Data and Molecular Fingerprint Driven Machine Learning Approaches to Halogen Bonding

Daniel P. Devore and Kevin L. Shuford*

Department of Chemistry and Biochemistry, Baylor University, One Bear Place #97348, Waco, TX 76798-7348, USA

E-mail: kevin_shuford@baylor.edu

Abstract

The ability to predict the strength of halogen bonds and properties of halogen bond (XB) donors has significant utility for medicinal chemistry and materials science. XBs are typically calculated through expensive *ab initio* methods. Thus, the development of tools and techniques for fast, accurate, and efficient property predictions has become increasingly more important. Herein, we employ three machine learning models to classify the XB donors and complexes by their principal halogen atom as well as predict the values of the maximum point on the electrostatic potential surface ($V_{S,max}$) and interaction strength of the XB complexes through a molecular fingerprint and data-based analysis. The fingerprint analysis produces a root mean square error of *ca*. 7.5 and *ca*. 5.5 kcal mol^{-1} while predicting the $V_{S,max}$ for the halo-benzene and halo-ethynyl benzene systems, respectively. However, the prediction of the binding energy between the XB donors and ammonia acceptor is shown to be within 1 kcal mol^{-1} of the DFT calculated energy. More accurate predictions can be made from the pre-calculated DFT data when compared to the fingerprint analysis.

Introduction

Halogen bonding, "the net attractive interaction between an electrophilic region on a halogen atom and a nucleophilic region on another atom,"[1] has become increasingly important in catalysis, materials science, and biology. This is due to the high tunability of the halogen bond (XB) interactions, which is affected by the polarization of the halogen atom,[2–5] hybridization of the atom covalently bonded to the halogen atom,[6–11] and electron donating/withdrawing groups contained in the XB donor molecule.[12–14] Each of these properties affect the redistribution of the electron density around the halogen atom, generating a smaller or larger area of electron depletion on the extension of the covalent bond between the halogen atom and an R-group, denoted as the σ-hole.[15–17]

The σ-hole on the "cap" of the halogen atom and the corresponding XB interactions are highly affected by the surrounding electronic environment. The magnitude of the σ-hole ($V_{S,max}$) being influenced by substituent groups has been shown throughout the literature,[18–23] and the interaction strength of the XB complexes has also been shown to correlate very well with electrostatic potential (ESP) of the XB donor and the XB acceptor.[24,25] Armed with these facts and studies conducted on halogen bonding using energy decomposition analysis and symmetry-adapted perturbation theory techniques stating that the XB interactions contain significant electrostatic character,[26–28] the prediction of $V_{S,max}$ becomes highly important. While the estimation of XB interactions and the ESP of the XB donor have typically been carried out through *ab initio* methods, these methods are not practical for large molecules and high-throughput studies associated with drug development or crystal structure prediction.

Multiple approaches to determining the ESP of the halogen atom in a molecule for quantitative structure-activity(property) relationship (QSAR/QSPR) predictions have been taken. Titov *et al.* observed how mono- and disubstituted halo-benzene XB donors affected the extra point charges placed on the extension of the C–X covalent bond and multipole expansion parameters, and fit those parameters to the ESP maps of the halo-benzene molecules through the Free-Wilson type QSPR model.[29] This investigation demonstrated the use of empirical molecular mechanics models for quick, accurate, and efficient predictions of the electronic environment for larger, drug-like molecules.[29] Heidrich *et al.* employed a support vector regression (SVR) machine learning (ML) technique on more than 16000 heterocycles to predict the MP2 calculated $V_{S,max}$ values.[30] Their results show that a reasonable prediction of the $V_{S,max}$ can be made for halogenated molecules with a significant speed up when compared to calculating the electrostatic surface of the molecule with *ab initio* methods, and that these initial predictions can be used as good estimates to be implemented into high-throughput docking calculations. The ability to predict the $V_{S,max}$ and interaction strengths of XB complexes before *ab initio* methods are implemented is highly beneficial for applying halogen bonds to multiple areas of study.

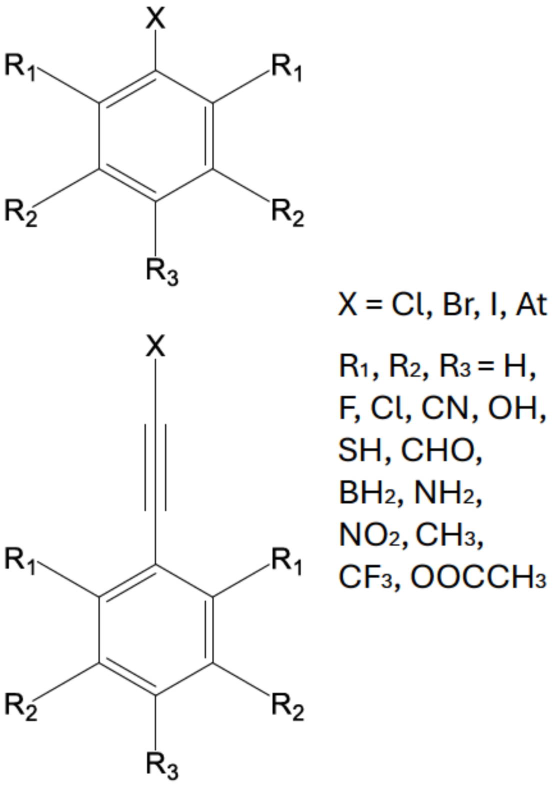

Figure 1: Schematic representation of the halo-benzene (top) and halo-ethynyl benzene (bottom) XB donors.

Herein, we present a proof-of-concept demonstrating the use of ML models for property determination of 1210 symmetric XB donor molecules (Figure 1) and their corresponding XB donor-acceptor complexes with ammonia. This study utilizes three ML models (Random Forest, boosted Decision Trees, and Support Vector Machines) to predict the $V_{S,max}$ of the halo-(ethynyl)benzene XB donors as well as the binding energy (E_{bind}) and $X \cdots N$ bond local force constant ($k^a_{X \cdots N}$) of the donor-ammonia complexes through data- and molecular fingerprint (MFP)-driven approaches. The XB donors used in this study incorporate electron-donating groups (EDGs) and electron-withdrawing groups (EWGs) examined previously by Devore *et al.*[12,31] These molecules were chosen to investigate the effects of hybridization, halogen identity, and different substituents on the $V_{S,max}$ of XB donors and the interaction strength of the corresponding complexes with ammonia. In addition, XB donors made with a mixture of EWGs and EDGs in the same donor, referred to as electron-donating-withdrawing group (EDWG) XB donors, were added to the dataset. The data-driven approach shows good agreement between the density functional theory (DFT) calculations, collected from previous studies[12,31] and this current study, and the ML model predictions for both the classification and regression cases. The MFP-driven approach revealed sizeable mean absolute error (MAE) and root mean-square error (RMSE) values for the prediction of the $V_{S,max}$ when the halo-(ethynyl)benzene donors were separated by the principal halogen atom, suggesting weakly accurate predictions. When the halo-(ethynyl)benzene donors were considered as a whole, the error values for predicting the $V_{S,max}$ were reduced. The binding energy and $X \cdots N$ bond local force constant were predicted to a more acceptable degree of accuracy. This result suggests that reliable predictions for the $V_{S,max}$, E_{bind}, and $k^a_{X \cdots N}$ properties can be obtained with decent accuracy from the molecular fingerprints.

Computational Methods

Calculations

All halogen bond donors and their corresponding complexes with ammonia were geometry optimized using the M06-2X global hybrid density functional[32] in conjunction with the correlation-consistent polarized valence double-ζ basis set augmented with diffuse functions on all atoms (aug-cc-pVDZ denoted as aVDZ)[33–35] using the Gaussian quantum chemical software.[36] Relativistic pseudopotentials were applied on all bromine, iodine, and astatine centers (i.e., aVDZ-PP for Br, I, and At).[37,38] In addition, harmonic vibrational frequency computations were implemented to ensure the stability of each structure and confirm that they represented a minima (i.e., $n_i = 0$) on the M06-2X/aVDZ-PP potential energy surface. Natural Bond Orbital (NBO) analyses were conducted for the determination of charge transfer between the Lewis acid-base pairs.[39–45] The binding energy of the XB complexes was determined by taking the difference between the energy of the complex and the energy of the fully relaxed monomers (i.e., $E^{complex} - (E^{donor} + E^{acceptor})$). In addition, Boys-Bernardi counterpoise corrections were executed to account for basis set superposition error.[46] The local mode force constants and dipoles for the X$\cdots$N, C–X, and C≡C bonds in the XB donors and corresponding complexes were extracted through the Local Mode Analysis code.[47–50] The topology of the electron density on the XB donors, acceptors, and complexes was analyzed by Bader's QTAIM algorithm,[51–53] employed through the Multiwfn software program.[54]

Models

The selection method chosen to identify the five most appropriate features for the classification and regression ML algorithms was the recursive feature elimination (RFE) technique in the sci-kit learn[55] library *Python* package. The decision tree classification and decision tree regression estimators were applied in the RFE process to eliminate the features that have the least importance to the identity of the principal halogen atom, magnitude of the σ-hole

($V_{S,max}$), binding energy of the XB complex (E_{bind}), and X$\cdots$N bond local force constant ($k^a_{X\cdots N}$). By reducing the number of features, the model is thus optimized due to reduction of possible sources of "noise." A complete list of the features collected and considered in the RFE algorithm for each XB donor and complex can be found in the SI (Table S1).

Three ML models, Random Forest (RF), XGBoost (XGB), and Support Vector Machines (SVM), from the scikit-learn[55] and XGBoost[56,57] library packages in *Python* were implemented throughout this article for the prediction of the electronic, energetic, and spectroscopic properties of aromatic XB donors and their corresponding complexes with ammonia. A brief explanation about the different ML models used can be found in the SI. The data employed in this article, introduced by Devore *et al.*[12,31] together with the additional XB structures calculated at the level of theory presented in this study, is publicly available on GitHub (`https://github.com/daniel-devore/XB-ML`). This study is split in two sections, classification and regression. The classification algorithms were utilized to predict the principal halogen atom based on the selected features from the RFE algorithm. The regression methods were further split into two sections, a data-based and molecular fingerprint-based approach.

The data-based predictions of $V_{S,max}$, E_{bind}, and $k^a_{X\cdots N}$ were found by using the five most important features identified using the RFE algorithm from the data gathered through DFT calculations. The molecular fingerprint (MFP)-based predictions utilized the structure of the XB donors themselves. The Simplified Molecular-Input Line-Entry System (SMILES)[58–60] code was used in conjunction with the RDkit[61] library package in *Python* to convert the structures to Morgan (circular) fingerprints of radius 2.[62] The MFPs are then used to train the ML algorithms. All data is split into a 80–20 ratio training and testing data set to train and test each model.

Separate metric systems were used to determine the accuracy of the classification and regression models. A confusion matrix, displayed as a heat map from the seaborn and matplotlib *Python* libraries,[63,64] is utilized to evaluate the quality of the output for the classification models. The diagonal elements show that the algorithm correctly predicted the true label, whereas the off-diagonal elements indicate when the ML model incorrectly predicted the true label. The metrics used to measure the accuracy of the regression models were the coefficient of determination (R^2), mean absolute error (MAE), and the root mean square error (RMSE). For further confirmation of the accuracy of the methods, 5-fold cross-validation was performed for each model. k-fold cross-validation is a resampling technique used to reduce selection bias and to gain insight of how a ML algorithm will generalize to an unseen dataset. The dataset is split into k folds (equal sized subsamples), where one of the folds is held as a validation set and the rest of the $k-1$ folds are utilized as the training set. The proccess is then repeated k times, with each of the subsamples being used as a validation set exactly once. The k results are then averaged.

Results and Discussions

Classification

The RFE technique was employed to find the five most important properties associated with the identity of the interacting halogen atom. The properties most associated with the identity of the halogen atom for the halo-benzene (BX) and halo-ethynyl benzene (BAX) systems are found in Table 1. The most important features for the BX donors are primarily associated with the energetic properties of the XB complex, namely the binding energy (E_{bind}) as well as formation energies of the donor (E_{def}^{donor}) and ammonia acceptor ($E_{def}^{NH_3}$). The structure of the XB donor in the corresponding complex also reflects the identity of the halogen atom, as indicated by the C–X bond length in the complex ($R_{C-X}^{complex}$) being one of the five most important features in determining the principal halogen atom. The magnitude of the charge transfer from the ammonia acceptor to the XB donor ($\Delta\rho$) is the fifth property most associated with the identity of the halogen atom in the BX systems.

The BAX systems, on the other hand, have more association with the electronic properties

Table 1: Five Most Important Features for Identifying the Principal Halogen Atom in the Halo-Benzene (BX) and Halo-Ethynyl Benzene (BAX) XB Donors.

Feature	BX	BAX
Feature 1	$R_{C-X}^{complex}$	$\nabla^2\rho^{BCP}$
Feature 2	E_{bind}	$H(r)^{BCP}$
Feature 3	E_{def}^{donor}	$\text{sign}(\lambda_2)*\rho$
Feature 4	$E_{def}^{NH_3}$	E_{bind}
Feature 5	$\Delta\rho$	E_{def}^{donor}

in the complex than energetic or structural features. This can be seen from three of the selected features being the Laplacian of the electron density ($\nabla^2\rho^{BCP}$), sign of the second eigenvalue of the electron density matrix multiplied by the electron density ($\text{sign}(\lambda_2)*\rho$), and the total energy density ($H(r)^{BCP}$) at the X$\cdots$N bond critical point. The remaining two features are the binding energy of the complex and donor deformation energy.

The RF, XGB, and SVM ML models were then employed to predict the identity of the principal (interacting) halogen atom for the BX and BAX systems based on these five most important features. Figure 2 shows the heat map of the confusion matrices applied to check the accuracy of the models. Each block in the heat map conveys the number of times a halogen atom was predicted to be the principal halogen atom (Predicted Halogen; x-axis) against the expected (or true) halogen atom (Expected Halogen; y-axis). The percentage displayed in each block is found by taking the number in that block (number of times the specific halogen atom was predicted) and dividing by the total number of predictions made (sum of total values in each block).

Table 2: Accuracy Score of *5*-fold Cross-Validation for the Prediction of the Principal Halogen Atom for the BX and BAX Systems.

System	RF	XGB	SVM
BX	100%	100%	100%
BAX	100%	100%	90%

The RF algorithm had over 99% accuracy when predicting the principal halogen atom for the halo-benzene XB donors (Figure 2a). The model predicted one structure to have chlorine

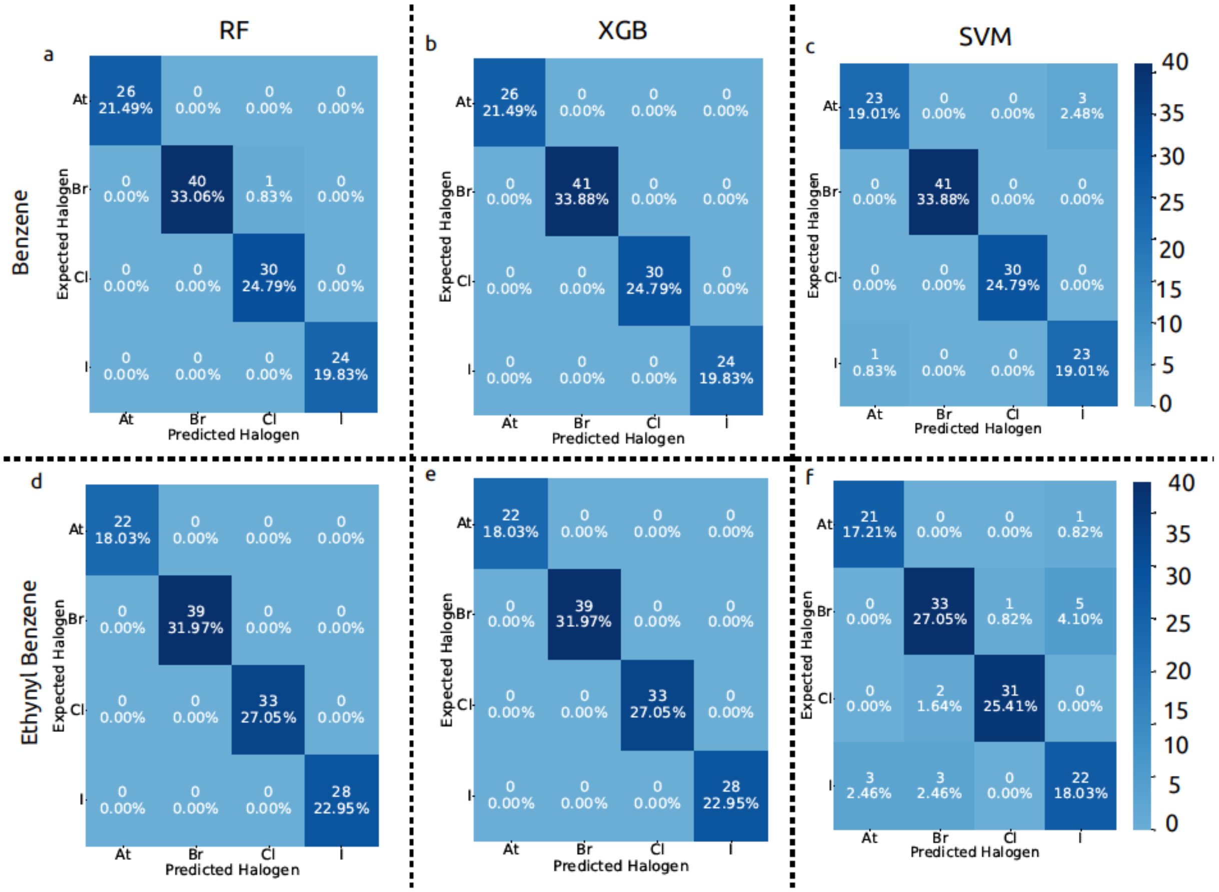

Figure 2: Heat map of the confusion matrix for the halo-benzene (top row) and halo-ethynyl benzene (bottom row) XB donors using the Random Forest (a and d), XGBoost (b and e), and Support Vector Machines (c and f) classification ML models.

as the principal halogen atom when bromine was the correct atom. The results showed excellent agreement with the 5-fold cross-validation that gave a 100% accuracy score for the RF model (Table 2). The XGB algorithm (Figure 2b) algorithm resulted in an accuracy score of 100%, agreeing with the cross-validation accuracy score. The SVM method (Figure 2c), however, showed a 96.69% accuracy for predicting the correct principal halogen atom. One structure was predicted to contain astatine when the correct halogen was iodine, and three structures were predicted to have an iodine atom present when the principal halogen atom was astatine. The SVM cross-validation predicted the halogen atoms with 100% accuracy for the BX systems.

Similar results can be seen for the BAX systems using the RF (Figure 2d) and XGB (Figure 2e) models, where both algorithms predicted the halogen atom with 100% accuracy. The SVM model (Figure 2f) predicted the principal halogen atom correctly 87.7% of the time. One structure was predicted to have an iodine atom when the principal halogen atom was astatine. Five halogen atoms were predicted as iodine while bromine was the appropriate halogen atom. One structure that contained a bromine halogen was predicted to have a chlorine atom. The model associated three structures with a bromine halogen when the structures contained an iodine halogen atom instead, and two structures were associated with a bromine atom when the principal halogen was chlorine. Finally, three structures were predicted to have astatine as the principal halogen when the appropriate principal halogen was iodine. The SVM algorithm with cross-validation gave a 90% prediction accuracy. This indicates that the RF and XGB models have a much better prediction accuracy for classifying the principal halogen atom in the XB donors/complex based on the data garnered from DFT calculations when compared to the SVM model. The ability of the three ML algorithms to separate the XB donors by their principal halogen atom highlights the significance of considering the identity (polarizability) of the interacting halogen atom in XB studies.

Regression

Molecular Fingerprint Prediction

In addition to predicting the principal halogen atom, the RF, XGB, and SVM models were utilized to predict the magnitude of the σ-hole ($V_{S,max}$), binding energy (E$_{bind}$), and the X$\cdots$N bond local force constant ($k^a_{X\cdots N}$) of the XB donor and its' corresponding complex with ammonia based on the XB donors' molecular fingerprint (MFP) and again with the properties calculated with DFT methods. We begin with applying the ML algorithms to the Morgan fingerprints of the XB donor molecules to demonstrate the important role that substitutions (i.e., electron-donating and electron-withdrawing groups) play in the determination of $V_{S,max}$, E$_{bind}$, and $k^a_{X\cdots N}$. The Morgan fingerprint is a type of hash fingerprint called the extended-connectivity fingerprint (ECFP). This fingerprint encodes fragments of a molecule as a binary vector through a hash function.[65] This gives a useful representation for identifying if an atom group (fragment) exists, thus distinguishing the different substituents in the XB donors and how they will affect the molecular properties.

In order to gain some understanding of the accuracy of the methods and observe how limiting the dataset will affect the accuracy of the ML models, the XB donors were split into separate groups by their principal halogen atom. Figures 3 and 4 display the DFT calculated vs ML predicted $V_{S,max}$, E$_{bind}$, and $k^a_{X\cdots N}$ properties based on the MFP for the chlorine (top row), bromine (second row), iodine (third row), and astatine (bottom row) halogen atoms in the BX and BAX systems, respectively. The coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE) are used to represent the accuracy of the models when predicting the selected properties. R^2 close to 1.00 and small RMSE/MAE values are evidence the ML algorithms are making predictions with a high degree of accuracy.

The BAX systems (Figure 4) are shown to have a smaller MAE and RMSE when predicting the $V_{S,max}$ (first column) based on the Morgan fingerprints compared to the BX systems

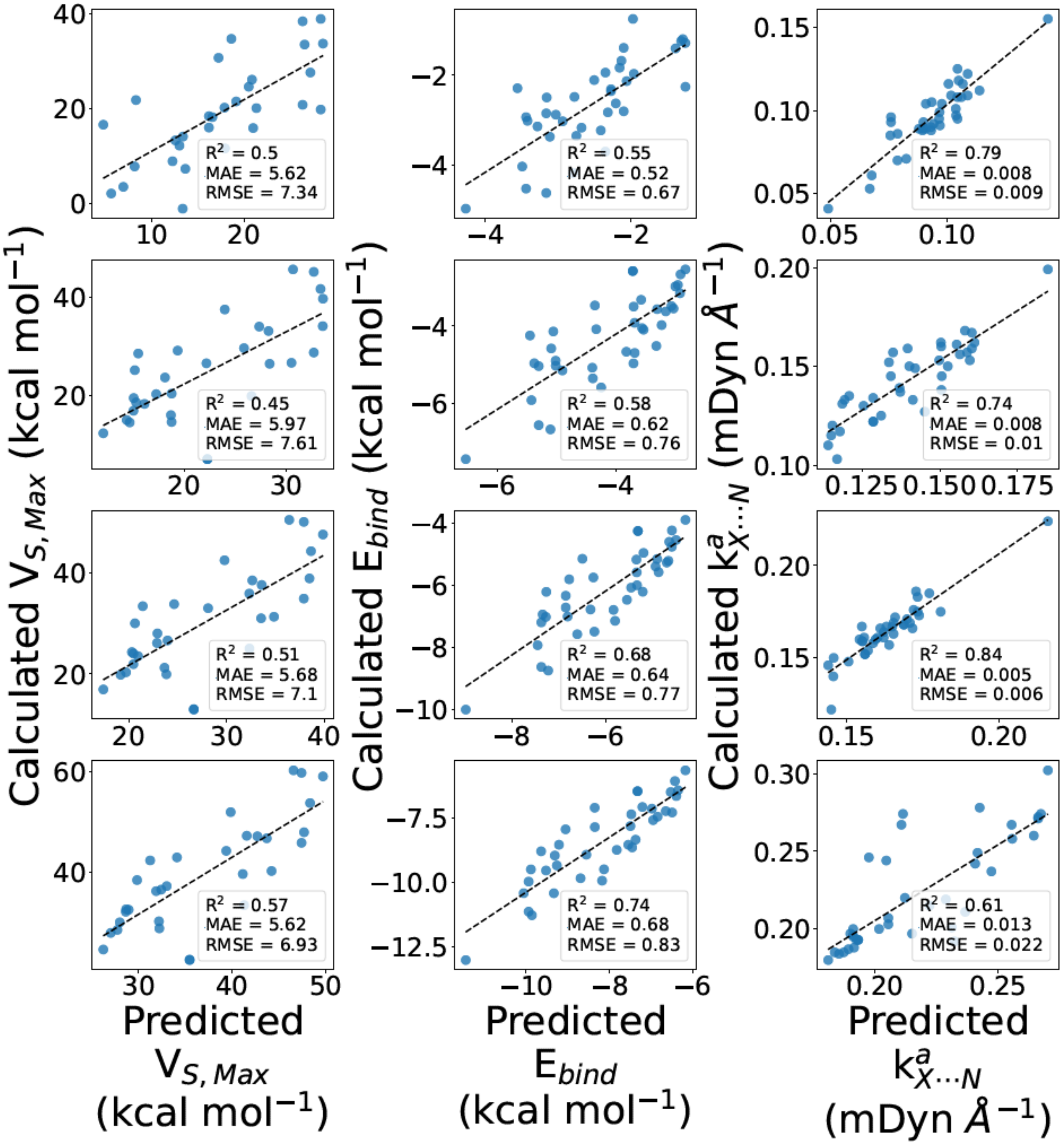

Figure 3: Regression plots of molecular fingerprint-based prediction of $V_{S,max}$ (left column), E$_{bind}$ (middle column), and $k^a_{X\cdots N}$ (right column) using the RF ML algorithm for the chlorine- (top row), bromine- (second row), iodine- (third row), and astatine-containing (bottom row) halo-benzene XB donors with their corresponding complexes.

(Figure 3). This is in part due to the BAX systems having a smaller range of $V_{S,max}$ values when compared to the BX systems because of the ethynyl linker serving as a steric factor, as reported in previous studies.[12] An interesting point to be made is that while the MAE and RMSE of $V_{S,max}$ for the BX XB donors is larger than for the BAX systems, R^2 for the BX systems is larger than for the BAX donors. This shows that the ML predicted vs DFT calculated $V_{S,max}$ is slightly more linear in the BX donors than the BAX systems. However, the MAE and RMSE for the $V_{S,max}$ is still quite large for both BX (> 5.6, >6.9) and BAX (> 4.2, > 5.4) XB donors. Similar results can be seen when using the XGB (Figures S3 and S4) and SVM (Figures S10 and S11) algorithms. These large errors for the $V_{S,max}$ prediction result from the difficulty of distinguishing conformers of the same molecule from the SMILES code and the small size of the dataset. The $V_{S,max}$ highly depends on the electronic environment, which can be altered in different conformers of the same molecule.[12,30] Consider a halo-benzene molecule with a hydroxyl group *ortho* to the halogen atom for example. One structure has the hydrogen atom in the hydroxyl group pointing toward the halogen, while another could have the hydrogen atom of the hydroxyl substituent pointing away from the halogen atom. Both molecules are conformers of one another (are local minima on the potential energy surface) and are represented by the same SMILES code. However, the electronic environment (especially around the halogen atom) is very different from one another. Therefore, the properties of the conformers (i.e., $V_{S,max}$) and even the properties of the XB complexes that are generated (i.e., E_{bind} and $k^a_{X \cdots N}$) will be vastly different.

The binding energy of the BX and BAX complexes with ammonia have an MAE and RMSE that range between 0.5 to 1.0 kcal mol^{-1} (second column of Figures 3 and 4). This remains true for all three algorithms used. The $X \cdots N$ bond local force constant has a range of 0.005 to 0.022 mDyn Å^{-1} for MAE and RMSE (third column of Figures 3 and 4). Similar to what was seen before with the $V_{S,max}$, the R^2 for E_{bind} and $k^a_{X \cdots N}$ is much larger in the BX systems than in the BAX complexes. The MAE and RMSE being within 1.0 kcal mol^{-1} for E_{bind} and 0.022 mDyn Å^{-1} for $k^a_{X \cdots N}$ shows that the interaction strength between the

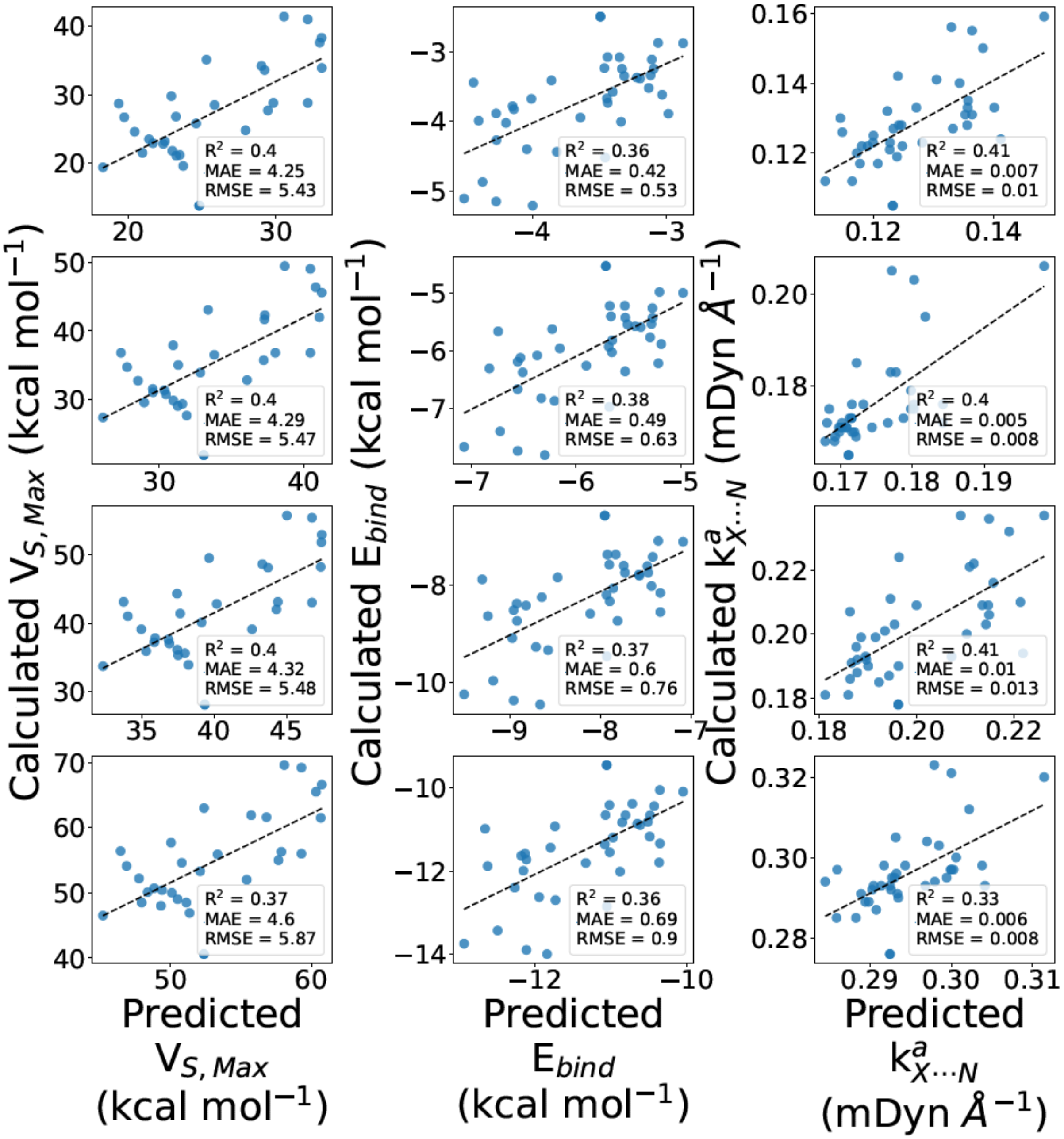

Figure 4: Regression plots of molecular fingerprint-based prediction of $V_{S,max}$ (left column), E$_{bind}$ (middle column), and $k^a_{X \cdots N}$ (right column) using the RF ML algorithm for the chlorine- (top row), bromine- (second row), iodine- (third row), and astatine-containing (bottom row) halo-ethynyl benzene XB donors with their corresponding complexes.

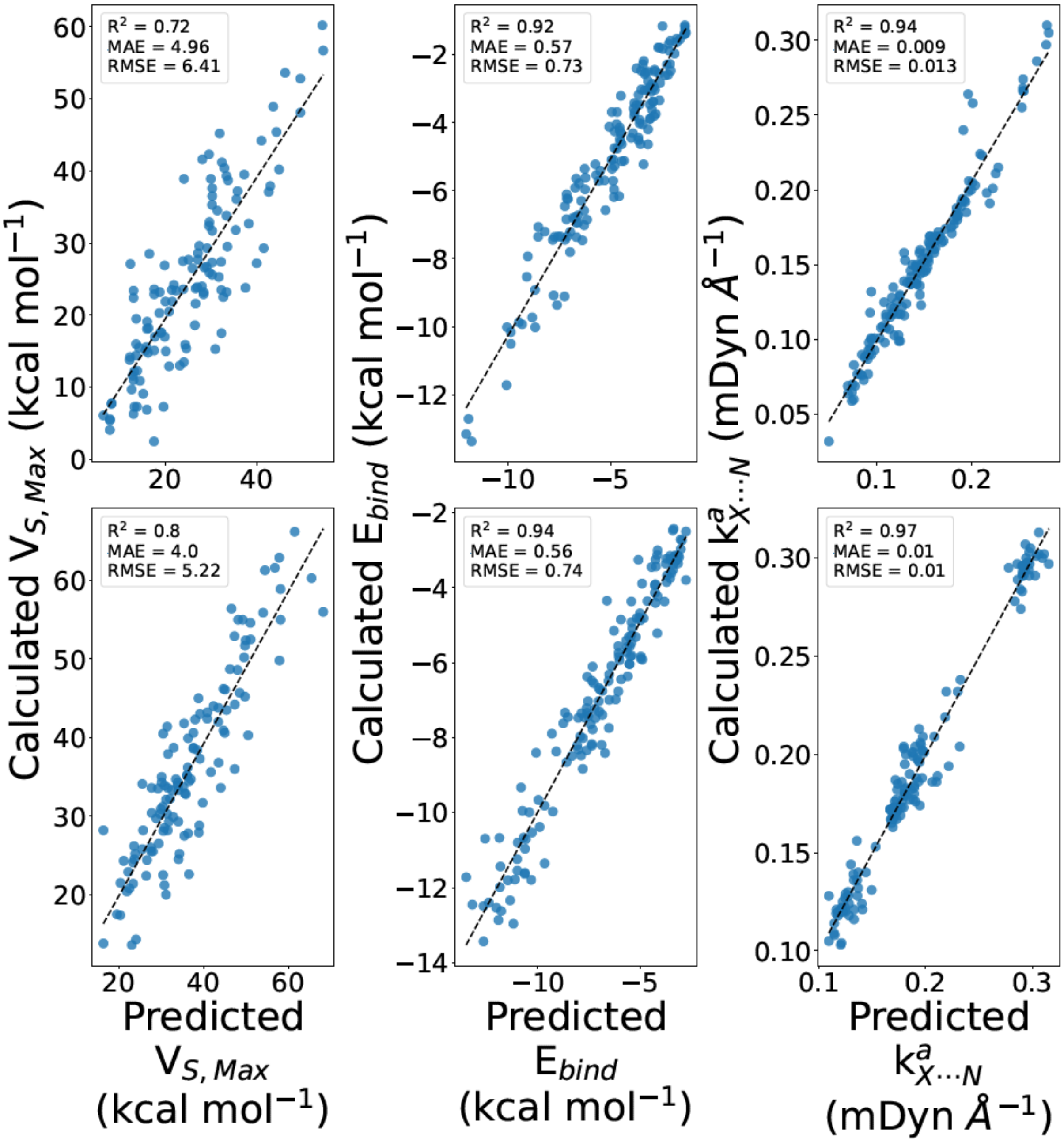

Figure 5: Regression plots of molecular fingerprint-based prediction of $V_{S,max}$ (left column), E_{bind} (middle column), and $k^{a}_{X \cdots N}$ (right column) using the RF ML algorithm for the all halo-benzene (top row) and halo-ethynyl benzene (bottom row) XB donors with their corresponding complexes.

symmetric XB donors and ammonia acceptor can be predicted from the MFP with good accuracy, even if the ML algorithms had a more difficult time predicting the $V_{S,max}$.

The small R^2 and large MAE and RMSE values when predicting the $V_{S,max}$ for the BX and BAX systems when separated by their principal halogen atom show that while the ML models can make decent predictions based on the MFPs in these smaller datasets, more data is required to make more accurate predictions. This fact can be seen by the smaller MAE/RMSE and larger R^2 values for the $V_{S,max}$ when all BX (top row) and BAX (bottom row) systems are taken into consideration (Figures 5, S5, and S12). The MAE and RMSE when predicting E$_{bind}$ for the total BX and BAX systems with the RF algorithm are slightly below the average MAE and RMSE produced when the XB complexes are separated by the principal halogen atom of the XB donor. The XGB (Figures S3 and S5) and SVM (Figures S10 and S12) algorithms show a vast improvement in lowering the prediction error for the E$_{bind}$ when grouping all the BX systems together into one dataset. The SVM (Figures S11 and S12) algorithm also shows the RMSE and MAE for the prediction of E$_{bind}$ to be much lower when grouping the BAX systems together compared to separating them by principal halogen atom, while the XGB (Figures S4 and S5) method has an RMSE and MAE of just below the average RMSE and MAE for E$_{bind}$ when separating the data by principal halogen atom. Tables 3 and 4, which display the metric values for each algorithm after performing *5*-fold cross-validation, are in agreement with these findings. This result shows that the prediction accuracy of the ML algorithms can be improved with additional data. These findings also start to show a distinction in the prediction accuracy between algorithms when more data is added to the dataset.

In an effort to expand the dataset a little more, and observe how it would affect the accuracy of the ML models, the BX and BAX datasets were conjoined into a single dataset (Figure 6). The MAE and RMSE for the prediction of the $V_{S,max}$ from the complete XB donor dataset (top left plot of Figure 6) is slightly larger than the average of the MAE and RMSE of the BX and BAX datasets. The E$_{bind}$ and $k^a_{X\cdots N}$ (top middle and right of Figure 6),

Table 3: R^2, MAE, and RMSE Accuracy Metrics for the MFP-based 5-fold Cross-Validation on the RF, XGB, and SVM models for the Chloro-Benzene (BCl), Bromo-Benzene (BBr), Iodo-Benzene (BI), Astato-Benzene (BAt), and All Halo-Benzene (BX) XB Donors.

Metric	Property	BCl	BBr	BI	BAt	BX
RF						
R^2	$V_{S,max}$	0.48	0.51	0.56	0.63	0.72
	E$_{bind}$	0.61	0.72	0.77	0.81	0.92
	$k^a_{X\cdots N}$	0.72	0.82	0.76	0.83	0.95
MAE	$V_{S,max}$	5.02	4.89	4.70	4.76	5.01
	E$_{bind}$	0.45	0.52	0.55	0.59	0.55
	$k^a_{X\cdots N}$	0.008	0.007	0.006	0.009	0.008
RMSE	$V_{S,max}$	6.42	6.36	6.02	5.90	6.34
	E$_{bind}$	0.59	0.66	0.71	0.79	0.70
	$k^a_{X\cdots N}$	0.012	0.010	0.010	0.014	0.011
XGB						
R^2	$V_{S,max}$	0.31	0.37	0.44	0.55	0.72
	E$_{bind}$	0.49	0.66	0.79	0.83	0.94
	$k^a_{X\cdots N}$	0.74	0.81	0.85	0.87	0.98
MAE	$V_{S,max}$	5.47	5.40	5.05	4.94	4.49
	E$_{bind}$	0.49	0.57	0.53	0.56	0.46
	$k^a_{X\cdots N}$	0.008	0.007	0.005	0.008	0.006
RMSE	$V_{S,max}$	7.39	7.23	6.79	6.52	6.33
	E$_{bind}$	0.67	0.73	0.68	0.74	0.61
	$k^a_{X\cdots N}$	0.012	0.010	0.007	0.012	0.008
SVM						
R^2	$V_{S,max}$	0.58	0.58	0.62	0.70	0.83
	E$_{bind}$	0.59	0.72	0.79	0.85	0.94
	$k^a_{X\cdots N}$	0.83	0.87	0.82	0.89	0.97
MAE	$V_{S,max}$	4.29	4.33	4.17	4.10	3.48
	E$_{bind}$	0.42	0.47	0.50	0.51	0.42
	$k^a_{X\cdots N}$	0.007	0.006	0.005	0.008	0.006
RMSE	$V_{S,max}$	5.70	5.84	5.53	5.28	4.95
	E$_{bind}$	0.60	0.65	0.68	0.68	0.61
	$k^a_{X\cdots N}$	0.009	0.008	0.008	0.012	0.008

*units for $V_{S,max}$ and E$_{bind}$ are in kcal mol^{-1} and $k^a_{X\cdots N}$ in mDyn Å^{-1}

Table 4: R^2, MAE, and RMSE Accuracy Metrics for the MFP-based *5*-fold Cross-Validation on the RF, XGB, and SVM models for the Chloro-Ethynl Benzene (BACl), Bromo-Ethynl Benzene (BABr), Iodo-Ethynl Benzene (BAI), Astato-Ethynl Benzene (BAAt), and All Halo-Ethynl Benzene (BAX) XB Donors.

Metric	Property	BACl	BABr	BAI	BAAt	BAX
			RF			
R^2	$V_{S,max}$	0.39	0.38	0.39	0.40	0.81
	E$_{bind}$	0.50	0.42	0.48	0.49	0.95
	$k^a_{X\cdots N}$	0.57	0.67	0.54	0.49	0.98
MAE	$V_{S,max}$	3.62	3.67	3.65	3.78	3.83
	E$_{bind}$	0.36	0.46	0.52	0.60	0.51
	$k^a_{X\cdots N}$	0.007	0.004	0.009	0.005	0.006
RMSE	$V_{S,max}$	4.75	4.79	4.76	4.96	4.99
	E$_{bind}$	0.46	0.61	0.67	0.78	0.67
	$k^a_{X\cdots N}$	0.008	0.005	0.011	0.007	0.009
			XGB			
R^2	$V_{S,max}$	0.21	0.21	0.22	0.24	0.81
	E$_{bind}$	0.39	0.30	0.35	0.36	0.95
	$k^a_{X\cdots N}$	0.50	0.63	0.42	0.43	0.98
MAE	$V_{S,max}$	3.99	4.05	3.97	4.09	3.48
	E$_{bind}$	0.38	0.49	0.55	0.64	0.46
	$k^a_{X\cdots N}$	0.007	0.004	0.009	0.006	0.006
RMSE	$V_{S,max}$	5.38	5.44	5.39	5.58	5.06
	E$_{bind}$	0.51	0.67	0.74	0.87	0.68
	$k^a_{X\cdots N}$	0.009	0.006	0.012	0.008	0.009
			SVM			
R^2	$V_{S,max}$	0.41	0.41	0.41	0.42	0.88
	E$_{bind}$	0.37	0.30	0.35	0.36	0.97
	$k^a_{X\cdots N}$	0.47	0.58	0.42	0.40	0.99
MAE	$V_{S,max}$	3.33	3.36	3.32	3.43	2.20
	E$_{bind}$	0.37	0.47	0.53	0.61	0.34
	$k^a_{X\cdots N}$	0.007	0.004	0.009	0.005	0.005
RMSE	$V_{S,max}$	4.63	4.66	4.64	4.82	3.99
	E$_{bind}$	0.52	0.67	0.75	0.87	0.53
	$k^a_{X\cdots N}$	0.009	0.006	0.012	0.008	0.007

*units for $V_{S,max}$ and E$_{bind}$ are in kcal mol^{-1} and $k^a_{X\cdots N}$ in mDyn Å^{-1}

on the other hand, show a slight improvement in prediction accuracy when compared to the separate BX and BAX systems. Similar results can be seen from the XGB (Figures S5 and middle row of Figure 6) and SVM (Figures S12 and bottom row of Figure 6) models. A comparison between the three ML models employed throughout this study can also be done based on these results. XGB has better prediction accuracy for the $V_{S,max}$, E$_{bind}$, and $k^a_{X \cdots N}$ properties of the XB donors and corresponding complexes compared to the RF model (top row of Figure 6). The SVM and XGB models give comparable results for prediction accuracy of the E$_{bind}$ and $k^a_{X \cdots N}$ of the XB complexes, while SVM performs better than both the XGB and RF techniques for predicting the $V_{S,max}$ based on the MFP of the XB donors.

Figure 6 shows that the RMSE for predicting the $V_{S,max}$ of the combined BX/BAX dataset is 5.36 – 5.92 kcal mol^{-1} for any of the three algorithms. In comparison to the ML study on σ-holes conducted by Heidrich *et al.*, where they found the RMSE for predicting the $V_{S,max}$ on an 0.001 au electron isodensity to be 0.0140 au (8.79 kcal mol^{-1}),[30] we seem to have a slightly better prediction accuracy. This could be due to our XB donors all having the same carbon backbone, whereas Heidrich *et al.* used a vast assortment of XB donors with differing carbon backbones and a significantly larger number of molecules (16,000 molecules compared to our 1,210 XB donors). When performing cross-validation, however, Heidrich *et al.* find the RMSE for predicting the $V_{S,max}$ to be 0.0061 au (3.83 kcal mol^{-1}). This is notably lower than RMSE we find for the BX and BAX systems after running *5*-fold cross-validation (Tables 3 and 4).

Data-Based Prediction

In conjunction with the molecular fingerprint-based predictions, we also implemented DFT data-based predictions. The features selected through the RFE procedure for the halo-benzene and halo-ethynyl benzene systems are displayed in Tables S2 and S3, respectively. The majority of the most important features for predicting the $V_{S,max}$, E$_{bind}$, and $k^a_{X \cdots N}$ properties are primarily properties found in the XB donor-acceptor complex. Very few of

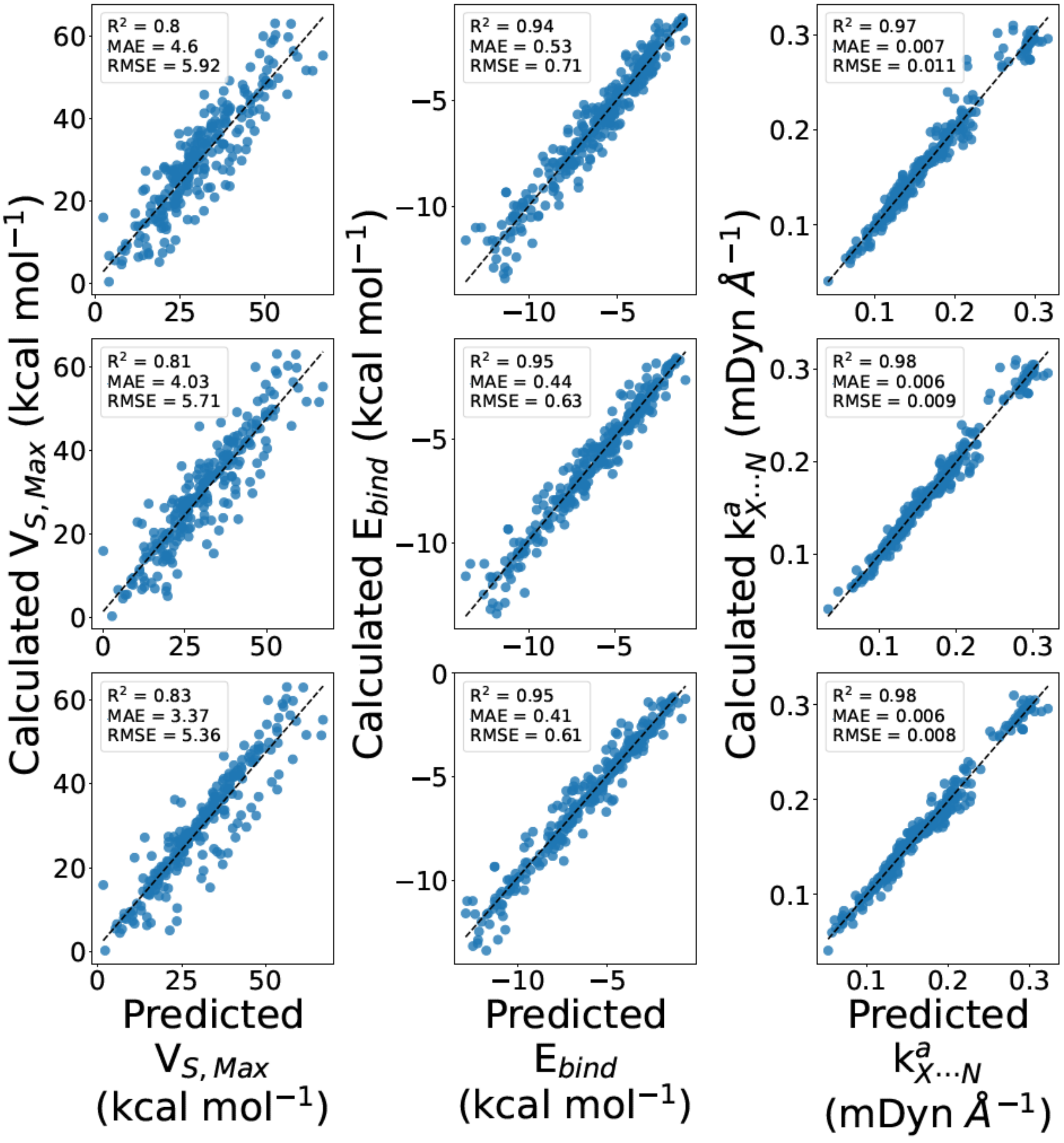

Figure 6: Regression plots for the molecular fingerprint-based prediction of all halo-benzene and halo-ethynyl benzene XB donors with the RF (top row), XGB (middle row), and SVM (bottom row) algorithms.

the features are found to be properties of the XB donor alone. E$_{bind}$ and $k^{a}_{X\cdots N}$ represent how strong the interaction in the XB complex is. Therefore, a more accurate prediction of the strength of the complex would be made with additional details from the complex itself, whereas the features from the XB donor alone would not properly portray how significant the interaction between the XB donor and acceptor might be. The $V_{S,max}$ is a measure of how attractive the XB donor is to a Lewis base. While details solely about the XB donor might lead to a relatively accurate estimation of the $V_{S,max}$ for the donor, features about the complex will provide a better prediction for how attractive the Lewis acid is to the nucleophile.

When predicting the $V_{S,max}$ systems, the energetic features (e.g., E$_{bind}$, E$^{def}_{donor}$, and E$^{def}_{NH_3}$) are found to have the highest prevalence (Tables S2 and S3). This is in part due to the RFE algorithm finding that the $V_{S,max}$ is very highly correlated to these energetic properties, as shown in our previous studies.[12,31] The spectroscopic, electronic, and structural properties have a varied importance, depending on the principal halogen atom of the XB donor. The electronic terms, especially the electron or energy density at the X$\cdots$N bond critical point, have a comparable (and potentially related) importance to the energetic terms when predicting the $V_{S,max}$ in the complete BX or BAX dataset, regardless of the identity of the principal halogen atom. Similar trends can be seen when predicting the E$_{bind}$ or $k^{a}_{X\cdots N}$ of the XB complex.

Figures 7 and 8 display regression plots for data-based prediction using the RF algorithm. These figures have a much lower MAE and RMSE, in addition to a much larger R^2, for the prediction of all properties ($V_{S,max}$, E$_{bind}$, and $k^{a}_{X\cdots N}$) compared to Figures 3 and 4. This shows that when certain properties of the complex are known, the remaining properties can be predicted accurately. Figures S6, S7, S13, and S14 also present lower MAE and RMSE and larger R^2 values compared to Figures S3, S4, S10, and S11, respectively. Thus, the XGB and SVM models also predict the $V_{S,max}$, E$_{bind}$, and $k^{a}_{X\cdots N}$ more accurately when other properties of the complex are known. These results are agreed upon when the 5-fold cross-

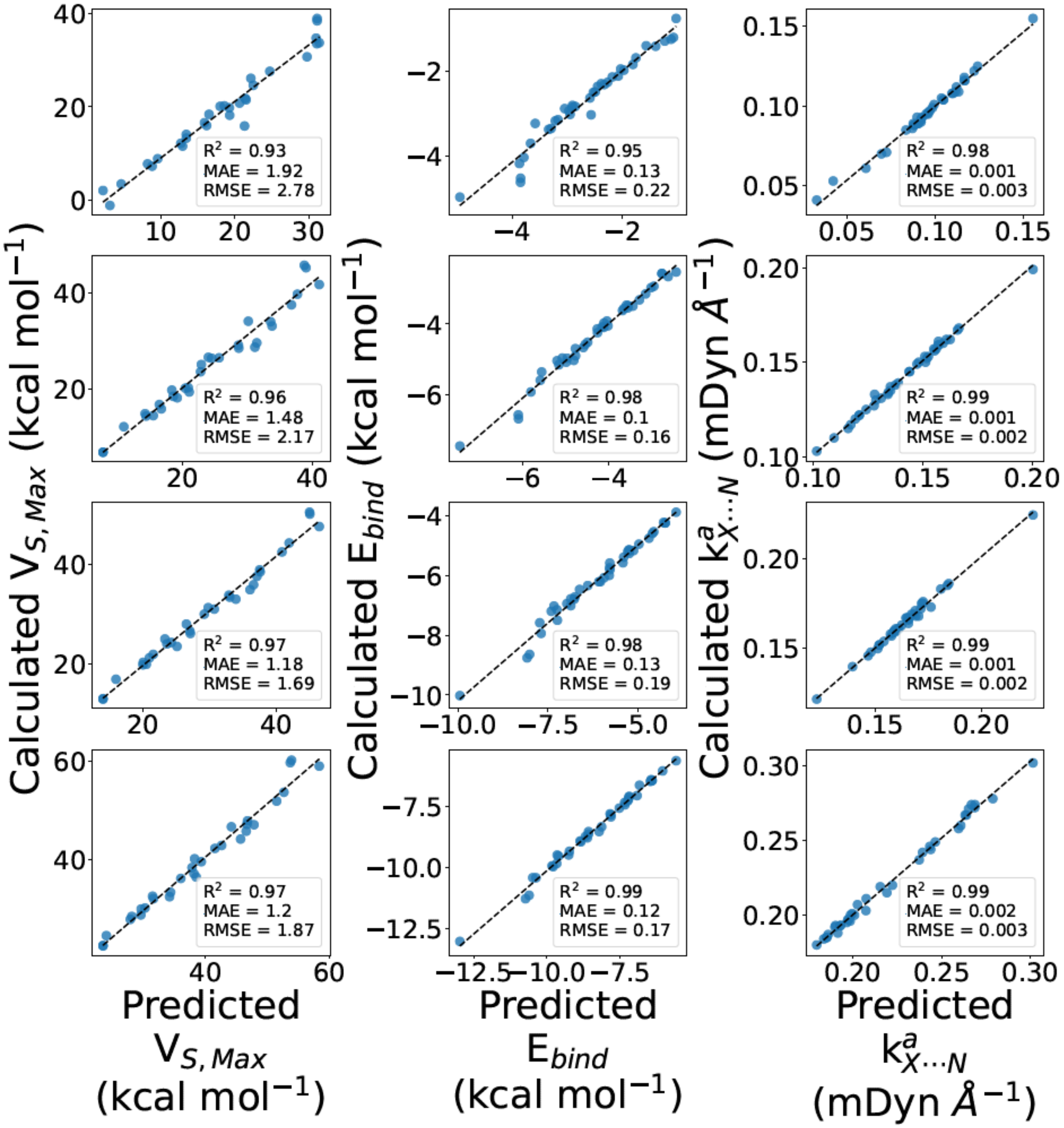

Figure 7: Regression plots of data-based prediction of $V_{S,max}$ (left column), E_{bind} (middle column), and $k^a_{X \cdots N}$ (right column) using the RF ML algorithm for the chlorine- (top row), bromine- (second row), iodine- (third row), and astatine-containing (bottom row) halobenzene XB donors with their corresponding complexes.

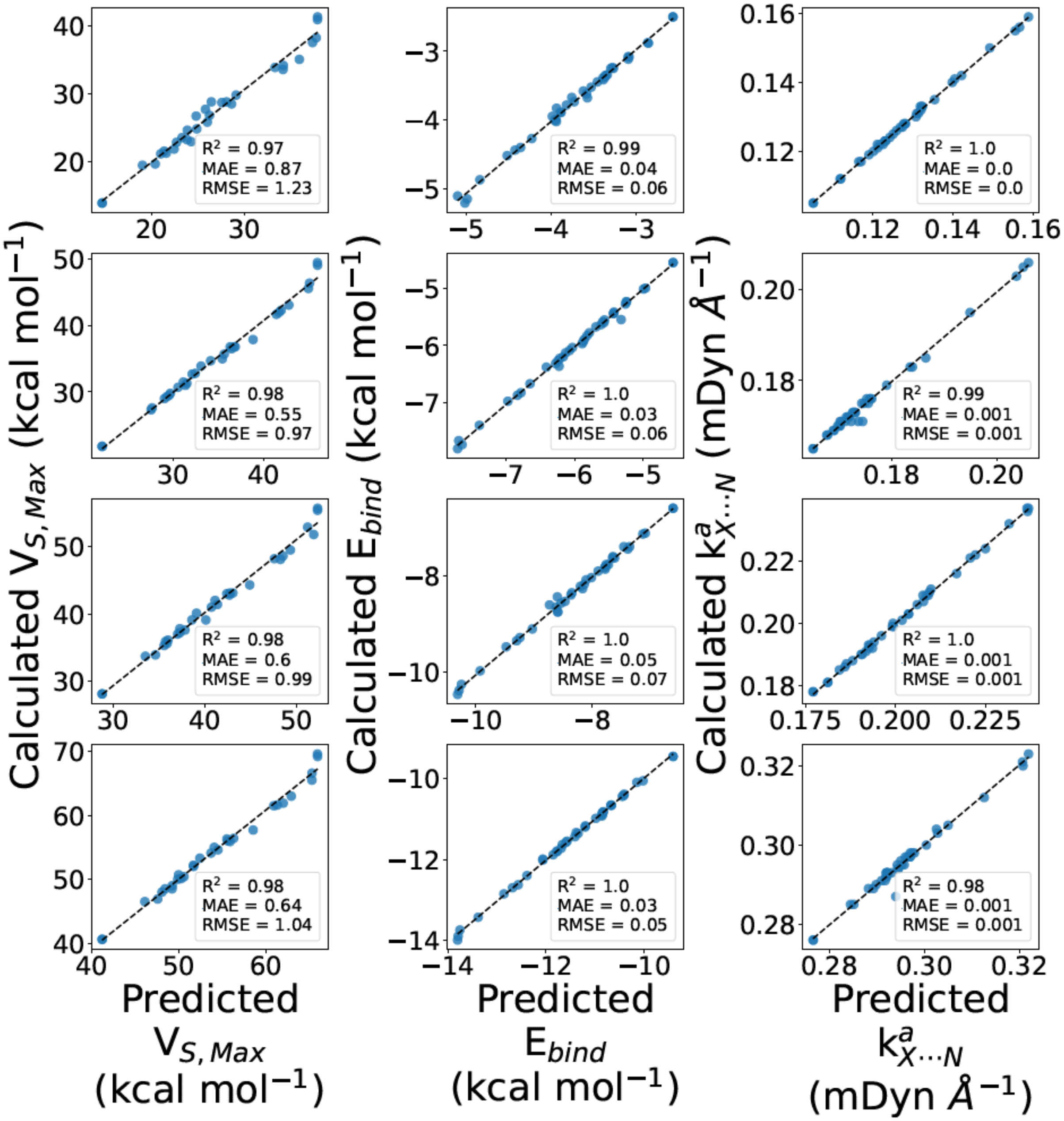

Figure 8: Regression plots of data-based prediction of $V_{S,max}$ (left column), E$_{bind}$ (middle column), and $k^a_{X\cdots N}$ (right column) using the RF ML algorithm for the chlorine- (top row), bromine- (second row), iodine- (third row), and astatine-containing (bottom row) halo-ethynyl benzene XB donors with their corresponding complexes.

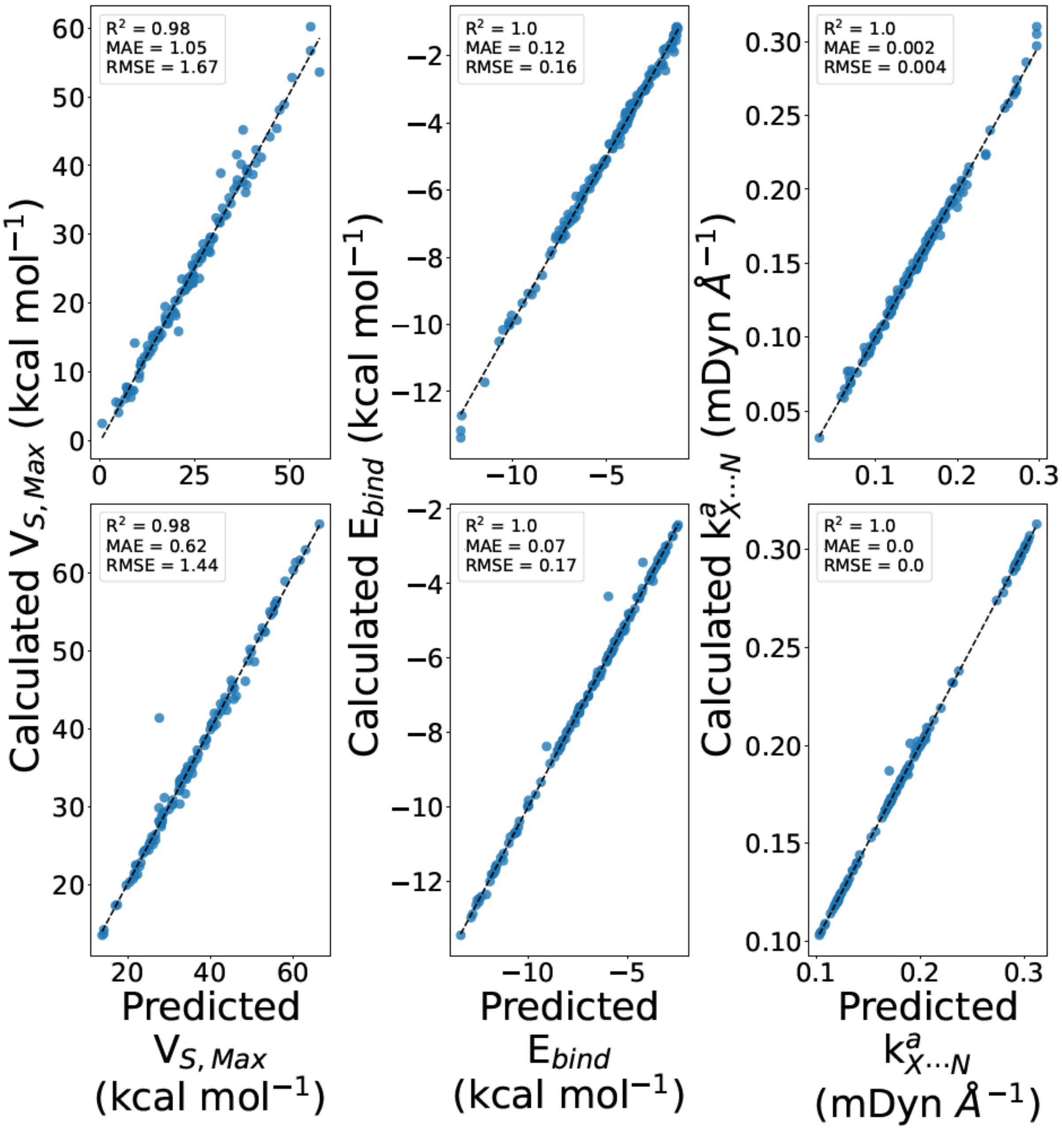

Figure 9: Regression plots of data-based prediction of $V_{S,max}$ (left column), E_{bind} (middle column), and $k^a_{X\cdots N}$ (right column) using the RF ML algorithm for the all halo-benzene (top row) and halo-ethynyl benzene (bottom row) XB donors with their corresponding complexes.

Table 5: R^2, MAE, and RMSE Accuracy Metrics for the data-based 5-fold Cross-Validation on the RF, XGB, and SVM models for the Chloro-Benzene (BCl), Bromo-Benzene (BBr), Iodo-Benzene (BI), Astato-Benzene (BAt), and All Halo-Benzene (BX) XB Donors.

Metric	Property	BCl	BBr	BI	BAt	BX
RF						
R^2	$V_{S,max}$	0.96	0.97	0.97	0.98	0.98
	E$_{bind}$	0.98	0.99	0.99	0.98	1.00
	$k^a_{X\cdots N}$	0.97	0.99	0.99	0.99	0.99
MAE	$V_{S,max}$	1.30	1.16	1.05	0.96	1.13
	E$_{bind}$	0.10	0.07	0.10	0.15	0.11
	$k^a_{X\cdots N}$	0.002	0.001	0.001	0.002	0.002
RMSE	$V_{S,max}$	1.71	1.63	1.46	1.40	1.69
	E$_{bind}$	0.14	0.11	0.15	0.22	0.16
	$k^a_{X\cdots N}$	0.004	0.002	0.002	0.003	0.004
XGB						
R^2	$V_{S,max}$	0.96	0.98	0.96	0.97	0.98
	E$_{bind}$	0.97	0.99	0.99	0.99	1.00
	$k^a_{X\cdots N}$	0.95	0.99	0.99	0.99	1.00
MAE	$V_{S,max}$	1.25	1.02	1.19	1.02	1.15
	E$_{bind}$	0.11	0.08	0.11	0.14	0.12
	$k^a_{X\cdots N}$	0.002	0.002	0.001	0.003	0.002
RMSE	$V_{S,max}$	1.72	1.40	1.89	1.52	1.65
	E$_{bind}$	0.16	0.11	0.16	0.19	0.17
	$k^a_{X\cdots N}$	0.005	0.003	0.002	0.004	0.004
SVM						
R^2	$V_{S,max}$	0.91	0.90	0.90	0.91	0.86
	E$_{bind}$	0.93	0.91	0.96	0.89	0.91
	$k^a_{X\cdots N}$	0.95	0.92	0.87	0.93	0.84
MAE	$V_{S,max}$	2.10	2.17	2.29	2.36	3.66
	E$_{bind}$	0.17	0.27	0.22	0.46	0.60
	$k^a_{X\cdots N}$	0.004	0.005	0.005	0.007	0.013
RMSE	$V_{S,max}$	2.60	2.90	2.98	3.02	4.51
	E$_{bind}$	0.24	0.38	0.31	0.59	0.74
	$k^a_{X\cdots N}$	0.005	0.006	0.007	0.009	0.020

units for $V_{S,max}$ and E$_{bind}$ are in kcal mol^{-1} and $k^a_{X\cdots N}$ in mDyn Å^{-1}

Table 6: R^2, MAE, and RMSE Accuracy Metrics for the data-based *5*-fold Cross-Validation on the RF, XGB, and SVM models for the Chloro-Ethynyl Benzene (BACl), Bromo-Ethynyl Benzene (BABr), Iodo-Ethynyl Benzene (BAI), Astato-Ethynyl Benzene (BAAt), and All Halo-Ethynyl Benzene (BAX) XB Donors.

Metric	Property	BACl	BABr	BAI	BAAt	BAX
			RF			
R^2	$V_{S,max}$	0.97	0.99	0.99	0.99	0.99
	E$_{bind}$	0.98	0.94	0.99	1.00	1.00
	$k^a_{X \cdots N}$	0.99	0.99	0.99	0.95	1.00
MAE	$V_{S,max}$	0.64	0.48	0.45	0.42	0.56
	E$_{bind}$	0.04	0.05	0.04	0.04	0.05
	$k^a_{X \cdots N}$	0.001	0.001	0.001	0.001	0.001
RMSE	$V_{S,max}$	0.93	0.72	0.59	0.58	0.88
	E$_{bind}$	0.08	0.16	0.06	0.06	0.10
	$k^a_{X \cdots N}$	0.001	0.001	0.002	0.002	0.001
			XGB			
R^2	$V_{S,max}$	0.97	0.99	0.99	0.99	0.99
	E$_{bind}$	0.98	0.94	0.99	0.99	1.00
	$k^a_{X \cdots N}$	0.99	0.98	0.99	0.94	1.00
MAE	$V_{S,max}$	0.76	0.49	0.46	0.51	0.63
	E$_{bind}$	0.05	0.07	0.05	0.06	0.05
	$k^a_{X \cdots N}$	0.001	0.001	0.001	0.001	0.001
RMSE	$V_{S,max}$	1.09	0.67	0.60	0.69	0.91
	E$_{bind}$	0.08	0.17	0.08	0.12	0.10
	$k^a_{X \cdots N}$	0.001	0.001	0.002	0.003	0.001
			SVM			
R^2	$V_{S,max}$	0.95	0.95	0.98	0.98	0.92
	E$_{bind}$	0.17	0.94	0.01	0.98	0.97
	$k^a_{X \cdots N}$	0.97	0.54	0.98	0.91	0.88
MAE	$V_{S,max}$	0.80	0.76	0.65	0.64	2.60
	E$_{bind}$	0.50	0.10	0.77	0.11	0.41
	$k^a_{X \cdots N}$	0.001	0.004	0.002	0.002	0.017
RMSE	$V_{S,max}$	1.26	1.28	0.94	0.92	3.28
	E$_{bind}$	0.62	0.18	0.95	0.16	0.50
	$k^a_{X \cdots N}$	0.002	0.006	0.002	0.003	0.021

units for $V_{S,max}$ and E$_{bind}$ are in kcal mol^{-1} and $k^a_{X \cdots N}$ in mDyn Å^{-1}

validation is performed on the data (Tables 5 and 6). Data of the electron density, crystal structure, and dissociation energy determined through experimental methods[66–68] could also potentially be used to improve predictions of XB properties like the $V_{S,max}$, E_{bind}, or $k^{a}_{X\cdots N}$.

When the complete BX or BAX dataset are taken into consideration upon running the ML models, the MAE and RMSE are observed to be comparable if not slightly lower than when the dataset is split into smaller batches by the structures' principal halogen atom for the RF (Figure 9) and XGB (Figure S8) models. However, the SVM method (Figure S15) shows the opposite. The SVM is able to predict the $V_{S,max}$, E_{bind}, and $k^{a}_{X\cdots N}$ properties of the XB donor and its' corresponding complex with a higher degree of accuracy when the dataset is split by the principal halogen atom. This may be because the SVM algorithm struggles to separate densely packed data points and operates more effectively in sparse datasets, whereas the RF and XGB algorithms are able to more easily distinguish groups or trends in tightly compacted datasets. By adding the datasets of each individual principal halogen atom together, the data points become less sparse, decreasing the overall accuracy of the SVM model. The DFT-based predictions are more accurate than the MFP-based approach for all three ML algorithms implemented. However, the MFP predictions provide reasonable initial values for the electronic environment of the XB donor and the interaction strength of the corresponding complex with ammonia while avoiding the time-consuming and complicated DFT quantum chemical calculations.

Conclusion

In this proof-of-concept study, we have employed three ML modelling techniques for the classification of the identity of the principal halogen atom in XB donors and complexes based on DFT collected data. We have shown that the RF, XGB, and SVM models have comparable results when identifying the principal halogen atom in the halo-(ethynyl)benzene systems, however the SVM model suffers in accuracy when predicting the halo-ethynyl benzene systems. The ML algorithms having $> 90\%$ prediction accuracy for separating the XB donors by the principal halogen atom emphasizes the importance of considering the identity of the interacting halogen atom when performing XB studies.

In addition, we have demonstrated that the RF, XGB, and SVM regression algorithms can predict the $V_{S,max}$, E_{bind}, and $k^a_{X\cdots N}$ of the symmetric XB donor and corresponding complex with ammonia based on the XB donor's molecular fingerprints and DFT calculated properties. The prediction of the $V_{S,max}$ of the XB donor suffers due to limited data and difficulty in distinguishing molecular conformers based on the SMILES code; however, the $V_{S,max}$ can still be predicted within 5 kcal mol^{-1}, or within 36%, 17%, 21%, and 11% error (from MAE) of the mean for chloro-benzene, chloro-ethynyl benzene, iodo-benzene, and iodo-ethynyl benzene donors, respectively. The interaction strength of the XB complex (i.e., E_{bind} and $k^a_{X\cdots N}$) is predicted to a much higher degree of accuracy (within 1.0 kcal mol^{-1} and 0.022 mDyn Å^{-1}) from the molecular fingerprint alone. In terms of relative comparison, this falls within 22%, 12%, 11%, and 7% error (from MAE) of the mean for the binding energy in the chloro-benzene, chloro-ethynyl benzene, iodo-benzene, and iodo-ethynyl benzene XB complexes, respectively. The accuracy of the ML models estimating $V_{S,max}$ of the XB donors and the interaction strength of the complexes can be further improved upon when using pre-calculated DFT data. Future work will involve 3D and 4D descriptors to account for conformational differences in the electronic environment of the XB donor.

Acknowledgement

The authors were supported by the National Science Foundation (NSF) under award number 2147956. The authors thank Baylor University High Performance Research Computing services for their generous allocation of computational resources on the Kodiak HPC cluster, and the CATCO Research group for the use of the Local Mode Analysis code.

Supporting Information Available

The additional figures and explanation of the models can be found in the Supporting Information PDF. The data and code used throughout this study, including all figures made, can be found in the GitHub page: `https://github.com/daniel-devore/XB-ML`.

References

(1) Desiraju, G. R.; Ho, P. S.; Kloo, L.; Legon, A. C.; Marquardt, R.; Metrangolo, P.; Politzer, P.; Resnati, G.; Rissanen, K. Definition of the halogen bond (IUPAC Recommendations 2013). *Pure and Applied Chemistry* **2013**, *85*, 1711–1713.

(2) Parisini, E.; Metrangolo, P.; Pilati, T.; Resnati, G.; Terraneo, G. Halogen bonding in halocarbon–protein complexes: a structural survey. *Chemical Society Reviews* **2011**, *40*, 2267–2278.

(3) Metrangolo, P.; Meyer, F.; Pilati, T.; Resnati, G.; Terraneo, G. Halogen Bonding in Supramolecular Chemistry. *Angewandte Chemie International Edition* **2008**, *47*, 6114–6127.

(4) Politzer, P.; Murray, J. S.; Clark, T. Halogen bonding: an electrostatically-driven highly directional noncovalent interaction. *Physical Chemistry Chemical Physics* **2010**, *12*, 7748–7757.

(5) Sedlak, R.; Kolář, M. H.; Hobza, P. Polar Flattening and the Strength of Halogen Bonding. *Journal of Chemical Theory and Computation* **2015**, *11*, 4727–4732.

(6) Bent, H. A. An Appraisal of Valence-bond Structures and Hybridization in Compounds of the First-row elements. *Chemical Reviews* **1961**, *61*, 275–311.

(7) Szczęśniak, M. M.; Chałasinski, G. Reassessing the Role of σ Holes in Noncovalent Interactions: It is Pauli Repulsion that Counts. *Frontiers in Chemistry* **2022**, *10*.

(8) Sakai, T.; Torii, H. Substituent Effect and Its Halogen-Atom Dependence of Halogen Bonding Viewed through Electron Density Changes. *Chemistry – An Asian Journal* **2023**, *18*, e202201196.

(9) Xu, Y.; Hao, A.; Xing, P. X$\cdots$X Halogen Bond-Induced Supramolecular Helices. *Angewandte Chemie International Edition* **2022**, *61*, e202113786.

(10) Cunha, A. V.; Havenith, R. W. A.; van Gog, J.; De Vleeschouwer, F.; De Proft, F.; Herrebout, W. The Halogen Bond in Weakly Bonded Complexes and the Consequences for Aromaticity and Spin-Orbit Coupling. *Molecules* **2023**, *28*, 772.

(11) Nziko, V. d. P. N.; Scheiner, S. Comparison of π-hole tetrel bonding with σ-hole halogen bonds in complexes of XCN (X = F, Cl, Br, I) and NH3. *Physical Chemistry Chemical Physics* **2016**, *18*, 3581–3590.

(12) Devore, D. P.; Ellington, T. L.; Shuford, K. L. Elucidating the Role of Electron-Donating Groups in Halogen Bonding. *The Journal of Physical Chemistry A* **2024**,

(13) Riley, K. E.; Murray, J. S.; Fanfrlík, J.; Řezáč, J.; Solá, R. J.; Concha, M. C.; Ramos, F. M.; Politzer, P. Halogen bond tunability I: the effects of aromatic fluorine substitution on the strengths of halogen-bonding interactions involving chlorine, bromine, and iodine. *Journal of Molecular Modeling* **2011**, *17*, 3309–3318.

(14) Riley, K. E.; Murray, J. S.; Fanfrlík, J.; Řezáč, J.; Solá, R. J.; Concha, M. C.; Ramos, F. M.; Politzer, P. Halogen bond tunability II: the varying roles of electrostatic and dispersion contributions to attraction in halogen bonds. *Journal of Molecular Modeling* **2013**, *19*, 4651–4659.

(15) Clark, T. σ-Holes. *WIREs Computational Molecular Science* **2013**, *3*, 13–20.

(16) Politzer, P.; Murray, J. S.; Clark, T.; Resnati, G. The σ-hole revisited. *Physical Chemistry Chemical Physics* **2017**, *19*, 32166–32178.

(17) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. Halogen bonding: the σ-hole. *Journal of Molecular Modeling* **2007**, *13*, 291–296.

(18) Cavallo, G.; Metrangolo, P.; Milani, R.; Pilati, T.; Priimagi, A.; Resnati, G.; Terraneo, G. The Halogen Bond. *Chemical Reviews* **2016**, *116*, 2478–2601.

(19) Quiñonero, D. Sigma-hole carbon-bonding interactions in carbon–carbon double bonds: an unnoticed contact. *Physical Chemistry Chemical Physics* **2017**, *19*, 15530–15540.

(20) Alaminsky, R. J.; Seminario, J. M. Sigma-holes from iso-molecular electrostatic potential surfaces. *Journal of Molecular Modeling* **2019**, *25*, 160.

(21) Donald, K. J.; Pham, N.; Ravichandran, P. Sigma Hole Potentials as Tools: Quantifying and Partitioning Substituent Effects. *The Journal of Physical Chemistry A* **2023**, *127*, 10147–10158.

(22) Yu, S.; Rautiainen, J. M.; Kumar, P.; Gentiluomo, L.; Ward, J. S.; Rissanen, K.; Puttreddy, R. Ortho-Substituent Effects on Halogen Bond Geometry for N-Haloimide $\cdots$ 2-Substituted Pyridine Complexes. *Advanced Science* **2024**, *11*, 2307208.

(23) Rautiainen, J. M.; Valkonen, A.; Lundell, J.; Rissanen, K.; Puttreddy, R. The Geometry and Nature of C–I $\cdots$ O–N Interactions in Perfluoroiodobenzene-Pyridine N-oxide Halogen-Bonded Complexes. *Advanced Science* **2024**, *11*, 2403945.

(24) Alkorta, I.; Sánchez-Sanz, G.; Elguero, J. Linear free energy relationships in halogen bonds. *CrystEngComm* **2013**, *15*, 3178–3186.

(25) Politzer, P.; Murray, J. S. Halogen Bonding: An Interim Discussion. *ChemPhysChem* **2013**, *14*, 278–294.

(26) Herrmann, B.; Svatunek, D. Directionality of Halogen-Bonds: Insights from 2D Energy Decomposition Analysis. *Chemistry – An Asian Journal* **2024**, *19*, e202301106.

(27) Stone, A. J. Are Halogen Bonded Structures Electrostatically Driven? *Journal of the American Chemical Society* **2013**, *135*, 7005–7009.

(28) Ai-Guo, Z. Dissecting the nature of halogen bonding interactions from energy decomposition and wavefunction analysis. *Monatshefte für Chemie - Chemical Monthly* **2017**, *148*, 1259–1267.

(29) Titov, O. I.; Shulga, D. A.; Palyulin, V. A.; Zefirov, N. S. Perspectives of Halogen Bonding Description in Scoring Functions and QSAR/QSPR: Substituent Effects in Aromatic Core. *Molecular Informatics* **2015**, *34*, 404–416.

(30) Heidrich, J.; Exner, T. E.; Boeckler, F. M. Predicting the Magnitude of σ-Holes Using VmaxPred, a Fast and Efficient Tool Supporting the Application of Halogen Bonds in Drug Discovery. *Journal of Chemical Information and Modeling* **2019**, *59*, 636–643.

(31) Devore, D.; Ellington, T.; Shuford, K. Illuminating the Performance of Electron Withdrawing Groups in Halogen Bonding. *ChemPhysChem* **2024**, e202400607.

(32) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts* **2008**, *120*, 215–241.

(33) Dunning, T. H., Jr. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *The Journal of Chemical Physics* **1989**, *90*, 1007–1023.

(34) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *The Journal of Chemical Physics* **1992**, *96*, 6796–6806.

(35) Woon, D. E.; Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. *The Journal of Chemical Physics* **1993**, *98*, 1358–1371.

(36) Frisch, M. J. et al. Gaussian˜16 Revision C.01. 2016.

(37) Peterson, K. A.; Figgen, D.; Goll, E.; Stoll, H.; Dolg, M. Systematically convergent basis sets with relativistic pseudopotentials. II. Small-core pseudopotentials and correlation consistent basis sets for the post-d group 16–18 elements. *The Journal of Chemical Physics* **2003**, *119*, 11113–11123.

(38) Peterson, K. A.; Shepler, B. C.; Figgen, D.; Stoll, H. On the Spectroscopic and Thermochemical Properties of ClO, BrO, IO, and Their Anions. *The Journal of Physical Chemistry A* **2006**, *110*, 13877–13883.

(39) Foster, J. P.; Weinhold, F. Natural hybrid orbitals. *Journal of the American Chemical Society* **1980**, *102*, 7211–7218.

(40) Reed, A. E.; Weinhold, F. Natural bond orbital analysis of near-Hartree–Fock water dimer. *The Journal of Chemical Physics* **1983**, *78*, 4066–4073.

(41) Reed, A. E.; Weinhold, F. Natural localized molecular orbitals. *The Journal of Chemical Physics* **1985**, *83*, 1736–1740.

(42) Reed, A. E.; Weinstock, R. B.; Weinhold, F. Natural population analysis. *The Journal of Chemical Physics* **1985**, *83*, 735–746.

(43) Carpenter, J. E.; Weinhold, F. Analysis of the geometry of the hydroxymethyl radical by the "different hybrids for different spins" natural bond orbital procedure. *Journal of Molecular Structure: THEOCHEM* **1988**, *169*, 41–62.

(44) Reed, A. E.; Curtiss, L. A.; Weinhold, F. Intermolecular interactions from a natural bond orbital, donor-acceptor viewpoint. *Chemical Reviews* **1988**, *88*, 899–926.

(45) Weinhold, F.; Carpenter, J. E. In *The Structure of Small Molecules and Ions*; Naaman, R., Vager, Z., Eds.; Springer US: Boston, MA, 1988; pp 227–236.

(46) Boys, S.; Bernardi, F. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Molecular Physics* **1970**, *19*, 553–566.

(47) Kraka, E.; Zou, W.; Tao, Y. Decoding chemical information from vibrational spectroscopy data: Local vibrational mode theory. *WIREs Computational Molecular Science* **2020**, *10*, e1480.

(48) Kraka, E.; Cremer, D. Dieter Cremer's contribution to the field of theoretical chemistry. *International Journal of Quantum Chemistry* **2019**, *119*, e25849.

(49) Tao, Y.; Zou, W.; Nanayakkara, S.; Kraka, E. LModeA-nano: A PyMOL Plugin for Calculating Bond Strength in Solids, Surfaces, and Molecules via Local Vibrational Mode Analysis. *Journal of Chemical Theory and Computation* **2022**, *18*, 1821–1837.

(50) Zou, W.; Tao, Y.; Freindorf, M.; Makos, M.; Verma, N.; Kraka, E. LMODEA2020. 2022.

(51) Bader, R. F. W. Bond Paths Are Not Chemical Bonds. *The Journal of Physical Chemistry A* **2009**, *113*, 10391–10396.

(52) Bader, R. F. W.; Carroll, M. T.; Cheeseman, J. R.; Chang, C. Properties of atoms in molecules: atomic volumes. *Journal of the American Chemical Society* **1987**, *109*, 7968–7979.

(53) Bader, R. F. W.; Nguyen-Dang, T. T. In *Advances in Quantum Chemistry*; Löwdin, P.-O., Ed.; Academic Press, 1981; Vol. 14; pp 63–124.

(54) Lu, T.; Chen, F. Multiwfn: A multifunctional wavefunction analyzer. *Journal of Computational Chemistry* **2012**, *33*, 580–592.

(55) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(56) dmlc: XGBoost. `https://xgboost.ai/`.

(57) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA, 2016; pp 785–794.

(58) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.

(59) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences* **1989**, *29*, 97–101.

(60) Weininger, D. SMILES. 3. DEPICT. Graphical depiction of chemical structures. *Journal of Chemical Information and Computer Sciences* **1990**, *30*, 237–243.

(61) RDKit: Open-source cheminformatics. `https://www.rdkit.org`.

(62) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **1965**, *5*, 107–113.

(63) Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source Software* **2021**, *6*, 3021.

(64) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* **2007**, *9*, 90–95.

(65) Zhong, S.; Guan, X. Count-Based Morgan Fingerprint: A More Efficient and Interpretable Molecular Representation in Developing Machine Learning-Based Predictive

Regression Models for Water Contaminants' Activities and Properties. *Environmental Science & Technology* **2023**, *57*, 18193–18202.

(66) Smith, G. T.; Mallinson, P. R.; Frampton, C. S.; Farrugia, L. J.; Peacock, R. D.; ; Howard, J. A. K. Experimental Determination of the Electron Density Topology in a Non-centrosymmetric Transition Metal Complex: [Ni(H_3L)][NO_3][PF_6] [H_3L = N,N",N"-Tris(2-hydroxy-3-methylbutyl)-1,4,7-triazacyclononane]. *ACS Publications* **1997**,

(67) Frey, J. A.; Holzer, C.; Klopper, W.; Leutwyler, S. Experimental and Theoretical Determination of Dissociation Energies of Dispersion-Dominated Aromatic Molecular Complexes. *Chemical Reviews* **2016**, *116*, 5614–5641.

(68) Gibbs, G. V.; Downs, R. T.; Cox, D. F.; Rosso, K. M.; Ross, N. L.; Kirfel, A.; Lippmann, T.; Morgenroth, W.; Crawford, T. D. Experimental Bond Critical Point and Local Energy Density Properties Determined for Mn-O, Fe-O, and Co-O Bonded Interactions for Tephroite, Mn_2SiO_4, Fayalite, Fe_2SiO_4, and Co_2SiO_4 Olivine and Selected Organic Metal Complexes: Comparison with Properties Calculated for Non-Transition and Transition Metal M-O Bonded Interactions for Silicates and Oxides. *The Journal of Physical Chemistry A* **2008**, *112*, 8811–8823.

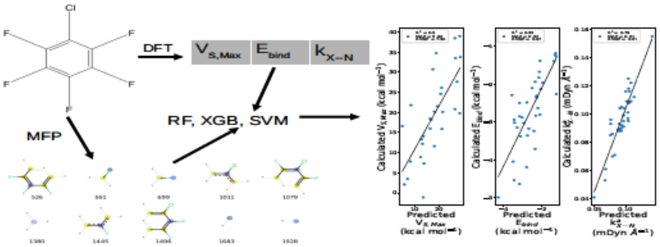
Cl
F
F
F
F
F
DFT
V S,Max
E bind
k X--N
RF, XGB, SVM
MFP
526
561
699
1011
1079
1380
1445
1496
1683
1926
Calculated V S,Max (kcal mol⁻¹)
Predicted V S,Max (kcal mol⁻¹)
Calculated E bind (kcal mol⁻¹)
Predicted E bind (kcal mol⁻¹)
Calculated k X--N (mDyn Å⁻¹)
Predicted k X--N (mDyn Å⁻¹)
0
5
10
15
20
25
30
35
40
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.05

Table of Contents Only