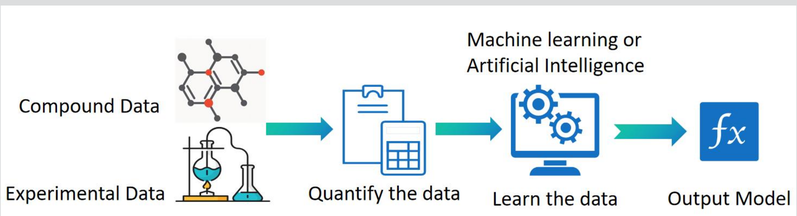
Data-Driven Smart Chemistry：Quantitative Structure Activity Relationship Modeling of Chemical Reactions

Student: 张一超[1]
Mentor: 洪鑫[2]
[1]浙江省杭州第二中学,Hangzhou No.2 High School of Zhejiang Province,Hangzhou,Zhejiang.
[2]浙江大学,Zhejiang University,Hangzhou,Zhejiang.

About Data-driven models

Data driven models, including QSAR (quantitative structure-activity relationship) and QSPR (quantitative structure-activity relationship), are modeling tools based on statistical analysis of chemical data to predict the biological activity and physicochemical properties of substances. In the field of chemistry, they are used for drug design, catalyst discovery, synthesis pathway optimization, etc., to improve reaction efficiency and selectivity, optimize chiral drug synthesis, promote green chemistry and sustainable synthesis, and provide new perspectives for chemical research. I will conduct research in the field of chemical modeling based on this concept.

Taking a review paper published in Nature Reviews Chemistry as an example[1]. This article explores the application of quantitative prediction techniques in the discovery of small molecule chiral catalysts, evaluates the accuracy and reliability of different methods in predicting catalyst performance, and fully demonstrates the important role of data-driven models in the field of chemistry.Similarly, another review paper published in Accounts of Chemical Research[2] explored the application of data-driven models in designing novel catalysts, which have the potential to improve reaction efficiency and selectivity, with a particular focus on the application of data-driven models in chiral synthesis. Both reviews demonstrate the significant applications of data-driven models in the field of chemistry.



How to establish a Data-driven Model

Symbolic Regression

Symbolic Regression (SR) is a machine learning method capable of inferring mathematical expressions from data, revealing the underlying patterns. Compared to traditional numerical regression, SR not only provides predictive models but also offers interpretability, aiding in understanding the relationships between data. SR searches through the space of possible mathematical expressions to find the best fit for the data, with the ability to autonomously select operational symbols and exhibiting good adaptability.

In the field of chemistry, SR technology has a wide range of applications, such as predicting the physical and chemical properties of materials in materials chemistry, guiding the development of new materials and the improvement of existing material performance; analyzing reaction conditions and product data in chemical synthesis pathway optimization, proposing improvements to reduce side reactions, and increase product yield and purity; playing a role in ligand selection, targeted drug design, and more. The SISSO algorithm developed by Professor Ouyang Runhai from Shanghai University has achieved significant results in the field of chemistry, being used for model building and new material prediction for various materials, such as perovskite materials, topological insulators, catalytic materials, superconductors, two-dimensional materials, polymers, etc. SR technology, with its strong interpretive and predictive capabilities, shows great potential in chemical research and applications.[4]

The differences between Linear Regression, Logistic Regression, and Symbolic Regression

For **Logistic Regression**:
$f_{w,b}(x) = \sigma\left(\sum_i w_i x_i + b\right)$
The output value is between 0 and 1.
For **Linear Regression**:
$f_{w,b}(x) = \sum_i w_i x_i + b$
The output value can be any number.
For **Symbolic Regression**:
Discovers the underlying mathematical relationships in the data.
The form of the model is not fixed and can include a variety of mathematical operators and functions.
Outputs can be any mathematical expression that best fits the data, providing both predictive power and interpretability.

Main References

1. Reid, J. P., & Sigman, M. S. (2018). Comparing Quantitative Prediction Methods for the Discovery of Small-Molecule Chiral Catalysts. Nature Reviews Chemistry, 2(10), 290–305. https://doi.org/10.1038/s41570-018-0040-8
2. Crawford, J. M., Kingston, C., Toste, F. D., & Sigman, M. S. (2021). Data Science Meets Physical Organic Chemistry. Accounts of Chemical Research, 54(6), 3136–3148. https://doi.org/10.1021/acs.accounts.1c00285
3. The first experimental data used in this paper comes from: Wang, Y. The Study on the Synthesis Process of Anti-AIDS Drug Efavirenz [D]. Zhejiang University of Technology, 2019.
4. Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M., & Ghiringhelli, L. M. (2018). SISSO: A compressed sensing approach to identify optimal low-dimensional descriptors from a large pool of candidates. Physical Review Materials, 2, 083802. DOI: 10.1103/physrevmaterials.2.083802
5. Zi-Jing Zhang, Matthias M. Simon, Shuang Yu, Shu-Wen Li, Xinran Chen, Silvia Cattani, Xin Hong, and Lutz Ackermann, "Nickel-Catalyzed Atroposelective C−H Alkylation Enabled by Bimetallic Catalysis with Air-Stable Heteroatom-Substituted Secondary Phosphine Oxide Preligands," J. Am. Chem. Soc., DOI: https://doi.org/10.1021/jacs.3c14600.
6. R. Ouyang et al.， Phys. Rev. Mater. 2，083802 （2018）

Key URLs

SISSO：https://github.com/rouyang2017/SISSO
SPSS：https://www.ibm.com/spss
The second experimental data used comes from：https://github.com/zju-ys/Nickel-MLR

Multiple Linear Regression Modeling for the Synthesis of Efavirenz

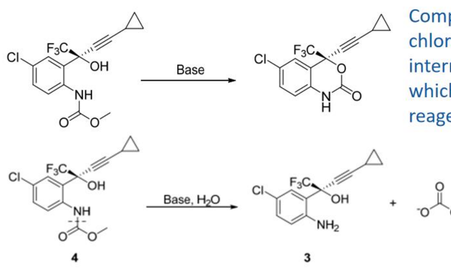
This study employed a multiple linear regression model to analyze the impact of various basic catalysts on the asymmetric amide formation reaction in the synthesis process of Efavirenz. Quantitative property data of the catalysts were retrieved from the RDKit database, and these, along with the mass and concentration of the catalysts, were used as independent variables. The reaction's conversion rate, the content of by-products, and selectivity served as dependent variables to construct the regression model. The results indicated a high degree of model fit, suggesting that the model could effectively explain the variations in the dependent variables. Given the small sample size, caution is needed regarding the risk of overfitting, but the study still demonstrated the capability of linear regression technology to accurately describe chemical reactions.

Multiple Linear Regression Model Equations

ConversionRate = -1.5825 * MolecularWeight + 10.4653 * TPSA - 5.285e-14 * NumRotatableBonds + 42.8935 * HeavyAtomCount - 168.7190 * NumHeteroatoms + 25.7903 * FractionCSP3 - 31.2300 * LogP + 0.1000

Compound = -0.0289 * MolecularWeight + 0.1827 * TPSA - 6.321e-16 * NumRotatableBonds + 0.5051 * HeavyAtomCount - 1.9183 * NumHeteroatoms + 0.0355 * FractionCSP3 - 0.0033 * LogP + 0.0700

Selectivity = -1.5635 * MolecularWeight + 10.2981 * TPSA - 5.236e-14 * NumRotatableBonds + 42.6172 * HeavyAtomCount - 167.1690 * NumHeteroatoms + 25.6843 * FractionCSP3 - 31.2956 * LogP + 2.915e-14



Compound 3 reacts with methyl chloroformate to form intermediate amide compound 4, which is then catalyzed by a basic reagent to produce Efavirenz (1).

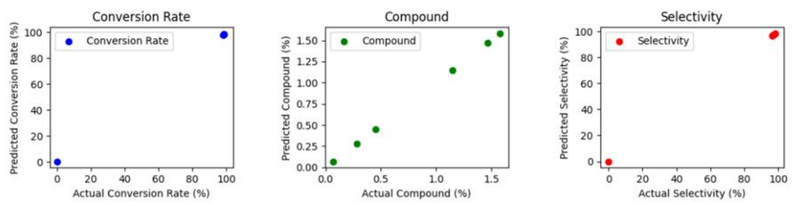
During the process of generating the target product 1 from the intermediate, side reactions occur, leading to the formation of ineffective product 3.

The basic catalytic process of Efavirenz

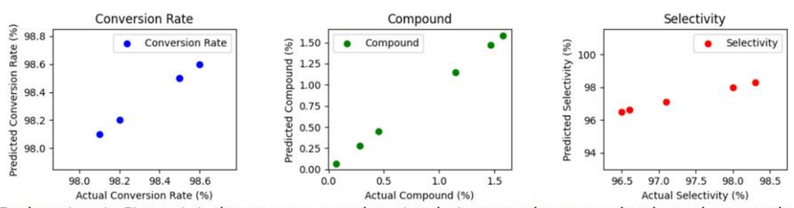
We integrated experimental data[3] and quantitative property data to construct multiple linear regression models using Python and the ordinary least squares function from the statsmodels library. By adding a constant term to the independent variables and analyzing three dependent variables, we obtained models with high R-squared values, indicating a good fit to the data. Although the small sample size necessitates caution regarding overfitting, it still demonstrates the modeling capability of multiple linear regression for chemical reactions.

Display the model fitting effect through predict-actual plots

Picture 1-1



Picture 1-2



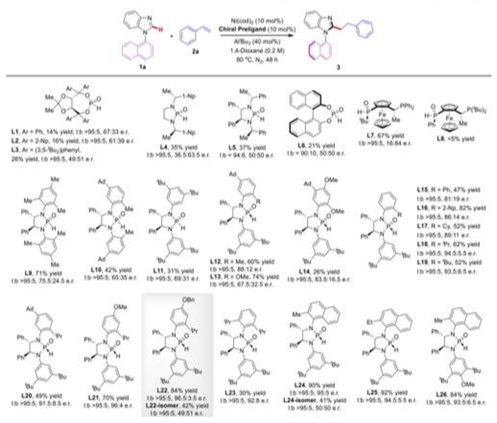
Explanation: In Figure 1-1, due to some sample points being very close to each other, a larger scale is chosen to draw the overall graph; in Figure 1-2, the area with dense samples is zoomed in to more clearly demonstrate the high fitting degree of the multiple linear regression.

These models aid in understanding and predicting the role of different catalysts in the synthesis of Efavirenz, providing guidance for future experimental design and catalyst optimization.

The Enantioselective Model of HASPO Preligand in Nickel-Catalyzed C−H Alkylation Reaction

This research project focuses on the model development for nickel-catalyzed C−H alkylation reactions, particularly the enantioselectivity when using HASPO (heteroatom-substituted secondary phosphine oxide) preligrands, with data support provided by *Professor Xin Hong*'s research group and *Yu Shuang*. The study compares the strengths and weaknesses of linear regression and symbolic regression techniques through the model-building process. Based on the paper published by *Zi-Jing Zhang* et al. in the *Journal of the American Chemical Society*, the paper[5] explores efficient C−H alkylation reactions facilitated by a Ni−Al bimetallic catalyst and HASPO ligands for the construction of C−N axially chiral compounds. The research team optimized reaction conditions and identified HASPO ligands with high enantioselectivity, significantly enhancing the reaction's yield and stereoselectivity.

The importance of HASPO ligands in nickel-catalyzed atroposelective C−H alkylation reactions

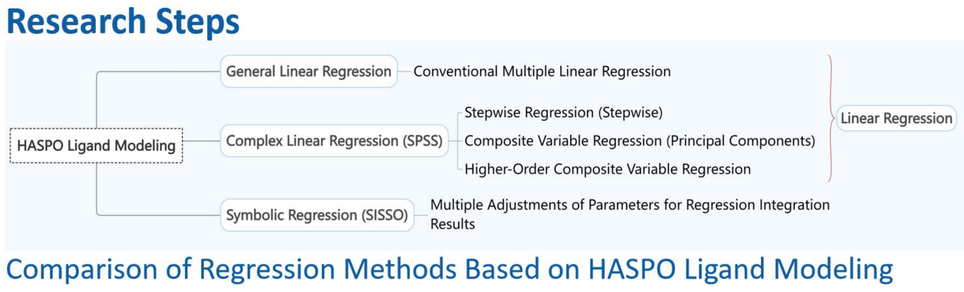


The HASPO ligand is pivotal in nickel-catalyzed atroposelective C−H alkylation, ensuring high enantioselectivity by providing a chiral environment. It forms complexes with nickel, influencing catalytic properties through coordination. The ligand's stereochemistry is crucial for reaction stereocontrol, and its structure aids in stabilizing substrates and intermediates, enhancing selectivity. HASPO ligands also participate in the catalyst's formation and product release, and their optimization can boost reaction efficiency and yield. They enable mild reaction conditions, minimizing side reactions and increasing the selectivity of the target product, thus offering an effective route for synthesizing axially chiral compounds with application potential.

The image used in here is from the paper cited as reference 5

This study outlines the steps for modeling HASPO ligands, beginning with general linear regression to analyze the linear relationship between one or more independent variables and the dependent variable. Subsequently, complex linear regression is conducted using SPSS software, which includes stepwise regression to select significant variables, composite variable regression (principal components) to reduce data dimensions and enhance model interpretability, and higher-order composite variable regression to account for the higher-order terms and interactions of variables, capturing more complex relationships. Then, symbolic regression is applied using the SISSO software, a flexible method that automatically discovers the underlying mathematical relationships in the data. This process involves adjusting parameters multiple times to integrate and optimize the model. Finally, a comparison of the performance of linear and symbolic regression is conducted to evaluate their applicability and effectiveness in modeling HASPO ligands, thereby providing a deeper understanding and a superior strategy for chemical modeling.

Research Steps

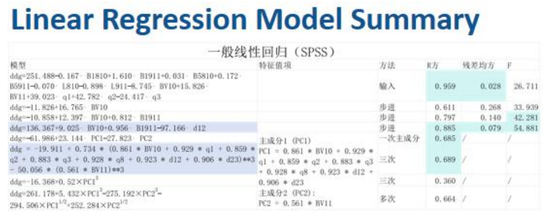


Comparison of Regression Methods Based on HASPO Ligand Modeling

By comparing the model-building processes and results of both types of models, we have significantly identified the strengths and suitable scenarios for both linear and symbolic regression methods. Linear regression is suitable for simple linear relationship modeling, easy to operate, but struggles to accurately fit complex relationships; symbolic regression, on the other hand, has an advantage in handling nonlinear relationships with better fitting effects but requires more computational power and a more complex modeling process.

In this study, symbolic regression modeling was conducted using the SISSO tool developed by *Professor Runhai Ouyang*.[6]

Linear Regression Model Summary



Symbolic Regression Model Summary

*Due to the limited space on the poster, it is not possible to clearly display the summary of the model data. I would highly recommend **scanning the QR code in the top right corner of the poster to read my full report for complete data.***

By comparing the model-building processes and results of both types of models, we have significantly identified the strengths and suitable scenarios for both linear and symbolic regression methods. Linear regression is suitable for simple linear relationship modeling, easy to operate, but struggles to accurately fit complex relationships; symbolic regression, on the other hand, has an advantage in handling nonlinear relationships with better fitting effects but requires more computational power and a more complex modeling process.

Advantages and Disadvantages of Linear Regression and Symbolic Regression in Chemical Reaction Modeling

Tech	Advantages	Disadvantages
Linear Regression	Easy to operate, convenient to use	Dealing with overfitting issues is troublesome, and it is not effective for modeling complex relationships (inefficient in modeling non-linear relationships).
Symbolic Regression	Automatically constructs models with a variety of forms Efficient in handling non-linear relationships, with broad application potential in complex multivariate fields	There is a certain level of operational difficulty, and it requires significant computational power.