Synergy of Machine Learning and Density Functional Theory Calculations for Predicting Experimental Lewis Base Affinity and Lewis Polybase Binding Atoms

Hieu Huynh[1], Khanh Le[1], Linh Vu[1], Trang Nguyen[1], Matthew Holcomb[2], Stefano Forli[2], Hung Phan[1,3]

Correspondence to: Hung Phan (E-mail: hphan@soka.edu)

[1] *Hieu Huynh, Khanh Le, Linh Vu, Trang Nguyen, Hung Phan*

Fulbright University Vietnam, Ho Chi Minh city, Vietnam, Ho Chi Minh City 700000

[2] *Matthew Holcomb, Stefano Forli*

Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037 USA.

[3] *Hung Phan*

Soka University of America, Aliso Viejo, California, United States, CA 92656

Abstract

Investigation of Lewis acid-base interactions has been conducted by *ab initio* calculations and Machine Learning (ML) models. This study aims to resolve two critical tasks that have not been quantitatively investigated. First, ML models developed from Density Functional Theory (DFT) calculations predict experimental BF_3 affinity with Pearson correlation coefficients around 0.9 and mean absolute errors around 10 kJ mol^{-1}. The ML models are trained by DFT-calculated BF_3 affinity of more than 3000 adducts, with input features readily obtained by rdkit. Second, the ML models have the capability of predicting the relative strength of Lewis base binding atoms in Lewis polybases, which is either an extremely challenging task to conduct experimentally or a computationally expensive task for *ab initio* methods. The study demonstrates and solidifies the potential of combining DFT calculations and ML models to predict experimental properties, especially those that are scarce and impractical to empirically acquire.

Introduction

Lewis acid-base interactions have been shown to have a wide range of applications. Examples include the Frustrated Lewis pair in organic synthesis, the surface passivation of perovskite solar cells with Lewis bases, and color tuning in organic electronics.[1–7] A Lewis base (LB) contains atoms with lone electron pairs (hereinafter called LB atoms). In contrast, a Lewis acid (LA) contains atoms that have empty orbitals that can be used to bind with LB atoms to form a product called a Lewis acid-base adduct. For complex Lewis acid-base adducts, the LA-LB binding depends on a variety of intercorrelated properties such as electronegativity of the binding atoms, the energetic availability and steric accessibility of the lone pairs of the LB atom and empty orbitals of the LA atom, and the geometrical adjustment of either LA or LB upon binding.[8] Therefore, predictive controls and rational designs of Lewis acid-base adducts require chemistry fundamentals, experimental experience, and computational simulations. Among computational methods, DFT has been demonstrated to be a powerful prediction tool for binding energy, electron population, excited states, and other properties of Lewis acid-base adducts.[7,9–13]

The Lewis bases with multiple LB atoms (hereinafter called Lewis polybases)[14] have been recognized in numerous promising applications of Lewis interaction, especially organic semiconducting materials.[15–18]

Understanding the binding propensity and related properties of each LB atom in a Lewis polybase would contribute to the theoretical framework of Lewis interaction as well as improve their effectiveness in different applications. Those properties were reported to be predictable using *ab initio* methods in previous studies.[7,8,19,20] However, due to the employment of quantum mechanical wavefunctions, *ab initio* calculations are relatively costly, especially for large adducts.[21,22] Furthermore, the computational expense increases significantly with Lewis polybases containing numerous LB atoms since there would be numerous feasible adducts.

As a less computationally expensive alternative to quantum computation, machine learning (ML) and deep learning (DL) have demonstrated the potential in seeking solutions in various issues of materials science as well as chemistry. The diversity in model algorithms contributes to the high performance of ML and DL in predicting physical and chemical properties such as molecular polarizabilities,[23] molecular energies,[24,25] dipole moments,[25] and a variety of other properties.[26] Furthermore, the combination of quantum computation and ML/DL models has resulted in the remarkable power of predicting experimental properties. For example, a study of Sahu et al. has employed quantum calculations and gradient boosting to predict efficiency of organic solar cells with the Pearson coefficient of 0.79.[27] A recent study of Bauer et al. demonstrated that ML has the capability of predicting experimental hydrogen bonding free energies with RMSE below 4 kJ mol^{-1}.[28] Noteworthily, the model from Bauer et al. was developed from quantum-chemically computed free energies instead of directly training with experimental data. Besides, some models have the capability of filtering, screening a large dataset of molecules to a small number of most promising ones, and even generating new structures with favorable properties.[29–31]

As a typical characteristic of ML, many models are designed to perform effectively with tabular data. In the fields of chemistry and materials science, tabular data often includes eligible descriptors representing molecular structures as well as molecular properties. Among them, molecular descriptors are popular descriptors that can be computed readily from molecular structure, for example the molecular weight, the number of Carbon atoms or the mean atomic polarizability. The low computational cost promotes the molecular descriptors to be widely employed in ML models.[32–34] For instance, our recent study has demonstrated the capability of ML in using the readily obtained molecular descriptors to predict DFT-calculated properties of Lewis adducts that are challenging to obtain experimentally such as charge transfers from an LB to an LA upon binding.[22] Another type of popular descriptors is fingerprint descriptors with the exceptional performance in different tasks such as predicting bandgap,[35] atomic force[36] and polymer properties.[37] Some studies also constructed their own new descriptors.[27,28] Different from the tabular data type, deep learning is more flexible in the input types of the datasets. Among different DL models, graph neural network (GNN) is unique in utilizing object graphs for its training process. The resemblance of a molecular structure to a graph contributes to the accomplishment of GNN as the state-of-the-art model in predicting some molecular properties such as potential energy,[38,39] force field[39] and optoelectronic properties.[40]

In this study, we present the potential of machine learning in two critical aspects. Firstly, ML models developed from DFT calculations of BF_3 affinity (i.e., the magnitude of enthalpy change in a 1:1 complex formation of a Lewis base and BF_3) can predict experimental BF_3 affinity. Secondly, the ML models are capable of predicting the LB binding atoms of Lewis polybases in 1:1 complex formation with BF_3, which are extremely challenging to obtain with experiments. To the best of our knowledge, there has not been a comprehensive quantitative investigation of the competitiveness of different LB atoms in Lewis polybases. Features used in the ML models include descriptors representing atomic properties of a specific LB atom in an LB, and molecular descriptors of LB molecules. Additionally, the weight of descriptors is investigated

to provide more insights into the performance of the ML models. Besides, the ML models are briefly compared with GNN, which is one of the state-of-the-art DL models in learning molecular structures.[38–41]

Methods

Experimental dataset

The experimental dataset contains BF_3 affinity of 347 Lewis bases (LBs), which were assembled and reported by Laurence et al. (excluded one LB containing iodine).[14] It should be noted that most BF_3 affinities are primary values measured in dichloromethane, while some are secondary values calculated from measurements in other solvents. All the reported measurements were conducted at the temperature of 298 K and the pressure of 1 atm using the calorimetry method (i.e., a standard technique for measuring the amount of heat involved in a chemical reaction).

In silico dataset

The *in silico* dataset was constructed to be the training dataset for ML models. To ensure that the training set covers the chemical space of the experimental dataset, we used the chemical moieties from the experimental dataset to build new LBs for the *in silico* dataset. Thus, the molecules in the experimental dataset were manually broken down into 98 chemical moieties (**Figure S1**). All those moieties were combined to make 1000 new LBs using a customized algorithm based on rdkit.[42] The algorithm regulated several properties of the *in silico* dataset to be comparable to the experimental ones: the molecular weight distribution (**Figure S2a**), the ratio between Carbon and other heavy atoms, and the number of LB atoms from two to seven LB atoms. This quantity of LB atoms is a reasonable number of electron-withdrawing atoms for the range of the molecular weight of LBs in the *in silico* dataset as well as the experimental dataset. Each generated LB was also checked to be present in Pubchem library with pubchempy for a rapid assessment of chemical feasibility.[43] Additionally, the generated LB must not be included in the experimental dataset, and not be a duplicate. The SMILE strings of molecules in the *in silico* and experimental datasets are provided in the Supplementary Materials.

All the LBs in the *in silico* dataset as well as the experimental dataset were bound with BF_3 to form 1:1 adducts. For a Lewis polybase, one BF_3 molecule was connected to each LB atom respectively to build up different 1:1 adducts. In this study, the investigated LB atoms include Nitrogen, Oxygen, Phosphorus, and Sulfur with at least one lone pair of electrons. As a result, the 1000 molecules in the *in silico* dataset produced 3109 1:1 adducts, and the 347 LBs in the experimental dataset produced 648 adducts. Those adducts went through a quick geometrical optimization with the Merck molecular force field (MMFF94) before being optimized with DFT.

DFT methods

Diverse functionals and basis sets commonly used for organic molecules were examined with a small number of adducts reported recently.[20] They include the G4 compound method[44] (**Table S1**) and a few common DFT methods using APFD for functionals, Aug-cc-pVTZ and 6-311+g(2d,p) for basis sets, and polarizable continuum model (PCM) as well as SMD model for solvent models. The selected models were reported with high accuracy in previous studies.[7,20] Furthermore, APFD/6-311+g(2d,p) was recommended by Gaussian for typical organic molecules.[45] The APFD functionals, with built-in dispersion functions, has been tested for weak interactions with comparable accuracy to CCSD(T)/aug-cc-pVTZ calculations, which are recognized as an accurate and costly method.[46] The performance of each method is presented in **Table S1** and **Figure S3**. Considering both the accuracy and the computational cost, the APFD/6-311+g(2d,p) with either PCM or SMD are reasonable for the calculations in the study. We chose APFD/6-311+g(2d,p)

with PCM due to the popularity of PCM and the availability of our data from previous studies for comparison purpose.[7,22,45] All calculations were conducted with dichloromethane, which was the solvent used in the most of experimental measurements.

Subsequently, all the adducts, LBs, and BF_3 were optimized using the aforementioned APFD/6-311+g(2d,p) with PCM. Thermodynamic properties, including enthalpies, of the optimized structures were then calculated at the temperature of 298 K and the pressure of 1 atm, which were the temperature and pressure used in the experimental measurements. The optimized structures were also confirmed by the absence of negative vibrational frequencies. Finally, BF_3 affinity was calculated as the negative of the enthalpy of formation for Lewis acid-base adducts (i.e., $BF_3\ Affinity = -\Delta H_{formation} = H_{LB} + H_{BF_3} - H_{adduct}$). For the *in silico* dataset, BF_3 affinity was calculated for 3086 out of 3109 adducts whose DFT calculations were properly converged. For the experimental dataset, BF_3 affinity was calculated for all the possible 648 adducts.

In order to compare with the experimental affinity in the experimental dataset, the representative DFT-calculated affinity of LBs was calculated based on the Boltzmann average method. For an LB with one LB atom, the method simply took the only DFT-calculated affinity to be the representative one. For a polybase, the Boltzmann distribution was applied to derive the probabilities of the complex formation between the LB atoms and BF_3 from their corresponding affinity values. Then, the probabilities and the affinities of the LB atoms were multiplied and summed up to result in a representative affinity of the complex formation of all LB atoms. The calculated representative values (hereinafter called Boltzmann-average affinity) were employed to statistically reflect experimental affinity.

Descriptors

Descriptors are used as inputs to train ML models of molecules in a training set, and then to predict target properties of molecules in a test set. We utilized three types of descriptors including LB-atom descriptors, radial charge descriptors and molecular descriptors. While the LB-atom descriptors and radial charge descriptors are used to represent the LB atoms, the molecular descriptors provide the information of the whole LBs. Rdkit was employed to calculate descriptors describing an LB atom that bonds with BF_3. They include atomic number (Atomic_Number), atomic Gasteiger charge (Atomic_Charge), number of bonded neighbors (Degree), and atomic free solvent-accessible surface area (Free_SASA).[42] Inspired by radial atomic descriptors in studies reported by Göller et al.,[28,47] radial charge descriptors were calculated using Gasteiger charge as the core property. The Gasteiger charge is chosen because this type of charge can be computed readily from molecular structures with rdkit. For an LB atom, the radial charge descriptor numbered n (Radial_Charge_n) is the total Gasteiger charge of the LB atom and all atoms within its radius of n bonds. The radial charge descriptor numbered 1 (Radial_Charge_1) is the sum of the Gasteiger charge of the LB atom itself and all atoms directly binding to that LB atom. In this study, nine radial charge descriptors numbered from 1 to 9 were calculated for each LB atom. This amount of radial charge descriptors is suitable for LBs in the range of molecular weight in the *in silico* and experimental datasets. Molecular descriptors were extracted from the Mordred calculator.[48] 15 descriptor groups from Mordred were selected by the level of insights they could inform molecular design. They include AcidBase, Aromatic, AtomCount, BondCount, CarbonTypes, Constitutional, FragmentComplexity, Framework, HydrogenBond, Polarizability, RingCount, RotatableBond, SLogP, VdwVolumeABC and Weight. After removing the descriptors whose values were the same for more than 95% of LBs, 81 molecular descriptors remained. The three types of descriptors were combined into a set of 94 descriptors.

Model evaluation and testing

In this study, four ML models were selected from Scikit-learn based on their versatility and applicability in the fields of chemistry and materials science.[49] They include Linear Regression (LR) and Ridge Regression (Ridge) as the representatives of linear-based models, while Random Forest (RF) and Gradient Boosting (GB) are chosen for tree-based models. The *in silico* dataset is the training set, and the experimental dataset is the final test set for the ML models. It should be noted that none of the molecules in the *in silico* dataset exists in the experimental dataset.

We employed the standard three steps in developing and testing ML models including tuning hyperparameters, training models, and testing models. In order to tune hyperparameters, we applied cross-validation grid-searches using Stratified shuffle splitting strategy. This splitting approach is used to split the training set (i.e., the *in silico* dataset) into five separate groups with comparable DFT-calculated BF_3 affinity distributions. This is to avoid the sampling bias towards any certain range of affinity.[50] Then, the models with the tuned hyperparameters were also trained with cross-validation strategy, resulting in 30 values of Pearson correlation coefficient (R) and mean absolute error (MAE) to reflect the training performance. For testing the accuracy and applicability of the models, two validation approaches were utilized in this study. First, the models were used to predict the experimental BF_3 affinity in the experimental dataset. As a result, R and MAE between Boltzmann-average ML-predicted affinity and experimental data were calculated. Second, rankings of LB atoms within Lewis polybases between ML prediction and DFT calculations of both datasets were compared and assessed. The ML performance was also briefly compared with a GNN model from deepchem, a library with high quality tools to democratize the use of deep learning in the sciences.[51] More details of the GNN model can be found in the caption of **Figure S6**.

Results & Discussion

Correlation of DFT-calculated and experimental BF_3 affinity

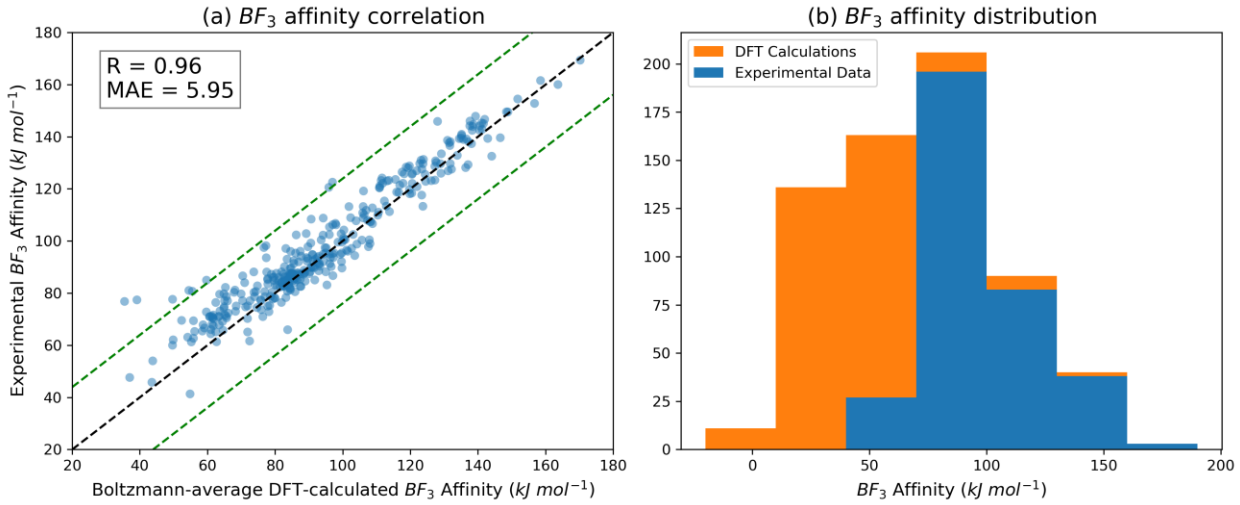

Figure 1. (a) Correlation of Boltzmann-average DFT-calculated affinity versus experimental BF_3 affinity of 347 LBs in the experimental dataset. The diagonal line is a plot of y = x as a helpful visual guide; the two green lines are error boundaries of one standard deviation of the experimental data from the y = x line. (b) BF_3 affinity histograms of experimental data (blue) and DFT calculations of all LB atoms (orange) of

347 LBs in the experimental dataset. In these histograms, each LB has one experimental BF_3 affinity, while having multiple values of DFT-calculated BF_3 affinity corresponding to multiple LB atoms.

The assessment of DFT calculations on the experimental dataset is displayed in **Figure 1**. After conducting DFT calculations for all LB atoms, Boltzmann-average affinity of each LB was computed to reflect the experimental BF_3 affinity. DFT simulations demonstrate a strong correlation with experimental BF_3 affinity with an R value of 0.96 and MAE of 5.95 kJ mol^{-1}, presented in **Figure 1a**. The figure also shows that most of the data points stay within the error boundaries of one standard deviation of experimental data (23.93 kJ mol^{-1}). The results confirm the reliability of the selected method (APFD/6-311+g(2d,p) with PCM). In contrast to one BF_3 affinity per one LB for experimental data, DFT simulations have the merit of calculating BF_3 affinity for each LB atom, which is immensely tough to measure by experiments. It includes significant weak LB atoms with affinity close to zero, such as the Nitrogen of the Amide groups. This leads to the difference between histograms of experimental data and DFT calculations in the experimental dataset in **Figure 1b**. Noticeably, the DFT calculations of all LB atoms (orange histogram) have numerous values of BF_3 affinity below 50 kJ mol^{-1}, which is the range with almost no experimental data (blue histogram). This capability of DFT calculations is a crucial foundation for ML to learn the BF_3 affinity corresponding to different LB atoms of a Lewis polybase and identify the binding one with BF_3 in a 1:1 Lewis acid-base adduct.

Exploratory data analysis

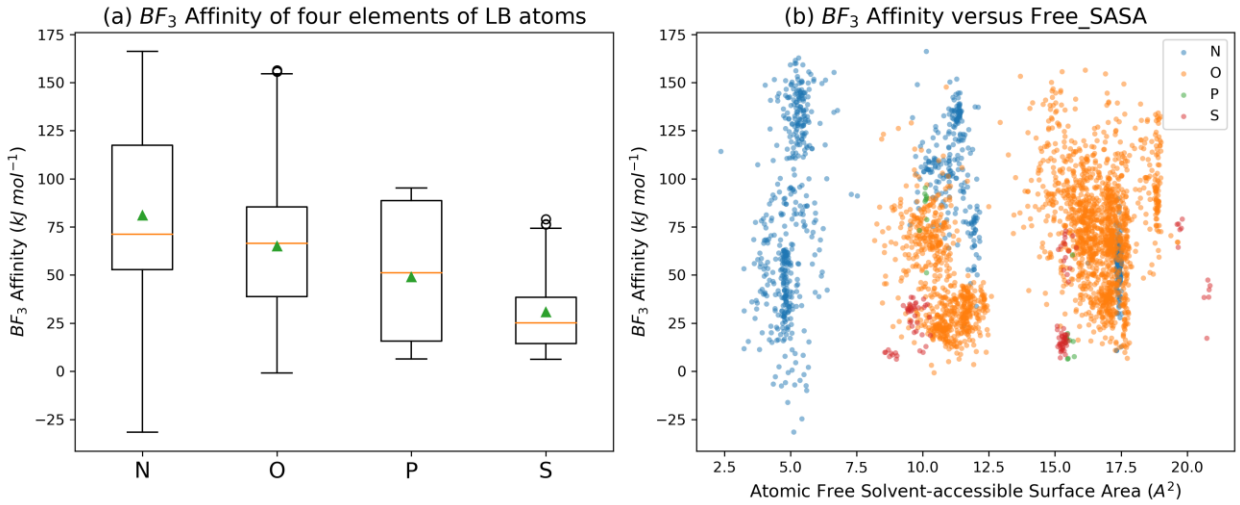

Figure 2. (a) Box plots of DFT-calculated BF_3 affinity of four different elements of LB atoms in the *in silico* dataset. In box plots, the line in the box represents the median, the green triangle represents the arithmetic mean, the box covers from the first quartile to the third quartile of the data, and the whisker extends from the lowest data point within the distance of 1.5 times the interquartile range below the first quartile to the highest data point within the same distance above the third quartile, the circle represents the outliers. (b) DFT-calculated BF_3 affinity of the *in silico* dataset plot against the atomic free solvent-accessible surface area (Free_SASA) as an example of the descriptors.

We explore the descriptive statistics of the dataset to get insights into the relationship between BF_3 affinity and descriptors. **Figure 2a** demonstrates the DFT-calculated BF_3 affinity of four elements of LB atoms in the *in silico* dataset. Nitrogen is recognized as the LB element with the highest average affinity (~ 80 kJ mol^{-1}). The Oxygen element is in the second place with an average affinity of approximately 65 kJ mol^{-1}. A

similar pattern is also applied to the third-row elements with nearly 50 kJ mol^{-1} for Phosphorus and over 30 kJ mol^{-1} for Sulfur. This order of N, O, P, S binding affinity could be generally explained by chemical intuition. The higher affinity of Nitrogen compared to Oxygen, and Phosphorus compared to Sulfur can be rationalized with the higher electronegativity of Oxygen and Sulfur, respectively. The higher affinity of the second-row elements (N, O) compared to those of the third-row (P, S) can be attributed to the higher orbital overlap between the lone pair orbital of the second-row LB atoms and the empty orbital of Boron (also in the second row). This trend has also been qualitatively explained by hard-soft acid-base theory.[52–54]

The Pearson correlation coefficients between DFT-calculated BF_3 affinity and 94 descriptors in the *in silico* dataset are presented in **Table S2**. The low average of absolute R values (0.16 ± 0.11) implies that the correlation is not significant. The relationship of BF_3 affinity and most descriptors is rather complex, which is represented in **Figure 2b** with the atomic free solvent-accessible surface area (Free_SASA) as an example. Noticeably, the figure indicates three vertical clusters around Free_SASA of 5.0, 11.0 and 16.5 A^2, respectively. Each cluster contains a full range of BF_3 affinity from below zero to more than 150 kJ mol^{-1}. This complex relation makes it improbable to interpret and predict DFT-calculated BF_3 affinity with just Free_SASA. Although steric accessibility is considered as an important factor in Lewis acid-base interactions, the descriptor representing this effect, Free_SASA, does not show a direct correlation to BF_3 affinity. This might be due to the diverse chemical space of the large dataset containing different LB atoms. Besides, **Figure S2b** presents the histograms of DFT-calculated BF_3 affinity in both the *in silico* dataset (training set) and the experimental dataset (test set). Along with the comparable molecular weight distributions of the two datasets presented in **Figure S2a**, the similarity of BF_3 affinity distribution can be observed in the training and test set. It indicates the suitability of the *in silico* dataset as the training set for ML models.

ML validation on the *in silico* dataset

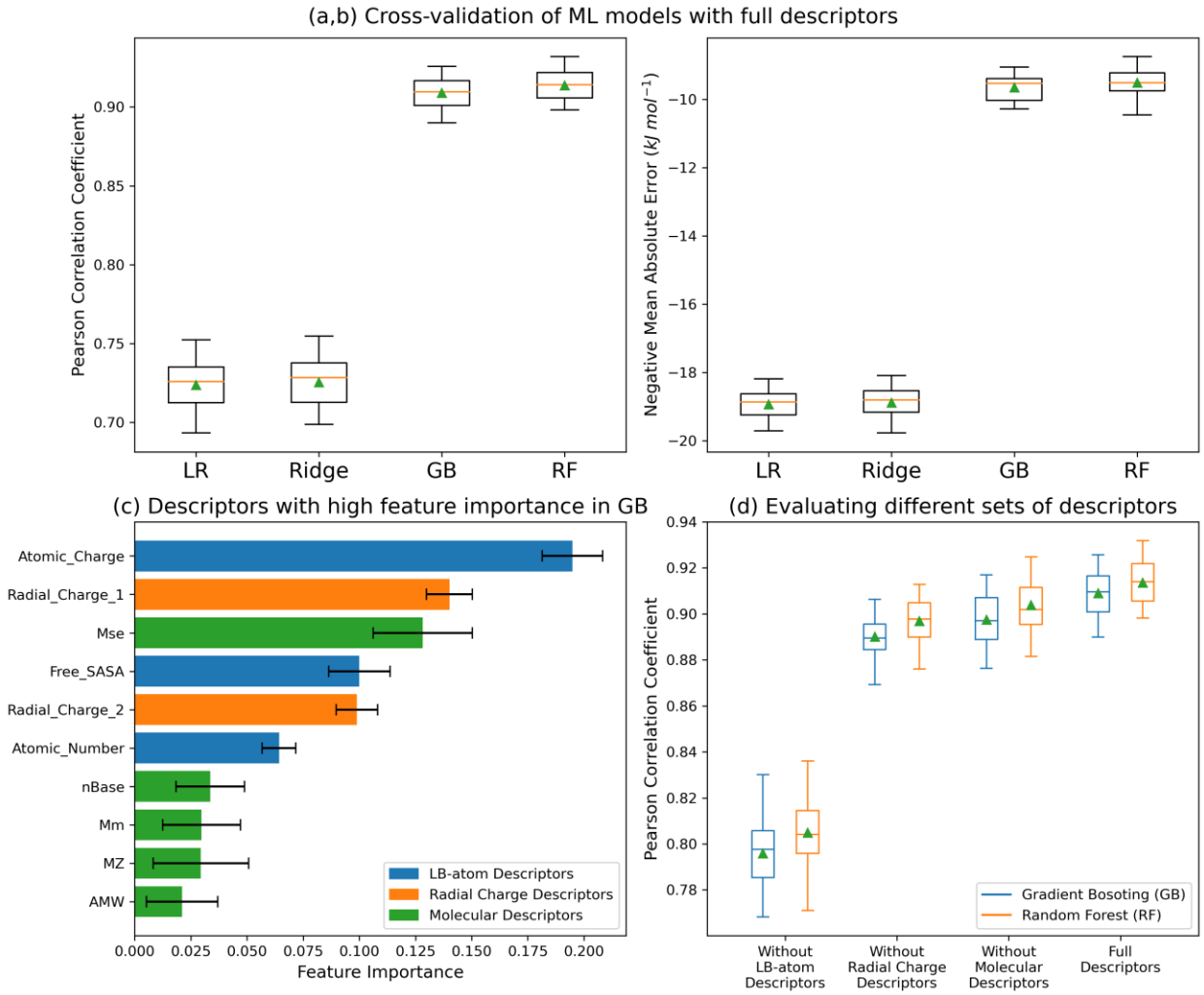

Figure 3. (a) Pearson Correlation Coefficient and (b) Mean Absolute Error of different ML models with full descriptors in cross validation on the *in silico* dataset. (c) Top ten descriptors with highest feature importance in GB model (average values from the cross-validation). (d) Pearson Correlation Coefficient of ML models either without LB-atom descriptors, radial charge descriptors or molecular descriptors in cross validation on the *in silico* dataset. In box plots, the line in the box represents the median, the green triangle represents the arithmetic mean, the box covers from the first quartile to the third quartile of the data, and the whisker extends from the lowest data point within the distance of 1.5 times the interquartile range below the first quartile to the highest data point within the same distance above the third quartile.

Figure 3 presents the performance of ML models in learning DFT-calculated BF_3 affinity in the *in silico* dataset. Box plots in **Figure 3a,b** are Pearson Correlation Coefficients (R) and negative Mean Absolute Errors (MAE) of four different models in the cross validation. The performance is divided into two clear clusters of linear-based models and tree-based models. While GB and RF have high R values of 0.91 ± 0.01 and small MAE of 9.58 ± 0.39 kJ mol^{-1}, LR and Ridge models demonstrate much worse results with 0.72 ± 0.02 for R and 18.92 ± 0.45 kJ mol^{-1} for MAE. The dominance of the tree-based models shows that GB and RF are suitable ML models for learning the patterns of BF_3 affinity. This dominance has also been observed in previous studies,[22,55,56] and might be resulted from the aforementioned low correlations between

BF_3 affinity and descriptors.[22] Consequently, we choose the GB and RF models for further analysis of the prediction.

The feature importance of the tree-based models is analyzed to comprehend how each group of descriptors affects the prediction. In GB and RF models, the feature importance quantifies how much a descriptor contributes to reducing the uncertainty of prediction.[57–59] The descriptor with higher feature importance has more influence on the model prediction. Additionally, the sum of all values of feature importance in a model is normalized to be 1.00. As a result, the average feature importance of 94 descriptors is approximately 0.01. **Figure 3c** and **Figure S4a** present ten descriptors with the highest average feature importance in the GB and RF models, respectively. Noticeably, the ten descriptors include all three distinct types of descriptors (radial charge descriptors, LB-atom descriptors, and molecular descriptors), which confirms the robustness in the selection of descriptors. In both tree-based models, Atomic_Charge dominates with the highest feature importance of 0.195 for GB and 0.181 for RF. For other LB-atom descriptors, the feature importance of Free_SASA is higher than that of Atomic_Number. This implies that, at least for those models, the prediction of BF_3 affinity is influenced by the properties of LB atoms in the descending order of the electronegativity, the steric accessibility, and the elements of those atoms. Among the two dominant radial charge descriptors, Radial_charge_1 is recorded with higher feature importance than Radial_Charge_2 in both models, which demonstrates that the more localized the radial charge is to the LB atoms, the more influence it has on the prediction. For molecular descriptors, we notice that the ones with high feature importance also have relatively high correlation coefficients (0.396 - 0.485) with DFT-calculated BF_3 affinity, as compared to the average coefficients of all molecular descriptors (0.16 ± 0.11, presented in **Table S2**).

To further investigate the influence of each descriptor type, we build ML models without either LB-atom descriptors, radial charge descriptors, or molecular descriptors, respectively. The performance of these models is presented in the Pearson correlation coefficient boxplots in **Figure 3d** and negative mean absolute error boxplots in **Figure S4b**. Without LB-atom descriptors, the GB and RF models demonstrate noticeably inferior accuracy with a decrease of R from 0.91 to 0.80 on average, and an increase of MAE from around 10 kJ mol^{-1} to 14.5 kJ mol^{-1}. This indicates the importance of the LB-atom descriptors for ML in learning BF_3 affinity, hence implies that the BF_3 affinity mostly depends on the local environment surrounding the LB atoms. The models without one of the two remaining descriptor types do not show significant performance degradation. Interestingly, the difference in MAEs between models without molecular descriptors and with full descriptors is negligible. This behavior could be further explored in the future investigation.

ML prediction and application

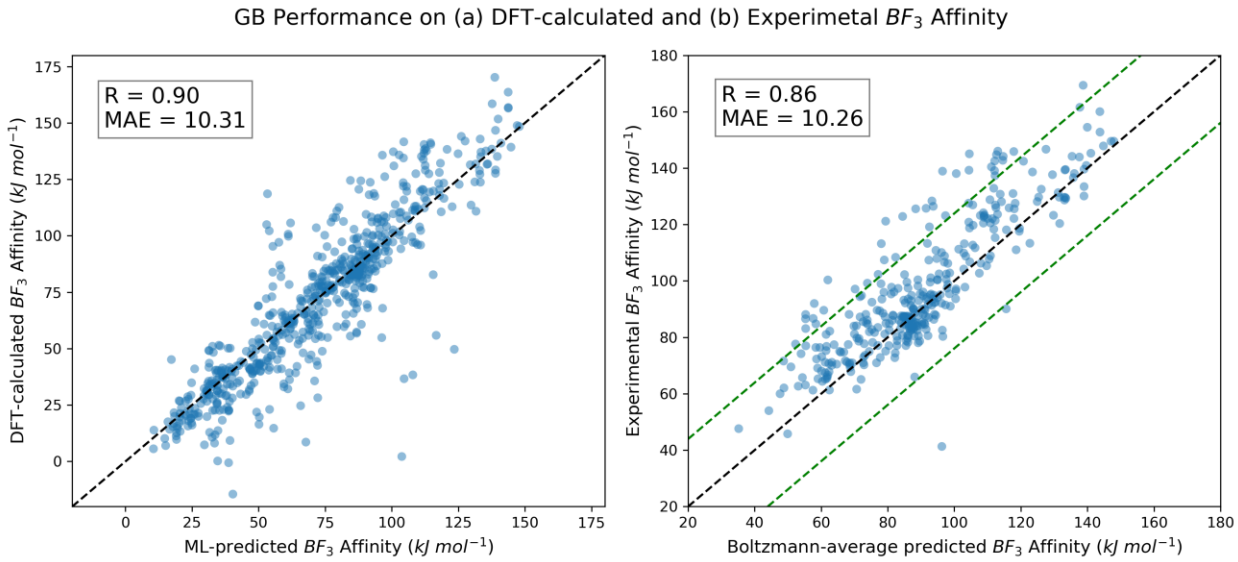

Figure 4. GB-predicted BF_3 affinity versus (a) DFT-calculated and (b) experimental BF_3 affinity in the experimental dataset. The diagonal lines are simply a plot of y = x as a helpful visual guide; the green lines in (b) are error boundaries of one standard deviation of experimental data from the y = x line. Although each LB has one experimental BF_3 affinity, it can have multiple values of DFT-calculated and ML-predicted BF_3 affinity corresponding to multiple LB atoms. While ML-predicted BF_3 affinity of all LB atoms are compared to their corresponding DFT-calculated ones, the Boltzmann-average ML-predicted BF_3 affinity is used to compare with the experimental BF_3 affinity.

With confirmed performance in the cross validation, GB and RF models are applied to predict BF_3 affinity in the experimental dataset. The correlation between ML prediction and DFT-calculated BF_3 affinity of the experimental dataset is shown in **Figure 4a** for GB, with R of 0.9 and MAE of 10.31 kJ mol^{-1}, which are comparable to RF model **(Figure S4c).** It should be noted that **Figure 4a and S4c** show ML-predicted and DFT-calculated BF_3 affinity of all LB atoms. The results from the two tree-based models validate the capability of ML in learning and predicting DFT-calculated BF_3 affinity in the test set. Then, the models are examined with experimental BF_3 affinity. Performance of GB and RF models on experimental BF_3 affinity (**Figure 4b** and **Figure S4d**) are comparable with a slight superiority for GB (R value of 0.86 for GB and 0.84 for RF, MAE value of 10.26 kJ mol^{-1} for GB and 11.40 kJ mol^{-1} for RF). The comparable MAE of the tree-based performance on DFT calculations and experimental data (10.62 kJ mol^{-1} for DFT calculations and 10.83 kJ mol^{-1} for experimental data in average of both models) validates the capability of the models built from DFT-calculated affinity in predicting experimental BF_3 affinity. Furthermore, the robustness of the ML models is further assessed with a slight superiority in comparison with the results of the GNN model (R of 0.81 and MAE of 12.68 kJ mol-1, **Figure S6**). This demonstrates the combined potential of ML and DFT in predicting experimental properties.

Moreover, the prediction of BF_3 affinity for all LB atoms (**Figure 4a**) allows us to predict the relative strength of LB atoms of Lewis polybases in 1:1 adducts with BF_3. **Table 1** and **Figure 5a** show the analysis of the binding strength of LB atoms, via affinity ranking, between DFT-calculated and ML-predicted BF_3 affinity for an example polybase in the *in silico* dataset. The LB atom with the highest affinity in a polybase is ranked first, and this atom is regarded as the binding atom in the 1:1 adduct of the polybase with BF_3. The prediction accuracy in affinity order is calculated by dividing the number of correct rankings over all

the rankings. The prediction accuracy in highest-affinity atoms is calculated by dividing the correct 1st rankings over all the 1st rankings. For the example polybase, the ML model presents the accurate prediction of the highest-affinity atom and four out of six LB atoms in the affinity order, which results in the prediction accuracy in affinity order and highest-affinity atoms of 0.67 (4/6) and 1.00, respectively.

Table 1. BF_3 affinity of an example molecule in *in silico* dataset

LB atom	DFT-calculated		ML-predicted	
	BF_3 Affinity (kJ mol^{-1})	Affinity Ranking	BF_3 Affinity (kJ mol^{-1})	Affinity Ranking
N1	45.23	3	48.27	3
O1	23.19	4	21.93	5
O2	88.36	1	87.53	1
O3	3.49	6	14.44	6
O4	20.81	5	22.20	4
O5	66.37	2	64.83	2

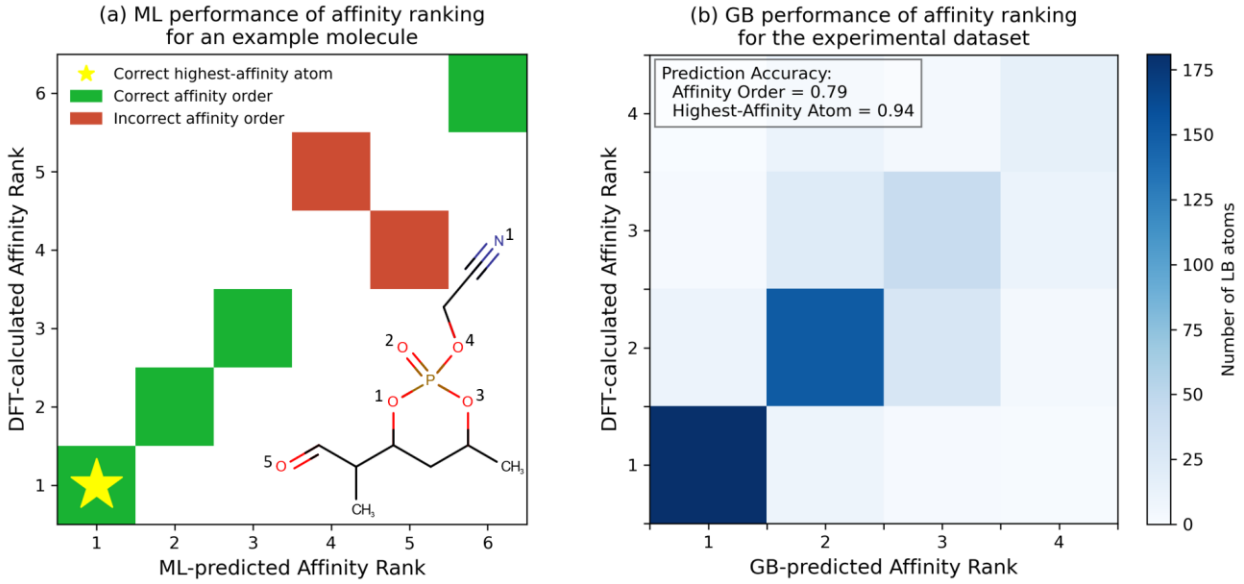

Figure 5. (a) ML performance of affinity orders and highest-affinity atoms for the example molecule in Table 1 with six LB atoms. (b) GB performance of affinity orders and highest-affinity atoms for the whole experimental dataset where color scale represents the number of LB atoms. In the experimental dataset, the highest number of LB atoms in an LB is four.

Figure 5b and **Figure S5a** display the ML prediction accuracy of the affinity orders and the highest-affinity atoms in the experimental dataset and the *in silico* dataset, respectively. The GB and RF performance on the experimental dataset comprising of 193 Lewis polybases (out of 347 LBs) demonstrates the prediction accuracy of 0.79 and 0.75 for the affinity orders and 0.94 and 0.91 for the highest-affinity atoms. The

prediction of 1000 Lewis polybases in the *in silico* dataset is revealed with an accuracy of 0.90 for the affinity orders and 0.93 for the highest-affinity atoms in both tree-based models (**Figure S5b,c**). Remarkably, the prediction accuracy of the highest-affinity atoms for the experimental dataset and the *in silico* dataset is comparable. The affinity order accuracy for the experimental dataset, however, is lower compared to the *in silico* dataset. It should be noted that the prediction of the highest-affinity atoms might be critical to empirical data because it reflects the LB binding atoms in 1:1 experimental adducts, while the affinity orders are an illustration of the relative strength of LB atoms in Lewis polybases. The performance of both tree-based models confirms the capability of ML in predicting the LB binding atoms of Lewis polybases in 1:1 adducts with BF_3. The validation of the ML models is further reinforced with the comparable performance of the GNN model (accuracy of affinity order of 0.79 and highest-affinity atom of 0.94, **Figure S6**). As an application, our models can be applied in identifying LB binding atoms of an LB where properties of adducts depend on the specific LB binding atoms. This will provide more tools for chemists and material scientists to increase the effectiveness of a *priori* design of Lewis acid-base interactions and adducts.

Finally, we assess the outliers in the predictions of the experimental BF_3 affinity in the ML and DL models to investigate the reasons behind the suboptimal prediction for certain structural motifs. The molecules with the prediction error larger than one standard deviation of the experimental data (green lines in **Figure 4b, S4d and S6a**) were defined as the outliers. We selected common outliers in GB, RF and GNN models for the assessment to avoid certain biases of each model. The structures and affinity data of the outliers are presented in **Figure S7** and **Table S3**. Some noticeable structural motifs in the outlier molecules are N,N-disubstituted aniline groups, 3-aminocyclohex-2-en-1-one groups and phosphonate groups. The suboptimal performance of the models in predicting 3-aminocyclohex-2-en-1-one groups is possibly because the descriptors and models cannot capture the electron donation from Nitrogen to Oxygen via the conjugated system in this structural motif. Similarly, phosphonate outliers might come from the inconsistence of the model performance in learning the pattern of the electron donation to the dominant Oxygen from other Oxygens via Phosphorus atom. The issue with those motifs might be attributed to the input descriptors, which are not designed for capturing the quantum effects, leading to the outlier performance of them. The reason for N,N-disubstituted aniline outliers is unclear to us since several other N,N-disubstituted anilines are not outliers. Although we highlight some structural motifs whose BF_3 affinity are challenging to predict by the presented machine learning and deep learning models, a more comprehensive investigation into those behaviors remains for future studies.

Conclusion

In summary, we have demonstrated the potential of combining *ab initio* methodology and ML models in accurately predicting experimental values. By employing LB-atom descriptors, radial charge descriptors, and molecular descriptors, which can be rapidly calculated, different ML models are able to predict experimental BF_3 affinity. Furthermore, the models have the capability of identifying the binding LB atom in a 1:1 adduct of a Lewis polybase with BF_3, which is either an intensively challenging task to conduct experimentally or a computationally expensive task to calculate by *ab initio* methods. The descriptors that highly influence the tree-based model prediction are also investigated via their feature importance and performance degradation of models without each descriptor type, providing more insights into the impact of each descriptor type on the prediction. Additionally, the ML models are compared with GNN, and the outliers are assessed in detail, which reinforces the robustness and the interpretability of the models. This study expands and strengthens the power of ML as well as *ab initio* methodology in analyzing and predicting Lewis acid-base interaction at the atomic level, which can be applied to a *priori* design of functional adducts used in organic electronics and catalysis. Moreover, it demonstrates the power of

augmenting experimental data with high-quality calculations from simulation to create a dataset capable of encoding complex features beyond the original experiment, in this case the behavior of individual atoms in a Lewis polybase. In a broader context, we show that the approach of combining ML and DFT is a potential solution for diverse problems with limited experimental data.

Acknowledgements

All the graphs were conducted by Matplotlib.[60] Hieu Huynh acknowledges the support of TPBank scholarship and the high-performance computing server of Fulbright University Vietnam. Hung Phan acknowledges the support of Soka University of America via the Faculty Research Development Fund and the high-performance computing server of The Scripp Research Institute. This work was supported by the National Institutes of Health Grant R01GM069832 (S.F.) and U54AI170855.

Keywords: Lewis acid-base adducts, Density Functional Theory, Machine Learning, Graph Neural Network, Lewis base affinity

References

(1) Zhu, M.; Li, C.; Li, B.; Zhang, J.; Sun, Y.; Guo, W.; Zhou, Z.; Pang, S.; Yan, Y. Interaction Engineering in Organic–Inorganic Hybrid Perovskite Solar Cells. *Mater. Horiz.* **2020**, *7* (9), 2208–2236. https://doi.org/10.1039/D0MH00745E.

(2) Li, C.; Wang, X.; Bi, E.; Jiang, F.; Park, S. M.; Li, Y.; Chen, L.; Wang, Z.; Zeng, L.; Chen, H.; Liu, Y.; Grice, C. R.; Abudulimu, A.; Chung, J.; Xian, Y.; Zhu, T.; Lai, H.; Chen, B.; Ellingson, R. J.; Fu, F.; Ginger, D. S.; Song, Z.; Sargent, E. H.; Yan, Y. Rational Design of Lewis Base Molecules for Stable and Efficient Inverted Perovskite Solar Cells. *Science* **2023**, *379* (6633), 690–694. https://doi.org/10.1126/science.ade3970.

(3) Stephan, D. W. Frustrated Lewis Pairs. *J. Am. Chem. Soc.* **2015**, *137* (32), 10018–10032. https://doi.org/10.1021/jacs.5b06794.

(4) Welch, G. C.; Coffin, R.; Peet, J.; Bazan, G. C. Band Gap Control in Conjugated Oligomers via Lewis Acids. *J. Am. Chem. Soc.* **2009**, *131* (31), 10802–10803. https://doi.org/10.1021/ja902789w.

(5) Welch, G. C.; Bazan, G. C. Lewis Acid Adducts of Narrow Band Gap Conjugated Polymers. *J. Am. Chem. Soc.* **2011**, *133* (12), 4632–4644. https://doi.org/10.1021/ja110968m.

(6) Zalar, P.; Henson, Z. B.; Welch, G. C.; Bazan, G. C.; Nguyen, T.-Q. Color Tuning in Polymer Light-Emitting Diodes with Lewis Acids. *Angew. Chem.* **2012**, *124* (30), 7613–7616. https://doi.org/10.1002/ange.201202570.

(7) Phan, H.; Kelly, T. J.; Zhugayevych, A.; Bazan, G. C.; Nguyen, T.-Q.; Jarvis, E. A.; Tretiak, S. Tuning Optical Properties of Conjugated Molecules by Lewis Acids: Insights from Electronic Structure Modeling. *J. Phys. Chem. Lett.* **2019**, *10* (16), 4632–4638. https://doi.org/10.1021/acs.jpclett.9b01572.

(8) Bessac, F.; Frenking, G. Why Is BCl3 a Stronger Lewis Acid with Respect to Strong Bases than BF3?,. *Inorg. Chem.* **2003**, *42* (24), 7990–7994. https://doi.org/10.1021/ic034141o.

(9) Khorief Nacereddine, A.; Merzoud, L.; Morell, C.; Chermette, H. A Computational Investigation of the Selectivity and Mechanism of the Lewis Acid Catalyzed Oxa-Diels–Alder Cycloaddition of Substituted Diene with Benzaldehyde. *Journal of Computational Chemistry* **2021**, *42* (18), 1296–1311. https://doi.org/10.1002/jcc.26547.

(10) Zhang, Z.-F.; Su, M.-D. The Reactivity of the Trapping Reaction of the Benzene-Bridged Boron/Phosphorus-Based Frustrated Lewis Pair with Difluorocarbene and Its Group 14 Analogs: A Theoretical Investigation. *Journal of Computational Chemistry* **2022**, *43* (26), 1783–1792. https://doi.org/10.1002/jcc.26980.

(11) Mebs, S.; Grabowsky, S.; Förster, D.; Kickbusch, R.; Hartl, M.; Daemen, L. L.; Morgenroth, W.; Luger, P.; Paulus, B.; Lentz, D. Charge Transfer via the Dative N−B Bond and Dihydrogen Contacts. Experimental and Theoretical Electron Density Studies of Small Lewis Acid−Base Adducts. *J. Phys. Chem. A* **2010**, *114* (37), 10185–10196. https://doi.org/10.1021/jp100995n.

(12) Baruah, T.; Pederson, M. R. DFT Calculations on Charge-Transfer States of a Carotenoid-Porphyrin-C60 Molecular Triad. *J. Chem. Theory Comput.* **2009**, *5* (4), 834–843. https://doi.org/10.1021/ct900024f.

(13) Vlček, A.; Záliš, S. Modeling of Charge-Transfer Transitions and Excited States in D6 Transition Metal Complexes by DFT Techniques. *Coordination Chemistry Reviews* **2007**, *251* (3), 258–287. https://doi.org/10.1016/j.ccr.2006.05.021.

(14) Laurence, C.; Gal, J.-F. *Lewis Basicity and Affinity Scales: Data and Measurement*; John Wiley & Sons, 2009.

(15) Kang, I.; Yun, H.-J.; Chung, D. S.; Kwon, S.-K.; Kim, Y.-H. Record High Hole Mobility in Polymer Semiconductors via Side-Chain Engineering. *J. Am. Chem. Soc.* **2013**, *135* (40), 14896–14899. https://doi.org/10.1021/ja405112s.

(16) Pouliot, J.-R.; Sun, B.; Leduc, M.; Najari, A.; Li, Y.; Leclerc, M. A High Mobility DPP-Based Polymer Obtained via Direct (Hetero)Arylation Polymerization. *Polym. Chem.* **2014**, *6* (2), 278–282. https://doi.org/10.1039/C4PY01222D.

(17) Wienk, M. M.; Turbiez, M.; Gilot, J.; Janssen, R. A. J. Narrow-Bandgap Diketo-Pyrrolo-Pyrrole Polymer Solar Cells: The Effect of Processing on the Performance. *Adv. Mater.* **2008**, *20* (13), 2556–2560. https://doi.org/10.1002/adma.200800456.
(18) Zhang, X.; Richter, L. J.; DeLongchamp, D. M.; Kline, R. J.; Hammond, M. R.; McCulloch, I.; Heeney, M.; Ashraf, R. S.; Smith, J. N.; Anthopoulos, T. D.; Schroeder, B.; Geerts, Y. H.; Fischer, D. A.; Toney, M. F. Molecular Packing of High-Mobility Diketo Pyrrolo-Pyrrole Polymer Semiconductors with Branched Alkyl Side Chains. *J. Am. Chem. Soc.* **2011**, *133* (38), 15073–15084. https://doi.org/10.1021/ja204515s.
(19) Jonas, V.; Frenking, G.; Reetz, M. T. Comparative Theoretical Study of Lewis Acid-Base Complexes of BH3, BF3, BCl3, AlCl3, and SO2. *J. Am. Chem. Soc.* **1994**, *116* (19), 8741–8753. https://doi.org/10.1021/ja00098a037.
(20) Gal, J.-F.; Maria, P.-C.; Yáñez, M.; Mó, O. Enthalpies of Adduct Formation between Boron Trifluoride and Selected Organic Bases in Solution: Toward an Accurate Theoretical Entry to Lewis Basicity. *Molecules* **2021**, *26* (21), 6659. https://doi.org/10.3390/molecules26216659.
(21) Qu, X.; Latino, D. A.; Aires-de-Sousa, J. A Big Data Approach to the Ultra-Fast Prediction of DFT-Calculated Bond Energies. *Journal of Cheminformatics* **2013**, *5* (1), 34. https://doi.org/10.1186/1758-2946-5-34.
(22) Huynh, H.; Kelly, T. J.; Vu, L.; Hoang, T.; Nguyen, P. A.; Le, T. C.; Jarvis, E. A.; Phan, H. Quantum Chemistry–Machine Learning Approach for Predicting Properties of Lewis Acid–Lewis Base Adducts. *ACS Omega* **2023**, *8* (21), 19119–19127. https://doi.org/10.1021/acsomega.3c02822.
(23) Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; DiStasio, R. A.; Ceriotti, M. Accurate Molecular Polarizabilities with Coupled Cluster Theory and Machine Learning. *Proceedings of the National Academy of Sciences* **2019**, *116* (9), 3401–3406. https://doi.org/10.1073/pnas.1816132116.
(24) Lubbers, N.; Smith, J. S.; Barros, K. Hierarchical Modeling of Molecular Energies Using a Deep Neural Network. *The Journal of Chemical Physics* **2018**, *148* (24), 241715. https://doi.org/10.1063/1.5011181.
(25) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J Chem Theory Comput* **2019**, *15* (6), 3678–3693. https://doi.org/10.1021/acs.jctc.9b00181.
(26) Jørgensen, P. B.; Jacobsen, K. W.; Schmidt, M. N. Neural Message Passing with Edge Updates for Predicting Properties of Molecules and Materials. arXiv June 8, 2018. https://doi.org/10.48550/arXiv.1806.03146.
(27) Sahu, H.; Rao, W.; Troisi, A.; Ma, H. Toward Predicting Efficiency of Organic Solar Cells via Machine Learning and Improved Descriptors. *Advanced Energy Materials* **2018**, *8* (24), 1801032. https://doi.org/10.1002/aenm.201801032.
(28) Bauer, C. A.; Schneider, G.; Göller, A. H. Machine Learning Models for Hydrogen Bond Donor and Acceptor Strengths Using Large and Diverse Training Data Generated by First-Principles Interaction Free Energies. *Journal of Cheminformatics* **2019**, *11* (1), 59. https://doi.org/10.1186/s13321-019-0381-4.
(29) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nat Biotechnol* **2019**, *37* (9), 1038–1040. https://doi.org/10.1038/s41587-019-0224-x.
(30) Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-like Molecules. *Nat Rev Drug Discov* **2005**, *4* (8), 649–663. https://doi.org/10.1038/nrd1799.
(31) Batra, R.; Chan, H.; Kamath, G.; Ramprasad, R.; Cherukara, M. J.; Sankaranarayanan, S. K. R. S. Screening of Therapeutic Agents for COVID-19 Using Machine Learning and Ensemble Docking Studies. *J. Phys. Chem. Lett.* **2020**, *11* (17), 7058–7065. https://doi.org/10.1021/acs.jpclett.0c02278.

(32) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative Structure–Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev.* **2012**, *112* (5), 2889–2919. https://doi.org/10.1021/cr200066h.
(33) Le, T. C.; Ballard, M.; Casey, P.; Liu, M. S.; Winkler, D. A. Illuminating Flash Point: Comprehensive Prediction Models. *Molecular Informatics* **2015**, *34* (1), 18–27. https://doi.org/10.1002/minf.201400098.
(34) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; John Wiley & Sons, Ltd, 2009. https://doi.org/10.1002/9783527628766.fmatter.
(35) Pilania, G.; Mannodi-Kanakkithodi, A.; Uberuaga, B. P.; Ramprasad, R.; Gubernatis, J. E.; Lookman, T. Machine Learning Bandgaps of Double Perovskites. *Sci Rep* **2016**, *6* (1), 19375. https://doi.org/10.1038/srep19375.
(36) Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R. Machine Learning Force Fields: Construction, Validation, and Outlook. *J. Phys. Chem. C* **2017**, *121* (1), 511–522. https://doi.org/10.1021/acs.jpcc.6b10908.
(37) Xu, C.; Wang, Y.; Barati Farimani, A. TransPolymer: A Transformer-Based Language Model for Polymer Property Predictions. *npj Comput Mater* **2023**, *9* (1), 1–14. https://doi.org/10.1038/s41524-023-01016-5.
(38) Schmidt, J.; Pettersson, L.; Verdozzi, C.; Botti, S.; Marques, M. A. L. Crystal Graph Attention Networks for the Prediction of Stable Materials. *Science Advances* **2021**, *7* (49), eabi7948. https://doi.org/10.1126/sciadv.abi7948.
(39) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet – A Deep Learning Architecture for Molecules and Materials. *The Journal of Chemical Physics* **2018**, *148* (24), 241722. https://doi.org/10.1063/1.5019779.
(40) St. John, P. C.; Phillips, C.; Kemper, T. W.; Wilson, A. N.; Guan, Y.; Crowley, M. F.; Nimlos, M. R.; Larsen, R. E. Message-Passing Neural Networks for High-Throughput Polymer Screening. *The Journal of Chemical Physics* **2019**, *150* (23), 234111. https://doi.org/10.1063/1.5099132.
(41) Fung, V.; Zhang, J.; Juarez, E.; Sumpter, B. G. Benchmarking Graph Neural Networks for Materials Chemistry. *npj Comput Mater* **2021**, *7* (1), 1–8. https://doi.org/10.1038/s41524-021-00554-0.
(42) *RDKit*. https://www.rdkit.org/ (accessed 2023-09-29).
(43) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 Update. *Nucleic Acids Research* **2023**, *51* (D1), D1373–D1380. https://doi.org/10.1093/nar/gkac956.
(44) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 Theory. *The Journal of Chemical Physics* **2007**, *126* (8), 084108. https://doi.org/10.1063/1.2436888.
(45) Foresman, J.; Frisch, A. *Exploring Chemistry With Electronic Structure Methods, 3rd Edition*; 2015.
(46) Austin, A.; Petersson, G. A.; Frisch, M. J.; Dobek, F. J.; Scalmani, G.; Throssell, K. A Density Functional with Spherical Atom Dispersion Terms. *J. Chem. Theory Comput.* **2012**, *8* (12), 4989–5007. https://doi.org/10.1021/ct300778e.
(47) Bauer, C. A.; Schneider, G.; Göller, A. H. Gaussian Process Regression Models for the Prediction of Hydrogen Bond Acceptor Strengths. *Molecular Informatics* **2019**, *38* (4), 1800115. https://doi.org/10.1002/minf.201800115.
(48) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *Journal of Cheminformatics* **2018**, *10* (1), 4. https://doi.org/10.1186/s13321-018-0258-y.
(49) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; others. Scikit-Learn: Machine Learning in Python. *Journal of machine learning research* **2011**, *12* (Oct), 2825–2830.
(50) *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition [Book]*. https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/ (accessed 2020-03-24).
(51) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V.; Leswing, K.; Wu, Z. *Deep Learning for the Life Sciences*; O'Reilly Media, 2019.

(52) Ayers, P. W. The Physical Basis of the Hard/Soft Acid/Base Principle. *Faraday Discuss.* **2006**, *135* (0), 161–190. https://doi.org/10.1039/B606877D.
(53) Ayers, P. W. An Elementary Derivation of the Hard/Soft-Acid/Base Principle. *The Journal of Chemical Physics* **2005**, *122* (14), 141102. https://doi.org/10.1063/1.1897374.
(54) Cárdenas, C.; Ayers, P. W. How Reliable Is the Hard–Soft Acid–Base Principle? An Assessment from Numerical Simulations of Electron Transfer Energies. *Phys. Chem. Chem. Phys.* **2013**, *15* (33), 13959–13968. https://doi.org/10.1039/C3CP51134K.
(55) Sahu, H.; Rao, W.; Troisi, A.; Ma, H. Toward Predicting Efficiency of Organic Solar Cells via Machine Learning and Improved Descriptors. *Advanced Energy Materials* **2018**, *8* (24), 1801032. https://doi.org/10.1002/aenm.201801032.
(56) Lee, M.-H. Robust Random Forest Based Non-Fullerene Organic Solar Cells Efficiency Prediction. *Organic Electronics* **2020**, *76*, 105465. https://doi.org/10.1016/j.orgel.2019.105465.
(57) Scornet, E. Trees, Forests, and Impurity-Based Variable Importance in Regression. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **2023**, *59* (1), 21–52. https://doi.org/10.1214/21-AIHP1240.
(58) Sandri, M.; Zuccolotto, P. Analysis and Correction of Bias in Total Decrease in Node Impurity Measures for Tree-Based Algorithms. *Stat Comput* **2010**, *20* (4), 393–407. https://doi.org/10.1007/s11222-009-9132-0.
(59) Cassidy, A. P.; Deviney, F. A. Calculating Feature Importance in Data Streams with Concept Drift Using Online Random Forest. In *2014 IEEE International Conference on Big Data (Big Data)*; 2014; pp 23–28. https://doi.org/10.1109/BigData.2014.7004352.
(60) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **2007**, *9* (3), 90–95. https://doi.org/10.1109/MCSE.2007.55.

Graphical Abstract

Author Names

Hieu Huynh, Khanh Le, Linh Vu, Trang Nguyen, Matthew Holcomb, Stefano Forli, Hung Phan

Title

Synergy of Machine Learning and Density Functional Theory Calculations for Predicting Experimental Lewis Base Affinity and Lewis Polybase Binding Atoms

Text

The objective of this study is to resolve two critical tasks in the realm of Lewis acid-base interactions that have not been quantitatively investigated. First, machine learning developed from simulation is successful at predicting experimental properties of Lewis adducts. Second, the models have the capability of identifying the binding atoms of Lewis polybases in 1:1 adducts, which is either improbable to obtain with experiments, or computationally expensive for quantum simulation.

Graphical Abstract Figure