



Investigating the Correlation between Molecular Descriptors and BF₃ Affinity with Machine Learning

Huynh Tam Minh Hieu – 190025

Pham Hoang Lan – 190042



Introduction

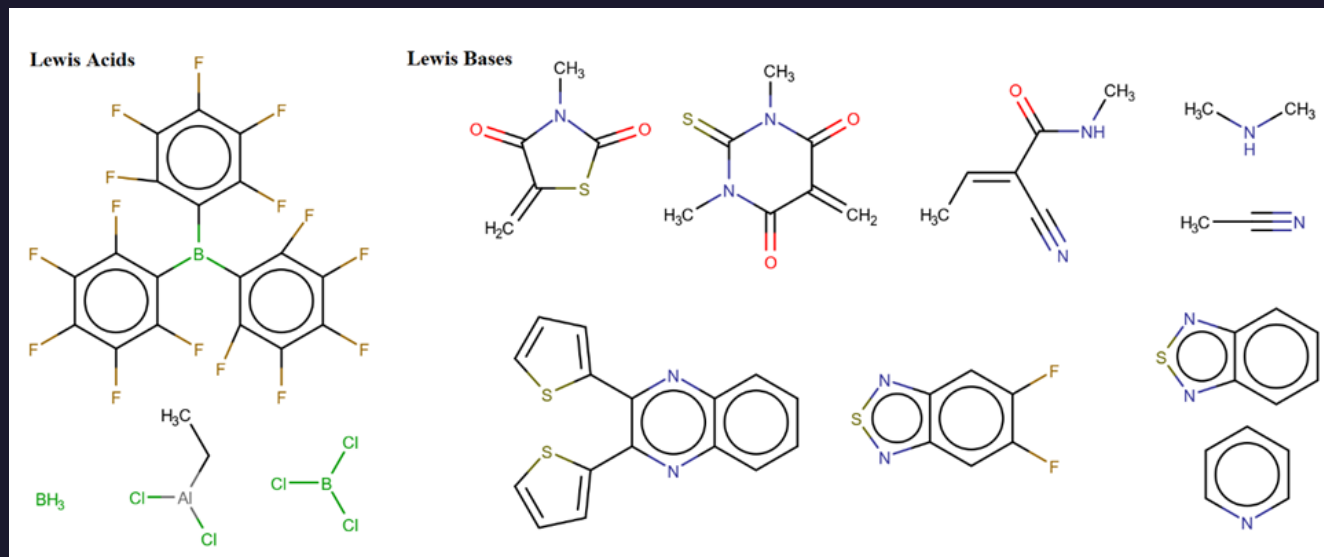
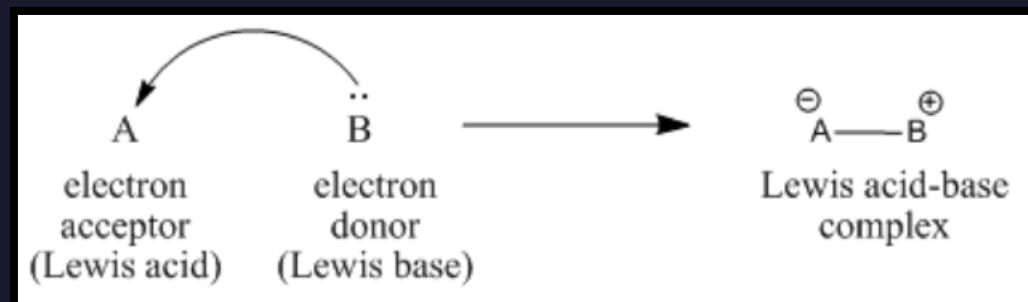


Semiconductor Applications



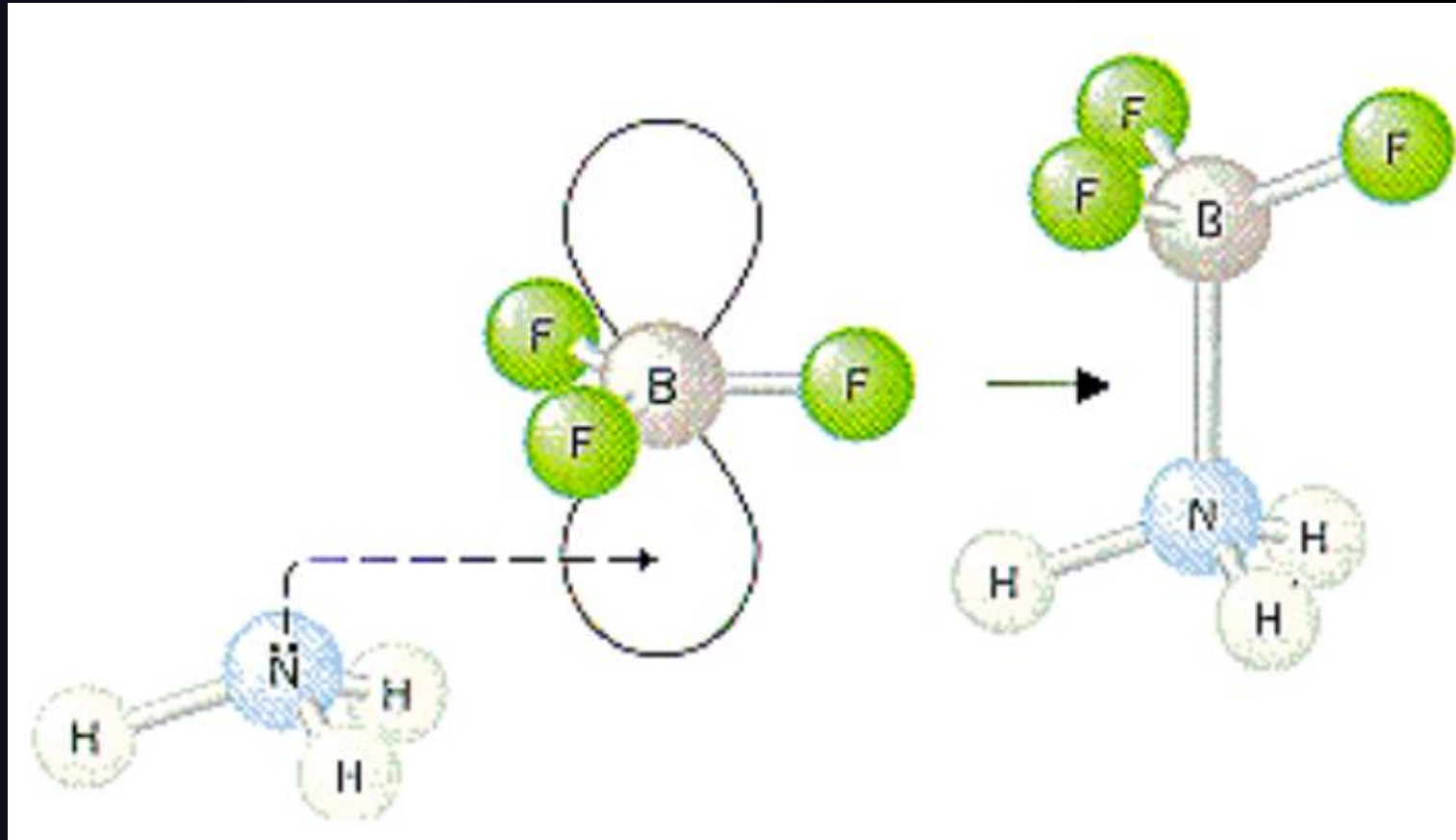
Semiconductor & Application fields

Lewis Acid – Lewis base Interaction



Lewis acid-base interaction is an intermolecular bond formed when one molecule donates a pair of electrons (base) to another molecule that accepts the pair (acid).

Lewis Acid – Lewis base Interaction



Molecular Descriptor

Molecular descriptors are defined as numerical representations of molecules' chemical information generated by a logical and mathematical process or collected from atomic experimental data

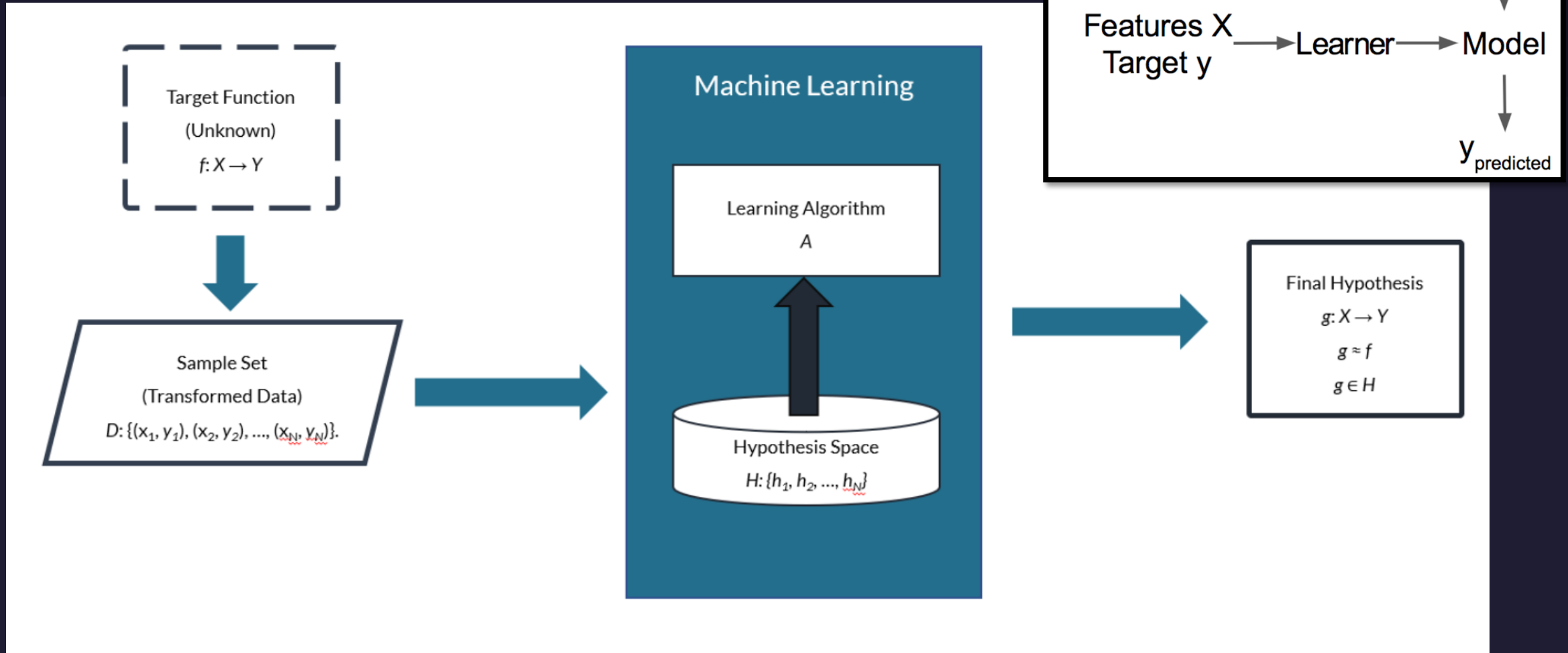
Pros: Reduce reliance on costly experiments

Dragon Software



Molecular descriptors calculation - Dragon - Talete srl.
http://www.talete.mi.it/products/dragon_description.htm

Machine Learning



Project Objectives


1. Investigate the correlation between molecular descriptors and BF3 affinity using statistical techniques.
2. Identify and select appropriate machine learning algorithms for predicting BF3 affinity.
3. Evaluate the performance of the machine learning models through hypothesis testing, specifically:
 - a. Quantify the goodness of predictions made by the models.
 - b. Compare the performance between models.

Statistical Approach and Results

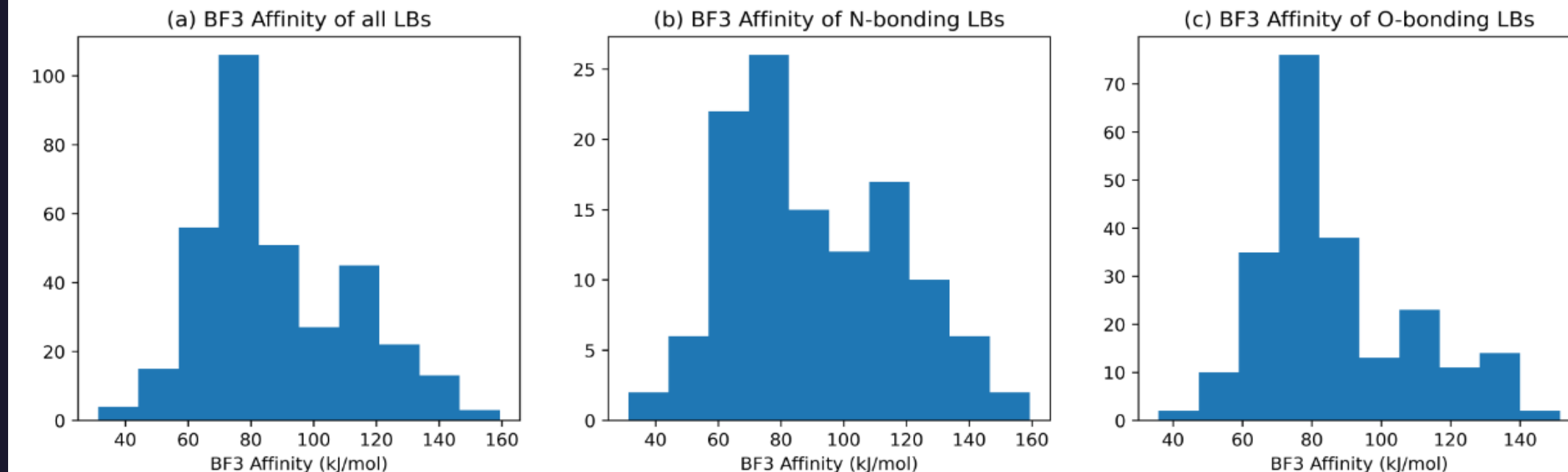




BF3 Affinity Dataset

- Dataset: 344 Lewis adducts (BF3 Lewis acid + 344 Lewis bases) from Christian Laurence's Chapter 3 of 'Lewis basicity and affinity scales'.
 - Response Variable: Experimental values of Lewis bases' BF3 affinity
 - Molecular Descriptors: Extracted using Dragon software.
 - Four Groups of Descriptors:
 - Constitutional descriptors
 - Atom-centered fragments
 - Functional group count descriptors
 - Molecular properties
 - A total of 141 molecular descriptors used as explanatory variables to study BF3 affinity of the Lewis bases.
- 

Exploratory Data Analysis



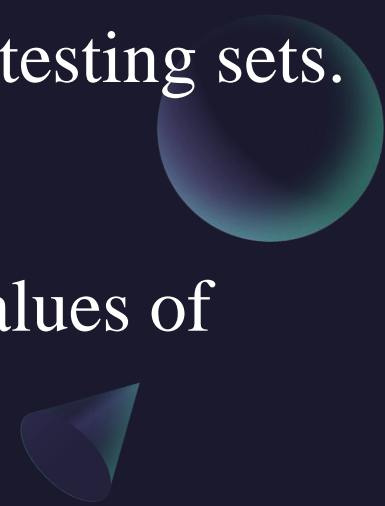
- Similar pattern observed for all types of Lewis bases regarding BF3 affinity.
- Histograms show a two-peak shape with peaks at approximately 75 and 115 kJ/mol.
- Suggest focusing on BF3 affinity of all Lewis bases as the main target of the study.

Exploratory Data Analysis

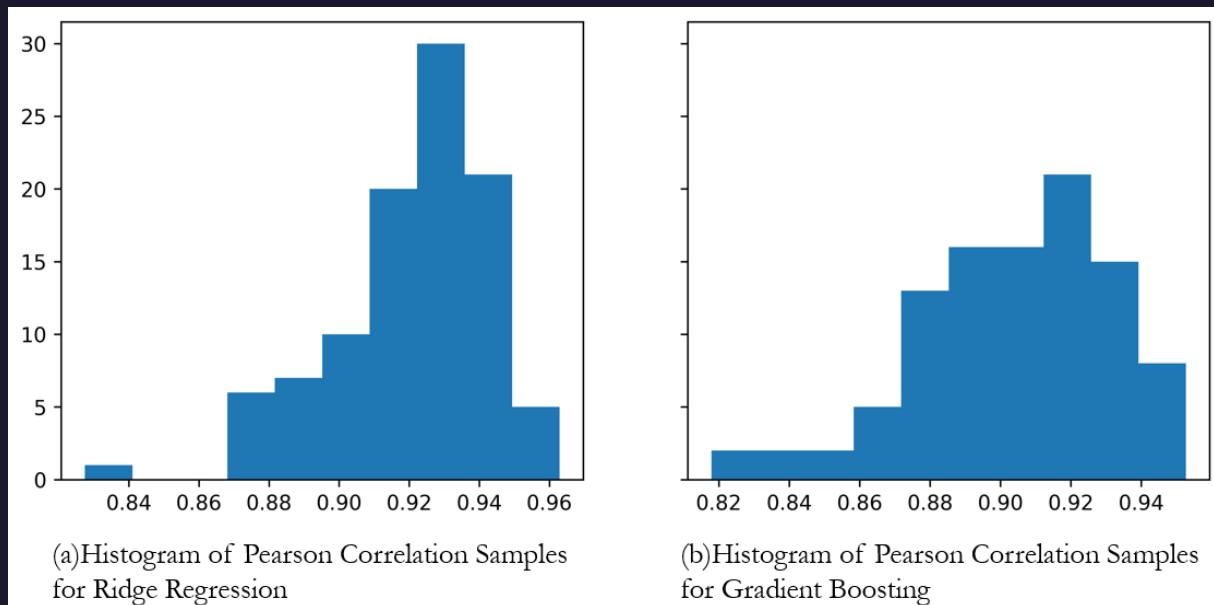
Descriptor Type	Quantity	Absolute Pearson Correlation with Response Variable			
		Average	Standard Deviation	Minimum	Maximum
Continuous	17	0.115	0.101	0.000	0.495
Discrete	124	0.156	0.090	0.007	0.411
All	141	0.119	0.100	0.000	0.495

- Pearson correlation between molecular descriptors and experimental BF3 affinity is small, requiring a machine learning algorithm to combine descriptors effectively.
- Imbalance between discrete and continuous data in molecular descriptors requires a non-linear regression model or linear regression with regularization for accurate continuous value predictions.

Machine Learning Analysis

- **Gradient Boosting:** A tree-based non-linear regression approach.
 - **Ridge Regression:** A linear-based regression method.
 - **Validation Procedure:**
 - Set hyper-parameters for each model.
 - Randomly split the dataset into 80% training and 20% testing sets.
 - Train the models using Scikit-learn library.
 - Repeat steps 2 and 3 for 100 iterations to obtain 100 values of Pearson correlation coefficients (r) for each model.
- 

Normality Test



- Ridge model's Chi-square test statistic = 21.339.
- GB model's Chi-square test statistic = 15.162.
- Degrees of freedom = 19
- Significance value = 0.01.

→ Both test statistics (21.339 and 15.162) are larger than the critical Chi-square value (36.191), confirming that the Pearson correlation samples from the two models are normally distributed.

Hypothesis Testing – Hypothesis I

- One-Sample Hypothesis Testing: to quantify the goodness of predictions made by the chosen models.
 - Null hypothesis: population mean of $r \leq 0.8$. Compare sample mean of r to 0.8 to determine model performance.
 - 99% one-tail confidence intervals for models:
 - Ridge Regression: $(0.92, +\infty)$
 - Gradient Boosting: $(0.90, +\infty)$
- Both models are statistically significant and have "good" predictive capabilities. Confidence intervals suggest robustness in predictions up to positive infinity, with high correlation coefficients.

Hypothesis Testing – Hypothesis II

- Two-Sample Hypothesis Testing: Compare two models' performance
 - Null hypothesis: One model will perform better than the other
 - 99% one-tail confidence interval for performance differences:
 - Ridge Regression vs. Gradient Boosting: $(0.02, +\infty)$
- Ridge Regression slightly outperforms Gradient Boosting. Critical value 0 lies outside the interval, showing Ridge Regression is statistically better for the dataset



Conclusion

- **EDA:** BF3 affinity chosen as primary target. Molecular descriptors showed small correlation, requiring ML algorithms.
- **Normality Test:** Chi-square confirmed normal distribution of Pearson correlation samples for Ridge Regression and Gradient Boosting.
- **Hypothesis Testing:** 99% confidence intervals for r showed both models' favorable performance (>0.8). Ridge Regression slightly outperformed Gradient Boosting.
- **Conclusion:** Both models effectively predict BF3 affinity. Ridge Regression recommended due to slightly better performance. Both models suitable for practical applications.

Thank You

