香港中文大學
The Chinese University of Hong Kong

Journal Club

Constructing and explaining machine learning models for chemistry: example of the exploration and design of boron-based Lewis acids

Zihan Li

https://doi.org/10.48550/arXiv.2501.01576

25th Jul 2025

Content

1. Background for Research

2. Results and Discussion

3. Conclusion

1. Background for Research

Machine Learning

↓

drug discovery, molecular simulations, chemical reaction prediction, synthesis planning, etc.

Black box model Eg. deep neural networks

highly accurate ✅

lack of interpretability ❌

↓

limits the ability to extract scientific knowledge

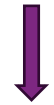
Solution

explainable artificial intelligence (XAI)

- elucidating what ML algorithms have learned
- fostering scientific knowledge
- inspiring new concepts and ideas

Applied on QSAR

Quantitative Structure–Activity Relationship

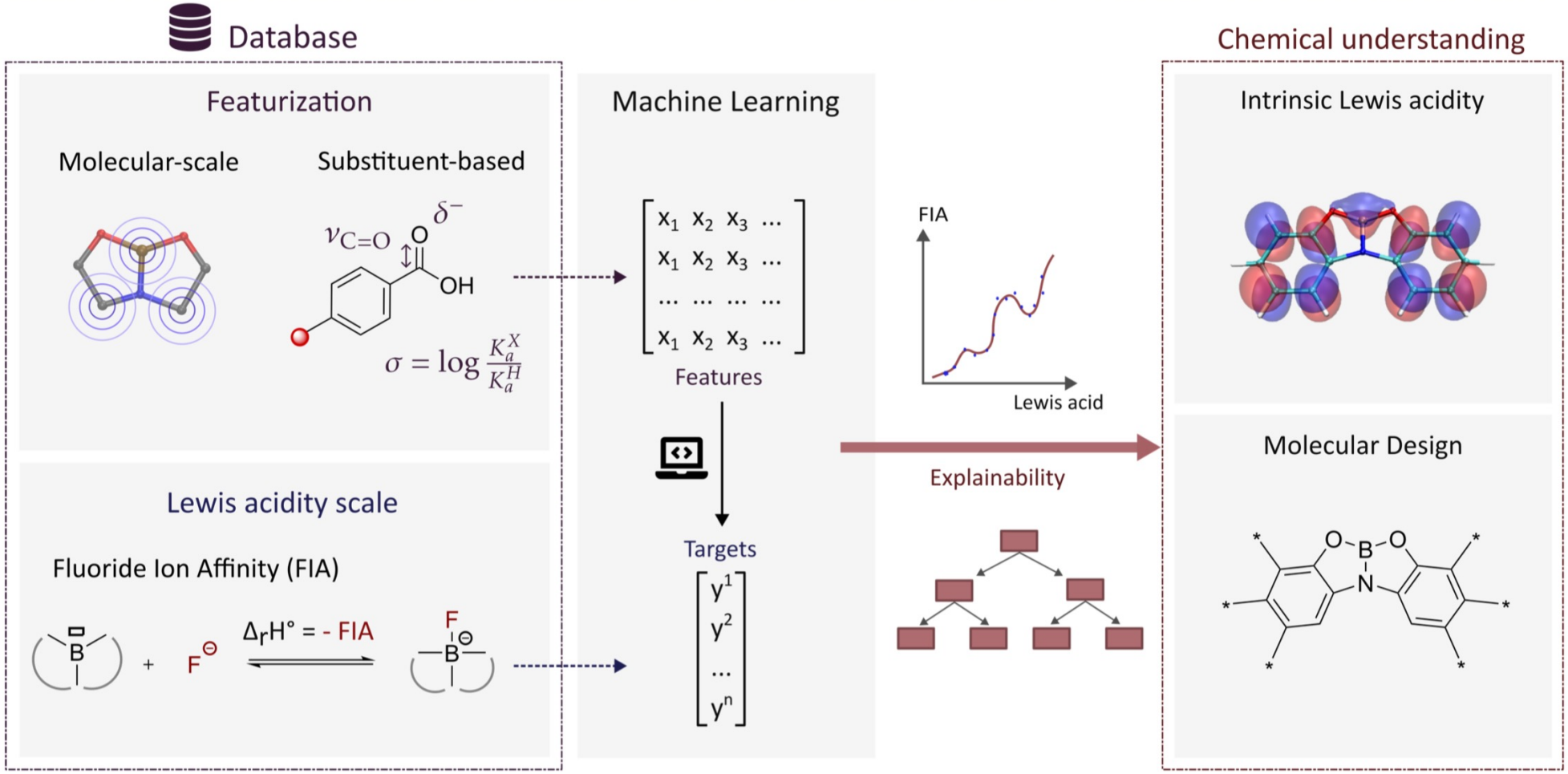


Predict Lewis acidity

current job: Greb group's GNN model

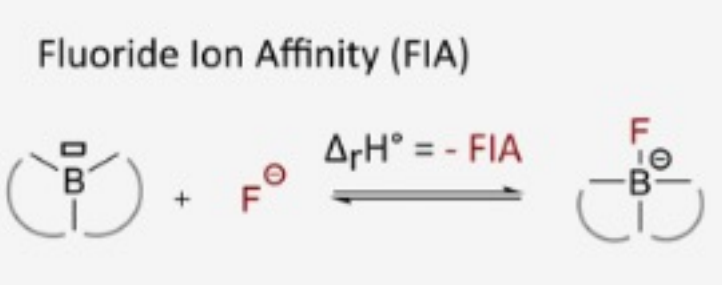
Angew Chem Int Ed 2024, e202401084.

This Work


Database
Featurization
Molecular-scale
Substituent-based
$\nu_{C=O}$
δ^-
O
OH
$\sigma = \log \frac{K_a^X}{K_a^H}$
Lewis acidity scale
Fluoride Ion Affinity (FIA)
$\Delta_r H° = -$ FIA
F
Machine Learning
x_1 x_2 x_3 ...
Features
Targets
y^1
y^2
...
y^n
FIA
Lewis acid
Explainability
Chemical understanding
Intrinsic Lewis acidity
Molecular Design
O
B
N

2. Results and Discussion

Lewis acidity scale

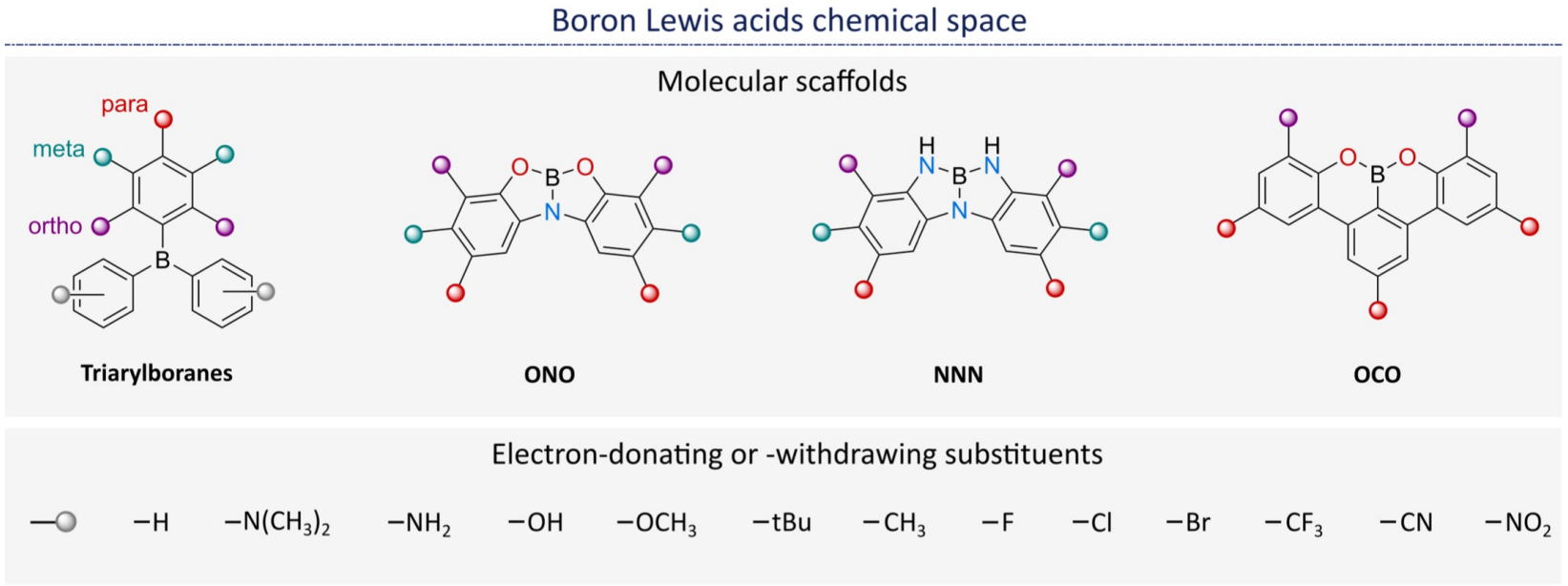


relevant, consistent and accessible quantity

use DFT at the M062X/6-31G(d) level of theory in isodesmic calculations as a compromise between efficiency and precision to provide reliable FIA data.

SMILES	δ(^{31}P) (solvent)	Δδ(^{31}P)	average Δδ(^{31}P)	FIA (kJ.mol-1)	HIA (kJ.mol-1)	Reorganization energy (kJ.mol-1)	GEI (eV)
CCOB(OCC)OCC	48.7 (neat)[6]	7.7[6]	7.7	248.69	371.27	208.95	0.69
COB(OC)OC	48.1 (C6D6)[10]	1.3[10]	1.3	234.63	359.50	188.53	0.73
ClCCOB(OCCCl)OCCCl	55.1 (neat)[11]	14.1[11]	14.1	318.12	409.35	214.23	1.15
ClCCCOB(OCCCCl)OCCCCl	56.3 (neat)[11]	15.3[11]	15.3	312.98	431.19	221.68	1.39

Chemical space

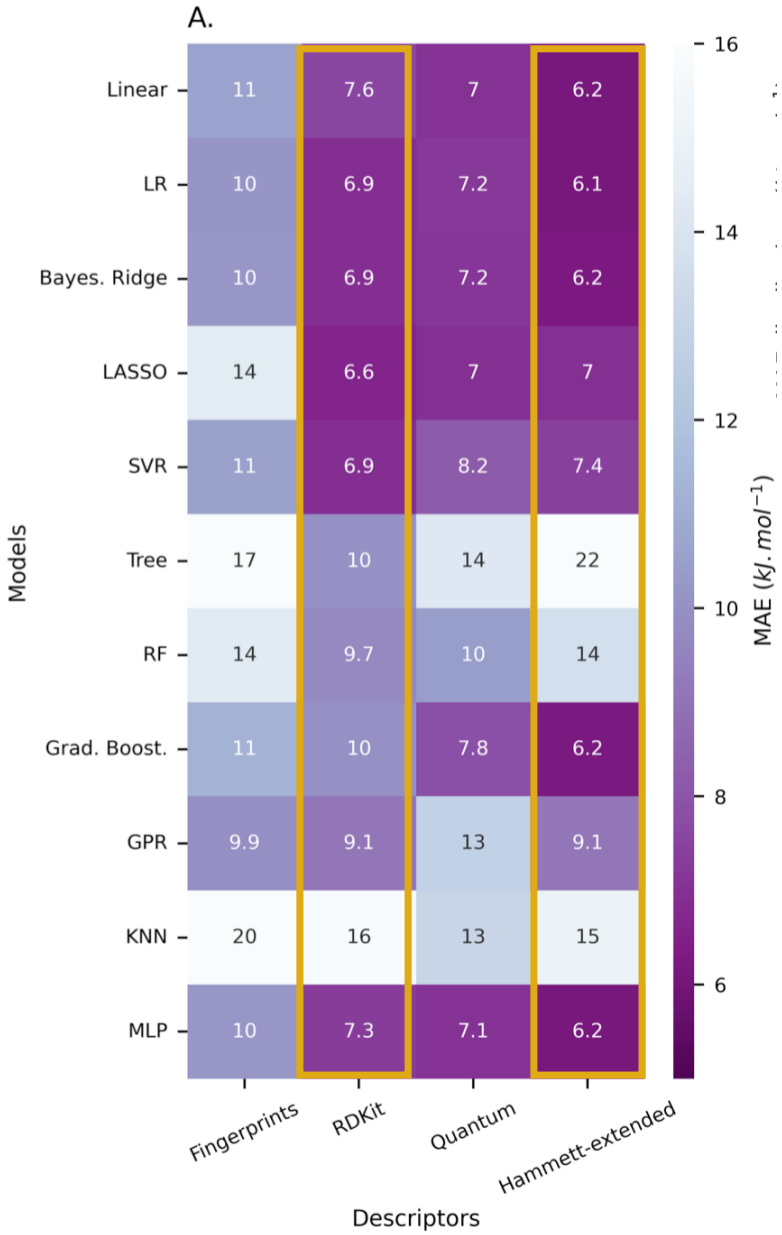


Use k-means + Morgan fingerprint to enhance diversity and chemical space

↓

Get more samples

Constructing models for ONO



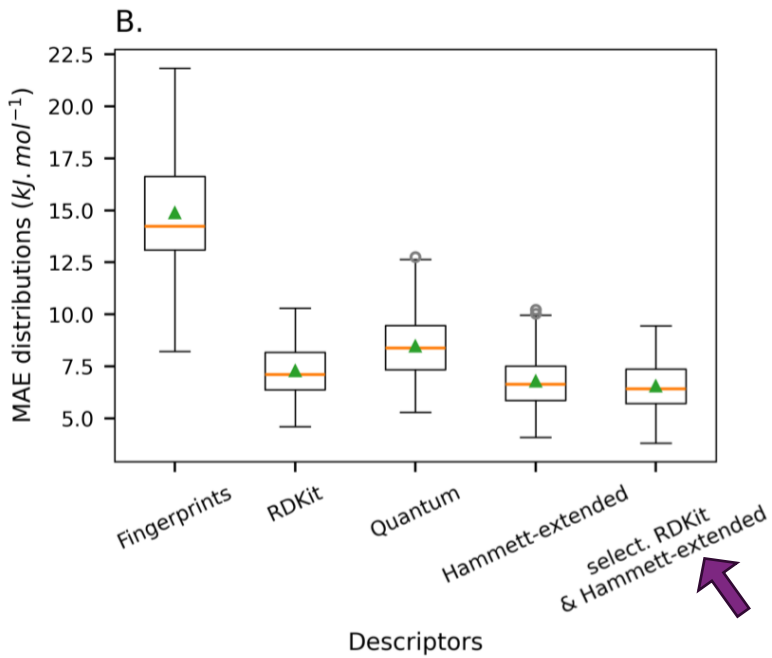
RDKit

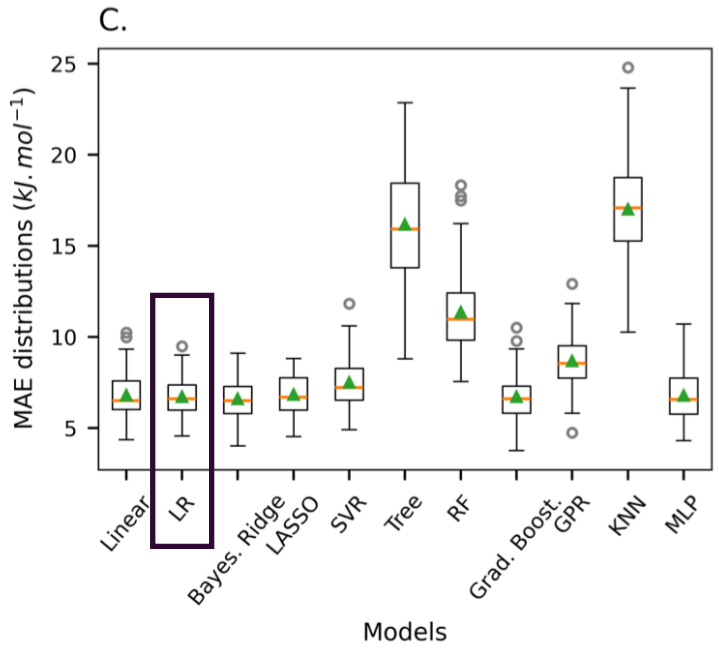
+ F-statistic

Hammett-extended

Mean Absolute Error

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

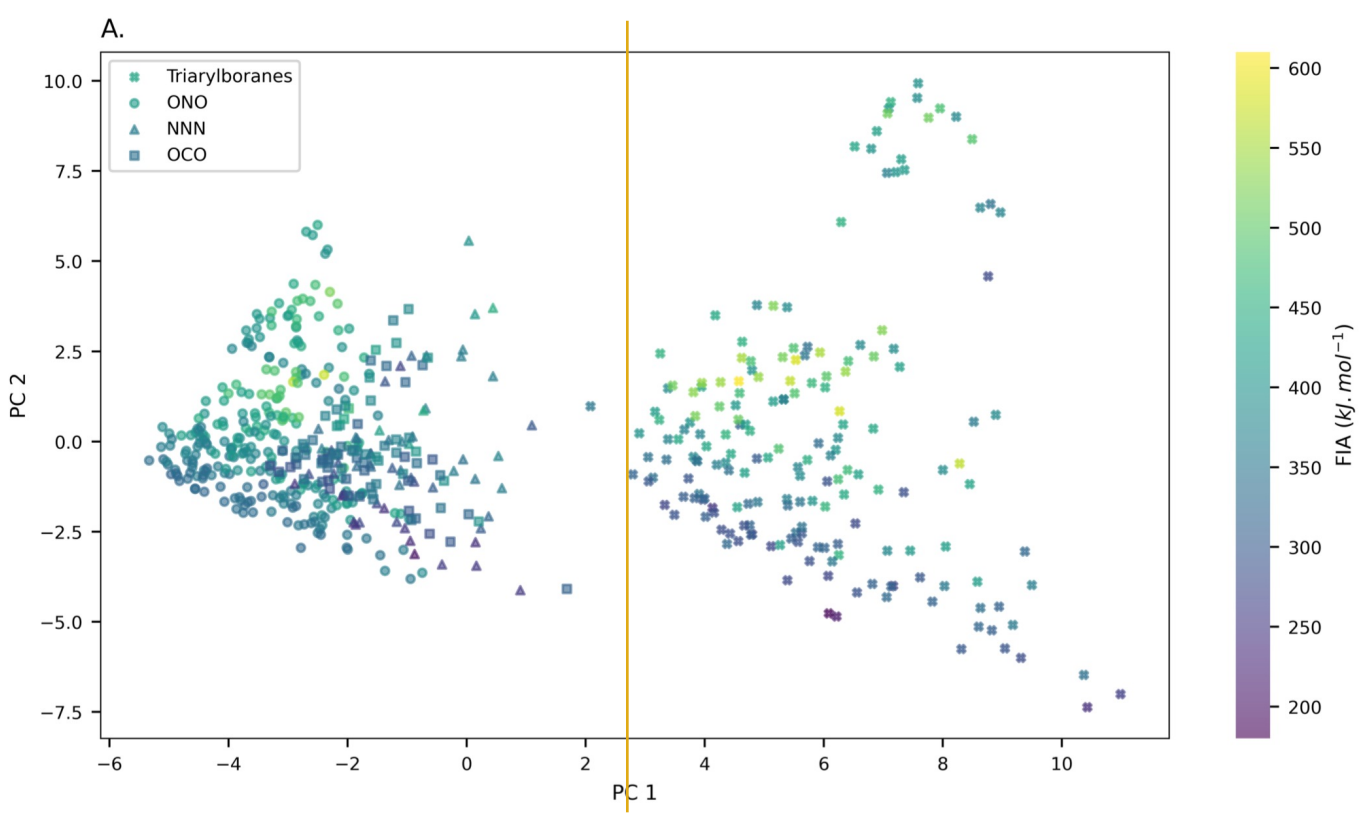


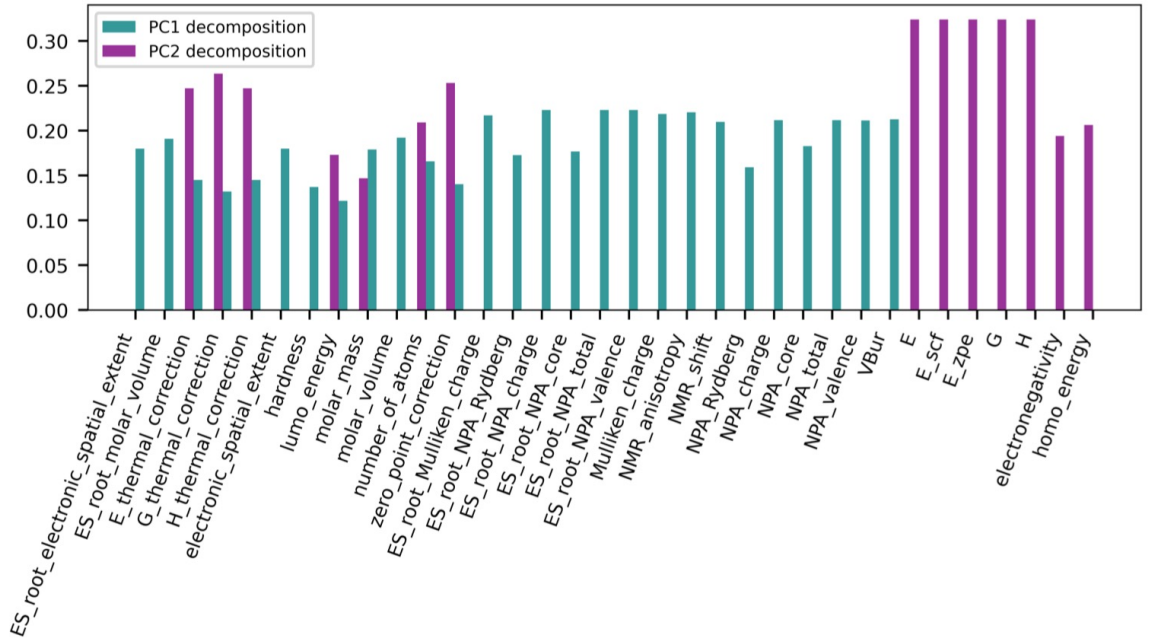


Interpretability

Insights in the Lewis acidity

Principal Component Analysis (PCA) reduces high-dimensional correlated data into a few uncorrelated components, capturing the main variance for easier visualization and analysis.





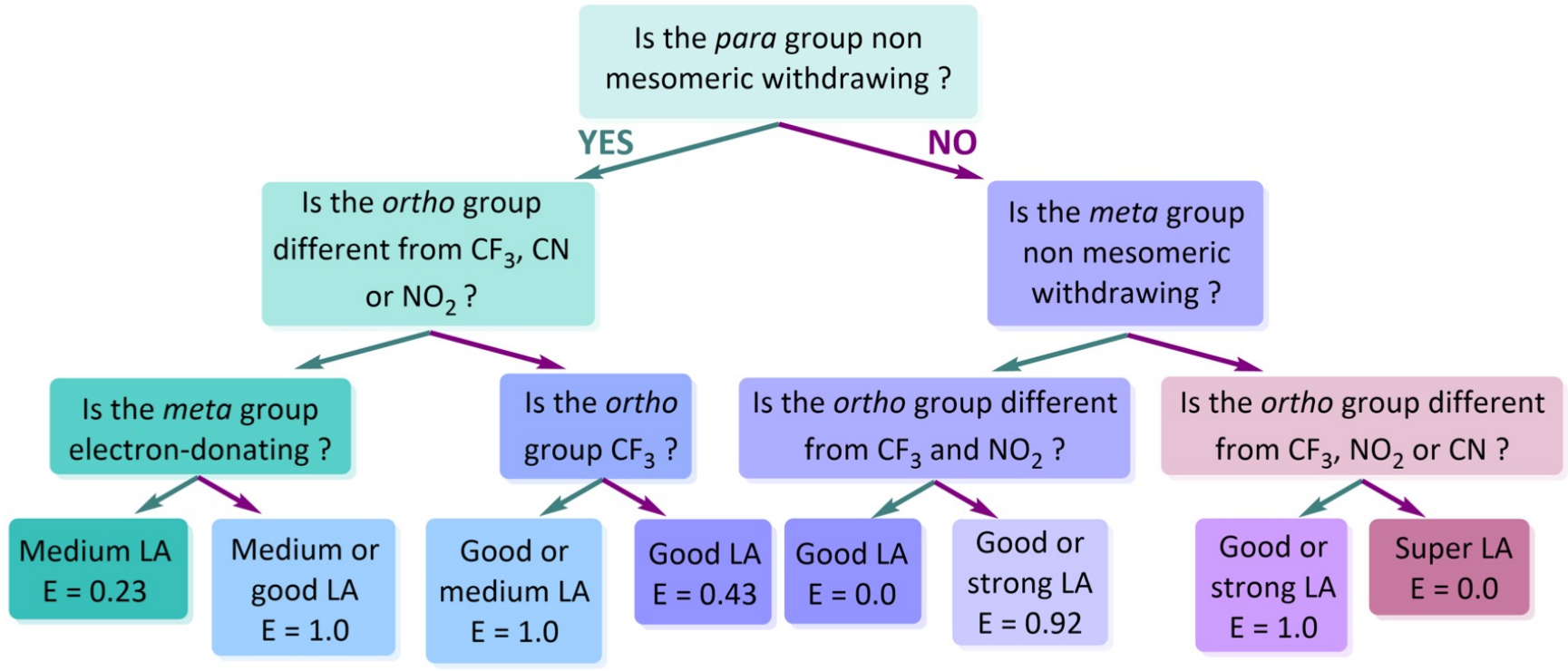
Interpretability

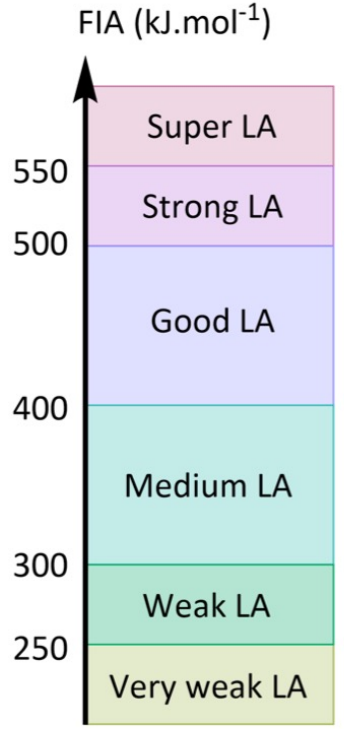
ONO Molecular design

para-σ

meta-σ

ortho-σ





3. Conclusion

Conclusion

The authors combined RDKit descriptors with Hammett-extended descriptors and built a high-performance predictive model, oracle, using linear regression. FIA was used as the index to predict the Lewis acidity of boron compounds, achieving a mean absolute error (MAE) of less than 6 $kJ \cdot mol^{-1}$.

- Clear Process
- Interpretability (PCA)

Thank You