Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Intrinsic Bond Strength Index as a halogen bond interaction energy predictor†

Ona Šivickytė,[a] and Paulo J. Costa,[a*]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Halogen bonds (XBs) have become increasingly popular over the past few years with numerous applications in catalysis, material design, anion recognition, and medicinal chemistry. To avoid a *post factum* rationalization of XB trends, descriptors can be tentatively employed to predict the strength of potential halogen bonds. These typically comprise the electrostatic potential maximum at the tip of the halogen, $V_{S,max}$, or properties based on the topological analysis of the electronic density. However, such descriptors either can only be used with confidence for specific families of halogen bonds or require intense computations and, therefore, are not particularly attractive for large datasets with diverse compounds or biochemical systems. Therefore, the development of a simple, widely applicable, and computationally feasible descriptor remains a challenge as it would facilitate the discovery of new XB applications while also improving the existing ones. Recently, the Intrinsic Bond Strength Index (IBSI) has been proposed as a new tool to evaluate any bond strength, however, it has not been extensively explored in the context of halogen bonding. In this work, we show that IBSI values linearly correlate with the interaction energy of diverse sets of halogen-bonded complexes and therefore, can be used to quantitatively predict halogen bond strength. The linear fit models based on quantum-mechanics-based electron density provided MAEs typically below 1 kcal mol^{-1}. Moreover, we also explored the exciting possibility to use a promolecular density approach (IBSI^{PRO}), which only requires the complex geometry as an input which is computationally cheap. Surprisingly, the performance was comparable to the QM-based methods, thus opening the door for the usage of IBSI^{PRO} as a fast, yet accurate, XB strength descriptor in large datasets but also in biomolecular systems such as protein–ligand complexes.

1 Introduction

A halogen bond (XB) is a directional non-covalent interaction between a Lewis base (B), *e.g.*, the lone pairs on an N-, O-containing molecule, and a halogen atom (X) in a molecular entity acting as a Lewis acid [1–3]. Indeed, while typically halogens are perceived as electron-rich electronegative species behaving as nucleophiles, the picture is more complicated when they are covalently bound to another atom (R–X) as the electrons are anisotropically distributed, forming regions of higher and lower electron density (ED). The region of lower ED located at the tip of the halogen opposite to the covalent bond corresponds to the so-called σ-hole [4]. This site is typically electropositive and can interact with nucleophiles, thus offering an electrostatic explanation for the formation of XBs (R–X$\cdots$B). A larger polarizability of X corresponds to a larger σ-hole and consequently to a stronger XB, and therefore, the XB strength typically increases along the halogen series: Cl < B < I [5,6]. A seemingly opposing view describes XBs as charge-transfer (CT) complexes explained by the existence of electron transfer from a filled donor orbital of the Lewis base to the accepting R–X σ^* orbital of the halogenated molecule [3,7,8], following the same trends as mentioned above. However, while both views might reveal different sides of a dual XB nature [9], it has been argued that both essentially describe the same phenomenon [10,11] or that CT is practically negligible for the overall interaction [12].

XBs are seen as hydrophobic counterparts for hydrogen bonds (HBs), but they are often considered to be more versatile [13] since halogen atoms can act as both a Lewis base (HB acceptor) and a Lewis acid (XB donor). This versatility also arises from their directionality and tunability as the XB length, the R–X$\cdots$B angle, and the magnitude of the σ-hole largely depend on the halogen, the existence of other substituents on the XB donor, and the nature of the Lewis base [13–15]. All these factors can easily be chosen or adjusted to meet a set of unique specifications thus, XBs span a wide range of interaction energies [13,16,17]. In principle, such factors could be represented by a combination of descriptors of

[a] *BioISI - Instituto de Biosistemas e Ciências Integrativas, Faculdade de Ciências, Universidade de Lisboa, 1749-016, Lisboa, Portugal; E-mail: pjcosta@ciencias.ulisboa.pt*

† Electronic Supplementary Information (ESI) available: Supporting Figures and Tables and Outlier removal discussion. See DOI: 10.1039/cXCP00000x/

electronic and/or electrostatic effects and, therefore, these could be used to estimate the strength of XBs[7,18]. However, it is also admitted that the applicability of XBs is often rationalized *post factum* as it remains a challenge to accurately predict the outcomes and their strength in complex systems (*e.g.* protein-ligand complexes) and thus, we are still far from taking full advantage of XBs in the rational design of new systems[19–21], even though various potential applications are constantly emerging[22–25]. There have been attempts to overcome this challenge and provide a solid basis for designing new halogen-bonded structures[8,21,26–28], but accurate modelling of XBs is still not straightforward. This issue is paramount given the increased attention put on XBs and and their broad application in catalysis[19,29–32], material design[33–35], supramolecular[36–38] and medicinal[39–41] chemistry, among other areas.

Several approaches allow the estimation of the XB strength and this topic is tightly related to the discussion regarding the nature of this interaction and the importance of various bonding components to the overall XB stability[12,42]. The most commonly employed XB interaction energy descriptor is the electrostatic potential maximum at the tip of the halogen ($V_{S,max}$), *i.e.* the magnitude of the σ-hole. This simple descriptor encompasses only the electrostatic component of XBs and does not always adequately predict the interaction strength[43–45]. This occurs mainly due to its static nature as it is computed in the absence of the base, thus neglecting the contribution from the XB acceptor. It can be corrected by adding polarization to the static $V_{S,max}$, evaluating its magnitude in the presence of a negative point charge, yielding an extended electrostatic model[46–48]. It has also been proposed that the minima of the local attachment energy, analogous to the average ionization energy but reflecting the susceptibility towards nucleophiles, can be used to complement $V_{S,max}$ or used as an independent descriptor to predict XB energies in methyl- and aryl halides[49]. Alternatively, some authors approached the incompleteness of the $V_{S,max}$ by combining it with CT descriptors such as the charge transfer energy[45,50] or the C–X σ^* orbital energy[8], often leading to improved XB energy predictions[8]. Other attempts to predict XB strength include the usage of ED properties such as the kinetic, potential, and total energy density[51,52], also its Laplacian and curvature[53] evaluated at the bond critical point.

All the mentioned approaches typically require *ab initio* or DFT calculations to obtain the descriptors and therefore, could be computationally demanding, hindering their application in large datasets or large molecules such as protein-ligand systems[54]. Machine-learning (ML) approaches could offer an alternative, as highlighted by a statistical model trained against high–accuracy *ab initio* calculations, which depends on only two fitted parameters along with the equilibrium distance. This model, whose computational cost is negligible, outperforms some of the best density functionals[55]. However, the physical interpretation of fitted ML parameters is often not straightforward.

The above considerations indicate that the need to develop more straightforward and easily accessible XB energy estimators persists. In this context, the Intrinsic Bond Strength Index (IBSI)[56], emerging from the Independent Gradient Model (IGM)[57,58], evaluates the strength of the interaction between a given pair of atoms. It is a score that allows us to quantitatively compare interactions and estimate their nature, *i.e.*, distinguish covalent from non-covalent interactions, based on threshold values. Although methods relying on topological analysis of the ED are common in identifying and characterizing chemical bonds, *e.g.*, electron localization function, these are often not able to quantify interactions[59]. In contrast, with IBSI, the quantification is outstandingly easy to interpret and is becoming a common tool to evaluate other types of interactions[60–62]. However, despite a few XB complexes being included in the original study[56], IBSI has not yet been systematically explored in the context of these interactions. Herein we report an exploratory study on how IBSI can be used to fairly predict XB interaction energies. Most strikingly, we will show that IBSI values calculated using a promolecular approach that does not require any QM calculation, also linearly correlate with interaction energies while providing similar accuracy. These exploratory results open the door for the development of fast methods to estimate XB energies in large datasets and/or biomolecular systems, and also for the usage of IBSI as a fast-obtainable XB feature for ML models.

2 Methods

2.1 IGM and IBSI

Herein, a succinct overview of the IBSI approach is given. Further details can be found in the original publications[56–58]. The concept of IBSI originates from the Independent Gradient Model (IGM)[57,58] which can be viewed as an extension of the NCI analysis method[63]. NCI is based on the topological analysis of the reduced density gradient s (also called RDG). However, the NCI approach has a semiquantitative character since the integration of quantities over NCI regions is not trivial[57]. On the contrary, the IGM approach, by providing a non-interacting reference system[58], allows quantification of the interactions.

For a system with interacting fragments A and B, the norm of the ED gradient $|\nabla\rho^{pair}|$, defined as

$$|\nabla\rho^{pair}| = |\nabla\rho_A + \nabla\rho_B| \tag{1}$$

is attenuated in the region between the interacting fragments. The sum of the absolute value of the density gradient of each fragment, denoted $|\nabla\rho^{IGM,pair}|$

$$|\nabla\rho^{IGM,pair}| = |\nabla\rho_A| + |\nabla\rho_B| \tag{2}$$

is introduced by the IGM approach as a non-interacting reference. Since the sign of the individual gradients is ignored in the summation, and thus, the density gradient originating from different fragments will not cancel with each other, $|\nabla\rho^{IGM,pair}|$ is the upper limit of the true ED gradient. From these, the δg^{pair} descriptor emerges

$$\delta g^{pair} = |\nabla\rho^{IGM,pair}| - |\nabla\rho^{pair}| \tag{3}$$

which is a unique bond signature that precisely quantifies the net ED gradient collapse due to the interaction between any pair of interacting atoms. Additionally, δg can be plotted against the ED multiplied by the second eigenvalue of the ED hessian matrix, $sign(\lambda_2)\rho$, producing plots analogous to those obtained in NCI

analyses, allowing to discriminate if δg^{pair} occurs in attractive ($\lambda_2 < 0$) or repulsive ($\lambda_2 > 0$) regions. To get a global score for a given bond, the integral of δg^{pair} over the interaction volume divided by the square of the internuclear distance d is taken

$$\Delta g^{pair} = \int_V \frac{\delta g^{pair}}{d^2} dV \quad (4)$$

Δg^{pair} is a bond index by itself, however, in order to obtain a score comparable between bond indices and molecules, it has to be normalized for the H_2 molecule:

$$IBSI = \frac{\Delta g^{pair}}{\Delta g^{H_2}} \quad (5)$$

IBSI is dimensionless value that does not correspond to a bond order, but reflects the bond strength[56].

2.2 ED and partition schemes

IGM and IBSI are dependent on the ED (ρ) and the partition scheme used to assign it to atoms/fragments. Originally, IGM was developed specifically for promolecular ED-based calculations[57] and this promolecular density is obtained from a sum of simple exponential atomic functions fitted to averaged *ab initio* atomic electron densities. Even though the obtained gradient is approximate as it lacks relaxation, the accuracy is reasonable as long as it is used in the non-covalent regime[57]. This approach is very attractive since minimal computational resources are required and only the geometry is required as input. Given its simplicity, the partition of the total ED gradient into atomic contributions is straightforward. The calculation of IBSI values from promolecular densities (here denoted as $IBSI^{PRO}$) was not considered in the original implementation[56,64] which used QM-based densities and a Gradient-Based Partitioning scheme (see below). However, such calculation is implemented in the MultiWFN package[65] (see Computational Details).

Another approach takes advantage of the ED obtained from QM calculations. In this case, the ED is in principle more accurate but the partition of the total gradient is not trivial. The Gradient-Based Partitioning (GBP) scheme was introduced in the context of IGM[58] and is implemented in IGMplot[64]. This method proposes that each individual gradient element $\partial\rho_i/\partial_x$ can be assigned to an atomic orbital φ_i and IBSI values calculated within this approach are henceforth termed $IBSI^{GBP}$.

Recently[66], it was argued that the isosurfaces of δg are too bulgy leading to erroneous analysis conclusions. To tackle this, IGM based on an Hirshfeld partition of the ED was proposed[66] and implemented in the MultiWFN package[65]. Hirshfeld is a very common method to obtain atomic densities[67] and allows the calculation of all quantitative indices available under the framework of the original IGM, including IBSI (here denoted $IBSI^{H}$).

2.3 Data sets

To test a possible correlation of IBSI with XB interaction energies we used 3 data sets containing various X-bonded systems with available optimized equilibrium geometries and energies obtained from high-level QM calculations.

Set 1 was taken from reference 68 which revises and corrects some values earlier reported for the XB18 and XB51 benchmarking sets[69]. These benchmarks consist of 69 systems bearing only neutral fragments with Cl-, Br-, and I-containing molecules as XB donors, and N, O, P, and Cl as acceptor atoms. XB18 contains 18 systems with NCH and OCH_2 as acceptors. Here, the geometries were optimized at CCSD(T)/aVQZ, and the interaction energies were calculated at the CCSD(T)/CBS level of theory. XB18 was intentionally constructed using only small molecules so that highly accurate calculations could be easily performed. The XB51 is an extended version of XB18 and includes a wider range of both donor and acceptor fragments. The geometry optimizations for XB51 were performed at ωB97X/aVTZ level of theory with single-point energies computed using an MP2-based extrapolation of the CCSD(T) energy. Herein, we merged both XB18 and XB51 and in cases where binding energies and geometries were available from both, the data were taken from XB15, yielding a total of 64 complexes (see Table S1 in ESI†). The energy values reported for these data sets correspond to $-E_{int}$, meaning that more positive values show stronger interactions, however, in this work we used E_{int} values for consistency with other data sets.

Set 2 was taken from the Non-Covalent Interactions Atlas, a library containing accurate benchmark non-covalent interaction energies[70]. It comprises halogen-bonded systems, optimized at the B3LYP-D3(BJ)/def2-QZVP level, containing small molecules with Cl, Br, and I as XB donors, and various XB acceptors bearing O, N, P, and S, such as acetonitrile, pyrazine, acetone, thiacetone, and molecular halogens. The X-bonded compounds in this library were chosen to cover a wide range of σ-hole magnitudes and each fragment contains no more than 13 atoms. The benchmark interaction energies were calculated using a composite CCSD(T)/CBS scheme based on MP2 and CCSD(T) calculations with very large basis sets. Herein we excluded X$\cdots\pi$ bonds because they cannot be unambiguously described by a single IBSI value, therefore yielding a final set of 99 complexes (see Table S2 in ESI†).

Set 3 consists of A–X$\cdots$B systems, where A = H, F, Cl, Br, I, and X = F, Cl, Br, I taken from reference 55. The data contained originally 140 high-accuracy *ab initio* benchmark interaction energies (CCSD(T)-F12b/CBS) calculated on CCSD(T)-F12b/VTZ-(PP)-F12 optimized structures whose geometry is available. In this work, only 124 of those systems were used (see Table S3 in ESI†) since 10 complexes containing X$\cdots\pi$ contacts were removed for the same reason mentioned above for **Set 2** along with those containing F_2 as a XB donor. Notice that fluorine is typically not considered a XB donor and fluorine interactions are fundamentally different from typical XBs[71].

2.4 Computational details

All QM calculations were performed using the Gaussian 09 program package[72]. Since optimized geometries were available, to obtain the ED for the IBSI estimation ($IBSI^{GBP}$ and $IBSI^{H}$), single-point calculations were performed at the DFT M06-2X/def2-TZVP level of theory[73] in the gas phase with the associated effective core potential for iodine. This functional is commonly applied in XB studies[8,31,74] with good performances[69,70,75] and is also

recommended by the IBSI method[56]. Additional calculations using def2-SVPD[73,76] and def2-QZVP[77] were performed in order to evaluate the significance of the basis set (see below). An ultrafine integration grid was applied in all the calculations. Checkpoint (chk) or wave function files (wfn/wfx) were stored for further analysis and calculation of IBSI.

IBSI^{PRO} and IBSI^{H} were calculated with MultiWFN 3.7[65]. As mentioned above, IBSI^{PRO} only required the optimized geometry while for IBSI^{H}, the M06-2X/def2-TZVP wave function file was provided. In both cases, the reported values are normalized to H_2 by the Δg^{H_2} value obtained in the same conditions. IBSI^{GBP} values were obtained with IGMplot 2.6.9b[64] using the same wave function files. Herein, the values are internally normalized for H_2 at the M06-2X/6-31G** level of theory and no re-normalization was performed for M06-2X/def2-TZVP values. Notice that this does not have any impact on statistics of the obtained linear-fit models.

2.5 Statistical analysis

The data in the three sets were fitted separately to the following equation

$$E_{int} = m \times IBSI + b \quad (6)$$

via the m and b parameters using an ordinary least squares (OLS) regression model. The quality of the fit of the data was analyzed by evaluating the coefficient of determination R^2, the Pearson correlation coefficient r, the Spearman's Rank Correlation Coefficient ρ, and the Kendall rank correlation coefficient (τ) using in-house python scripts. The Mean Absolute Error (MAE) was employed as a performance metric of each model. Before the fitting stage, an explanatory data analysis (EDA) was performed to characterize each set. Multivariate outliers, *i.e.* the unusual combination of E_{int} and IBSI values, were discarded using the minimum covariance determinant (MCD) method[78,79] with a significance level threshold of 0.001 using in-house python scripts. This is a highly robust estimator of multivariate location and scatter as the MCD is computed using only a subset of the sample, thus, the outlying points will have a small impact on the MCD location or shape estimate. Further information can be found in the section Outlier Removal in ESI†.

3 Results and discussion

3.1 Basis set influence on IBSI

In the original IBSI implementation based on GBP the authors showed that IBSI^{GBP} values are typically independent of the method and basis set, therefore, stable results are expected as long as the same method is used for comparative studies[56]. Regarding the IBSI^{H} approach[66], although it was claimed that a low sensitivity to wave function quality was observed, the data was not disclosed. Given the novelty of this partition scheme, and the unprecedented application to XBs, we selected 3 complexes from **Set 1** featuring a strong (FI$\cdots$pyr, -20.34 kcal mol^{-1}), a mild (FI$\cdots$OPH$_3$, -13.36 kcal mol^{-1}) and a week (FI$\cdots$PCH, -2.74 kcal mol^{-1}) XB, and calculated IBSI^{H} values with an increasing basis set size (def2-SVPD, def2-TZVP, and def2-QZVP). IBSI^{GBP} values were also calculated for comparison and the results are presented in Table 1). When comparing IBSI^{H} and IBSI^{GBP}, an obvious difference is observed in the magnitude of the values, with IBSI^{GBP} yielding larger IBSI values. This will be further discussed below, nonetheless, we highlight the fact that comparative studies also require that the same scheme is used. Within the same partition scheme, the values obtained vary little with the basis set. Strikingly, IBSI^{H} is even less sensitive to the size of the basis set while the larger deviation is found for the stronger XBs, especially for IBSI^{GBP}, although still acceptable.

3.2 IBSI^{H} and IBSI^{GBP} linearly correlate with XB interaction energies

Although XBs were explored in the original IBSI reference[56], a real systematic study for this type of non-covalent bond is yet to be performed. Herein, we explore if such a "simple" index linearly correlates with the interaction energy (E_{int}) for large and diverse sets of halogen-bonded systems taken from the literature. Since two methods based on QM EDs exist, namely, the original GBP formulation (IBSI^{GBP}) and the recently proposed Hirshfeld partitioning (IBSI^{H}), in the next sections we will compare the performance of both for each set individually.

3.2.1 Set 1.

In **Set 1** the interaction energies span a wide range of values, from very weak (FI$\cdots$FCCH, -0.29 kcal mol^{-1}) to strong (FI$\cdots$HLi, -33.79 kcal mol^{-1}) XBs. However, the distribution of the energies is slightly skewed (Figure S1 in ESI†) and with a data gap between the very strong XBs and the remaining values. Indeed, **Set 1** is actually more representative of weak to moderate XBs (median = -4.17 kcal mol^{-1}). Although plotting E_{int} as a function of IBSI for the full dataset shows a fair linear correlation between the two properties (Figure S19 in ESI†), two outliers, easily identified by visual inspection, were identified by the MCD method. They correspond to Br$_2\cdots$HLi (-23.11 kcal mol^{-1}) and FI$\cdots$HLi (-33.79 kcal mol^{-1}), the stronger XBs in the set, both possessing an hydride as the XB acceptor atom B (see the Outlier Removal section in ESI† for further discussion). Curiously, an analysis of the IBSI values beyond the X$\cdots$B pair shows that the bonding pattern is odd (Figure 1) for these two complexes. Indeed, for Br$_2\cdots$HLi, the covalent Br–Br bond is much weaker (0.075 and 0.189 for IBSI^{H} and IBSI^{GBP}, respectively) than the supposedly non-covalent Br$\cdots$H contact (0.185 and 0.382). A less pronounced difference was found for FI$\cdots$HLi where the IBSI^{H} value for the I–F bond is lower (0.120) than the H$\cdots$I contact (0.128) ever so slightly. IBSI^{GBP} values, on the contrary, yield a lower value for the H$\cdots$I XB pair (0.259), although very similar

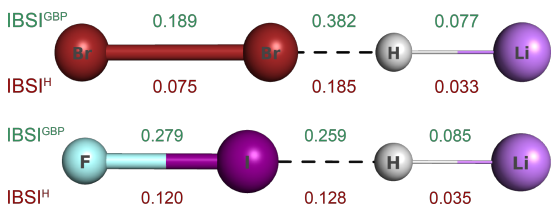

Fig. 1 Calculated IBSI^{H} and IBSI^{GBP} values for all interacting atoms in Br$_2\cdots$HLi (top) and FI$\cdots$HLi (bottom).

Table 1 Calculated IBSI^{H} and IBSI^{GBP} values for 3 halogen bonded systems taken from Set 1. The ED was obtained at the M06-2X/b (b = def2-SVPD, def2-TZVP, and def2-QZVP) level. The reported interaction energies (E_{int}) are CCSD(T)/CBS values from reference 68

		IBSI^{H}			IBSI^{GBP}		
System	E_{int} / kcal mol^{-1}	def2-SVPD	def2-TZVP	def2-QZVP	def2-SVPD	def2-TZVP	def2-QZVP
FI···PCH	-2.74	0.016	0.016	0.016	0.053	0.051	0.052
FI···OPH$_3$	-13.36	0.038	0.041	0.041	0.109	0.102	0.107
FI···pyr	-20.34	0.057	0.062	0.061	0.175	0.154	0.141

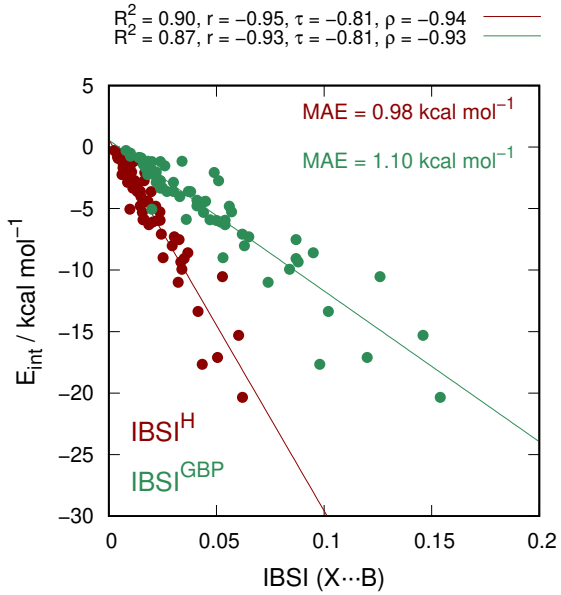

Fig. 2 E_{int} as a function of IBSI^{H} and IBSI^{GBP} for the final Set 1 (see Tables S1 and S5 in ESI†).

to the one observed for the I–F bond (0.279). Notice also that the IBSI^{GBP} values (X···B pair) are larger than the 0.15 threshold for non-covalent interactions defined in reference 56. The above considerations highlight the importance of an explanatory data analysis along with the minimum covariance determinant (MCD) method to confidently discard outliers and this approach was followed in all subsequent analyses. Further discussion regarding outlier removal can be found in ESI†.

Using an outlier-free **Set 1**, the plot of E_{int} as a function of IBSI (Figure 2) shows a strong linear correlation between the variables ($R^2 \approx 0.9$ and $|r| > 0.93$ for both IBSI^{H} and IBSI^{GBP}). Additionally, ρ and τ clearly indicate a monotonic association between the variables. The final fitted parameters can be found in Table S4 in ESI†. Noticeably, the intercepts b are ≈ 0 for both IBSI^{H} and IBSI^{GBP} while the slopes are very different reflecting the different ranges of the IBSI scales. Thus, it seems that the indicative threshold of the non-covalent domain (0.15) for IBSI^{GBP} [56] is not applicable to IBSI^{H}, however, IBSI^{H} and IBSI^{GBP} correlate linearly (Figure S2 in ESI†). The performance of the model is acceptable with MAE values ≈ 1 kcal mol^{-1} and typically, larger deviations between the predicted and reference data are observed for stronger XBs (Figure S3 in ESI†) whereas the deviations are fairly distributed around zero (Figure S4 in ESI†) meaning that no obvious under or overestimation of the predicted values is observed although a very slight skewness is observed for IBSI^{GBP} towards negative deviations, probably leading the the slight larger MAE when compared with IBSI^{H}.

3.2.2 Set 2.

Set 2 contains systems featuring mostly weak XBs ($E_{int} > -10$ kcal mol^{-1}, median $= -2.97$ kcal mol^{-1}), with a few exceptions (see Table S2 and Figure S5 in ESI†). As in **Set 1**, there is also a data gap between the stronger XBs (max $= -17.14$ kcal mol^{-1}) and the remaining values. Apart from dihalogens and acetone, which were also present in **Set 1**, this set includes some cyclic acceptors (pyrazine, pyridine-N-oxide) and compounds containing sulfur (thioacetone, dimethylthioether). After the outlier removal (Table S5 in ESI†), and despite being a larger dataset, the correlation between IBSI and E_{int} is strong for both IBSI^{H} and IBSI^{GBP}, with $R^2 \approx 0.9$ and $|r| > 0.93$ (Figure 3). Again, the final fitted parameters can be found in Table S4 in ESI†. The intercepts b are ≈ 0 for both methods and the different slopes reflect a quite different scale of IBSI^{H} and IBSI^{GBP}, though a linear correlation between the two is observed (Figure S6 in ESI†) as observed earlier. There is no obvious performance difference between the methods, both providing a similar accuracy

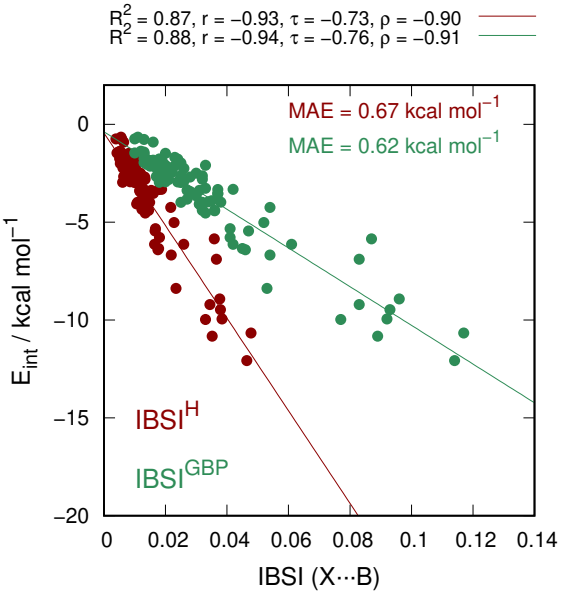

Fig. 3 E_{int} as a function of IBSI for the final Set 2 (see Tables S2 and S5 in ESI†).

(MAE < 0.70 kcal mol^{-1}) and larger deviations for stronger XBs (Figure S7 in ESI†). Again, no obvious under- or overestimation was found as the error values are close to normally distributed around zero for both partition schemes (Figure S8 in ESI†).

3.2.3 Set 3.

Set 3 comprises dihalogen and hydrogen halide XB donors paired up with common XB acceptors, mostly small molecules such as NH_3, CH_2O, and H_2O (Table S3 in ESI†). As earlier, the energies span a wide range of values, from -20.51 kcal mol^{-1} (FCl$\cdots$PH$_3$) to -1.28 kcal mol^{-1} HBr$\cdots$PH$_3$, the distribution being skewed towards negative values with a median of -5.52 kcal mol^{-1}, however, in this case, no obvious gaps exist in the energy values (Figure S9 in ESI†). The correlation between IIBSI^{H} and IBSI^{GBP} with E_{int} is shown in Figure 4 whereas the final fitted parameters can be found in Table S4 in ESI†. Both IBSI^{H} and IBSI^{GBP} provide

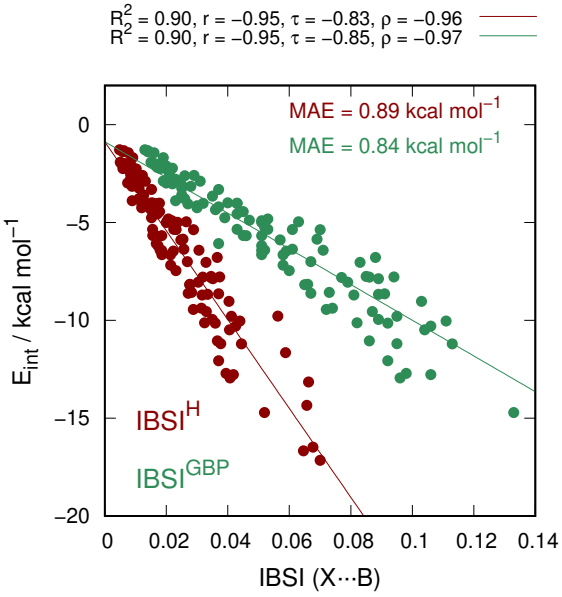

Fig. 4 E_{int} as a function of IBSI for the final Set 3 (see Tables S3 and S5 in ESI†).

strong linear correlations with E_{int} ($R^2 = 0.90$) and the MAE values are below 0.9 kcal mol^{-1} with no obvious systematic over- or underestimation of the predicted E_{int} values (Figure S11 and Figure S12 in ESI†). Both partition schemes perform similarly, the values being correlated (Figure S10 in ESI†).

3.3 IBSI^{PRO} as a fast XB strength descriptor

In the previous section, we showed that IBSI values obtained using QM-based electron densities (IBSI^{H} and IBSI^{GBP}) linearly correlate with XB interaction energies (E_{int}) for diverse sets of halogen-bonded complexes. This type of linear relationship can be useful, for instance, to estimate high-level *ab initio* E_{int} values using DFT geometries. However, such a task still requires the usage of QM-based electron density which could be unpractical not only for large datasets of small molecules, but also for biomolecular systems. Therefore, we wondered if IBSI values, calculated using a promolecular approach (IBSI^{PRO}) and therefore neglecting relaxation (among other terms), could also be used similarly. Notice that for the covalent regime, the promolecular ED underestimates the troughs of the ED gradient, hence not describing the bonds correctly[58]. Owing to the disputed varying degree of covalency involved in XBs, promolecular ED may not describe them correctly. However, it is also true that as long the complexes stay in the weak to mild non-covalent regime, the promolecular approach could be enough to capture the correct bond pattern. Indeed, in **Set 1** and after outlier removal (Table S6 in ESI†), the correlation between IBSI^{PRO} values and E_{int} is linear ($R^2 = 0.88$, $|r| = 0.94$) (see Figure 5 left and Table S4 in ESI† for the fitted parameters). Also, the other coefficients, ρ and τ, show a strong monotonic relationship between E_{int} and IBSI^{PRO}. In fact, comparing IBSI^{PRO} with IBSI^{GBP} or IBSI^{H}, the difference in linearity (R^2, $|r|$) is almost negligible while the MAE (≈ 1 kcal mol^{-1}) slightly outperforms the QM-based methods (IBSI^{H} and IBSI^{PRO}). The error is fairly normally distributed around zero (Figure S14 in ESI†), similarly to IBSI^{H} and IBSI^{GBP}, meaning that IBSI^{PRO} does not strongly over- or underestimate interaction energies, while larger errors are typically associated with stronger XBs (Figure Figure S13 in ESI†). This suggests that using approximate promolecular densities may result in similar accuracy compared to QM methods, even for moderate-strength XBs. Notice that here, the intercept of the plot is ≈ 1, and the slope is quite different from both other methods, indicating a different IBSI scale with IBSI^{PRO} values consistently larger than IBSI^{GBP} or IBSI^{H}.

Set 2 is larger than **Set 1** and contains a wider variety of acceptors. The linear correlation between E_{int} and IBSI^{PRO} (Figure 5 center) is again strong ($R^2 = 0.84$, $|r| = 0.92$) and equivalent to that found for IBSI^{H} and IBSI^{GBP}. The final fitted parameters can be found in Table S4 in ESI†. The MAE value (0.70 kcal mol^{-1}) is slightly larger than that for IBSI^{GBP}(0.62 kcal mol^{-1}) but similar to the one obtained for IBSI^{H} (0.67 kcal mol^{-1}). The difference between estimated and true E_{int} is close to normally distributed around zero (Figure S16 in ESI†), and, while the values are somewhat right-skewed, there is no significant tendency towards consistently over- or underestimating E_{int}.

Set 3 contains systems with only dihalides as donors, making it the most uniform dataset used in this work. The final linear correlation between IBSI^{PRO} and E_{int} is strong ($R^2 = 0.91$, $|r| = 0.95$, see Figure 5, right), with an MAE value of ≈ 0.8 kcal mol^{-1}, outperforming the QM-based methods. The values of error deviation are very close to normally distributed (Figure S18 in ESI†), the error increasing with increasing XB strength (Figure S17 in ESI†). The final fitted parameters are listed in Table S4 in ESI†.

The above results suggest that, overall, this quite simple model which uses promolecular density was able to adequately predict interaction energies in these fairly large and diverse datasets. It is also remarkable that the compounds that were poorly described by QM methods (outliers) are also recurrently observed as outliers with IBSI^{PRO}. Moreover, it could be expected that larger deviations are typically observed with increasing XB strength when using IBSI^{PRO} owing to the lack of relaxation of the ED which becomes important in the covalent regime. However, such a tendency (Figures S4, S10, S16 in ESI†) is also observed in QM-based

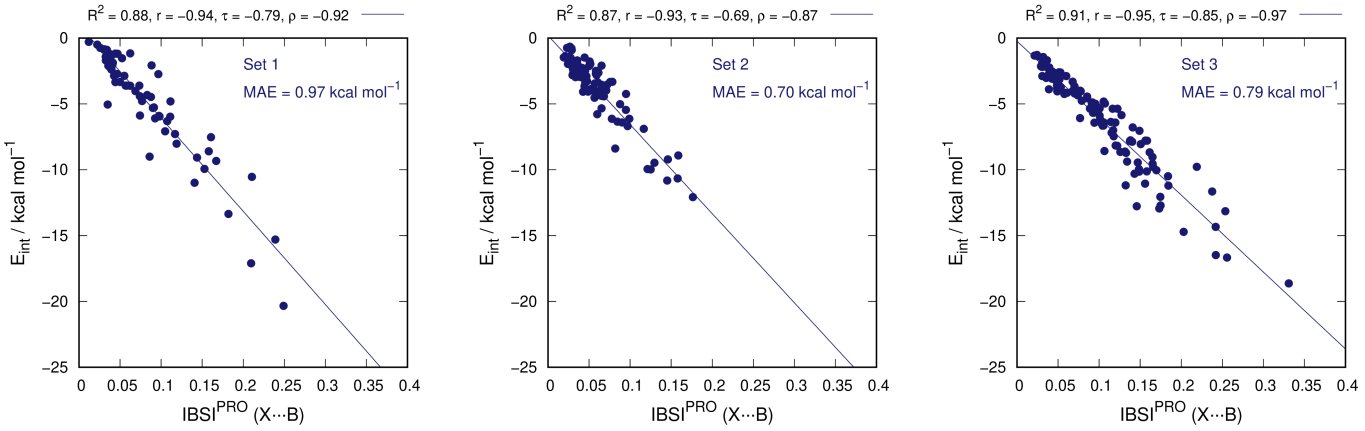

Fig. 5 E_{int} as a function of IBSI^{PRO} for Set 1, Set 2, and Set 3.

methods such as IBSI^{H} and IBSI^{GBP} (Figures S3, S7, S11 in ESI†). Therefore, it becomes evident that the accuracy of a simple linear regression estimator based on promolecular properties is at least very similar to that obtained by methods that require rigorous QM calculations for these types of halogen bonded complexes. This means that for large systems such as protein-ligand complexes where high-level QM calculations are often not feasible, IBSI^{PRO} can be a fast and reliable solution, provided that proper calibration curves are available.

4 Conclusions

Predicting the trends and interaction energies of halogen bonding interactions using simple and computationally cheap molecular descriptors has been recursively addressed in the literature. In this scope, the usage of $V_{S,max}$ of the halogen atom has been an example of such an approach, however, this single descriptor cannot be directly applied to diverse datasets such as **Set 1**–**Set 3**. In this exploratory work, we tested the possibility of using the Intrinsic Bond Strength Index (IBSI) as halogen bond strength descriptor for three different datasets containing highly accurate QM-based interaction energies. Notice that XBs were mentioned in the original IBSI reference [56], however, this is the first systematic study regarding the usage of IBSI in halogen bonding. We first addressed two ED partition methods that rely on QM calculations (IBSI^{GBP} and IBSI^{H}). Both yielded IBSI values that were insensitive to the basis set size for 3 complexes featuring strong, mild, and weak XBs. When applied to the **Set 1**-**Set 3**, both IBSI^{GBP} and IBSI^{H} linearly correlated with the interaction energy with the linear models providing MAEs typically below 1 kcal mol^{-1}, reaching 0.62 kcal mol^{-1} for IBSI^{GBP} in **Set 2**. We did not observe any systematic differences in the performance of the two different partition schemes apart from the different IBSI scale and both IBSI^{GBP} and IBSI^{H} produced consensual outliers, typically corresponding to a few complexes featuring the strongest XBs in each set. Thus, both IBSI^{GBP} and IBSI^{H} can in principle be used as a qualitative index to compare the halogen bond strength in complexes, but also can be used to provide quantitative estimates of the interaction energy. Despite these exciting results, the usage of QM-based electron density could still hinder applications in large datasets or biomolecular systems. Therefore, we also explored the possibility of obtaining a quantitative model that predicts the interactions energies based in IBSI^{PRO} which relies on the so-called promolecular approach which is based on tabulated data and hence, it only requires the geometry as an input. In spite of its simplicity, the performance of IBSI^{PRO} was comparable to the QM-based methods, actually outperforming IBSI^{GBP} and IBSI^{H} for **Set 3**, suggesting that computationally demanding calculations are not necessary in order to achieve reasonable accuracy, as long as one stays in the non-covalent regime, which is often the case in halogen-bonded protein-ligand systems systems [54]. Our exploratory work can open the door to the usage of IBSI^{PRO}as a fast and reliable XB strength descriptor in large systems, *e.g.* proteins, provided that proper calibration curves are available.

Author Contributions

Conceptualization: P.J.C.; Methodology: O.S. and P.J.C.; Validation: O.S. and P.J.C.; Investigation: O.S.; Writing – original draft: O.S.; Writing – review & editing: P.J.C.; Funding Acquisition: P.J.C.; Supervision: P.J.C.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors thank Fundação para a Ciência e a Tecnologia (FCT), Portugal for grants UIDB/04046/2020 and UIDP/04046/2020 (to BioISI) and for Individual Call to Scientific Employment Stimulus grant 2021.00381.CEECIND (P. J. Costa). PJC thanks Diogo Vila–Viçosa for discussions regarding this work.

Notes and references

1 G. Cavallo, P. Metrangolo, R. Milani, T. Pilati, A. Priimagi, G. Resnati and G. Terraneo, *Chem. Rev.*, 2016, **116**, 2478–2601.

2 G. R. Desiraju, P. S. Ho, L. Kloo, A. C. Legon, R. Marquardt,

P. Metrangolo, P. Politzer, G. Resnati and K. Rissanen, *Pure Appl. Chem.*, 2013, **85**, 1711–1713.
3 P. J. Costa, *Phys. Sci. Rev.*, 2017, **2**, 20170136.
4 T. Clark, M. Hennemann, J. S. Murray and P. Politzer, *J. Mol. Model.*, 2007, **13**, 291–296.
5 P. Politzer, J. S. Murray and T. Clark, *Phys. Chem. Chem. Phys.*, 2010, **12**, 7748–7757.
6 K. E. Riley and K.-A. Tran, *Faraday Discuss.*, 2017, **203**, 47–60.
7 L. P. Wolters, P. Schyman, M. J. Pavan, W. L. Jorgensen, F. M. Bickelhaupt and S. Kozuch, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2014, **4**, 523–540.
8 E. Engelage, D. Reinhard and S. M. Huber, *Chem. – Eur. J*, 2020, **26**, 3843–3861.
9 J. M. Holthoff, R. Weiss, S. V. Rosokha and S. M. Huber, *Chem. – Eur. J*, 2021, **27**, 16530–16542.
10 J. S. Murray and P. Politzer, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2017, **7**, e1326.
11 J. S. Murray and P. Politzer, *ChemPhysChem*, 2021, **22**, 1201–1207.
12 J. Řezáč and A. De La Lande, *Phys. Chem. Chem. Phys.*, 2017, **19**, 791–803.
13 L. Mendez, G. Henriquez, S. Sirimulla and M. Narayan, *Molecules*, 2017, **22**, 1397.
14 Y. Li, B. Guo, Z. Xu, B. Li, T. Cai, X. Zhang, Y. Yu, H. Wang, J. Shi and W. Zhu, *Sci. Rep.*, 2016, **6**, 1–10.
15 J. Lapp and S. Scheiner, *J. Phys. Chem. A*, 2021, **125**, 5069–5077.
16 A. Priimagi, G. Cavallo, P. Metrangolo and G. Resnati, *Acc. Chem. Res.*, 2013, **46**, 2686–2695.
17 V. Oliveira, E. Kraka and D. Cremer, *Phys. Chem. Chem. Phys.*, 2016, **18**, 33031–33046.
18 M. H. Kolar and P. Hobza, *Chem. Rev.*, 2016, **116**, 5155–5187.
19 R. L. Sutar, E. Engelage, R. Stoll and S. M. Huber, *Angew. Chem. Int. Ed.*, 2020, **59**, 6806–6810.
20 M. Kaasik and T. Kanger, *Front. Chem.*, 2020, **8**, 958.
21 E. Margiotta, S. C. C. van der Lubbe, L. de Azevedo Santos, G. Paragi, S. Moro, F. M. Bickelhaupt and C. Fonseca Guerra, *J. Chem. Inf. Model*, 2020, **60**, 1317–1328.
22 L. Schifferer, M. Stinglhamer, K. Kaur and O. G. Macheño, *Beilstein J. Org. Chem.*, 2021, **17**, 2270–2286.
23 C. Curiac, L. A. Hunt, M. A. Sabuj, Q. Li, A. Baumann, H. Cheema, Y. Zhang, N. Rai, N. I. Hammer and J. H. Delcamp, *J. Phys. Chem C*, 2021, **125**, 17647–17659.
24 S. An, A. Hao and P. Xing, *ACS Nano*, 2021, **15**, 15306–15315.
25 M. H. H. Voelkel, E. Engelage, M. Kondratiuk and S. M. Huber, *Eur. J. Org. Chem.*, **2022**, e202200211.
26 J.-W. Zou, M. Huang, G.-X. Hu and Y.-J. Jiang, *RSC Advances*, 2017, **7**, 10295–10305.
27 M. R. Scholfield, M. C. Ford, A.-C. C. Carlsson, H. Butta, R. A. Mehl and P. S. Ho, *Biochemistry*, 2017, **56**, 2794–2802.
28 J. Heidrich, L. E. Sperl and F. M. Boeckler, *Front. Chem.*, 2019, **7**, 9.
29 S. H. Jungbauer and S. M. Huber, *J. Am. Chem. Soc.*, 2015, **137**, 12110–12120.
30 P. M. J. Szell, S. Zablotny and D. L. Bryce, *Nat. Commun.*, 2011, **10**, 916.
31 H. Yang and M. W. Wong, *Molecules*, 2020, **25**, 1045.
32 C. Xu, V. U. B. Rao, J. Weigen and C. C. J. Loh, *Nat. Commun.*, 2020, **11**, 4911.
33 H. Wang, H. K. Bisoyi, A. M. Urbas, T. J. Bunning and Q. Li, *Chem. – Eur. J*, 2019, **25**, 1369–1378.
34 X. Miao, Z. Cai, H. Zou, J. Li, S. Zhang, L. Ying and W. Deng, *J. Mater. Chem. C*, 2022, **10**, 8390–8399.
35 R. Shi, D. Yu, F. Zhou, J. Yu and T. Mu, *Chem. Commun.*, 2022, **58**, 4607–4610.
36 A. Vanderkooy and M. S. Taylor, *Faraday Discuss.*, 2017, **203**, 285–299.
37 M. S. Alvarez, C. Houzé, S. Groni, B. Schöllhorn and C. Fave, *Org. Biomol. Chem.*, 2021, **19**, 7587–7593.
38 A. Singh, A. Torres-Huerta, T. Vanderlinden, N. Renier, L. Martínez-Crespo, N. Tumanov, J. Wouters, K. Bartik, I. Jabin and H. Valkenier, *Chem. Commun.*, 2022, **58**, 6255–6258.
39 M. Kokot, M. Weiss, I. Zdovc, M. Hrast, M. Anderluh and N. Minovski, *ACS Med. Chem. Lett.*, 2021, **12**, 1478–1485.
40 S. Jena, J. Dutta, K. D. Tulsiyan, A. K. Sahu, S. S. Choudhury and H. S. Biswal, *Chem. Soc. Rev.*, 2022, **51**, 6255–6258.
41 R. S. Nunes, D. Vila-Viçosa and P. J. Costa, *J. Am. Chem. Soc.*, 2021, **143**, 4253–4267.
42 S. M. Huber, E. Jimenez-Izal, J. M. Ugalde and I. Infante, *Chem. Commun.*, 2012, **48**, 7708–7710.
43 I. Nicolas, F. Barriere, O. Jeannin and M. Fourmigue, *Cryst. Growth Des.*, 2016, **16**, 2963–2971.
44 J. Thirman, E. Engelage, S. M. Huber and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2018, **20**, 905–915.
45 B. Inscoe, H. Rathnayake and Y. Mo, *J. Phys. Chem. A*, 2021, **125**, 2944–2953.
46 T. Clark and A. Heßelmann, *Phys. Chem. Chem. Phys.*, 2018, **20**, 22849–22855.
47 T. Clark, J. S. Murray and P. Politzer, *ChemPhysChem*, 2018, **19**, 3044–3049.
48 T. Brinck and A. N. Borrfors, *J. Mol. Model.*, 2019, **25**, null.
49 T. Brinck, P. Carlqvist and J. H. Stenlid, *J. Phys. Chem. A*, 2016, **120**, 10023–10032.
50 R. Nunes and P. J. Costa, *Chem. – Asian J.*, 2017, **12**, 586–594.
51 M. L. Kuznetsov, *Int. J. Quantum Chem.*, 2019, **119**, e25869.
52 M. L. Kuznetsov, *Molecules*, 2019, **24**, 2733.
53 M. L. Kuznetsov, *Molecules*, 2021, **26**, 2083.
54 P. J. Costa, R. Nunes and D. Vila-Viçosa, *Expert. Opin. Drug. Discov.*, 2019, **14**, 805–820.
55 R. A. Shaw and J. G. Hill, *Inorganics*, 2019, **7**, 19.
56 J. Klein, H. Khartabil, J.-C. Boisson, J. Contreras-García, J.-P. Piquemal and E. Hénon, *J. Phys. Chem. A*, 2020, **124**, 1850–1860.
57 C. Lefebvre, G. Rubez, H. Khartabil, J.-C. Boisson, J. Contreras-García and E. Hénon, *Phys. Chem. Chem. Phys.*,

2017, **19**, 17928–17936.

58 C. Lefebvre, H. Khartabil, J.-C. Boisson, J. Contreras-García, J.-P. Piquemal and E. Hénon, *ChemPhysChem*, 2018, **19**, 724–735.

59 J. Contreras-García and Y. Weitao, *Acta Physico-Chimica Sinica*, 2018, **34**, 567.

60 T. O. Unimuke, H. Louis, E. A. Eno, E. C. Agwamba and A. S. Adeyinka, *ACS Omega*, 2022, **7**, 13704–13720.

61 E. A. Katlenok, A. V. Rozhkov, O. V. Levin, M. Haukka, M. L. Kuznetsov and V. Y. Kukushkin, *Cryst. Growth Des.*, 2020, **21**, 1159–1177.

62 M. Ponce-Vargas, J. Klein and E. Hénon, *Dalton Trans.*, 2020, **49**, 12632–12642.

63 E. R. Johnson, S. Keinan, P. Mori-Sánchez, J. Contreras-García, A. J. Cohen and W. Yang, *J. Am. Chem. Soc.*, 2010, **132**, 6498–6506.

64 J. Klein, E. Pluot, G. Rubez, J. C. Boisson and E. Hénon, *IGMPlot (Revision 2.6.9b)*, `http://igmplot.univ-reims.fr/`.

65 T. Lu and F. Chen, *J. Comput. Chem.*, 2011, **33**, 580–592.

66 T. Lu and Q. Chen, *J. Comput. Chem.*, 2022, **43**, 539–555.

67 F. L. Hirshfeld, *Theor. Chimi. Acta*, 1977, **44**, 129–138.

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

69 S. Kozuch and J. M. L. Martin, *J. Chem. Theory Comput.*, 2013, **9**, 1918–1931.

70 K. Kriz and J. Řezáč, *Phys. Chem. Chem. Phys.*, 2022, **24**, 14794–14804.

71 K. Eskandari and M. Lesani, *Chem. – Eur. J*, 2015, **21**, 4739–4746.

72 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, *Gaussian 09 Revision A.2*, 2009.

73 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.

74 D. Bulfield, E. Engelage, L. Mancheski, J. Stoesser and S. M. Huber, *Chem. – Eur. J*, 2020, **26**, 1567–1575.

75 A. Otero-De-La-Roza, E. R. Johnson and G. A. DiLabio, *J. Chem. Theory Comput.*, 2014, **10**, 5436–5447.

76 D. Rappoport and F. Furche, *J. Chem. Phys.*, 2010, **133**, 134105.

77 F. Weigend, *Phys. Chem. Chem. Phys.*, 2006, **8**, 1057–1065.

78 M. Hubert, M. Debruyne and P. J. Rousseeuw, *Wiley Interdiscip. Rev. Comput. Stat.*, 2018, **10**, e1421.

79 J. Hardin and D. M. Rocke, *J. Comput. Graph. Stat.*, 2005, **14**, 928–946.