ISA, and also how even more of the short-range parameters may be derived rather than fitted.

However, there are issues with the models we have presented here. Second virial coefficients are well reproduced using our isotropic and anisotropic potentials, though all three models give $B(T)$ somewhat too negative. This indicates that the models are somewhat too attractive on the average. We have established that there are indeed regions of configuration space where all potentials systematically overbind and these are associated with stacked-like configurations. While we do not fully understand the origin of the problem, it is possible that the additivity assumption we have made in the definition of ρ_{ab} in eq. (28) is inappropriate, and also that the SAPT(DFT) interaction energies are themselves too attractive for these configurations due to the known problems with the $\delta_{\mathrm{int}}^{\mathrm{HF}}$ term for dispersion-bound systems [108, 109]. We are actively engaged in understanding these issues.

XIV. ACKNOWLEDGEMENTS

AJM would like to thank Prof Sally Price for initiating and motivating this project and supporting it, particularly in its early stages, Dr Richard Wheatley for useful discussions related to the ISA, and Lauretta Schwartz for preliminary work on rank reduction of the multipole model. We would like to thank Mary J. Van Vleet for useful comments on the manuscript. AJM would also like to thank Queen Mary University of London for computing resources, the Thomas Young Centre for a stimulating environment, and the Cambridge University Library for generous resources.

This work was partially funded by EPSRC grant EP/C539109/1.

XV. SUPPLEMENTARY INFORMATION

The SI included with this paper contains details of the three potentials derived in this paper. Additionally, plots referenced but not included in this paper are provided in the SI.

Appendices

Appendix A: Programs

Many of the theoretical methods described in this paper are implemented in programs available for download. Some of these, together with their main uses in the present work, are:

- CamCASP 5.9 [36]: Calculation of WSM polarizabilities, the dispersion models, the SAPT(DFT) energies, and overlap models.
- Orient 4.8 [107]: Localization of the distributed polarizabilities, calculation of dimer energies using the electrostatic, polarization and dispersion models, visualization of the energy maps, and fitting to obtain the analytic atom–atom potentials.
- Dalton 2.0 [63]: DFT calculations. A patch [64] is needed to enable Dalton 2.0 to work with CamCASP.

Appendix B: CamCASP

Many of the algorithmic details of the electronic structure methods implemented in the CamCASP suite of programs have been described in previous publications. Rather than provide an exhaustive list, we will indicate those algorithms and methods of importance for potential development, as well as some numerical techniques that are particularly important for accuracy and computational efficiency.

Some of the capabilities of the CamCASP suite of programs are as follows:

- *SAPT(DFT)*: Interaction energies to second-order can be calculated using SAPT(DFT) [7–10]. Infinite-order effects may be approximated using the $\delta_{\rm int}^{\rm HF}$ correction.
- *Distributed multipole models*: These may be evaluated using both the GDMA algorithms [44, 45], or directly from a density-fitting-based partitioning using a variety of constraints (see the CamCASP User's Guide for details), or from the recently implemented ISA algorithm [47].
- *Distributed frequency-dependent polarizabilities*: These may be calculated in non-local form using constrained density-fitting-based partitioning schemes [48], which include the SRLO method [80] as a special case. Localised models may be obtained using the Williams–Stone–Misquitta (WSM) model [43, 49].
- *Distributed dispersion models*: These may be evaluated directly using the non-local frequency-dependent models [77], or from localised polarizability models obtained using the WSM procedure [50]. Models may be isotropic or anisotropic.
- *Linear-response kernel*: The code is able to evaluate the linear-response kernel using the ALDA, CHF and hybrid, ALDA+CHF, kernels. These integrals are evaluated internally.
- *Interfaces*: CamCASP can use molecular orbitals calculated from the Dalton program (versions from 2006 to 2015 are supported), the NWChem 6.x program and GAMESS(US) .

These are the major features of the CamCASP program, and the code additionally includes other algorithms that are important for model development. These include the ability to calculate distributed density-overlap integrals and, from these, develop density overlap models for the short-range intermolecular interaction energy, and interfaces to the Orient program[107] to aid in visualisation of the interaction energy models and fitting of intermolecular potentials.

[1] Podeszwa, R.; Bukowski, R.; Szalewicz, K. *J. Chem. Theory Comput.* **2006**, *2*, 400–412.

[2] Hesselmann, A.; Jansen, G.; Schütz, M. *J. Chem. Phys.* **2005**, *122*, 014103.

[3] Hesselmann, A.; Jansen, G.; Schütz, M. *J. Am. Chem. Soc.* **2006**, *128*, 11730–11731.

[4] Podeszwa, R.; Bukowski, R.; Rice, B. M.; Szalewicz, K. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5561–9.

[5] Hesselmann, A.; Jansen, G.; Schutz, M. *J. Am. Chem. Soc.* **2006**, *128*, 11730–11731.

[6] Fiethen, A.; Jansen, G.; Hesselmann, A.; Schutz, M. *J. Am. Chem. Soc.* **2008**, *130*, 1802–1803.

[7] Misquitta, A. J.; Szalewicz, K. *Chem. Phys. Lett.* **2002**, *357*, 301–306.

[8] Misquitta, A. J.; Jeziorski, B.; Szalewicz, K. *Phys. Rev. Lett.* **2003**, *91*, 33201.

[9] Misquitta, A. J.; Szalewicz, K. *J. Chem. Phys.* **2005**, *122*, 214109.

[10] Misquitta, A. J.; Podeszwa, R.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **2005**, *123*, 214103.

[11] Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *357*, 464–470.

[12] Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2002**, *362*, 319–325.

[13] Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.* **2003**, *367*, 778–784.

[14] Hodges, M. P.; Stone, A. J.; Xantheas, S. S. *J. Phys. Chem. A* **1997**, *101*, 9163–9168.

[15] Mas, E. M.; Bukowski, R.; Szalewicz, K. *J. Chem. Phys.* **2003**, *118*, 4386–4403.

[16] Mas, E. M.; Bukowski, R.; Szalewicz, K. *J. Chem. Phys.* **2003**, *118*, 4404–4413.

[17] Bukowski, R.; Szalewicz, K.; Groenenboom, G.; van der Avoird, A. *J. Chem. Phys.* **2006**, *125*, 044301.

[18] Welch, G. W. A.; Karamertzanis, P. G.; Misquitta, A. J.; Stone, A. J.; Price, S. L. *J. Chem. Theory Comput.* **2008**, *4*, 522–532.

[19] Podeszwa, R.; Szalewicz, K. *J. Chem. Phys.* **2007**, *126*, 194101.

[20] Stone, A. J. *The Theory of Intermolecular Forces*, 2nd ed.; Oxford University Press, Oxford, 2013.

[21] Jiang, H.; Jordan, K. D.; Taylor, C. E. *J. Phys. Chem. B* **2007**, *111*, 6486–6492.

[22] Illingworth, C. J.; Domene, C. *Proc. R. Soc. A* **2009**, *465*, 1701–1716.

[23] Bukowski, R.; Sadlej, J.; Jeziorski, B.; Jankowski, P.; Szalewicz, K.; Kucharski, S. A.; Williams, H. L.; Rice, B. M. *J. Chem. Phys.* **1999**, *110*, 3785–3803.

[24] Bukowski, R.; Szalewicz, K. *J. Chem. Phys.* **2001**, *114*, 9518.

[25] Kim, H.-Y.; Sofo, J. O.; Velegol, D.; Cole, M. W.; Lucas, A. A. *J. Chem. Phys.* **2006**, *124*, 074504–4.

[26] Korona, T.; Williams, H. L.; Bukowski, R.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.* **1997**, *106*, 5109.

[27] Bukowski, R.; Szalewicz, K.; Chabalowski, C. *J. Phys. Chem. A* **1999**, *103*, 7322.

[28] Misquitta, A. J.; Bukowski, R.; Szalewicz, K. *J. Chem. Phys.* **2000**, *112*, 5308–5319.

[29] Vissers, G. W. M.; Hesselmann, A.; Jansen, G.; Wormer, P. E. S.; van der Avoird, A. *J. Chem. Phys.* **2005**, *122*, 054306.

[30] Podeszwa, P.; Bukowski, R.; Szalewicz, K. *J. Phys. Chem. A* **2006**, *110*, 10345–10354.

[31] Misquitta, A. J.; Welch, G. W. A.; Stone, A. J.; Price, S. L. *Chem. Phys. Lett.* **2008**, *456*, 105–109.

[32] Totton, T.; Misquitta, A. J.; Kraft, M. *J. Chem. Theory Comput.* **2010**, *6*, 683–695.

[33] Yu, K.; McDaniel, J. G.; Schmidt, J. R. *J. Phys. Chem. B* **2011**, *115*, 10054–10063.

[34] McDaniel, J. G.; Schmidt, J. *J. Phys. Chem. A* **2013**, *117*, 2053–2066.

[35] Schmidt, J. R.; Yu, K.; McDaniel, J. G. *Acc. Chem. Res.* **2015**, *48*, 548–556.

[36] Misquitta, A. J.; Stone, A. J. CamCASP: a program for studying intermolecular interactions and for the calculation of molecular properties in distributed form. University of Cambridge, 2016; `http://www-stone.ch.cam.ac.uk/programs.html\#CamCASP`, Accessed: Mayy 2016.

[37] Jansen, G. 2015; Notation for second-order energies. Private communication.

[38] Jeziorska, M.; Jeziorski, B.; Cizek, J. *Int. J. Quantum Chem.* **1987**, *32*, 149–164.

[39] Moszynski, R.; Heijmen, T. G. A.; Jeziorski, B. *Mol. Phys.* **1996**, *88*, 741–758.

[40] Stone, A. J.; Misquitta, A. J. *Int. Revs. Phys. Chem.* **2007**, *26*, 193–222.

[41] Tang, K. T.; Toennies, J. P. *Surf. Sci. Lett.* **1992**, *279*, 203–206.

[42] Misquitta, A. J. *J. Chem. Theory Comput.* **2013**, *9*, 5313–5326.

[43] Misquitta, A. J.; Stone, A. J. *J. Chem. Theory Comput.* **2008**, *4*, 7–18.

[44] Stone, A. J.; Alderton, M. *Mol. Phys.* **1985**, *56*, 1047–1064.

[45] Stone, A. J. *J. Chem. Theory Comput.* **2005**, *1*, 1128–1132.

[46] Lillestolen, T. C.; Wheatley, R. J. *J. Chem. Phys.* **2009**, *131*, 144101–6.

[47] Misquitta, A. J.; Stone, A. J.; Fazeli, F. *J. Chem. Theory Comput.* **2014**, *10*, 5405–5418.

[48] Misquitta, A. J.; Stone, A. J. *J. Chem. Phys.* **2006**, *124*, 024111–14.

[49] Misquitta, A. J.; Stone, A. J.; Price, S. L. *J. Chem. Theory Comput.* **2008**, *4*, 19–32.

[50] Misquitta, A. J.; Stone, A. J. *Mol. Phys.* **2008**, *106*, 1631 – 1643.

[51] Stone, A. J. *J. Phys. Chem. A* **2011**, *115*, 7017–7027.

[52] Van Vleet, M. J.; Misquitta, A. J.; Stone, A. J.; Schmidt, J. R. **2015**, submitted.

[53] Shoemake, K. In *Graphics Gems III*; Kirk, D., Ed.; Academic Press, 1992; pp 124–132.

[54] Kita, S.; Noda, K.; Inouye, H. *J. Chem. Phys.* **1976**, *64*, 3446–3449.

[55] Kim, Y. S.; Kim, S. K.; Lee, W. D. *Chem. Phys. Lett.* **1981**, *80*, 574–575.

[56] Nobeli, I.; Price, S. L.; Wheatley, R. J. *Mol. Phys.* **1998**, *95*, 525–537.

[57] McDaniel, J. G.; Yu, K.; Schmidt, J. R. *J. Phys. Chem. C* **2012**, *116*, 1892–1903.

[58] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004.

[59] Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

[60] Dunning, Jr., T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

[61] Williams, H. L.; Mas, E. M.; Szalewicz, K.; Jeziorski, B. *J. Chem. Phys.* **1995**, *103*, 7374–7391.

[62] Woon, D. E.; T. H. Dunning, J. *J. Chem. Phys.* **1994**, *100*, 2975–2889.

[63] Helgaker, T.; Jensen, H. J. A.; Joergensen, P.; Olsen, J.; Ruud, K.; Aagren, H.; Auer, A.; Bak, K.; Bakken, V.; Christiansen, O.; Coriani, S.; Dahle, P.; Dalskov, E. K.; Enevoldsen, T.; Fernandez, B.; Haettig, C.; Hald, K.; Halkier, A.; Heiberg, H.; Hettema, H.; Jonsson, D.; Kirpekar, S.; Kobayashi, R.; Koch, H.; Mikkelsen, K. V.; Norman, P.; Packer, M. J.; Pedersen, T. B.; Ruden, T. A.;

Sanchez, A.; Saue, T.; Sauer, S. P. A.; Schimmelpfennig, B.; Sylvester-Hvid, K. O.; Taylor, P. R.; Vahtras, O. DALTON, a molecular electronic structure program, Release 2.0. 2005; See http://www.kjemi.uio.no/software/dalton/dalton.html.

[64] Bukowski, R.; Cencek, W.; Jankowski, P.; Jeziorski, B.; Jeziorska, M.; Lotrich, V.; Kucharski, S.; Misquitta, A. J.; Moszynski, R.; Patkowski, K.; Podeszwa, R.; Rybak, S.; Szalewicz, K.; Williams, H.; Wheatley, R. J.; Wormer, P. E. S.; Zuchowski, P. S. SAPT2008: an ab initio program for many-body symmetry-adapted perturbation theory calculations of intermolecular interaction energies. University of Delaware and University of Warsaw, 2008; `http://www.physics.udel.edu/~szalewic/`, Accessed: Oct 2013.

[65] Tozer, D. J.; Handy, N. C. *J. Chem. Phys.* **1998**, *109*, 10180–10189.

[66] Weigend, F.; Haser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143–152.

[67] Weigend, F.; Kohn, A.; Hattig, C. *J. Chem. Phys.* **2002**, *116*, 3175–3183.

[68] Akin-Ojo, O.; Bukowski, R.; Szalewicz, K. *J. Chem. Phys.* **2003**, *119*, 8379–8396.

[69] Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244–13249.

[70] Sadlej, A. J. *Coll. Czech Chem. Commun.* **1988**, *53*, 1995–2016.

[71] Jeziorski, B.; Szalewicz, K. In *Handbook of Molecular Physics and Quantum Chemistry*; Wilson, S., Ed.; Wiley, 2002; Vol. 8; Chapter 8, pp 37–83.

[72] Winn, P. J.; Ferenczy, G. G.; Reynolds, C. A. *J. Phys. Chem. A* **1997**, *101*, 5437–5445.

[73] Ferenczy, G. G.; Winn, P. J.; Reynolds, C. A. *J. Phys. Chem. A* **1997**, *101*, 5446–5455.

[74] Totton, T.; Misquitta, A. J.; Kraft, M. *Chem. Phys. Lett.* **2011**, *510*, 154–160.

[75] Patkowski, K.; Jeziorski, B.; Szalewicz, K. *J. Mol. Struct. (Theochem)* **2001**, *547*, 293–307.

[76] Sebetci, A.; Beran, G. J. O. *J. Chem. Theory Comput.* **2010**, *6*, 155–167.

[77] Misquitta, A. J.; Spencer, J.; Stone, A. J.; Alavi, A. *Phys. Rev. B* **2010**, *82*, 075312–7.

[78] Le Sueur, C. R.; Stone, A. J. *Mol. Phys.* **1994**, *83*, 293–308.

[79] Lillestolen, T. C.; Wheatley, R. J. *J. Phys. Chem. A* **2007**, *111*, 11141–11146.

[80] Rob, F.; Szalewicz, K. *Chem. Phys. Lett.* **2013**, *572*, 146–149.

[81] Williams, G. J.; Stone, A. J. *J. Chem. Phys.* **2003**, *119*, 4620–4628.

[82] Stone, A. J. *The Theory of Intermolecular Forces*; International Series of Monographs in Chemistry; Clarendon Press: Oxford, 1996.

[83] Tang, K. T.; Toennies, J. P. *J. Chem. Phys.* **1984**, *80*, 3726–3741.

[84] Soderhjelm, P.; Karlstrom, G.; Ryde, U. *J. Chem. Phys.* **2006**, *124*, 244101–10.

[85] Mitchell, J. B. O.; Price, S. L. *J. Phys. Chem. A* **2000**, *104*, 10958–10971.

[86] Patil, S. H.; Tang, K. T. In *Asymptotic methods in quantum mechanics*; Scäfer, F. P., Toennies, J. P., Zinth, W., Eds.; Springer, 2000.

[87] Nyeland, C.; Toennies, J. P. *Chem. Phys. Lett.* **1986**, *127*, 172–177.

[88] Rosen, N. *Phys. Rev.* **1931**, *38*, 255–276.

[89] Smirnov, B. M.; Chibisov, M. I. *Sov. Phys.-JETP* **1965**, *21*, 624.

[90] Andreev, E. A. *Theor. Chim. Acta* **1973**, *28*, 235–239.

[91] Kleinekathöfer, U.; Tang, K. T.; Toennies, J. P.; Yiu, C. L. *J. Chem. Phys.* **1995**, *103*, 6617–.

[92] Zemke, W. T.; Stwalley, W. C. *J. Chem. Phys.* **1999**, *111*, 4962–4965.

[93] Grüning, M.; Gritsenko, O. V.; van Gisbergen, S. J. A.; Baerends, E. J. *J. Chem. Phys.* **2001**, *114*, 652–660.

[94] Anghel, A. T.; Day, G. M.; Price, S. L. *CrystEngComm.* **2002**, *4*, 348 – 355.

[95] Hodges, M. P.; Wheatley, R. J. *Chem. Phys. Lett.* **2000**, *326*, 263–268.

[96] Piquemal, J.-P.; Cisneros, G. A.; Reinhardt, P.; Gresh, N. *J. Chem. Phys.* **2006**, *124*, 104101–12.

[97] Wales, D. J. *Energy Landscapes: with applications to clusters, biomolecules and glasses*; Cambridge University Press, Cambridge, 2003.

[98] Piacenza, M.; Grimme, S. *ChemPhysChem* **2005**, *6*, 1554 – 1558.

[99] Hohenstein, E. G.; Sherrill, C. D. *J. Phys. Chem. A* **2009**, *113*, 878–886.

[100] Jmol: an open-source Java viewer for chemical structures in 3D. 2015; `http://www.jmol.org/`, Accessed: Sep 2015.

[101] Andon, R. J. L.; Cox, J. D.; Herington, E. F. G.; Martin, J. F. *Trans. Faraday Soc.* **1957**, *53*, 1074–1082.

[102] Cox, J. D.; Andon, R. J. L. *Trans. Faraday Soc.* **1958**, *54*, 1622–1629.

[103] Stone, A. J. *J. Chem. Theory Comput.* **2005**, *1*, 1128–1132.

[104] Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanisc, P. G.; Day, G. M. *Phys. Chem. Chem. Phys.* **2010**, *12*, 8478–8490.

[105] Eastman, P.; Friedrichs, M. S.; Chodera, J. D.; Radmer, R. J.; Bruns, C. M.; Ku, J. P.; Beauchamp, K. A.; Lane, T. J.; Wang, L.-P.; Shukla, D.; Tye, T.; Houston, M.; Stich, T.; Klein, C.; Shirts, M. R.; Pande, V. S. *J. Chem. Theory Comput.* **2013**, *9*, 461–469.

[106] Todorov, I. T.; Smith, W.; Trachenko, K.; Dove, M. T. *J. Mater. Chem.* **2006**, *16*, 1911–1918.

[107] Stone, A. J.; Dullweber, A.; Engkvist, O.; Fraschini, E.; Hodges, M. P.; Meredith, A. W.; Nutt, D. R.; Popelier, P. L. A.; Wales, D. J. ORIENT: a program for studying interactions between molecules, version 4.8. University of Cambridge, 2016; `http://www-stone.ch.cam.ac.uk/programs.html\#Orient`, Accessed: May 2016.

[108] Podeszwa, R.; Szalewicz, K. *Chem. Phys. Lett.* **2005**, *412*, 488.

[109] Patkowski, K.; Szalewicz, K.; Jeziorski, B. *J. Chem. Phys.* **2006**, *125*, 154107.