A simple model for halogen bonds

Robert A. Shaw and J. Grant Hill*
Department of Chemistry, University of Sheffield
Sheffield S3 7HF, UK

Abstract

Halogen bonds are prevalent in many areas of chemistry, physics and biology. We present a statistical model for the interaction energies of halogen-bonded systems based on high-accuracy *ab initio* benchmark calculations for a range of complexes. Remarkably, the resulting model requires only two fitted parameters, X and B - one for each molecule - and optionally the equilibrium separation, R_e, between them, taking the simple form $E = XB/R_e^n$. For $n = 4$, it gives negligible root-mean-squared deviations of 0.14 and 0.28 kcal mol^{-1} over separate fitting and validation data sets of 60 and 74 systems, respectively. The simple model is shown to outperform some of the best density functionals for non-covalent interactions at, once parameters are available, essentially zero computational cost. Additionally, we demonstrate how it can be transferred to completely new, much larger complexes and still achieve accuracy within 0.5 kcal mol^{-1}. Utilising a principal component analysis and symmetry-adapted perturbation theory, we further show how the model can be used to predict the physical nature of a halogen bond, providing an efficient way to gain insight into the behaviour of halogen-bonded systems. This means that the model can be used to highlight cases where induction or dispersion significantly affect the underlying nature of the interaction.

1 Introduction

Halogen bonds are an important class of non-covalent interaction where a halogen-containing donor, AX, interacts with a Lewis base as acceptor, B. While examples of halogen bonds were recognised as early as 1814[1–3], it is only more recently with detailed X-ray diffraction[4–6] and spectroscopic[7–9] studies that they were found to be prevalent in both the gas and condensed phases[10, 11]. These investigations discovered a number of striking properties, in particular the strong preference for linear geometries[12, 13], where the AX···B angle is close to 180°, and interaction energies similar to those of hydrogen bonds[8, 14]. These factors give halogen bonds a high degree of tuneability, making them ideal for use in fields ranging from crystal engineering to nanomaterials and drug design[11, 15–23].

Halogens are conceptually seen as being electron rich, making their interaction with similarly electronegative bases counterintuitive. The most popular recent explanation is that of a σ-hole, first suggested in 2005 by Clark *et al.*[24] They posit that the attachment of a suitably electron-withdrawing group to a halogen atom results in withdrawal of electron density from the halogen along the σ-bond. This withdrawal results in a charge anisotropy such that there is a positive 'hole' on the face of the halogen atom opposite the bond. As such, a simple electrostatic argument can be made for how Lewis bases then interact with the halogen, and this explains the strong geometry dependence. It has been proposed that this is just an example of a wider class of such interactions, with similar effects seen for chalcogens, pnicogens, and tetrels[18, 25–28]. The electrostatic potential can be calculated, and measured experimentally, confirming

[*] E-mail: *grant.hill@sheffield.ac.uk*

[†] Supporting information for this article includes model parameters for all molecules, benchmark energies and separations for the fitting and validation sets, DFT results for all systems, additional figures, Cartesian coordinates of benchmark geometries, and a Jupyter notebook detailing some of the analysis and how to parametrise new molecules.

that the anisotropy does indeed exist[29–31]. It has been found that the σ-hole increases in size and intensity as one goes down the group, and that the strength of the interaction increases with the electron-withdrawing power attached to the halogen[32–36]. These factors lend support to the intuitive explanation.

The above points clearly indicate that electrostatics are important in such systems. However, the 'makeup' of halogen bonds has attracted considerable debate[37–45]. It has been comprehensively shown that electrostatics alone are not sufficient to fully describe these interactions[9, 46]. The IUPAC definition of the halogen bond emphasises that "the forces involved in the formation of the halogen bond are primarily electrostatic, but polarization, charge transfer, and dispersion contributions all play an important role"[47]. Indeed, in 1996 an analysis of the Cambridge Structural Database combined with intermolecular perturbation theory calculations came to the same conclusion as the IUPAC definition when considering interactions between carbon-bonded halogens and electronegative atoms.[48] As such, the σ-hole description, while conceptually very useful, is only part of the story. Numerous studies have demonstrated that dispersion is a very important component in differentiating halogen bonds from one another[12, 42, 49, 50]. Similarly, exchange-repulsion effects can have a large impact on the geometries of halogen-bonded systems[12, 42, 46], and charge transfer is argued to be a distinguishing factor in many cases[38, 43, 44, 51].

It has been pointed out that almost all such phenomena come under the umbrella of polarisation[39]. Certainly, significant charge transfer such as found in so-called 'Mulliken inner complexes' is often represented in the form $\mathrm{AX}^{-}\cdots\mathrm{B}^{+}$[52], suggesting an extreme form of polarisation. It is also true that dispersion, which is the purely quantum mechanical interaction due to the instantaneous fluctuations of electrons, can be formally derived from polarisabilities. Exchange-repulsion is somewhat distinct but necessarily contaminates all other terms. This does not mean that such decompositions are meaningless, however, rather that they should be treated with caution. It is sensible to distinguish 'local' polarisation and charge transfer, as this allows for a simple descriptor of when certain systems may behave substantially differently and utilises an idea that is well-established within both the experimental and theoretical communities. The local distinction in this context simply refers to distortions and anisotropies in the electron density of a molecule constrained primarily to that molecule, as opposed to distortions effected by the surrounding environment resulting in substantial transfer of density away from the original molecule. Several examples of such unusually strong interactions have been reported[43, 44, 53–56], and there has been experimental evidence for charge transfer, from both rotational[12, 57] and X-ray absorption[58] spectroscopy.

In a similar vein, the knowledge that dispersion is important implies that certain theoretical methods will not be useful. As this interaction is by definition due to the dynamical correlation of electrons, uncorrelated mean-field methods such as Hartree-Fock will not give accurate results. In particular, most density functionals are known to perform poorly on such systems[59–63]. The combination of this and the fact that non-covalent interactions involve small energy differences means that only the highest accuracy theoretical methods consistently give results in agreement with experiment. These are prohibitively expensive, however, and restricted to fairly small molecular complexes. The most important applications involve large, extended systems in the condensed phases, which are also difficult to study experimentally. As such, it is of considerable interest to find simple, reliable methods to accurately predict the strength of halogen bonds. Equally, quantitative measures for distinguishing when a system will behave substantially differently to other, similar examples could provide insight into the nature of these important interactions.

One interesting approach was used by Legon and Millen for hydrogen bonds[64]. They considered experimental spectroscopic force constants, k_σ, which are closely linked to the interaction energy, and parametrised the molecules involved to give a simple prediction of these force constants in new hydrogen-bonded systems. The model was

$$k_\sigma = cNE \tag{1}$$

where c is a proportionality constant, while N and E are termed the 'nucleophilicity' and 'electrophilicity' of the hydrogen-bond acceptor and donor respectively. Despite their name, no physical basis was suggested for these; they are empirically-derived parameters found by comparing to values of one for a model system, in this case $H_2O\cdots HCl$. Tests of this model showed remarkably small deviations from experiment of less than 0.5 kcal mol^{-1} in many cases, although unsurprisingly these errors increased upon extrapolation to new systems. More recent work has attempted to extend this approach to other types of noncovalent interaction, including halogen bonds[65, 66]. Such a model would be ideal for halogen-bonded systems as it requires minimal effort. High accuracy calculations or experiments would only be needed for a small number of 'standard' systems before the parameters so determined could be used to quickly predict interaction strengths in new complexes.

In light of this, the present study has three aims. Firstly, high-accuracy benchmark calculations are presented for a wide range of small molecular systems. These data will then be used to investigate various simple models, demonstrating in sections 3.1 to 3.3 that a similar approach to that in equation 1 very accurately describes many halogen-bonded complexes. Perhaps most importantly, the theoretical basis for the analysis is investigated in section 3.4, providing insight into the nature of halogen bonds and allowing for the development of criteria to distinguish substantially different subclasses of interaction. The approach is also tested on larger, more practically relevant systems, giving results that are at least as good as the best density functionals for non-covalent interactions.

2 Methodology

Explicitly correlated coupled cluster calculations with singles, doubles, and perturbative triples[67] were carried out in the MOLPRO suite of programs[68, 69] using the 3C(Fix) ansätz and approximation b, [CCSD(T)-F12b][70, 71], with a geminal Slater exponent of $1.0 a_0^{-1}$. The cc-pVnZ-F12 basis sets were used, with the exception of Br and I, which used the cc-pVnZ-PP-F12 sets with the Stuttgart-Cologne small-core relativistic pseudopotential[72–75]; we note that no discontinuities are seen in trends going from chlorine to bromine when the pseudopotentials are introduced. Although not apparent from the abbreviation, these basis sets include augmentation with diffuse s and p functions. Geometries were optimised at the n =T level, while single-point energies were calculated for n =T, Q, then used to extrapolate to the complete basis set (CBS) limit using the method described by Hill and coworkers[76]. The Fock and exchange matrices were density fitted using the cc-pVQZ/JKFit auxiliary basis for all atoms other than bromine and iodine, which used the def2-QZVPP/JKFit sets[77, 78]. All subsequent two-electron integrals were fitted using the aug-cc-pVQZ and cc-pVnZ-PP-F12 MP2Fit sets for the lighter and post-d elements, respectively[73, 79]. The CABS+ procedure was carried out using the auxiliary sets specifically matched to the orbital basis[67, 73, 80–83] and the CABS singles correction was applied to the Hartree-Fock reference energy. The full counterpoise correction of Boys and Bernardi was used for all interaction energies[84].

Calculations with the M06-2X[85] and ωB97X-D[86] density functionals were performed in Gaussian 09[87], with the UltraFine integration grid. These functionals have been shown to perform particularly well for non-covalent interactions[88]. Symmetry adapted perturbation theory[89] calculations at the SAPT2+(3)δMP2 truncation[90, 91] were carried out in the SAPT2012 program[92] interfaced to MOLPRO, with the so-called "chemist's grouping"[93]. The aug-cc-pV(T+d)Z basis sets (abbreviated here as aVTZ) were used in both cases[94–96].

All errors quoted in calculated values are deviations relative to the CCSD(T)-F12b/CBS limit value, which has previously been shown to closely follow the same trends as experimental intermolecular force constants.[45] The errors are assumed to be normally distributed in any statistical analyses. As such, models were fitted to the data by minimizing an ordinary least-squares loss function of the errors using a quasi-Newton-Raphson procedure. The variable step size of Snoek *et al.* was applied[97], as well as Tikhonov regularisation.

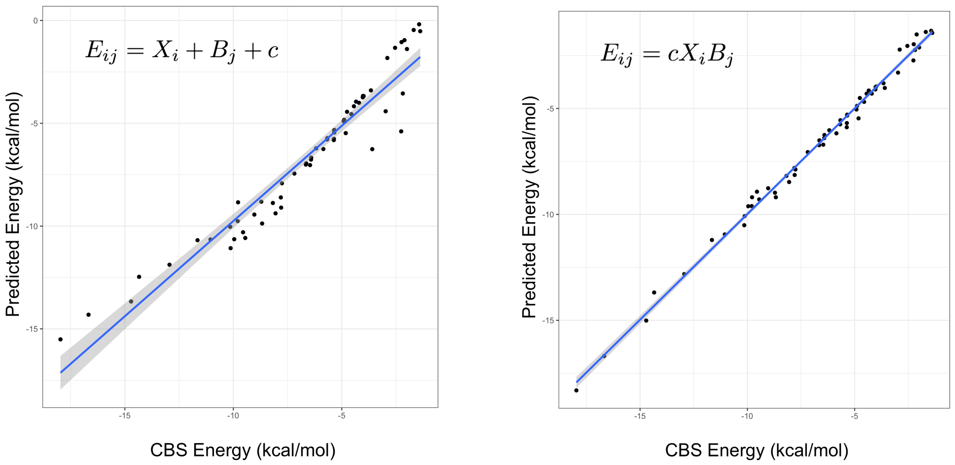

Figure 1: The predicted versus true interaction energies for the linear (left, equation 2) and $P0$ (right, equation 3) models. In the former, a non-linear trend is seen, suggesting non-normality of errors. The gradient and adjusted R^2 value of the line in the right-hand figure are 1.0 and 0.995, respectively. A perfect model would have unit gradient and zero intercept.

3 Results and discussion

In order to formulate a model for halogen bonds, systems of interest need to be selected and divided into two groups: fitting and validation sets. The criteria for which systems are to be considered are that they be tractable by the high-accuracy computational methods to be employed, and that they be representative of known halogen-bonded complexes. In practice, this restricts our pool of candidates to small molecules (less than 10 atoms) in the gas phase, many of which have been studied extensively using spectroscopy[7, 98–102].

For the fitting set, the halogen-bond donors were all chosen to be diatomics of the form AX, where A = H, F, Cl, or Br, and X= Cl, Br, or I. These have a broad range of electrostatic properties, with for example electric dipole moments ranging from weakly negative (from X to A, e.g. for HBr) to very strong positive dipole moments, as in FI. The σ-hole model intuitively predicts that the size of the positive hole on the halogen acting as the halogen-bond donor should be larger the more positive this dipole moment, and that more polarisable atoms (such as iodine compared to bromine) will have larger holes. The halogen bond acceptors (Lewis bases) were chosen to be H_2O, CH_2O, H_2S, CH_2S, HCN, and H_3N, covering the most commonly found acceptor atoms (O, N, and S) in different environments.

The validation set, on the other hand, was purposefully chosen to have a more diverse selection of systems. The halogen-bond donors were F_2, Cl_2, and CF_3X where X = Cl, Br, or I; crucially, the latter three are no longer diatomics. Similarly, the acceptors were larger, comprising methanol, ethene, oxirane, thiirane, and phosphine. Of particular note is the inclusion of a π-to-halogen bond, and a different acceptor atom in phosphorous. Complete basis set (CBS) limit CCSD(T)-F12b counterpoise-corrected interaction energies and geometries for all systems can be found in the ESI. In agreement with previous investigations[12, 88], the interaction energies are found to be sensitive to small changes in geometries, and also to the size of the basis set. In particular, correctly identifying the extent to which the AX bond length increases on complex formation is vital in accurately determining the interaction strength. Notably the geometries agree well with spectroscopic data where available, and the predictive rules of Legon[103].

3.1 Model fitting

From a statistical viewpoint, the two simplest models that could be suggested for the interaction energy, E_{ij}, between a halogen-bond donor with parameter X_i and acceptor with parameter B_j

involve either a linear or product combination:

$$E_{ij} = X_i + B_j + c \tag{2}$$

$$E_{ij} = cX_iB_j \tag{3}$$

where c is a real constant setting the energy scale, the latter being of the same form as equation 1. However, we do not fit the parameters by arbitrarily choosing a single halogen-bond donor and acceptor to have unit parameters, as was done by Legon and Millen[64]; instead, we use an unbiased fitting over all molecules in the fitting set, as described earlier. These parameters are purely statistically-fitted values, which can be found in the ESI, and we ascribe them no specific physical meaning. A more physically motivated model might also include a distance dependence. Defining $R_{e,ij}$ to be the equilibrium separation between the donating halogen atom and the accepting atom on the base, a simple Coulombic model would suggest a dependence on $R_{e,ij}^{-1}$, whereas if the interaction were dispersive in nature, the classical dependence would be $R_{e,ij}^{-6}$. This could be included by modifying either of equations 2 and 3 by multiplying by $R_{e,ij}^{-n}$ for some integer n, or by adding a weighted correction depending upon it.The former will be particularly important, and we define the Pn model as being that of the form

$$E_{ij}^{Pn} = \frac{cX_iB_j}{R_{e,ij}^{n}} \tag{4}$$

Thus equation 3 would be the $P0$ model. There is an infinity of other possibilities, including allowing for multiple parameters per molecule; this risks severely overfitting, however, given the fitting set only has approximately four points per molecule. The restriction of n to integer values is motivated by analogy to standard expressions for the potential energy of interactions between stationary multipoles; however, as discussed below, the removal of this restriction to then allow non-integer values of n would have little impact on the performance of the model. We should stress at this point that we are categorically not suggesting that these models describe a geometric *dependence* of the energy, lest we wrongly be accused of making fallacious statements. Rather it is the total interaction energy *at equlibrium* that is being described, with the strength mediated by the intermolecular separation. However, it is also incorrect to say that this is entirely independent of any physical dependence of the energy on the separation: as the parameters are fitted across a set of molecules, the dependence on R_e cannot simply be absorbed into the parameters X and B, and must represent an independent factor in the model.

An important indicator of the validity of a fitting procedure is the distribution of the residuals, or equivalently, the correlation between the predicted and actual values. In Figure 1, the predicted vs. actual energies are plotted, demonstrating that the $P0$ model is a much better fit than equation 2. Crucially, it appears to abide by the assumptions of the fitting procedure, namely the assumption of normality of errors. This was not the case for the linear model, or any model with an added (rather than multiplied) R_e^{-n} correction term. The fitted parameters under both models can be found in the ESI. Summary statistics for several different models are given in Table 1, showing that the product models have by far the lowest errors and the most efficient use of information, as quantified by the Akaike information criterion[104]. In particular, a simple weighted dispersion model, i.e. $E_{ij} = kR_{e,ij}^{-6}$ model where k is optimised as a parameter fixed across all molecules, does not perform well.

3.2 Principal component analysis

The significance of the $P0$ model can be understood in its relation to a principal component analysis, a widely-used machine-learning technique for dimensionality reduction. In this case, if $\mathbf{E}$ is an $M \times N$ matrix of interaction energies E_{ij}, then a principal component analysis takes the form of a singular value decomposition:

$$\mathbf{E} = \mathbf{u}\Lambda\mathbf{v}^T$$

Table 1: Summary statistics for the linear, $P0$, $kR_{e,ij}^{-6}$, and $P4$ models. These include the root-mean-square, maximum, mean-signed, and mean-absolute errors in kcal mol^{-1}, and the Akaike information criterion (AIC). The latter statistic roughly equates to whether the increase in complexity prescribed by adding the parameters is justified in relation to the amount of data supplied[104]; a small number indicates the model is 'efficient' in its use of data. The AIC is not quoted for the kR^{-6} model, as it would not be directly comparable to the others.

Model	RMSE	Max.	MSE	MAE	AIC
Linear	1.13	3.13	0.00	0.75	204
$P0$	0.30	0.68	-0.01	0.24	35
kR^{-6}	2.99	7.11	0.50	2.32	—
$P4$	0.14	0.41	0.00	0.11	14

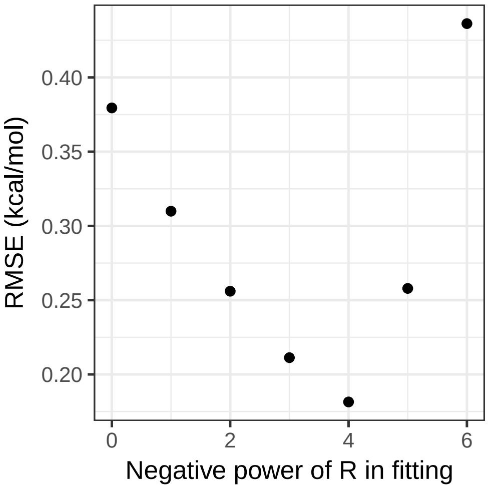

Figure 2: Root-mean-square error over the fitting set in kcal mol^{-1} for the Pn models, as a function of n.

where Λ is a diagonal matrix of N singular values (or 'components') λ_i, while $\mathbf{u}$ and $\mathbf{v}$ are $M \times N$ and $N \times N$ matrices of component vectors. In this way, any element of $\mathbf{E}$ can be written as

$$E_{ij} = \sum_{k=1}^{N} \lambda_k u_{ik} v_{kj} \tag{5}$$

If the principal component, λ_1, is much greater than all the other components, then we see that the sum in equation 5 simply reduces to the $P0$ model in equation 3, where $c = \lambda_1$, $X_i = u_{i1}$, and $B_j = v_{1j}$.

Performing this analysis on the matrix of interaction energies gives the principal component as being roughly 30 times larger than the second component, explaining 99.3 percent of the variance in the energies. A further 0.6 percent is explained by including the second component, with all further components being negligible. This explains both why the simple product model is strikingly successful, and a potential way to improve it by adding a second component in equation 5. Moreover, it provides an easy way to parametrise Pn models with $n > 0$, by forming a matrix with values $E_{ij} R_{e,ij}^{n}$ and performing a singular value decomposition. Figure 2 shows how the root-mean-squared error for the fitting set varies with n. Clearly, $n = 4$ provides the best results, and as can be seen in Table 1, is a substantial improvement on the $P0$ model, achieving accuracy beyond what can be achieved by using density-functional theory, as will be discussed shortly. As the $P4$ model has an RMSE of only 0.14 kcal mol^{-1} and Figure 2 demonstrates that the relationship between the error and the value of n is clearly discontinuous, attempting to include non-integer values of n would not substantially improve the model and would sacrifice simplicity, hence it has not been pursued.

While including a distance dependence clearly improves accuracy, it also introduces complications. Firstly, the functional form in equation 4 simply cannot describe geometries far removed from equilibrium, as it would diverge to negative infinity at short distances. Similarly, only $n = 6$ would correctly recover the expected long-range behaviour. Thus a more complete analysis over a range of displacements would reveal a changing dependence on n, with attractive and repulsive terms as seen for example in a Lennard-Jones potential. The second issue is that including a distance dependence requires an estimate for the separation to be available. However, in applications where a simple model such as this would be most useful, i.e. in estimating the strength of an interaction, the distance dependence is important and an estimate for the separation is readily available. We note that a small error in R_e of δ percent can easily be shown to give an error of roughly $n\delta$ percent in the Pn interaction energy, which for $n = 4$ remains small. Moreover, as can be seen from Figures 1 and 2, the $P0$ model still performs very well with an RMSE of 0.30 kcal mol^{-1}, and so could be used when no value for R_e was available.

3.3 Validation and comparison with other methods

The validation set comprises the five new halogen-bond donors described above paired with all six original acceptors, and the five new acceptors paired with the ten original donors. As such it constitutes a larger set (80 systems, as opposed to 60 in the fitting set). However, during the course of our investigations it became apparent that some of the systems behave markedly differently to any of the others. Specifically, those involving FCl, FBr, and FI interacting with phosphine and thiirane. These exceptional cases have been discussed elsewhere[44, 53], and were excluded from the validation set as their errors across all methods were over an order of magnitude larger than for any other systems. This was true also for the density functionals considered, both of which significantly under-bound the complexes, by as much as 8 kcal mol^{-1} in the case of FCl$\cdots$PH$_3$.

The relevant X_i and B_j parameters for all new molecules were found by calculating the CBS-limit CCSD(T)-F12b energy for the interaction with water for the halogen-bond donors and of the acceptors with BrI. These energies were then divided through by the known parameter (X_i or B_j) and the calculated $R_{e,ij}^{-n}$. This unnaturally results in ten systems having zero error, and as such these data were excluded from the subsequent error analysis. The bond lengths in the model for the remaining systems were taken to be those calculated using M06-2X, so as to give a fair and consistent comparison with later results where CCSD(T) level calculations are computationally intractable.

The error distributions for the $P4$ model over both the fitting and validation sets, along with those calculated using the M06-2X and ωB97xD density functionals at the aVTZ level, are shown in Figure 3, with the equivalent plot for the $P0$ model given in the ESI. These functionals were chosen as previous benchmarks have shown them to be particularly good for halogen bonding interactions[88]. From the figure, however, we see that the simple product model is performing at least as well, if not better than, even these functionals. In particular, both the fitting and validation data are centred around zero residual error, indicating no systematic bias, which is in contrast to the two functionals, which systematically over- and underestimate the energies slightly, for M06-2X and ωB97xD respectively. Moreover, the overall spread is narrow, and mostly concentrated around zero, staying consistently within nominal "chemical accuracy" of 1 kcal mol^{-1}. This is opposed to M06-2X, which shows a much more protruded density, significantly overpredicting some energies.

The mean-absolute errors for the validation set with the $P4$ model, M06-2X and ωB97xD are 0.28, 0.36, and 0.30 kcal mol^{-1}. These are all broadly similar, but it should be noted that a combination of Shapiro-Wilk and Kolmogorov-Smirnov tests[104] indicate that the error distributions for each are normally distributed, but drawn from distinct distributions, with $p < 0.01$ in each pairwise comparison. For reference, MP2/aVTZ results gave an MAE of 0.77 kcal mol^{-1}, almost three times that of the product model.

The performance of the statistical model is astonishing, as it gives better accuracy than high-

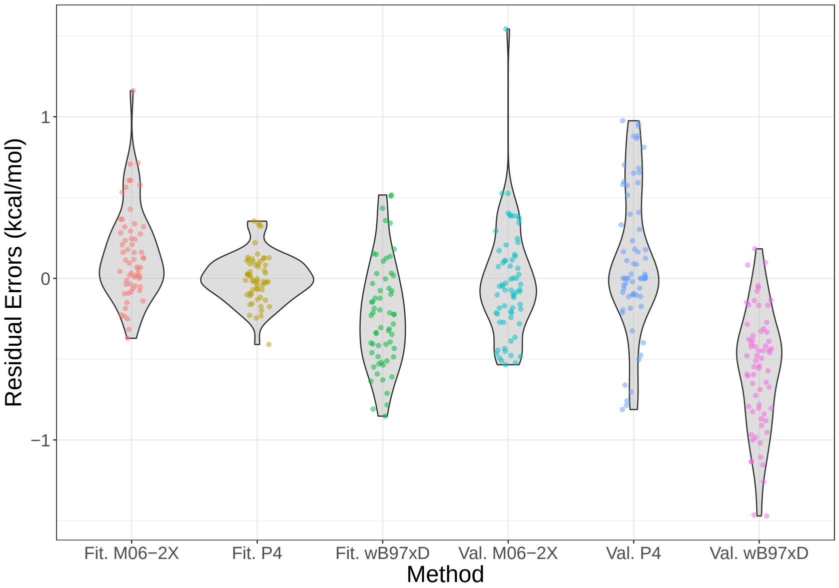

Figure 3: Violin plots of the error distributions of the *P*4 model, M06-2X/aVTZ, and ωB97xD/aVTZ, compared to CCSD(T)-F12b/CBS results. The model is split into data from the fitting (Fit.) and validation (Val.) sets. The shape of the violin shows where the density of errors is concentrated - i.e. the frequency with which errors are found in a small interval - such that an ideal distribution would be a very short, wide density centred on the origin. Note that the density is plotted symmetrically about the vertical axis, and the horizontal scale is relative (so it is the same for all the violins); the total area of a violin integrates to the number of points, the width representing a proportion of the total number. The individual data points have also been plotted, with a small amount of jitter added in the horizontal direction to aid visibility.

level quantum chemical methods at a fraction of the cost. For any new complex of interest, the relevant parameters can be determined from a single calculation with a reference molecule (water or BrI), and then reused in all other contexts.

Table 2: The energies for each pair of new halogen-bond acceptor and donor are given at the M06-2X/aVTZ level, along with the energies predicted by the model, in kcal/mol.

	C_6F_5Cl		C_6F_5Br		C_6F_5I	
	M06	Pred.	M06	Pred.	M06	Pred.
Sulphox.	−3.32	−4.06	−4.61	−4.95	−5.96	−6.48
Glycine	−2.49	−3.14	−3.66	−3.83	−5.18	−5.01
Valine	−3.75	−3.68	−4.58	−4.49	−6.22	−5.87
Leucine	−5.12	−4.04	−4.97	−4.94	−6.45	−6.46

To test this, calculations were performed on considerably larger molecules than those in the fitting or validation set, where using the high-level coupled cluster method would be unfeasible. Based on M06-2X producing an error distribution that is much more centred around zero for the validation set than ωB97xD (see Figure 3), M06-2X/aVTZ calculations were performed for the halogen-bond acceptors sulphoximine, glycine, valine, and leucine, and the donors C_6F_5X with $X = Cl$, Br, and I. The interacting atom on the acceptors were the nitrogen in sulphoximine and the carbonyl oxygen on the amino acids; geometries can be found in the ESI. The parameters for the model were determined with respect to the reference molecules at the same level of theory. Table 2 shows the results of these tests. Despite being extrapolated to calculations with different systems, not involving any of the original fitting data, the mean-absolute deviation is 0.49 kcal mol^{-1}. The mean-absolute deviation of the $P4$ model from the M06-2X results across the fitting and validation sets is 0.48 kcal mol^{-1}, suggesting that similar error levels have been maintained despite the significant increase in molecule size. The $P4$ model therefore represents a rapid and accurate approach to predicting the interaction energies of halogen-bonded systems.

3.4 The nature of the halogen bond

The principal component analysis has allowed for greater insight into the mechanics behind the product model, and for elucidation of the distance dependence of the interactions. It also suggests that a method to improve the performance of the model further would be to include the second component, λ_2, in equation 5:

$$E_{ij} \approx \lambda_1 X_i B_j + \lambda_2 u_{i2} v_{2j}$$

This is impractical for two main reasons: it would double the number of parameters, which we have seen leads to severe overfitting; it would complicate the determination of new parameters, as two reference calculations would be needed and a system of linear equations would have to be solved. However, the instances where the second component is important could serve as an indicator as to which systems behave differently to the norm.

To this end, symmetry-adapted perturbation theory (SAPT) calculations were carried out, providing a decomposition of the interaction energies in terms of the physically relevant quantities of electrostatics, exchange, induction, and dispersion. Charge transfer can be separated out from the induction energy[51, 105], as has been seen to be important for the phosphine systems[44], but we do not do that here as the best approach to doing so is not clear. The energy contributions for each system in the fitting set are given in the ESI, along with figures illustrating each component as a percentage of the total interaction. Figure 4 shows how the percentage error in predictions from the $P4$ model compares with the relative importance of induction and dispersion in the SAPT decomposition of the interaction energy, with the results split by halogen-bond donor to show how trends in each quantity correlate. The induction and dispersion terms are both presented as ratios relative to the SAPT electrostatic term. It is immediately apparent that complexes where the model displays the largest percentage errors

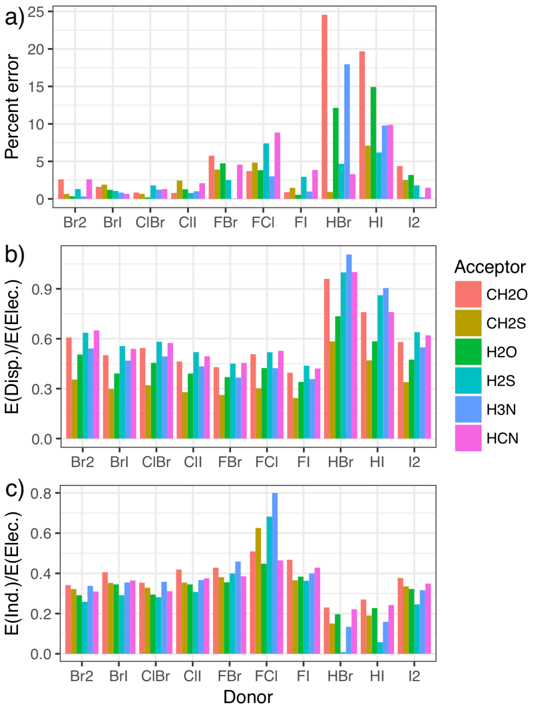

Figure 4: The error (relative to the CCSD(T)-F12b/CBS values) as a percentage of the overall interaction energy for the $P4$ model (a) compared with the ratio of the dispersion (b) and induction (c) contributions to the electrostatic component of the symmetry-adapted perturbation theory of the energy.

relative to the CCSD(T)-F12b/CBS data (Figure 4a) are those with significantly increased relative dispersion (Figure 4b) or induction (Figure 4c) contributions. An increase in induction (potentially charge transfer) also appears to be concomitant with a decrease in dispersion, and vice-versa. Perhaps most interestingly, it is the halogen-bond donors HBr and HI that show the largest percentage errors, and consequently the largest proportion of dispersion along with the smallest proportion of induction. This suggests that these interactions are predominantly dispersive rather than electrostatic, in line with what we would intuitively expect given the relative electronegativities of hydrogen and the halogen atoms.

Additionally, the induction contribution shown in Figure 4c) noticeably increases for the F–X donors, peaking for F–Cl. This is in agreement with trends noted for substantial charge transfer, namely the switching of the mode of binding to a Mulliken inner complex. This is again accompanied by a decrease in dispersion, and a pronounced increase in the errors from the simple $P4$ model. In both cases, the inclusion of the second component in the model almost entirely corrects for these differences, as can be seen in Figure 5. The second component reduces the strength of the interaction in situations where there is a large induction contribution (such as FCl in the top left), and increases the strength for those with a small proportion of induction (HI and HBr in the bottom right). Recalling that Figure 4 shows that a decrease in induction correlates with an increase in dispersion, this indicates that the single component $P4$ model underestimates the strength of interactions with a large dispersion contribution.

Moreover, it suggests that far from induction and dispersion being unimportant for the other systems, it is more that when combined they are of similar enough magnitude to one another that these effects are included in the fitting process. Inclusion of the second component in the model reduces the RMSE of the fitting data from 0.14 to 0.02 kcal mol^{-1}, a modest dimensionality reduction from six components to two. For practical use of the model, inclusion of these effects is irrelevant. The significance comes from the utility of deviation from the model in categorising the physical nature of the halogen bond. In particular, whether that deviation is an under- or overestimation of the interaction, or equivalently the importance of the second component,

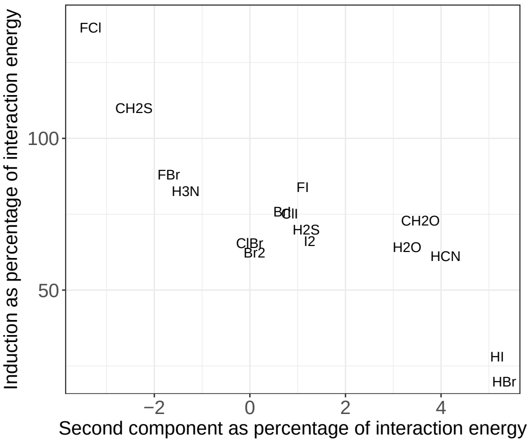

Figure 5: Comparison of induction energy with the energy due to the second component in the principal component analysis, each as a percentage of the total interaction energy, averaged over all systems containing the given molecule in the fitting set. Most systems fall in the middle, but those with larger dispersion (bottom right) or induction (top left) show a marked increase in the importance of the second component to the predicted energy.

indicates when a complex has changed from a ‘typical’ halogen bond, where electrostatics is seen to dominate, to one that is dispersive or induction (possibly charge transfer) based, respectively. It thus has the potential to provide insight with minimal effort.

To further demonstrate the simple model halogen bond and how it can be used, we have prepared an interactive Jupyter notebook that is available as part of the ESI and via GitHub.[106] This includes a walkthrough of a simplified version of the fitting and analysis of the model, and an example of how parameters for new halogen-bond donors and acceptors can be found. In addition to acting as an explanation of the analysis in the present investigation, it is intended that the Jupyter notebook could also act as template for attempting to find a simple statistical model for other types of σ-hole based interactions, such as chalcogen bonds.

4 Conclusions

We have presented a statistical model for the interaction energy of halogen-bonded systems that takes the simple form $X_iB_j/R^4_{e,ij}$ (denoted $P4$), where X_i and B_j are parameters for the halogen-bond donor and acceptor, while $R_{e,ij}$ is the separation between the two molecules. Using a regularised least-squares regression this model was fitted to benchmark quality data from the high-accuracy CCSD(T)-F12b method extrapolated to the complete basis set limit, for a set of 60 halogen-bonded complexes. Various alternative models were tested, but product models gave the best results. The mean-absolute and maximum errors in the calculated halogen-bond interaction energy over the fitting set for $P4$ were 0.11 and 0.41 kcal mol^{-1}, respectively. This represents greater accuracy than the M06-2X and ωB97xD density functionals, and is also the case when extended to 74 validation systems not in the original fitting set. The ease of parametrisation and speed of prediction inherent to using the product model make it potentially very useful for the rapid evaluation of interactions in, for example, atomistic simulations using force fields where quantum-chemical methods are unfeasible. Most promisingly, when extended to much larger and completely new complexes using a method (M06-2X) that is much less expensive than CCSD(T)-F12b, accuracy was maintained relative to the density-functional theory calculation, achieving root-mean-square deviations of less than half a kilocalorie per mole.

The effectiveness of the model was shown to be equivalent to the variance in the data being well explained by a single, principal component. In this way, we were able to determine that a distance dependence of R^{-4} best describes the interactions. Moreover, the cases where a second component became substantial were found to correlate with increases in dispersion or induction contributions to the energy. These correspond to under- and overestimation of the interaction energy by the principal component, respectively, and thus provide an indicator for changes in the underlying physical nature of the halogen bond. As σ-holes have been identified as playing a role in intermolecular interactions involving other p-block elements, such as chalcogens, pnicogens and tetrels, it is plausible that the applicability of the simple model is not restricted to halogen bonds. The current approach could easily be applied, perhaps elucidating both similarities and differences between many classes of non-covalent interaction. Particularly interesting would be to extend the analysis to off-equilibrium geometries, potentially leading to a simple model for the geometric dependence of the interaction strength. The emphasis is on the simplicity of the approach.

Acknowledgements

The authors thank Professor Fred Manby for helpful conversations and Professor Lee Brammer for comments on an earlier draft of the manuscript. We also thank the Engineering and Physical Sciences Research Council (UK) for a postgraduate studentship awarded to R.A.S.

References

(1) Colin, J. J.; Gaultier de Claubry, H. *Ann. Chim.* **1814**, *90*, 87–100.

(2) Colin, J. J. *Ann. Chim.* **1814**, *91*, 252–272.

(3) Guthrie, F. *J. Chem. Soc.* **1863**, *16*, 239.

(4) Benesi, H. A.; Hildebrand, J. H. *J. Am. Chem. Soc.* **1949**, *71*, 2703–2707.

(5) Hassel, O.; Rømming, C. *Q. Rev. Chem. Soc.* **1962**, *16*, 1.

(6) Hassel, O. *Science* **1970**, *170*, 497–502.

(7) Legon, A. C. *Angew. Chem. Int. Ed.* **1999**, *38*, 2686–2714.

(8) Legon, A. C. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7736–47.

(9) Cavallo, G.; Metrangolo, P.; Milani, R.; Pilati, T.; Priimagi, A.; Resnati, G.; Terraneo, G. *Chem. Rev.* **2016**, *116*, 2478–2601.

(10) Beale, T. M.; Chudzinski, M. G.; Sarwar, M. G.; Taylor, M. S. *Chem. Soc. Rev.* **2013**, *42*, 1667–1680.

(11) Aakeröy, C. B. et al. *Faraday Discuss.* **2017**, *203*, 227–244.

(12) Hill, J. G.; Legon, A. C. *Phys. Chem. Chem. Phys.* **2015**, *17*, 858–867.

(13) Ouvrard, C.; Le Questel, J.-Y.; Berthelot, M.; Laurence, C. *Acta Cryst. B* **2003**, *59*, 512–526.

(14) Politzer, P.; Murray, J. S.; Lane, P. *Int. J. Quantum Chem.* **2007**, *107*, 3046–3052.

(15) Mukherjee, A.; Tothadi, S.; Desiraju, G. R. *Acc. Chem. Res.* **2014**, *47*, 2514–2524.

(16) Brammer, L. *Chem. Soc. Rev.* **2004**, *33*, 476.

(17) Robertson, C. C.; Wright, J. S.; Carrington, E. J.; Perutz, R. N.; Hunter, C. A.; Brammer, L. *Chem. Sci.* **2017**, *8*, 5392–5398.

(18) Brammer, L. *Faraday Discuss.* **2017**, *203*, 485–507.

(19) Nunes, R.; Vila-Viçosa, D.; Machuqueiro, M.; Costa, P. J. In *Proceedings of MOL2NET 2017, International Conference on Multidisciplinary Sciences, 3rd edition*, MDPI: Basel, Switzerland, 2017, p 5075.

(20) Montaña, Á. M. *ChemistrySelect* **2017**, *2*, 9094–9112.

(21) Lu, Y.; Shi, T.; Wang, Y.; Yang, H.; Yan, X.; Luo, X.; Jiang, H.; Zhu, W. *J. Med. Chem.* **2009**, *52*, 2854–2862.

(22) Auffinger, P.; Hays, F. A.; Westhof, E.; Ho, P. S. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 16789–16794.

(23) Sirimulla, S.; Bailey, J. B.; Vegesna, R.; Narayan, M. *J. Chem. Inf. Model.* **2013**, *53*, 2781–2791.

(24) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. *J. Mol. Model.* **2007**, *13*, 291–6.

(25) Wang, W.; Ji, B.; Zhang, Y. *J. Phys. Chem. A* **2009**, *113*, 8132–8135.

(26) Scheiner, S. *Int. J. Quantum Chem.* **2013**, *113*, 1609–1620.

(27) Scheiner, S. *Acc. Chem. Res.* **2013**, *46*, 280–288.

(28) Legon, A. C. *Physical Chemistry Chemical Physics* **2017**, *19*, 14884–14896.

(29) Stevens, E. D. *Mol. Phys.* **1979**, *37*, 27–45.

(30) Stewart, R. F. *Chem. Phys. Lett.* **1979**, *65*, 335–342.

(31) Politzer, P.; Murray, J. S.; Clark, T. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7748.

(32) Alkorta, I.; Elguero, J.; Del Bene, J. E. *J. Phys. Chem. A* **2014**, *118*, 4222–4231.

(33) Murray, J. S.; Macaveiu, L.; Politzer, P. *J. Comput. Sci.* **2014**, *5*, 590–596.

(34) Kolár, M.; Hostaš, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2014**, *16*, 9987–9996.

(35) Karpfen, A In, Metrangolo, P., Resnati, G., Eds.; Springer: Berlin, Heidelberg, 2008, pp 1–15.

(36) Bundhun, A.; Ramasami, P.; Murray, J. S.; Politzer, P. *J. Mol. Model.* **2013**, *19*, 2739–46.

(37) Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2008**, *4*, 232–242.

(38) Wang, C.; Danovich, D.; Mo, Y.; Shaik, S. *J. Chem. Theory Comput.* **2014**, *10*, 3726–3737.

(39) Politzer, P.; Riley, K. E.; Bulat, F. A.; Murray, J. S. *Comput. Theor. Chem.* **2012**, *998*, 2–8.

(40) Politzer, P.; Murray, J. S.; Clark, T. *Phys. Chem. Chem. Phys.* **2013**, *15*, 11178–89.

(41) Politzer, P.; Murray, J. S. *Crystals* **2017**, *7*, 212.

(42) Anderson, L. N.; Aquino, F. W.; Raeber, A. E.; Chen, X.; Wong, B. M. *J. Chem. Theory Comput.* **2018**, *14*, 180–190.

(43) Thirman, J.; Engelage, E.; Huber, S. M.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2018**, *20*, 905–915.

(44) Shaw, R. A.; Hill, J. G.; Legon, A. C. *J. Phys. Chem. A* **2016**, *120*, 8461–8468.

(45) Hill, J. G.; Hu, X. *Chem. Eur. J.* **2013**, *19*, 3620–3628.

(46) Stone, A. J. *J. Am. Chem. Soc.* **2013**, *135*, 7005–9.

(47) Desiraju, G. R.; Ho, P. S.; Kloo, L.; Legon, A. C.; Marquardt, R.; Metrangolo, P.; Politzer, P.; Resnati, G.; Rissanen, K. *Pure Appl. Chem.* **2013**, *85*, 1711–1713.

(48) Lommerse, J. P. M.; Stone, A. J.; Taylor, R.; Allen, F. A. *J. Am. Chem. Soc.* **1996**, *118*, 3108–3116.

(49) Riley, K. E.; Murray, J. S.; Fanfrlík, J.; Rezáč, J.; Solá, R. J.; Concha, M. C.; Ramos, F. M.; Politzer, P. *J. Mol. Model.* **2013**, *19*, 4651–9.

(50) Riley, K. E.; Hobza, P. *Phys. Chem. Chem. Phys.* **2013**, *15*, 17742–17751.

(51) Stone, A. J. *J. Phys. Chem. A* **2017**, *121*, 1531–1534.

(52) Mulliken, R. S.; Person, W. B., *Molecular Complexes: A Lecture and Reprint Volume*; Wiley-Interscience: New York, 1969.

(53) Hill, J. G. *Phys. Chem. Chem. Phys.* **2014**, *16*, 19137–40.

(54) Řezáč, J.; de la Lande, A. *Phys. Chem. Chem. Phys.* **2017**, *19*, 791–803.

(55) Del Bene, J.; Alkorta, I.; Elguero, J. *Molecules* **2017**, *22*, 1955.

(56) Khanifaev, J.; Peköz, R.; Konuk, M.; Durgun, E. *Phys. Chem. Chem. Phys.* **2017**, *19*, 28963–28969.

(57) Rosokha, S. V.; Neretin, I. S.; Rosokha, T. Y.; Hecht, J.; Kochi, J. K. *Heteroat. Chem.* **2006**, *17*, 449–459.

(58) Mustoe, C. L.; Gunabalasingam, M.; Yu, D.; Patrick, B. O.; Kennepohl, P. *Faraday Discuss.* **2017**, *203*, 79–91.

(59) Grimme, S. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 211–228.

(60) Burns, L. A.; Mayagoitia, A. V.; Sumpter, B. G.; Sherrill, C. D. *J. Chem. Phys.* **2011**, *134*, 084107.

(61) Riley, K. E.; Pitonak, M.; Jurečka, P.; Hobza, P. *Chem. Rev.* **2010**, *110*, 5023–5063.

(62) Riley, K. E.; Hobza, P. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 3–17.

(63) Grimme, S.; Hansen, A.; Brandenburg, J. G.; Bannwarth, C. *Chem. Rev.* **2016**, *116*, 5105–5154.

(64) Legon, A. C.; Millen, D. J. *J. Am. Chem. Soc.* **1987**, *109*, 356–358.

(65) Legon, A. C. *Phys. Chem. Chem. Phys.* **2014**, *16*, 12415–12421.

(66) Alkorta, I.; Legon, A. C. *Molecules* **2017**, *22*, 1786.

(67) Adler, T. B.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2007**, *127*, 221106.

(68) Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 242–253.

(69) Werner, H.-J. et al. MOLPRO, version 2012.1, a package of ab initio programs., Cardiff, UK, 2012.

(70) Ten-no, S. *Chem. Phys. Lett.* **2004**, *398*, 56–61.

(71) Hättig, C.; Tew, D. P.; Köhn, A. *J. Chem. Phys.* **2010**, *132*, 231102.

(72) Peterson, K. A.; Adler, T. B.; Werner, H.-J. *J. Chem. Phys.* **2008**, *128*, 084102.

(73) Hill, J. G.; Peterson, K. A. *J. Chem. Phys.* **2014**, *141*, 094106.

(74) Peterson, K. A.; Figgen, D.; Goll, E.; Stoll, H.; Dolg, M. *J. Chem. Phys.* **2003**, *119*, 11113.

(75) Peterson, K. A.; Shepler, B. C.; Figgen, D.; Stoll, H. *J. Phys. Chem. A* **2006**, *110*, 13877–13883.

(76) Hill, J. G.; Peterson, K. A.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2009**, *131*, 194105.

(77) Weigend, F. *Phys. Chem. Chem. Phys.* **2002**, *4*, 4285–4291.

(78) Weigend, F. *J. Comput. Chem.* **2008**, *29*, 167–175.

(79) Hättig, C. *Phys. Chem. Chem. Phys.* **2005**, *7*, 59–66.

(80) Valeev, E. F. *Chem. Phys. Lett.* **2004**, *395*, 190–195.

(81) Knizia, G.; Werner, H.-J. *J. Chem. Phys.* **2008**, *128*, 154103.

(82) Yousaf, K. E.; Peterson, K. A. *J. Chem. Phys.* **2008**, *129*, 184108.

(83) Shaw, R. A.; Hill, J. G. *J. Chem. Theory Comput.* **2017**, *13*, 1691–1698.

(84) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.

(85) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(86) Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615.

(87) Frisch, M. J. et al. Gaussian 09 Revision D.01.

(88) Kozuch, S.; Martin, J. M. L. *J. Chem. Theory Comput.* **2013**, *9*, 1918–1931.

(89) Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.* **1994**, *94*, 1887–1930.

(90) Szalewicz, K. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 254–272.

(91) Parker, T. M.; Burns, L. A.; Parrish, R. M.; Ryno, A. G.; Sherrill, C. D. *J. Chem. Phys.* **2014**, *140*, 094106.

(92) Bukowski, R et al. SAPT2012: An *Ab Initio Program for Many-Body Symmetry-Adapted Perturbation Theory Calculations of Intermolecular Interaction Energies.*, University of Delaware and University of Warsaw, (accessed Feb 9, 2018), 2012.

(93) Hohenstein, E. G.; Sherrill, C. D. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 304–326.

(94) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(95) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

(96) Dunning, T. H.; Peterson, K. A.; Wilson, A. K. *J. Chem. Phys.* **2001**, *114*, 9244–9253.

(97) Snoek, J.; Larochelle, H.; Adams, R. P. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2012, pp 2951–2959.

(98) Legon, A. C.; Thumwood, J. M. A.; Waclawik, E. R. *J. Chem. Phys.* **2000**, *113*, 5278.

(99) Stephens, S. L.; Walker, N. R.; Legon, A. C. *J. Chem. Phys.* **2011**, *135*, 224309.

(100) Davey, J. B.; Legon, A. C. *Phys. Chem. Chem. Phys.* **2001**, *3*, 3006–3011.

(101) Davey, J. B.; Legon, A. C.; Waclawik, E. R. *Phys. Chem. Chem. Phys.* **2000**, *2*, 1659–1665.

(102) Legon, A. C.; Thumwood, J. M. A. *Phys. Chem. Chem. Phys.* **2001**, *3*, 2758–2764.

(103) Legon, A. C. *Struct Bond* **2008**, *126*, 17–64.

(104) Cox, D. R.; Hinkley, D. V., *Theoretical Statistics*, 1st ed.; Chapman and Hall Press: London, 1979.

(105) Stone, A. J.; Misquitta, A. J. *Chem. Phys. Lett.* **2009**, *473*, 201–205.

(106) Shaw, R. A.; Hill, J. G. A simple model for halogen bonds Jupyter notebook., (accessed Jun 10, 2018).