How to cite: Angew. Chem. Int. Ed. 2024, e202401084
doi.org/10.1002/anie.202401084

Machine Learning Hot Paper

Predicting Lewis Acidity: Machine Learning the Fluoride Ion Affinity of *p*-Block-Atom-Based Molecules

Lukas M. Sigmund, Shree Sowndarya S. V., Andreas Albers, Philipp Erdmann, Robert S. Paton,* and Lutz Greb**

Abstract: "How strong is this Lewis acid?" is a question researchers often approach by calculating its fluoride ion affinity (FIA) with quantum chemistry. Here, we present FIA49k, an extensive FIA dataset with 48,986 data points calculated at the RI-DSD-BLYP-D3(BJ)/def2-QZVPP//PBEh-3c level of theory, including 13 different *p*-block atoms as the fluoride accepting site. The FIA49k dataset was used to train FIA-GNN, two message-passing graph neural networks, which predict gas and solution phase FIA values of molecules excluded from training with a mean absolute error of 14 kJ mol^{-1} ($r^2 = 0.93$) from the SMILES string of the Lewis acid as the only input. The level of accuracy is notable, given the wide energetic range of 750 kJ mol^{-1} spanned by FIA49k. The model's value was demonstrated with four case studies, including predictions for molecules extracted from the Cambridge Structural Database and by reproducing results from catalysis research available in the literature. Weaknesses of the model are evaluated and interpreted chemically. FIA-GNN and the FIA49k dataset can be reached via a free web app (www.grebgroup.de/fia-gnn).

Introduction

Lewis acids are omnipresent in all branches of chemistry, including fundamental research and synthesis, industrial-scale production, and bio-related processes.[1] The fluoride ion affinity (FIA) has developed into one of the most common measures to quantitatively assess Lewis acidity in its thermodynamic sense (global Lewis acidity).[2] The FIA is defined as the negative reaction enthalpy of the binding between a fluoride anion and a given Lewis acid (Figure 1A). As the experimental determination of absolute FIA values requires highly sophisticated techniques,[2b,3] they are almost exclusively obtained through quantum chemical computations. To avoid the explicit treatment of the naked fluoride anion, the FIA is typically calculated with the help of a quasi-isodesmic anchoring scheme, such as the fluorotrimethylsilane system (Figure 1B).[2c,d]

Access to accurate FIA data is indeed the linchpin for the steady progress in the design and application of Lewis acids. For example, Kirschner et al. used the FIA of organoboranes to develop phase-transfer catalysts in nucleophilic fluorination reactions.[4] The List group correlated FIA values with the activity of their Lewis acidic organocatalysts.[5] Computed FIAs are also commonly

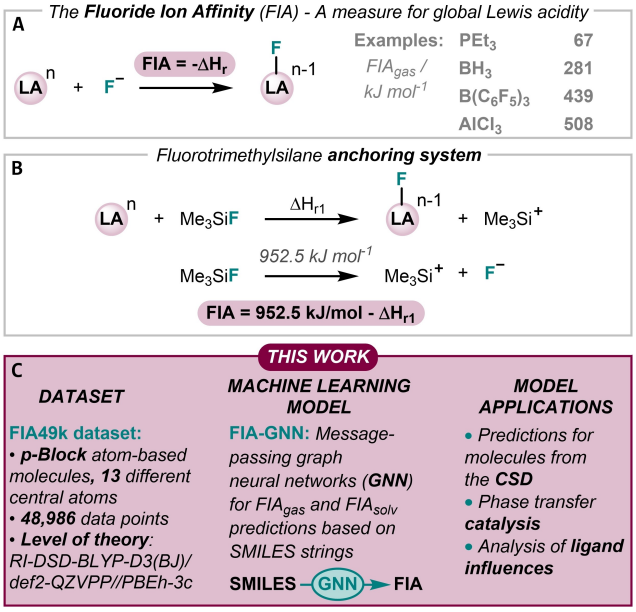


Figure 1. **A)** Definition of the fluoride ion affinity (FIA) and example values. **B)** Quantum chemical calculation scheme of FIA values with the help of the fluorotrimethylsilane anchoring system. The reaction enthalpy of the second reaction was calculated at the CCSD(T)/CBS level of theory and was taken from the literature.[2c] **C)** This work: compilation of an FIA dataset with 48,986 data points, training of machine learning models, and application in different case studies.

[*] L. M. Sigmund, A. Albers, P. Erdmann, Prof. Dr. L. Greb
Anorganisch-Chemisches Institut
Ruprecht-Karls-Universität Heidelberg
Im Neuenheimer Feld 270, 69120 Heidelberg, Germany
E-mail: lukas.sigmund@aci.uni-heidelberg.de
greb@uni-heidelberg.de
L. M. Sigmund, S. S. S. V., Prof. R. S. Paton
Department of Chemistry
Colorado State University
1301 Center Avenue, Fort Collins, CO 80523, USA
E-mail: robert.paton@colostate.edu

considered in studies on weakly coordinating anions[2d,6] and in the closely related field of Lewis superacidity;[7] not least by some of us.[8] The FIA has further been compared to multiple other Lewis acidity scales, such as the global electrophilicity index (GEI)[9] and others.[10]

In recent years, data-driven and statistical models have gained ever-increasing attention in chemical research.[11] Machine learning (ML) approaches are ideally suited to circumvent the execution of quantum chemical calculations of molecular or reaction properties.[12] An overwhelming number of ML regressors was trained for the prediction of various quantities. However, while predictors for Mayr's electrophilicity parameter[13] and BF_3 affinities of organic Lewis bases[14] exist, a statistical model for the prediction of Lewis acidity has not been reported. This lack might be explained by the high complexity of the task, which requires the description of a variety of chemical bond types while capturing deformation energy contributions that are extremely important for Lewis pair formation[15] and heavily depend on intricate ligand structures and constraints.

In this work (Figure 1C), we present broadly applicable ML models that can predict the gas and solution phase (dichloromethane) FIA values of a chemically diverse set of neutral *p*-block atom-based molecules comprising 13 different elements (B, Al, Ga, In, Si, Ge, Sn, Pb, P, As, Sb, Bi, and Te) with a mean absolute error of around 14 kJ mol^{-1} ($r^2 = 0.93$). A large dataset of 48,986 FIA values (the largest to date) was compiled to train and test the regressors. We demonstrate the applicability of the best models in four case studies. The entire dataset and all models are publicly available on GitHub (https://github.com/GrebGroup/fia-gnn) and figshare (https://figshare.com/projects/FIA-GNN/187050), respectively, and FIA-GNN can be used free of charge at www.grebgroup.de/fia-gnn.

Results and Discussion

Dataset Construction

As a sufficiently large collection of FIA values[2c] was not available previously, we compiled such a dataset according to the following principles:

- Only neutral Lewis acids were included. Atoms with formal charges were not considered.
- 13 different atoms from the *p*-block of the periodic table were included as the direct fluoride ion acceptors. Group 14 and 15 elements were used in two different oxidation states, resulting in 21 different central atom classes (Figure 2A).
- Ligands were built from H, C, N, O, S, F, Cl, Br, and I (Figure 2B). Bonds between heteroatoms within the ligand (e.g., O-O, N-O) were disallowed.
- Mono-, bi-, and tridentate ligands were considered. The bi-, and tridentate ligands were further categorized with respect to the size of the ring system formed when the ligand is installed at the central atom (Figure 2C). Bi- and tridentate ligands were not used within the same molecule.
- Cases of intramolecular precoordination with available neutral donor groups and bond rearrangements upon fluoride anion binding were excluded (Figure 2D).

With these guidelines at hand, we initially attempted to use the Cambridge Structural Database (CSD)[16] to acquire starting structures for the FIA dataset. However, large imbalances with respect to the central atom and ligand denticity classes were faced after data extraction (see Chapter S9 in the Supporting Information for details). As this would severely bias subsequent ML models toward certain regions of chemical space, we substantially aug-

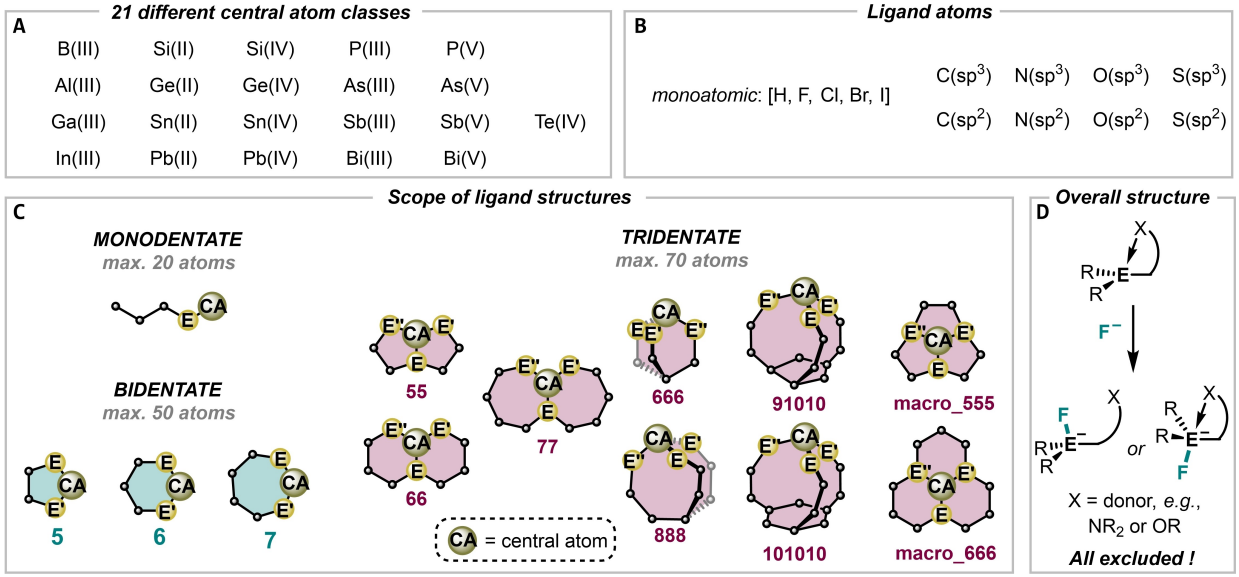


Figure 2. Design principles of the compiled FIA44k dataset. **A)** Included central atom classes. **B)** Atom types used to build up the ligands. **C)** Included types of ligand structures. The given maximum atom counts of the ligands include hydrogen atoms. **D)** Exclusion of intramolecularly saturated Lewis acidic centers and of bond rearrangements upon F^- binding.

mented the extracted CSD set with molecules constructed from scratch.

For that, we developed autoPAMS (**auto**matic generation of ***p***-block **a**tom-based **m**olecular **s**tructures, part of the fia-gnn Github repository), an RDKit-based implementation of a modular design strategy for the automatic generation of three-dimensional molecular structures of Lewis acids and respective FIA datasets.[17] Within the rules described above, autoPAMS sampled from over 2,500 unique building blocks, which were manually curated. By this procedure, we arrived at a dataset as balanced as possible with respect to the central atom, ligand denticity class, and further criteria. The three-dimensional starting structures of the Lewis acids were generated with RDKit.[18] The initial structures of the fluoride adducts were obtained geometrically from the optimized Lewis acid structures. This was done to avoid a second RDKit embedding step, which often tended to fail for the fluoride adducts. A detailed description of all computational routines is given in the Supporting Information (Chapters S3–S6).

The three-dimensional starting structures were submitted to CREST (GFN2-xTB).[19] The lowest energy conformer obtained was then used to calculate FIA_{gas} at the benchmarked[2c] RI-DSD-BLYP-D3(BJ)/def2-QZVPP//PBEh-3c level.[20] The conformer ensembles were spot-checked with DFT (see Chapter S4 in the Supporting Information). Solvent corrections for dichloromethane were calculated with the COSMO-RS scheme (BP86/TZP).[21] FIA_{gas} values were obtained as shown in Figure 1B, and from those, the solvation-corrected FIA_{solv} following

$$FIA_{solv} = FIA_{gas} - (\Delta H_{corr}(FA) - \Delta H_{corr}(LA) - \Delta H_{corr}(F^-)).$$

With the described computational workflow, 44,877 FIA_{gas} and the corresponding FIA_{solv} values were calculated. This part of the entire dataset is named FIA44k. Furthermore, the FIA values for 2,389 entries of the CSD were computed (FIA2k-CSD, see the Supporting Information for details).

Dataset Analysis I: Composition and Balance

At first, the composition of the FIA44k dataset was analyzed with respect to the design principles described above. The dataset is evenly distributed among all central atom classes (Figure 3A). The P(III) and Pb(IV) classes are slightly underrepresented, as FIA calculations had a much higher propensity to fail. P(III)-based molecules intrinsically show relatively low FIA values, which often resulted in dissociation of the fluoride adduct. Molecules with Pb(IV) atoms tended to undergo reductive elimination to give Pb(II). The low-valent group 14 molecules contribute less to FIA44k because they cannot have tridentate ligands. Also, the dataset is evenly distributed among the denticity classes "mono-", "bi-", and "tridentate" with overall relative shares of 32, 40, and 28 %, respectively (Figure 3B).

The distribution of atoms directly bound to the central atoms is shown in Figure 3C. An even distribution of atom

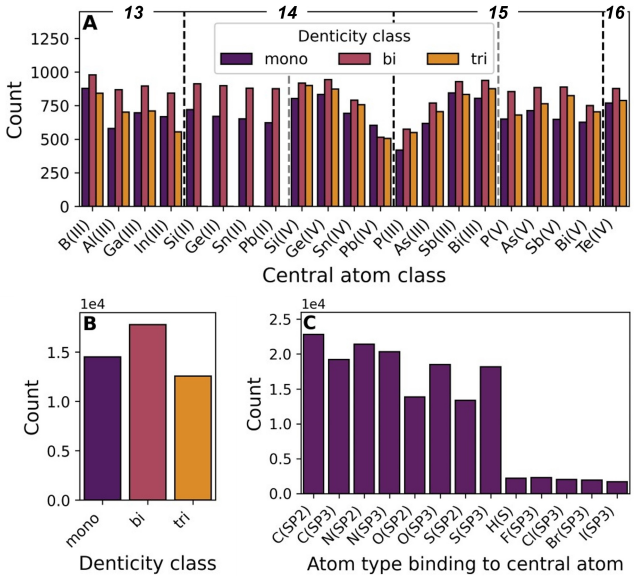


Figure 3. Composition of the FIA44k dataset with respect to **A)** the 21 different central atom and three ligand denticity classes, **B)** the overall distribution among the three denticity classes, and **C)** the atom types directly binding to the central atoms.

types of non-monoatomic ligands exists. The monoatomic ligands (hydrogen, halogens) have a lower contribution as they were treated within a general "monoatomic" class from which ligands were sampled during the generation of the dataset. This was done because monoatomic ligands bring much less potential for molecular complexity and diversity compared to polyatomic groups.

FIA44k has a diverse set of functional groups in the ligands. 5,265 different substructures and 2,987 different ring systems were identified. A more detailed analysis is given in the Supporting Information in Chapter S7. With respect to stereochemistry, fluoride adducts with pentavalent central atoms (e.g., Si(IV)-based molecules) have a (distorted) trigonal bipyramidal structure around the central atom and therefore, can adopt two different conformations (cf. Figure 9 and Chapter S8 in the Supporting Information). The fluoride can either be in apical or equatorial position. The FIA44k dataset includes both coordination modes in a 1:3 ratio in favor of structures with the added fluoride in apical position. This is of relevance (case study 4) since the mean FIA values of these two groups considerably differ by around 37 kJ mol^{-1}.

Dataset Analysis II: Trends and Correlations

The calculated FIA_{gas} values span a total range of 753 kJ mol^{-1}, with a minimum value of 10, a maximum value of 763, and a mean value of 309 kJ mol^{-1}. For FIA_{solv}, the range is essentially identical (744 kJ mol^{-1}), with the extrema at −201 and 543 kJ mol^{-1} shifted toward lower affinities due to solvation damping. The mean FIA_{solv} value is 124 kJ mol^{-1}. A more detailed analysis is given in Chapter

S14 in the Supporting Information. As FIA_{gas} and FIA_{solv} are strongly linearly correlated (see Figure S11A and S12 in the Supporting Information), the ensuing discussion is limited to FIA_{gas}.

Generally, the FIA distributions get sharper within a group of the periodic table with increasing atomic numbers (Figure 4A). The highest FIA values are found for group 13 molecules, with aluminum being the most Lewis acidic (on the FIA scale) with a mean FIA_{gas} of 471 kJ mol^{-1}. Boron, as the only central atom class of the second period of the periodic table, is clearly separated from the other group 13 molecules.[22]

For group 14 molecules, the low-valent central atom classes have higher FIA values compared to the high-valent cases, and in general, the FIA decreases with increasing atomic number. This is a notable finding, since previous research on neutral group 14 Lewis acids focused on the high-valent congeners. For molecules with central atoms from group 15, this situation is inverted. The low-valent central atom classes have a lower FIA compared to their high-valent counterparts. Also, the FIA increases with the atomic number. The analysis of the FIA values with respect to the three denticity classes shows an increase in FIA as well as an increasingly wider distribution with higher ligand denticity (Figure 4B). Accordingly, the highest FIA values are achieved with tridentate ligands.

As mentioned, the linear correlation between FIA_{gas} and FIA_{solv} is high ($r^2 = 0.921$, Figure 5A). On average, FIA_{solv} (calculated for dichloromethane) is 185 kJ mol^{-1} lower than FIA_{gas}, and a reasonable estimate of FIA_{solv} (± 16 kJ mol^{-1}) could be obtained from FIA_{gas} for neutral Lewis acids with

$$FIA_{solv} = 0.9077 \cdot FIA_{gas} - 157\ kJ\ mol^{-1}.$$

The correlation of the FIA with the intrinsic Lewis acidity descriptors LUMO energy and GEI was investigated (Chapter S7). In short, there is no correlation between any of these quantities and FIA_{gas}, which confirms that global and intrinsic Lewis acidity express two different aspects.[2c]

Machine Learning Models

The splitting of the FIA49k dataset in train, test, and validation portions is described in the Supporting Information in Chapter S15. As baseline models, random forest (RF) regressors[23] with 100 trees were trained using Morgan fingerprints (calculated for the whole molecules) of radius 3 and length 2048 as features. This resulted in a FIA_{gas} model that made predictions for the test set with an MAE of 32.2 kJ mol^{-1} ($r^2 = 0.660$). For FIA_{solv}, the model's accuracy was slightly higher (MAE = 29.7 kJ mol^{-1}, $r^2 = 0.665$). Changing to a LightGBM regressor[24] with 2000 boosting rounds and early stopping slightly increased the accuracy for both the FIA_{gas} (MAE = 28.3 kJ mol^{-1}, $r^2 = 0.738$) and FIA_{solv} model (MAE = 27.1 kJ mol^{-1}, $r^2 = 0.728$). These results indicate that molecular fingerprints (in the fashion they were applied) are features of only moderate quality for FIA prediction.

Next, we calculated all two-dimensional descriptors available in RDKit[17] and mordred[25] for the molecules of the FIA49k dataset. After feature reduction (see Chapter S16 in the Supporting Information for details), a set of 296 features was obtained. When the LightGBM regressor was retrained with these features, a significant improvement in accuracy was observed. The new FIA_{gas} model made predictions for the test set with an MAE of 17.4 kJ mol^{-1} ($r^2 = 0.899$), and the FIA_{solv} model with an MAE of 15.8 kJ mol^{-1} ($r^2 = 0.902$).

Reaching for higher predictive accuracy, message-passing graph neural networks (GNN) were considered.[26] For that, the two-dimensional molecular graph was initially transformed to an atom and bond token vector, respectively, by following a small number of classification rules: for

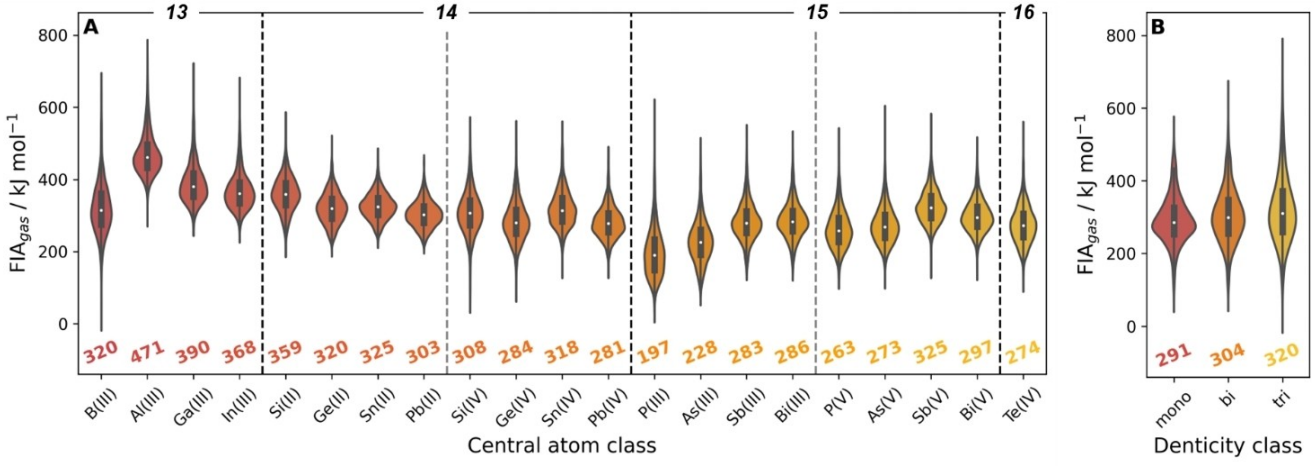


Figure 4. Distribution of calculated FIA_{gas} values of the FIA44k dataset with respect to **A**) the 21 different central atom classes and **B**) the three denticity classes. Low-valent group 14 molecules were omitted in B) as they cannot have tridentate ligands. The calculations were done at the RI-DSD-BLYP-D3(BJ)/def2-QZVPP//PBEh-3c level of theory. The numbers below the violins are the class-specific mean FIA values. See Figure S7 in the Supporting Information for the analogous plot with FIA_{solv} as target variable.

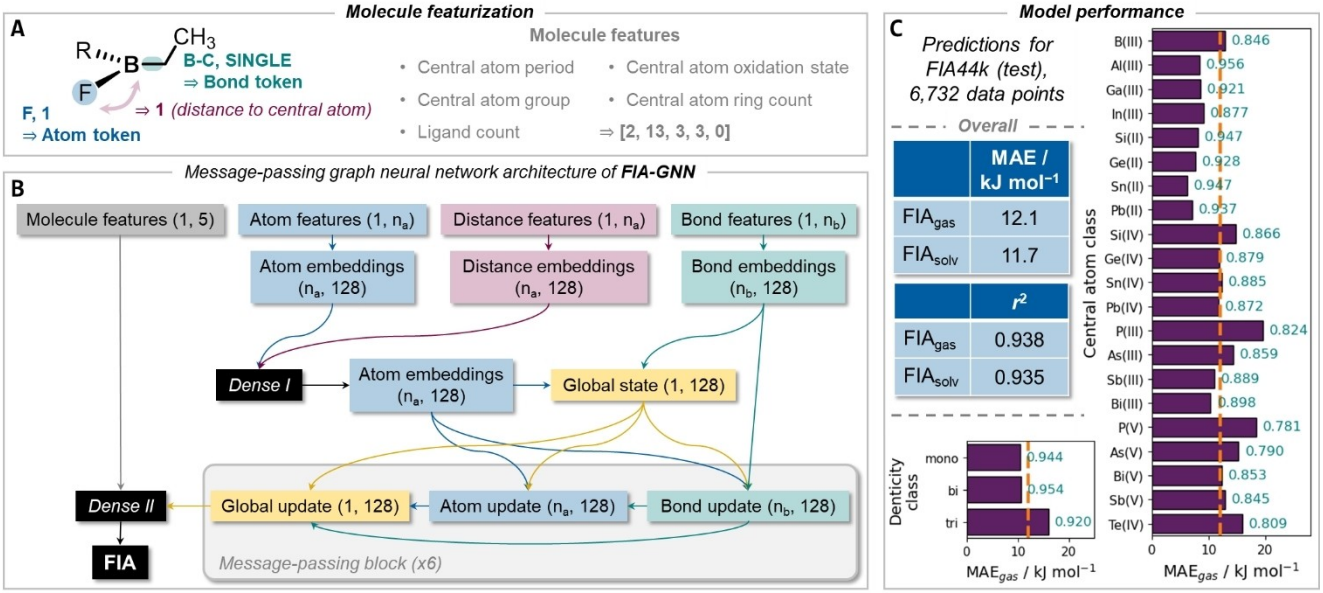


Figure 5. **A)** Featurization of molecular graphs (obtained from SMILES strings). Atom features for tokenization are atom symbol and degree, bond features participating atom symbols and bond order. **B)** FIA-GNN's model architecture. n_a means number of atoms, n_b number of bonds. "*Dense I*" is a set of three fully connected layers (512, 256, and 128 neurons). "*Dense II*" firstly expands the molecule features to 128 dimensions with a single layer and after concatenation provides a set of four fully connected layers for read-out (128, 64, 32, and 1 neuron(s)). **C)** FIA-GNN's performance in predicting FIA values of the test set evaluated overall, with respect to the 21 different central atom classes, and the three denticity classes. The dashed orange lines mark the overall MAE_{gas}. The numeric labels at the bars are the central atom and denticity class-specific r^2 values.

atoms, atom symbol and degree (neighbor count) were used; for bonds, participating atoms and the bond order were applied (Figure 5A).

Each token was then used to obtain initial atom and bond embeddings of length 128, which were sequentially updated during six rounds of message-passing. The GNNs were also provided with molecule features: the period and group of the central atom in the periodic table, its oxidation state, the number of rings the central atom is part of, and the number of ligands that are attached to it were used (Figure 5A). This resulted in a FIA_{gas} model which made predictions for the test set with an MAE of 13.1 kJ mol^{-1} ($r^2 = 0.931$), and in a FIA_{solv} model with an MAE of 12.3 kJ mol^{-1} ($r^2 = 0.927$).

Ultimately, the GNNs were extended by another input layer taking in a vector of the shortest bond path length between a given atom and the central atom. This informs the models on the concentricity of the problem at hand and potentially allows to learn the influence of a certain atom depending on its distance to the central atom. Indeed, this resulted in a further drop in MAE to 12.1 kJ mol^{-1} ($r^2 = 0.939$) for FIA_{gas} and 11.7 kJ mol^{-1} ($r^2 = 0.935$) for FIA_{solv}, the best FIA predictions we obtained. The respective model is called FIA-GNN, and its architecture is depicted in Figure 5B.

To put these results into perspective: the MAEs are less than 2 % of the FIA range across the entire dataset, which is around 750 kJ mol^{-1} (see above) and are on the same order of magnitude as the error of the quantum chemical method (RI-DSD-BLYP-D3(BJ)/def2-QZVPP//PBEh-3c) which was used to construct the dataset.[27] At the same time, FIA-GNN avoids quantum chemical calculations entirely and processes only SMILES strings as input.

Before FIA-GNN was applied in different case studies, its performance with respect to the central atom and ligand denticity classes was analyzed (Figure 5C). The FIA_{gas} values of data points with central atoms from group 13 (except for boron) and the low-valent group 14 atom-based molecules were most accurately predicted with MAEs below 10 kJ mol^{-1}. The largest absolute prediction errors showed P-based molecules (MAE of 19.6 and 18.4 kJ mol^{-1}). Lewis acids with only mono- or bidentate ligands had lower prediction MAEs (10.6 kJ mol^{-1}) compared to molecular structures with tridentate ligands (MAE = 16.1 kJ mol^{-1}).

Applications

As the first case study, FIA-GNN was employed to predict the FIA values of 1,200 molecules extracted from the CSD (second half of the FIA2k-CSD dataset, Figure 6). Importantly, these molecules were strictly excluded from any model selection or training steps. FIA_{gas} and FIA_{solv} were predicted with MAEs of 14.4 ($r^2 = 0.905$) and 13.5 kJ mol^{-1} ($r^2 = 0.895$), respectively, which is only slightly less accurate compared to the regressions made for the test set (cf. Figure 5C). This demonstrates the model's applicability to a broad variety of real-world examples, which are beyond the FIA44k dataset.

We inspected the molecules that showed the largest prediction errors. While it was not possible to detect systematic structural features across all poorly predicted

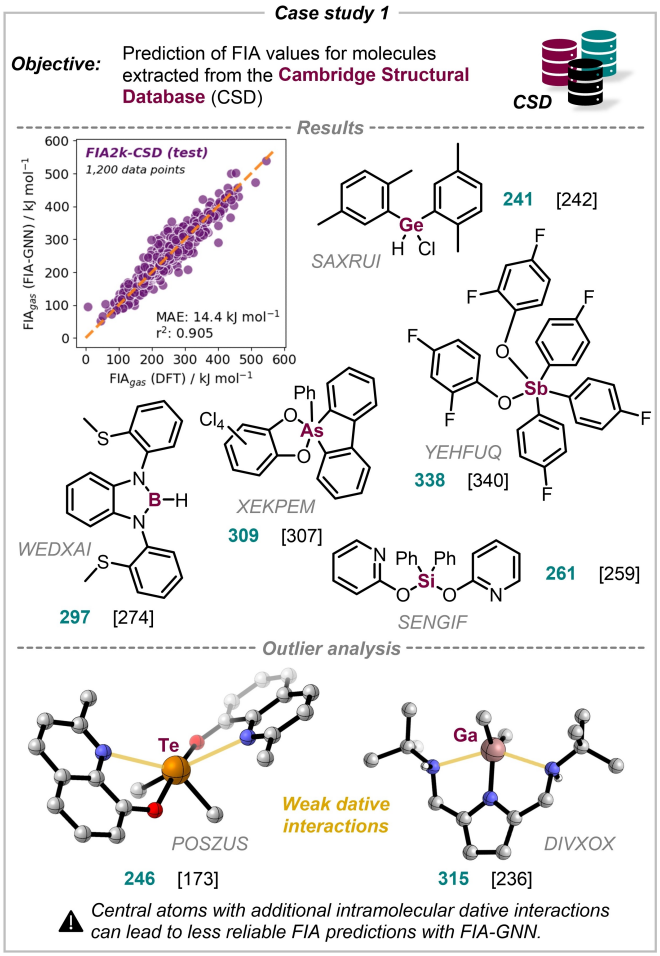


Figure 6. Application of FIA-GNN to predict FIA values of molecules from the Cambridge Structural Database (CSD). The explicitly shown molecules were randomly selected from the subset of data points that was added to the CSD in 2022. All numbers shown are FIA_{gas} values in kJ mol^{-1}. FIA-GNN predictions are given in bold and green, the DFT-calculated data in brackets. The carbon-bound hydrogen atoms of the shown molecular structures at the bottom were omitted for clarity.

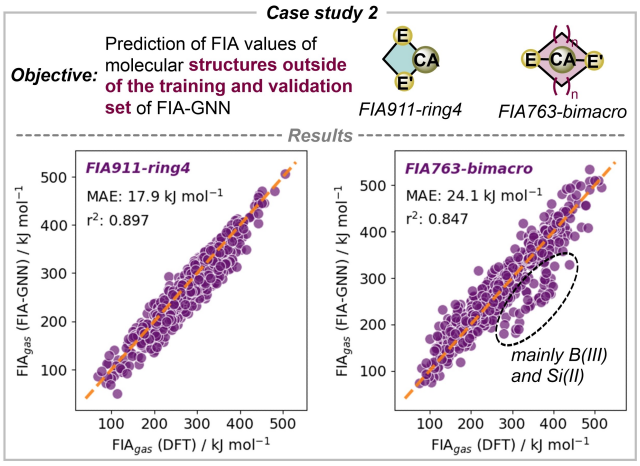


Figure 7. Application of FIA-GNN to predict FIA values of molecules with structures outside of the train and validation dataset of the model. For that, the FIA911-ring4 (molecules with the central atom embedded in a four-membered ring) and the FIA763-bimacro (molecules with a macrocyclic bidentate ligand) dataset were used.

cases, it was found that 5 of the 20 molecules with the highest error showed significant chemical bond rearrangement upon fluoride binding, a situation that was not explicitly considered during training (Figure 6, cf. Figure 2D). These hypercoordinations are within our cutoff values used for the transformation of xyz structures to molecular graphs with discrete edges (either bond or no bond) after structural optimization, however, it seems that they are not sufficiently represented in the training set to be adequately accounted for. This means that when substantial bond rearrangements occur after F^- binding, results from FIA-GNN could be less reliable.

As a second case study, we investigated how FIA-GNN performs on molecules with structural motifs that are systematically outside of the training dataset. As examples, molecules with their central atom embedded in a four-membered ring or which feature a macrocyclic bidentate ligand (e.g., 9-BBN) were chosen (Figure 7). Both structural characteristics are not present in the FIA44k dataset. A four-membered ring dataset with 911 members (FIA911-ring4) and a macrocyclic bidentate ligand dataset with 763 data points (FIA763-bimacro) were compiled (see Chapter S10 and S11 of the Supporting Information for details). FIA-GNN was more accurate when predicting the data of the FIA911-ring4 dataset ($MAE_{gas} = 17.9$ kJ mol^{-1}, $r^2_{gas} = 0.897$) compared to the FIA763-bimacro set of molecules ($MAE_{gas} = 24.1$ kJ mol^{-1}, $r^2_{gas} = 0.847$). Generally, a significant degree of predictive ability is conserved. However, when FIA values of molecular structures much different to the training and validation set are predicted, results should be considered with caution. When the prediction results for the FIA763-bimacro dataset were analyzed with respect to the central atom classes, it was found that especially the FIA values of B(III)- and Si(II)-based molecules were underestimated by the model (MAE_{gas} of 46.1 and 44.2 kJ mol^{-1}). For the FIA911-ring4 dataset, high-valent group 14 molecules showed the highest prediction errors (MAE_{gas} values of around 32 kJ mol^{-1}).

For the third application case study of FIA-GNN, the catalysis work of Kirschner et al., mentioned in the introduction, was consulted (Figure 8).[4] They found triethylborane and a fluoroaryl pinacolatoborane as effective phase transfer catalysts in their chosen applications and concluded that "[…] boranes with calculated fluoride affinity of 95–120 kJ mol^{-1} (vs. Me_3Si^+) appear to be suitable candidates as nucleophilic fluorination catalysts, […].".[4] We recalculated the FIA values for 15 boranes that were mentioned in the study (FIA15-PTCat dataset, see Chapter S13 in the Supporting Information). Despite a fair prediction error (28 kJ mol^{-1}), FIA-GNN identified BF_3 as the most fluorophilic molecule within the chosen scope, well beyond the desired FIA range (Figure 8). We tentatively assign this error to difficulties in correctly capturing the solvation stabilization when BF_4^- (fluoride adduct of BF_3) is transferred from vacuum to the solution phase. This stabilization

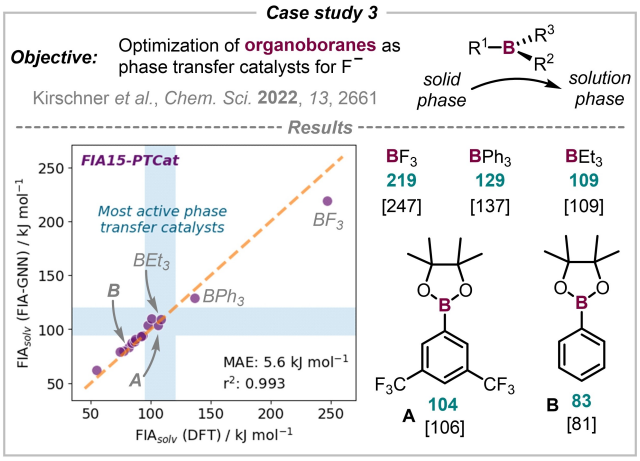


Figure 8. Application of FIA-GNN to predict FIA values of organoboranes, some of which were applied by Kirschner et al. as phase transfer catalysts for nucleophilic fluorination reactions.[4] All numbers shown are FIA_{solv} values in kJ mol^{-1}. FIA-GNN predictions are given in bold and green, the DFT-calculated data in brackets.

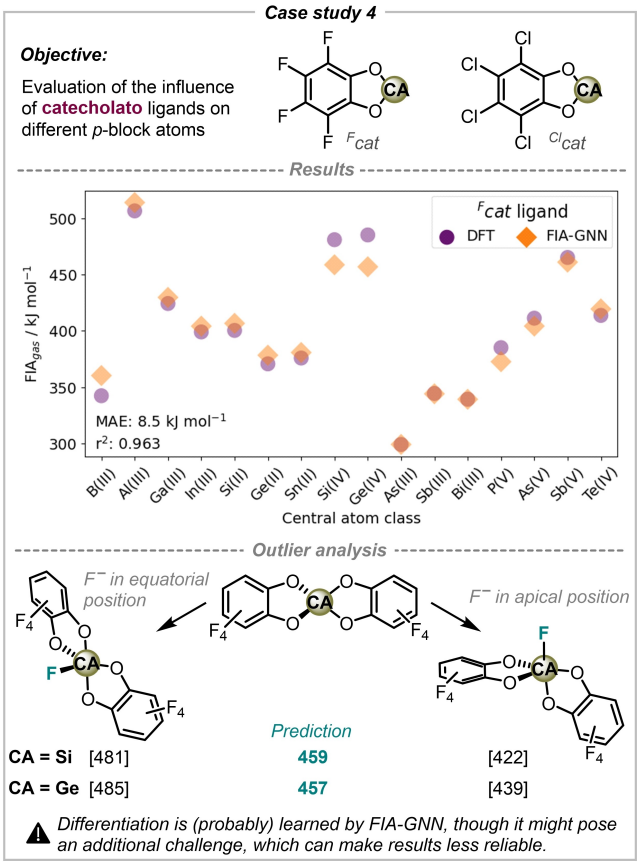


Figure 9. Application of FIA-GNN to predict FIA values of perfluoro- and perchlorocatecholato-substituted *p*-block atoms. All numbers shown are FIA_{gas} values in kJ mol^{-1}. FIA-GNN predictions are given in bold and green, the DFT-calculated data in brackets. For further details, see Chapter S8 and S12 in the Supporting Information.

is abnormally large for small anions such as BF_4^-. Our model correctly labeled BPh_3 as relatively strong and Ph-BPin (**B** in Figure 8) as a relatively weak Lewis acid on the FIA scale. Both are outside of the desired FIA span and accordingly, were found to give poor experimental results.[4] The FIA_{solv} values of two of the most effective phase transfer catalysts, BEt_3 and 3,5-$(CF_3)_2C_6H_3$-BPin (**A** in Figure 8) were predicted with high accuracy by the model. Importantly, FIA-GNN coped well with the difference between 40 % substrate conversion after 24 hours and 99 % conversion after 8 hours being less than 20 kJ mol^{-1} on the FIA_{solv} scale.

After FIA-GNN was used to study the influence of different ligands on one central atom (case study 3), it was employed to explore the impact of a single ligand system across different central atom classes (case study 4). For that, the perfluoro- (^{F}cat) and perchlorocatecholato (^{Cl}cat) ligand were chosen as they have been applied in multiple studies to prepare strong neutral *p*-block atom-based Lewis acids (Figure 9).[8a,c,d,f,28] In all cases, the maximum number of catecholato ligands was placed at the central atom, while free valences in molecules with an odd central atom oxidation state (e.g., in B(III)-based molecules) were saturated with methyl groups. FIA values were successfully calculated for 31 of the 42 different combinations (FIA31-cat dataset, see Chapter S12 in the Supporting Information). FIA-GNN predicted these accurately with MAE_{gas} values of 8.5 (^{F}cat, $r^2 = 0.963$) and 8.6 (^{Cl}cat, $r^2 = 0.968$) kJ mol^{-1}. FIA_{solv} was equally well predicted (MAE = 9.4 kJ mol^{-1}, $r^2 = 0.945$). It detected the aluminum-based molecules as the clearly strongest fluoride acceptors. FIA_{gas} values are on average 9.4 kJ mol^{-1} higher for molecules with the ^{Cl}cat compared to the ^{F}cat ligand and are also higher across the entire board of central atoms. When changing to FIA_{solv}, this difference decreases significantly ($\Delta = 0.8$ kJ mol^{-1}). Remarkably, FIA-GNN was able to qualitatively reproduce this subtle effect (from $\Delta = 6.0$ kJ mol^{-1} for FIA_{gas} to $\Delta = 3.2$ kJ mol^{-1} for FIA_{solv}). Also, for low-valent group 14 atom-based compounds, the minor change in the qualitative FIA trend (from Si > Sn > Ge > Pb for FIA_{gas} to Si > Ge > Sn > Pb for FIA_{solv}) when going from FIA_{gas} to FIA_{solv} was correctly modeled. This demonstrates the applicability of FIA-GNN to study the influence of a given ligand system on different central atoms and importantly, also in regard of potentially varying solvent influences.

Within the predictions of FIA-GNN for the FIA31-cat dataset, the FIA values of the high-valent group 14 atom-based molecules (Si(IV) and Ge(IV)) were systematically underestimated. In fact, pentavalent fluoride adducts can adopt two different conformations that differ in the location of the fluoride, being either in equatorial or apical position (Figure 9, cf. Chapter S8 in the Supporting Information). It is known from single crystal structure analyses that bis(catecholato)silanes and -germanes preferentially position the fluoride in equatorial position.[28a,29]

The "apical FIA_{gas}" of $(^{F}cat)_2Si$ and $(^{F}cat)_2Ge$ were approximately determined and were found to be significantly lower than the equatorial counterparts (Figure 9). We therefore hypothesize that FIA-GNN even learned the differentiation between equatorial and apical fluoride

acceptance while it was trained on the FIA regression task. Still, this differentiation most likely poses an additional challenge for the model, which can make predictions for molecules that result in fluoride adducts that can have multiple stereoisomers less straightforward.

To test this hypothesis, the final 32-dimensional features from FIA-GNN (FIA_{gas}) were used to train a binary linear discriminant analysis (LDA) classifier.[23] Indeed, a model with 79 % prediction accuracy for the relevant part of the test set of FIA44k was obtained. This points to the presence of information on the stereochemistry of high-valent group 14 atom-based fluoride adducts in the molecular representations learned by FIA-GNN. It should be emphasized that FIA-GNN was trained with the SMILES strings of the Lewis acids as input, and the described apical/equatorial problem does not apply to most of the training dataset. Interestingly, the LDA model correctly classified all four bis(catecholato)tetrels mentioned before as molecules that host fluoride anions in the equatorial position.

Conclusion

The fluoride ion affinity (FIA) is among the most popular descriptors for global (thermodynamic) Lewis acidity. However, its computation, including the often more meaningful solvation-corrected values, demands several steps with dedicated user input and computational requirements. Herein, we present a ML tool called FIA-GNN that allows the prediction of FIA_{gas} and the solvation-corrected FIA_{solv} (CH_2Cl_2) within seconds and with a mean absolute error of 14 kJ mol^{-1} ($r^2 = 0.93$) based on the Lewis structure of the compound of interest (in form of a SMILES string) as the only necessary input. This service, including a graphical structure editor, is provided at www.grebgroup.de/fia-gnn. The predictions are made by two message-passing graph-neural networks that have been trained with FIA49k, a newly compiled dataset with 48,986 fluoride ion affinities of neutral *p*-block atom-based molecules calculated at the RI-DSD-BLYP-D3(BJ)/def2-QZVPP//PBEh-3c level of theory for gas and solution phase. While spanning a FIA range of around 750 kJ mol^{-1}, the dataset covers 13 different *p*-block atoms as fluoride acceptors and features them in low and high-valent states, decorated with mono-, bi, or tridentate ligands.

FIA49k was further used to relate the FIA scale to other Lewis acidity descriptors and to explore the FIA space of *p*-block atom-based Lewis acids. It provides the most up-to-date general information source to describe Lewis acidity trends within *p*-block atom-based molecules.

Applications of FIA-GNN to molecules from the Cambridge Structural Database or catalysis research from the literature demonstrated the model's usefulness in real-world tasks. Molecules that possess intramolecular saturation of the Lewis acidic center with available donor atoms, can lead to larger prediction errors. Future work could address this problem by augmentation of FIA49k with respective data points – optimally with an active learning strategy.

To be noted, the major portion of FIA-GNN's training data was the FIA44k dataset. FIA44k is intended to cover an as general as possible chemical space of *p*-block atom-based molecules – without specifically addressing molecules, for example, with particularly large FIA values, high degree of halogenation, or minimal size. Extensions to these areas will be the objective of future research efforts. Keeping this in mind, we hope that the herein presented ML model will help to guide the development of new *p*-block atom-based Lewis acids by granting access to their FIA value within a fraction of a second.

Authors Contributions

L.M.S.: Conceptualization (lead), Investigation (equal), Methodology (lead), Project Administration (equal), Software (equal), Supervision (equal), Visualization (lead), Validation (equal), Writing – Original Draft Preparation (lead), Writing – Review & Editing (equal); **S.S.S.V.**: Investigation (equal), Methodology (supporting), Software (equal); **A.A.**: Investigation (supporting), Software (supporting), Validation (supporting), Visualization (supporting), Web design (lead); **P.E.**: Investigation (supporting); **R.S.P.**: Project Administration (equal), Resources (equal), Supervision (equal), Writing – Review & Editing (equal); **L.G.**: Project Administration (equal), Resources (equal), Supervision (equal), Writing – Review & Editing (equal).

Acknowledgements

The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 40/575-1 FUGG (JUSTUS 2 cluster) and GR5007/6-1. L.M.S. is grateful to the "Studienstiftung des deutschen Volkes" and the „Landesgraduiertenförderung" for scholarships.

R.S.P. and S.S.S.V. acknowledge the NSF under the CCI Center for Computer-Assisted Synthesis (CHE-2202693) for support, the Alpine high performance computing resource at the University of Colorado Boulder, which is jointly funded by the University of Colorado Boulder, the University of Colorado Anschutz, and Colorado State University, and the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) through allocation TG-CHE180056.

The authors acknowledge the contributions of Philipp Kollenz to the project. Open Access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available in the supplementary material of this article.

Keywords: Lewis acids · fluoride ion affinity · data science · machine learning · graph neural networks

[1] a) A. Corma, *Chem. Rev.* **1995**, *95*, 559–614; b) A. Corma, H. García, *Chem. Rev.* **2003**, *103*, 4307–4366; c) J. Mlynarski, *Chiral Lewis acids in organic synthesis*, Wiley-VCH, Weinheim, **2017**; d) H. Yamamoto, *Lewis acids in organic synthesis*, Wiley-VCH, Weinheim, **2008**.

[2] a) K. O. Christe, D. A. Dixon, D. McLemore, W. W. Wilson, J. A. Sheehy, J. A. Boatz, *J. Fluorine Chem.* **2000**, *101*, 151–153; b) J. C. Haartz, D. H. McDaniel, *J. Am. Chem. Soc.* **1973**, *95*, 8562–8565; c) P. Erdmann, J. Leitner, J. Schwarz, L. Greb, *ChemPhysChem* **2020**, *21*, 987–994; d) H. Böhrer, N. Trapp, D. Himmel, M. Schleep, I. Krossing, *Dalton Trans.* **2015**, *44*, 7489–7499.

[3] a) A. P. Altshuller, *J. Am. Chem. Soc.* **1955**, *77*, 6187–6188; b) H. D. B. Jenkins, H. K. Roobottom, J. Passmore, *Inorg. Chem.* **2003**, *42*, 2886–2893; c) K. A. G. MacNeil, J. C. J. Thynne, *J. Phys. Chem.* **1970**, *74*, 2257–2262; d) T. E. Mallouk, G. L. Rosenthal, G. Mueller, R. Brusasco, N. J. I. c. Bartlett, *Inorg. Chem.* **1984**, *23*, 3167–3173; e) J. A. Stockdale, D. R. Nelson, F. J. Davis, R. N. Compton, *J. Chem. Phys.* **2003**, *56*, 3336–3341.

[4] S. Kirschner, M. Peters, K. Yuan, M. Uzelac, M. J. Ingleson, *Chem. Sci.* **2022**, *13*, 2661–2668.

[5] a) D. Höfler, M. van Gemmeren, P. Wedemann, K. Kaupmees, I. Leito, M. Leutzsch, J. B. Lingnau, B. List, *Angew. Chem. Int. Ed.* **2017**, *56*, 1411–1415; b) L. Ratjen, M. van Gemmeren, F. Pesciaioli, B. List, *Angew. Chem. Int. Ed.* **2014**, *53*, 8765–8769.

[6] I. Krossing, I. Raabe, *Chem. Eur. J.* **2004**, *10*, 5017–5030.

[7] a) N. Bormann, J. S. Ward, A. K. Bergmann, P. Wenz, K. Rissanen, Y. Gong, W.-B. Hatz, A. Burbaum, F. F. Mulks, *Chem. Eur. J.* **2023**, *29*, e202302089; b) A. Hermannsdorfer, M. Driess, *Angew. Chem. Int. Ed.* **2020**, *59*, 23132–23136; c) C. Foroutan-Nejad, J. Vicha, R. Marek, *Chem. Eur. J.* **2014**, *20*, 11584–11590; d) J. F. Kögel, D. A. Sorokin, A. Khvorost, M. Scott, K. Harms, D. Himmel, I. Krossing, J. Sundermeyer, *Chem. Sci.* **2018**, *9*, 245–253; e) F. S. Tschernuth, T. Thorwart, L. Greb, F. Hanusch, S. Inoue, *Angew. Chem. Int. Ed.* **2021**, *60*, 25799–25803; f) A. Ben Saida, A. Chardon, A. Osi, N. Tumanov, J. Wouters, A. I. Adjieufack, B. Champagne, G. Berionni, *Angew. Chem. Int. Ed.* **2019**, *58*, 16889–16893; g) A. Osi, N. Niessen, D. Mahaut, B. Champagne, A. Chardon, N. Tumanov, J. Wouters, G. Berionni, *Z. Anorg. Allg. Chem.* **2023**, *649*, e202300009; h) J. Brzeski, *J. Comput. Chem.* **2023**, *44*, 1454–1463; i) L. O. Müller, D. Himmel, J. Stauffer, G. Steinfeld, J. Slattery, G. Santiso-Quiñones, V. Brecht, I. Krossing, *Angew. Chem. Int. Ed.* **2008**, *47*, 7659–7663; j) L. S. Warring, J. E. Walley, D. A. Dickie, W. Tiznado, S. Pan, R. J. Gilliard, Jr., *Inorg. Chem.* **2022**, *61*, 18640–18652; k) D. Duvinage, L. A. Malaspina, S. Grabowsky, S. Mebs, J. Beckmann, *Eur. J. Inorg. Chem.* **2023**, *26*, e202200482; l) K. F. Hoffmann, A. Wiesner, S. Steinhauer, S. Riedel, *Chem. Eur. J.* **2022**, *28*, e202201958; m) L. Greb, *Chem. Eur. J.* **2018**, *24*, 17881–17896; n) A. Hermannsdorfer, M. Driess, *Angew. Chem. Int. Ed.* **2021**, *60*, 13656–13660; o) Y. Gong, J. S. Ward, K. Rissanen, F. F. Mulks, *Molbank* **2023**, *29*, M1710.

[8] a) L. Greb, *Synlett* **2023**; b) M. Schorpp, R. Yadav, D. Roth, L. Greb, *Angew. Chem. Int. Ed.* **2022**, *61*, e202207963; c) D. Hartmann, M. Schädler, L. Greb, *Chem. Sci.* **2019**, *10*, 7379–7388; d) R. Maskey, M. Schädler, C. Legler, L. Greb, *Angew. Chem. Int. Ed.* **2018**, *57*, 1717–1720; e) D. Roth, J. Stirn, D. W. Stephan, L. Greb, *J. Am. Chem. Soc.* **2021**, *143*, 15845–15851; f) D. Roth, H. Wadepohl, L. Greb, *Angew. Chem. Int. Ed.* **2020**, *59*, 20930–20934; g) T. Thorwart, D. Roth, L. Greb, *Chem. Eur. J.* **2021**, *27*, 10422–10427.

[9] a) A. R. Jupp, T. C. Johnstone, D. W. Stephan, *Dalton Trans.* **2018**, *47*, 7029–7035; b) A. R. Jupp, T. C. Johnstone, D. W. Stephan, *Inorg. Chem.* **2018**, *57*, 14764–14771.

[10] a) J. R. Gaffen, J. N. Bentley, L. C. Torres, C. Chu, T. Baumgartner, C. B. Caputo, *Chem* **2019**, *5*, 1567–1583; b) R. J. Mayer, N. Hampel, A. R. Ofial, *Chem. Eur. J.* **2021**, *27*, 4070–4080; c) P. Erdmann, L. Greb, *Angew. Chem. Int. Ed.* **2022**, *61*, e202114550; d) L. Zapf, M. Riethmann, S. A. Föhrenbacher, M. Finze, U. Radius, *Chem. Sci.* **2023**, *14*, 2275–2288.

[11] a) N. Artrith, K. T. Butler, F.-X. Coudert, S. Han, O. Isayev, A. Jain, A. Walsh, *Nat. Chem.* **2021**, *13*, 505–508; b) P. M. Pflüger, F. Glorius, *Angew. Chem. Int. Ed.* **2020**, *59*, 18860–18865; c) A. Bender, N. Schneider, M. Segler, W. Patrick Walters, O. Engkvist, T. Rodrigues, *Nat. Chem. Rev.* **2022**, *6*, 428–442; d) A. Karthikeyan, U. D. Priyakumar, *J. Chem. Sci.* **2021**, *134*, 2; e) Y.-F. Shi, Z.-X. Yang, S. Ma, P.-L. Kang, C. Shang, P. Hu, Z.-P. Liu, *Engineering (Beijing)* **2023**; f) M. Meuwly, *Chem. Rev.* **2021**, *121*, 10218–10239; g) W. L. Williams, L. Zeng, T. Gensch, M. S. Sigman, A. G. Doyle, E. V. Anslyn, *ACS Cent. Sci.* **2021**, *7*, 1622–1637.

[12] a) F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, O. A. von Lilienfeld, *J. Chem. Theory Comput.* **2017**, *13*, 5255–5264; b) J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller, A. Tkatchenko, *Chem. Rev.* **2021**, *121*, 9816–9872; c) L. C. Gallegos, G. Luchini, P. C. St. John, S. Kim, R. S. Paton, *Acc. Chem. Res.* **2021**, *54*, 827–836.

[13] a) G. Hoffmann, M. Balcilar, V. Tognetti, P. Héroux, B. Gaüzère, S. Adam, L. Joubert, *J. Comput. Chem.* **2020**, *41*, 2124–2136; b) S. A. Cuesta, M. Moreno, R. A. López, J. R. Mora, J. L. Paz, E. A. Márquez, *J. Chem. Inf. Model.* **2023**, *63*, 507–521; c) Y. Liu, Q. Yang, J. Cheng, L. Zhang, S. Luo, J.-P. Cheng, *ChemPhysChem* **2023**, *24*, e202300162.

[14] H. Huynh, K. Le, L. Vu, T. Nguyen, M. Holcomb, S. Forli, H. Phan, **2023**, ChemRxiv preprint DOI: 10.26434/chemrxiv-22023-lcxn26430.

[15] D. Rodrigues Silva, L. de Azevedo Santos, M. P. Freitas, C. F. Guerra, T. A. Hamlin, *Chem. Asian J.* **2020**, *15*, 4043–4054.

[16] C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, *Acta Crystallogr. Sect. B* **2016**, *72*, 171–179.

[17] RDKit: Open-source cheminformatics. https://www.rdkit.org.

[18] a) J. M. Blaney, J. S. Dixon, *in Rev. Comput. Chem.* **1994**, pp. 299–335; b) S. Riniker, G. A. Landrum, *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.

[19] P. Pracht, F. Bohle, S. Grimme, *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

[20] F. Neese, *WIREs Comput. Mol. Sci.* **2022**, *12*, e1606.

[21] a) AMS 2021.102, SCM, Theoretical Chemistry, Vrije Universiteit, Amsterdam, The Netherlands, http://www.scm.com; b) G. te Velde, F. M. Bickelhaupt, E. J. Baerends, C. Fonseca Guerra, S. J. A. van Gisbergen, J. G. Snijders, T. Ziegler, *J. Comput. Chem.* **2001**, *22*, 931–967; c) C. C. Pye, T. Ziegler, E. van Lenthe, J. N. Louwen, *Can. J. Chem.* **2009**, *87*, 790–797.

[22] Z.-L. Wang, H.-S. Hu, L. von Szentpály, H. Stoll, S. Fritzsche, P. Pyykkö, W. H. E. Schwarz, J. Li, *Chem. Eur. J.* **2020**, *26*, 15558–15564.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. J. t. J. o. m. L. r. Dubourg, *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

[24] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, *Adv. Neural. Inf. Process. Syst.* **2017**, *30*.
[25] H. Moriwaki, Y.-S. Tian, N. Kawashita, T. Takagi, *J. Cheminf.* **2018**, *10*, 4.
[26] P. C. St. John, C. Phillips, T. W. Kemper, A. N. Wilson, Y. Guan, M. F. Crowley, M. R. Nimlos, R. E. Larsen, *J. Chem. Phys.* **2019**, *150*.
[27] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, S. Grimme, *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184–32215.
[28] a) A. L. Liberman-Martin, R. G. Bergman, T. D. Tilley, *J. Am. Chem. Soc.* **2015**, *137*, 5328–5331; b) A. T. Henry, D. A. R. Nanan, K. M. Baines, *Dalton Trans.* **2023**, *52*, 10363–10371.
[29] N. Ansmann, D. Hartmann, S. Sailer, P. Erdmann, R. Maskey, M. Schorpp, L. Greb, *Angew. Chem. Int. Ed.* **2022**, *61*, e202203947.

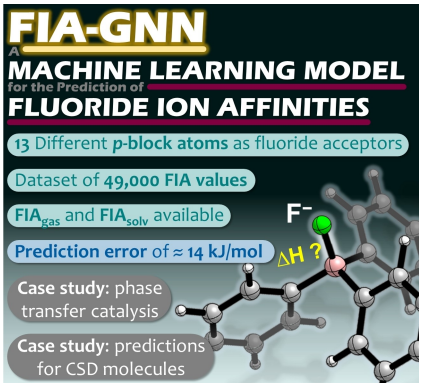
Manuscript received: January 19, 2024
Accepted manuscript online: March 7, 2024
Version of record online: ■■, ■■

Research Articles

Machine Learning

L. M. Sigmund,* S. S. S. V., A. Albers, P. Erdmann, R. S. Paton,* L. Greb* **e202401084**

Predicting Lewis Acidity: Machine Learning the Fluoride Ion Affinity of *p*-Block-Atom-Based Molecules



The fluoride ion affinity (FIA) is among the most prominent descriptors for Lewis acidity. This paper presents FIA-GNN, a machine learning model that predicts FIA values in gas and solution (CH_2Cl_2) phase within a fraction of a second on a standard personal laptop. This is several orders of magnitude faster compared to the conventional quantum chemical approach. FIA-GNN is applied in four different case studies including catalysis research.