Introduction to Data Analysis

Final Project Report

Investigating the Correlation between Molecular Descriptors and BF3 Affinity

Huynh Tam Minh Hieu UG190025

Pham Hoang Lan UG190042

Instructor: Prof. Tran Duy Hien

Fulbright University Vietnam

August 2nd, 2023

TABLE OF CONTENTS

1. Abstract

This report investigates the use of machine learning to predict BF3 affinity in Lewis bases. Molecular descriptors are used to characterize molecules for prediction. The dataset includes 344 Lewis Adducts, with 141 descriptors as explanatory variables. Exploratory data analysis reveals insights into BF3 affinity and its correlation with descriptors. Gradient Boosting and Ridge Regression models are employed and compared using hypothesis testing based on 99% one-tail confidence intervals. Results show both models effectively predict BF3 affinity, with Ridge Regression having a slight advantage. These findings have significant implications for chemistry research and experimental design, facilitating the development of simulation models to reduce reliance on costly experiments.

2. Introduction

In recent years, the application of machine learning in chemistry has increased significantly, resulting in different achievements in predicting HOMO-LUMO orbitals[1], lattice energies[2] and charge transfer integrals of organic crystals[2,3]. A crucial tool in this research is the use of molecular descriptors – numeric representations that help us grasp the properties of diverse molecules by describing their structures. Concurrently, researchers have been focusing on Lewis acid – Lewis base interaction, which plays a vital role in developing organic semiconductors, especially organic solar cells.[4,5]

In the realm of chemistry, Lewis bases play a fundamental role in the formation of chemical bonds and interactions between molecules. These species are electron donors, possessing lone pairs of electrons that are available for bonding. Conversely, Lewis's acids are electron acceptors, seeking to receive these electrons. One prominent property of this interaction is the red-shift phenomenon of a Lewis base's absorption spectrum when it forms a bond with a Lewis acid, which has been confirmed by various experiments.[4,6–11].

Boron trifluoride (BF3) is a well-known Lewis acid due to its electron deficiency. Consequently, BF3 has a strong affinity for Lewis bases. When a Lewis base encounters BF3, its electron-rich lone pair is attracted to the electron-deficient boron atom. The interaction results in the formation of a Lewis acid-based adduct, where the Lewis base donates its electron pair to BF3. This process illustrates the concept of electron pair donation and is central to various chemical reactions and coordination complexes.

Molecular descriptors are essential tools used to characterize the properties and behavior of molecules. These descriptors are numerical values that represent specific features of a molecule, such as its size, shape, polarity, or electronic properties. These descriptors are generated by a logical and mathematical process or collected from atomic experimental data.[12] This tool aid researchers in predicting diverse molecular properties like molecular reactivity, stability, and other crucial attributes.

The affinity of a Lewis base for BF3 can be correlated to certain molecular structures. For instance, the presence of a lone pair of electrons in a Lewis base is a crucial descriptor that determines its ability to interact with BF3. Molecules with more accessible and electron-rich lone pairs are likely to exhibit higher affinity for BF3. Those properties are expected to be feasibly demonstrated by different molecular descriptors.

In conclusion, understanding Lewis bases, their affinity for Lewis acids like BF3, and their correlation with molecular descriptors are vital in comprehending the principles governing chemical interactions and reactivity. Molecular descriptors provide valuable insights into the behavior of molecules, facilitating the prediction and design of chemical reactions and materials.

In this report, we present a comprehensive analysis of experimental data, using statistical techniques to explore the correlation between molecular descriptors and BF3 affinity. Based on our findings, we will select suitable machine learning algorithms to predict the BF3 affinity of Lewis Bases. Successful predictions will pave the way for constructing a simulation model, reducing the need for costly and time-consuming experiments. Moreover, we will evaluate the performance of these machine learning models, in terms of quantifying the goodness of their

predictions as well as comparing their performance to one another, offering valuable insights for future applications of computational simulations. Chemists can then use this knowledge to make informed decisions in their research and experimental designs, potentially advancing the field significantly.

3. Analytical Methodology

i. BF3 Affinity Dataset

The dataset used in this study consists of 344 Lewis adducts, which were constructed by combining the BF3 Lewis acid with a set of 344 Lewis bases from Chapter 3 of the book 'Lewis basicity and affinity scales' by Christian Laurence.[13] Due to that the focus of this research is to investigate the BF3 affinity of these Lewis bases, the experimental values of the Lewis bases' BF3 affinity, gathered in Chapter 3, serve as the response variable in this analysis.

To provide a better understanding of the molecular properties of the Lewis bases, the study employs the Dragon software,[14] which generates molecular descriptors. Four groups of descriptors were chosen based on their ability to offer valuable insights for molecular design. These groups are:

- Constitutional descriptors: These provide information about the molecular composition, including molecular weight (MW) and mean atomic Sanderson electronegativity scaled to C (Me).
- Atom-centered fragments: These descriptors are defined for specific atoms, such as hydrogen, carbon, and heteroatoms. Examples include the number of =CH2 fragments (C-015).
- Functional group count descriptors: These descriptors count various functional groups, like the number of non-aromatic conjugated C(sp2) (nCconj) or the number of imides (nN(CO)2).
- Molecular properties: This group includes descriptors like the Moriguchi octanol-water partition coefficient (logP) – MLOGP.

In total, 141 different molecular descriptors are gathered from these four groups and are used as explanatory variables in the dataset to study the BF3 affinity of the Lewis bases.

ii. Exploratory Data Analysis

Exploratory data analysis (EDA) is a crucial step in understanding the dataset and gaining valuable insights before implementing the predictive modeling process. In this study, EDA was applied to examine the dataset containing BF3 affinity values of different types of Lewis bases. The histograms of BF3 affinity for various Lewis's base types, including all Lewis bases, Lewis bases with Nitrogen as the bonding atom with BF3, and Lewis bases with Oxygen as the bonding atom with BF3. Additionally, statistical details (i.e., quantity, Pearson correlation with the response variable) of each molecular descriptor type (i.e., continuous and discrete) were analyzed.

The EDA performed on the dataset revealed essential insights, guiding the choice of modeling techniques and reinforcing the need for a robust machine learning approach to effectively predict the BF3 affinity of Lewis bases. The subsequent methodology section will elaborate on the chosen algorithm and the model construction process.

iii. Machine Learning Algorithms

Two machine learning models from the Scikit-learn library[15] have been selected to analyze the dataset:

- Gradient Boosting: This model is a tree-based non-linear regression approach.
- Ridge Regression: This model is a linear-based regression method.

For both models, we will compute the Pearson correlation coefficients (r), which range from 0 to 1, between the predicted and experimental values in the test set. These metrics will be extracted for hypothesis testing, where higher values of r indicate more accurate predictions.

The planned validation procedure for both models is as follows:

1. Set the hyper-parameters for each model.
2. Randomly split the dataset into an 80% training set and a 20% testing set.
3. Utilize the Scikit-learn library to train the models.
4. Repeat steps 2 and 3 for 100 iterations to generate 100 values of r for each model.

The above methodology will provide a comprehensive analysis of the two machine learning models and their effectiveness in predicting the target variable in the given dataset.

iv. Hypothesis Testing

In this section, we will conduct hypothesis testing to analyze the distribution of the generated samples of Pearson correlation coefficients (r) for each machine learning model and compare their performances. The following statistical techniques will be employed:

- **Normality Test:** We will assess whether the generated samples of r values for each machine learning model follow a normal distribution. This will be achieved by dividing the 100 generated Pearson correlations into 20 interval groups, each containing 5 observations. Normal probabilities corresponding to each interval will be computed. Subsequently, a Chi-square test will be conducted to confirm the normality of the distribution.
- **One-Sample Hypothesis Testing:** Once the normality of the data is confirmed, we will perform a one-sample hypothesis test to evaluate the significance and goodness of the results generated by the chosen machine learning models. The null hypothesis will assume that the population mean of r is not larger than 0.8. By comparing the sample mean of r to the hypothesized value of 0.8, we can determine if one's performance exceeds the specified threshold.
- **Two-Sample Hypothesis Testing:** In addition to the one-sample test, a two-sample hypothesis test will be conducted to compare the performance of the two machine learning models. The objective is to identify any statistically significant differences in predictive capabilities between the models. The null hypothesis will state that one model will perform better than the other. By comparing the sample means of r for both models, we can determine which model demonstrates superior performance.

These statistical analyses will provide us with valuable insights into the distribution of r values for each model and enable us to make informed conclusions regarding their predictive capabilities.

4. Results and Discussion

i. Exploratory data analysis

- **BF3 affinity of Lewis bases**

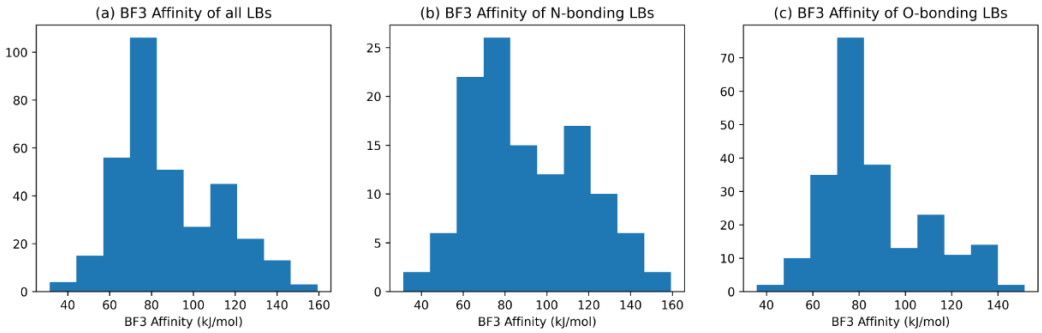


Figure 1. Histograms of experimental values of BF3 affinity of Lewis bases

Looking at the histograms of BF3 affinity of different types of Lewis bases (i.e., BF3 affinity of all Lewis bases; Lewis bases with Nitrogen as the bonding atom with BF3; Lewis bases with Oxygen as the bonding atom with BF3), it can be observed that there is a similar pattern of any type of Lewis bases regarding their BF3 affinity. All generated histograms have a two-peak shape with the peaks at around 75 and 115 kJ/mol. Thus, it is unnecessary to take the two types of Lewis bases into consideration separately, but to consider the BF3 of all Lewis bases as the main target of the study.

- **Molecular descriptors**

Descriptor Type	Quantity	Absolute Pearson Correlation with Response Variable			
		Average	Standard Deviation	Minimum	Maximum
Continuous	17	0.115	0.101	0.000	0.495
Discrete	124	0.156	0.090	0.007	0.411
All	141	0.119	0.100	0.000	0.495

Table 1. Statistical details (i.e., quantity, Pearson correlation with the response variable) of each molecular descriptor type (i.e., continuous and discrete)

There are two key observations in the above Table 1:

- Firstly, the Pearson correlation of each molecular descriptor with the target is significantly small. This indication demonstrates that by no means can the correlation between experimental BF3 affinity and the values of molecules be proven if the molecular descriptors are involved in the calculation individually. Therefore, there is a need for a machine learning algorithm to flexibly combine different groups of molecular descriptors.
- Secondly, it is noticeable that the quantity of molecular descriptors presented by discrete data outnumbers that presented by continuous data. This imbalance in molecular descriptor's data types can potentially lead to inaccuracy in the later prediction process since the prediction value is expected to be a continuous number. Thus, a simple linear regression model (least square regression) is unsuitable for the current dataset. A non-linear regression model or a linear regression model with an addition of a regularization function should be employed to increase the possibility of generating continuous values in the prediction.

ii. Normality Test

The normality test was performed using the Chi-square goodness-of-fit test to assess the distribution of the generated Pearson correlation samples for the Ridge Regression and Gradient Boosting models using a significance value of 0.01.

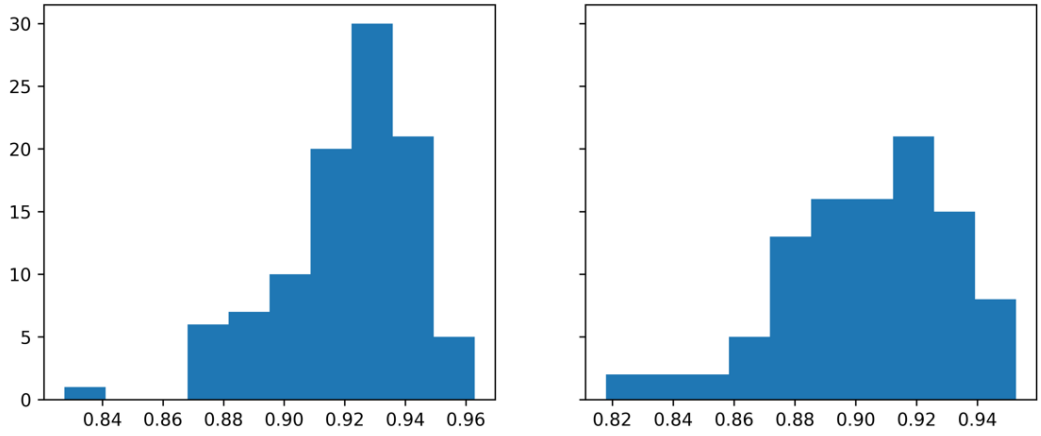
- For the Ridge Regression model, the squared difference between observed and expected values for Chi-square test statistic was calculated as 21.339.

- For the Gradient Boosting model, the squared difference between observed and expected values for Chi-square test statistic was calculated as 15.162.

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

Table 2. Critical values of the Chi-square distribution with degrees of freedom and significance values

With the degrees of freedom of 19 (20 interval groups) and a significance value of 0.01, we can compare the value of our test statistics (21.339 and 15.162) to the Chi-square value in Table 2 (36.191). Since our test statistics are larger than 36.191, it can be ascertained that the sample sets of Pearson correlations produced by the two machine learning models are normally distributed.



(a)Histogram of Pearson Correlation Samples for Ridge Regression

(b)Histogram of Pearson Correlation Samples for Gradient Boosting

Figure 3. Histograms of Pearson Correlation Samples for Ridge Regression and Gradient Boosting models

iii. Hypothesis Testing

In the report, we conducted two hypotheses testing for evaluating the performance of the Ridge Regression and Gradient Boosting machine learning models. The results are based on the analysis of 99% one-tail confidence intervals for Pearson correlation coefficients (r).

- **Hypothesis I**

  For the first hypothesis testing, the 99% one-tail confidence intervals for the two models are as follows:

  - Ridge Regression: (0.92, +∞)
  - Gradient Boosting: (0.90, +∞)

The critical value of 0.8, which represents the threshold for "good" performance, lies beyond the upper bounds of both confidence intervals. As a result, both models are statistically significant and exhibit "good" predictive capabilities. The confidence intervals provide strong evidence that the models' performance, with high correlation coefficients, extends to positive infinity, indicating robustness in predicting the target variable.

- **Hypothesis II**

  For the second hypothesis testing, the 99% one-tail confidence interval for the distribution of the performance differences between Ridge Regression and Gradient Boosting is (0.02, +∞). The interval encompasses positive values, suggesting that Ridge Regression slightly outperforms Gradient Boosting.

  The critical value of 0 lies beyond the confidence interval, indicating that the likelihood of the performance difference being equal to zero is relatively low. Therefore, Ridge Regression is statistically better than Gradient Boosting for the given dataset.

The hypothesis test results provide valuable insights into the effectiveness of the two machine learning models. Both Ridge Regression and Gradient Boosting demonstrate good predictive capabilities, with correlation coefficients exceeding 0.8. These high correlation values indicate a strong association between the predicted and experimental values, validating their reliability for the dataset at hand. Moreover, the second hypothesis testing reveals that Ridge Regression exhibits a statistically minor advantage over Gradient Boosting in terms of predictive performance.

5. Conclusion

The exploratory data analysis revealed crucial insights, guiding the selection of appropriate modeling techniques. The histograms for BF3 affinity showed a consistent two-peak pattern across all Lewis bases, so the overall BF3 affinity was chosen as the primary target. The analysis of molecular descriptors showed a small Pearson correlation with the experimental BF3 affinity, indicating the need for a machine learning algorithm to combine descriptors for better prediction accuracy. Non-linear regression models or linear regression models with regularization functions were recommended for continuous predictions.

The normality test using the Chi-square goodness-of-fit confirmed the normal distribution of generated Pearson correlation samples for Ridge Regression and Gradient Boosting, ensuring reliable subsequent statistical analyses.

Hypothesis testing with 99% one-tail confidence intervals for Pearson correlation coefficients (r) revealed both models' favorable performance. The calculated confidence intervals indicated significant correlation coefficients exceeding the threshold of 0.8, signifying "good" predictive capabilities. Ridge Regression showed a minor statistical advantage over Gradient Boosting in the second hypothesis test, supporting its suitability for the dataset.

In conclusion, both Ridge Regression and Gradient Boosting models effectively predicted the BF3 affinity of Lewis bases with strong correlation to experimental values, making them reliable for practical applications. Ridge Regression is suggested as a preferable choice due to its marginally better performance, but both models can be confidently employed for accurate predictions and valuable insights into the BF3 affinity of Lewis bases.

6. References

(1) Pereira, F.; Xiao, K.; Latino, D. A. R. S.; Wu, C.; Zhang, Q.; Aires-de-Sousa, J. Machine Learning Methods to Predict Density Functional Theory B3LYP Energies of HOMO and LUMO Orbitals. *J Chem Inf Model* 2017, *57* (1), 11–21. https://doi.org/10.1021/acs.jcim.6b00340.

(2) Musil, F.; De, S.; Yang, J.; Campbell, J. E.; Day, G. M.; Ceriotti, M. Machine Learning for the Structure–Energy–Property Landscapes of Molecular Crystals. *Chem. Sci.* 2018, *9* (5), 1289–1300. https://doi.org/10.1039/C7SC04665K.

(3) Wang, C.-I.; Braza, M. K. E.; Claudio, G. C.; Nellas, R. B.; Hsu, C.-P. Machine Learning for Predicting Electron Transfer Coupling. *J. Phys. Chem. A* 2019, *123* (36), 7792–7802. https://doi.org/10.1021/acs.jpca.9b04256.

(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) 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.

(6) Lee, J.-W.; Kim, H.-S.; Park, N.-G. Lewis Acid–Base Adduct Approach for High Efficiency Perovskite Solar Cells. *Acc. Chem. Res.* 2016, *49* (2), 311–319. https://doi.org/10.1021/acs.accounts.5b00440.

(7) Wang, Z.-F.; Yi, Z.; Ahmad, A.; Xie, L.; Chen, J.-P.; Kong, Q.; Su, F.; Wang, D.-W.; Chen, C.-M. Combined DFT and Experiment: Stabilizing the Electrochemical Interfaces via Boron Lewis Acids. *Journal of Energy Chemistry* 2021, *59*, 100–107. https://doi.org/10.1016/j.jechem.2020.10.041.

(8) Zalar, P.; Henson, Z. B.; Welch, G.; Bazan, G.; Nguyen, T.-Q. Color Tuning in Polymer Light-Emitting Diodes with Lewis Acids. *Angewandte Chemie* 2012. https://doi.org/10.1002/anie.201202570.

(9) Pang, B.; Tang, Z.; Li, Y.; Meng, H.; Xiang, Y.; Li, Y.; Huang, J. Synthesis of Conjugated Polymers Containing B←N Bonds with Strong Electron Affinity and Extended Absorption. *Polymers (Basel)* 2019, *11* (10), 1630. https://doi.org/10.3390/polym11101630.

(10) 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.

(11) Li, Y.; Meng, H.; Li, Y.; Pang, B.; Luo, G.; Huang, J. Adjusting the Energy Levels and Bandgaps of Conjugated Polymers via Lewis Acid–Base Reactions. *New J. Chem.* 2018, *42* (23), 18961–18968. https://doi.org/10.1039/C8NJ04453H.

(12) Frontmatter. In *Handbook of Molecular Descriptors*; John Wiley & Sons, Ltd, 2000; pp i–xxi. https://doi.org/10.1002/9783527613106.fmatter.

(13) The BF3 Affinity Scale. In *Lewis Basicity and Affinity Scales*; John Wiley & Sons, Ltd, 2009; pp 85–109. https://doi.org/10.1002/9780470681909.ch3.

(14) *Molecular descriptors calculation - Dragon - Talete srl.* http://www.talete.mi.it/products/dragon_description.htm (accessed 2022-12-30).

(15) *scikit-learn: machine learning in Python — scikit-learn 1.1.1 documentation*. https://scikit-learn.org/stable/ (accessed 2022-06-06).