Youth Talent Program
(2024 Annual Chemistry Group)
Research Report

Topic: Quantitative Structure-Activity Relationship Modeling of Chemical Reactions

Trainee: **Zhang Yichao** (Hangzhou No.2 High School of Zhejiang Province)

Supervisor:**Hong Xin** (Zhejiang University)

November 2024

Contents

Training Overview

Trainee Introduction

Zhang Yichao is a student at Hangzhou No.2 High School of Zhejiang Province.

In January 2024, he was selected as a participant of the Middle School Student Talent Program in Chemistry (only 13 students province-wide), under the supervision of Professor Hong Xin from the Department of Chemistry, Zhejiang University.

In July 2024, he participated in the "Talent Program" Summer Camp organized by Chu Kochen Honors College, Zhejiang University, and was awarded the title of "Outstanding Award of Trainees" (only 2 students in Chemistry).

On November 3, 2024, he delivered a research report at the Department of Chemistry, Zhejiang University, and represented the Chemistry Group in the cross-disciplinary research defense of the five subjects under Zhejiang University's Talent Program.

On November 29, 2024, he participated as a representative of Zhejiang Province (only 2 students in Chemistry) in the Chemistry Forum of the High School Student Talent Program held at the University of Science and Technology of China.

Research Topic Introduction

Centered on the research topic "Quantitative Structure-Activity Relationship Modeling of Chemical Reactions", this project integrates informatics, statistics, and other interdisciplinary foundations, comprehensively applying Machine Learning (ML) for chemical learning and research.

Key Research Directions

Interdisciplinary application in chemistry;
Application of machine learning in the field of chemistry.

Project Composition

- Literature Study Project 1: **Data-driven models for quantitative prediction of chemical reactions**
- Practical Application Project 2: **Construction and Analysis of a Multiple Linear Regression Model for the Impact of Basic Catalyst Types on the Asymmetric Amidation Reaction in the Synthesis Process of Efavirenz**
- Literature Study Project 3: **Mechanism of Symbolic Regression (SR) and Its Applications in the Chemical Field**
- Practical Application Project 4: **Enantioselectivity model of HASPO preligands in nickel-catalyzed C–H alkylation reactions**

Core Competencies Developed

- Interdisciplinary Learning Capability: Cultivate the ability to integrate and apply cross-disciplinary knowledge, with an emphasis on multi-field knowledge application.
- Comprehensive Chemical Competence:Master basic literature reading skills;Gain insights into cutting-edge chemical developments through multiple frontier lectures on chemistry organized by Zhejiang University;Develop basic chemical experimental skills through participation in several fundamental experimental activities during the study at Zhejiang University.
- Cheminformatics Operation Capability:Learn the basic principles of Machine Learning (ML);Master practical applications of cheminformatics tools such as Python, SPSS, and SISSO, as well as information technology including Symbolic

Regression.

- Statistical Modeling Capability:Grasp the basic operational concepts of QSAR (Quantitative Structure-Activity Relationship) and QSPR (Quantitative Structure-Property Relationship);Study and apply the modeling, validation, and application of various linear and nonlinear models;Attempt linear modeling methods such as principal component regression and stepwise regression, as well as Symbolic Regression technology based on SISSO.

Literature Study Project 1

Data-driven models for quantitative prediction of chemical reactions

Yichao Zhang,Hangzhou No.2 High School of Zhejiang Province

Data-driven models are playing an increasingly prominent role in the quantitative prediction of chemical reactions, especially in fields such as the discovery of novel chiral catalysts, understanding of reaction mechanisms, and design of new molecular materials, demonstrating strong application value. These models are typically constructed based on the principles of Quantitative Structure-Activity Relationship (QSAR) and Quantitative Structure-Property Relationship (QSPR).

QSAR (Quantitative Structure-Activity Relationship) is a modeling concept used to predict the biological activity or reaction activity of chemical substances. It correlates the structural characteristics of molecules (such as atomic number, functional groups, stereochemistry, etc.) with their biological or reaction activities through statistical methods. QSAR models are commonly applied in areas including drug design, agrochemicals, and risk assessment of environmental pollutants.

QSPR (Quantitative Structure-Property Relationship) is similar to QSAR but focuses on the physicochemical properties of chemical substances, such as solubility, boiling point, and reaction rate. QSPR models help understand the relationship between molecular structures and the physicochemical properties they exhibit. They possess the potential to feed back the integration of experimental data into the development and improvement of theories, playing a crucial role in gaining a deeper understanding of substances.

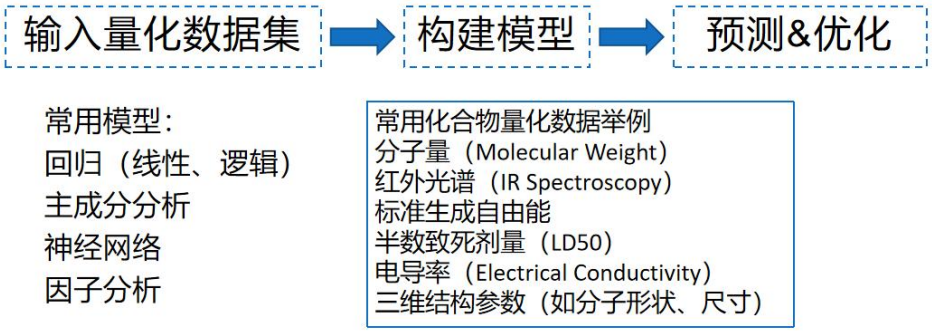


Figure 1 Preliminary Framework for Data-Driven Model Development

Data-driven models play a crucial role in the field of chemistry. By analyzing large volumes of chemical data, they predict and design new drugs, construct models for

drug toxicity or efficacy, thereby accelerating the drug discovery process and enhancing its safety. In the prediction of asymmetric reaction catalysts, these models can process quantitative data on selectivity and catalysts, identify key features influencing reaction selectivity, further optimize catalyst design, and improve the stereoselectivity of reactions. Additionally, data-driven models can plan synthetic reaction environments: by inputting quantitative characteristics of synthetic efficiency and the environment, they calculate the most efficient environmental conditions to achieve green chemistry and sustainable synthesis. In terms of ligand feature optimization, models can guide the adjustment of ligand structures and optimize enantioselectivity, which is particularly important for the synthesis of chiral drugs. Mechanistic research on physicochemical properties also benefits from data-driven models — they can refine and improve existing theories through modeling, providing deeper mechanistic insights.

Take a review article published in Nature Reviews Chemistry as an example (Reid, J. P., & Sigman, M. S. (2018). Comparing Quantitative Prediction Methods for the Discovery of Small-Molecule Chiral Catalysts.). This paper 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 chemistry.

Similarly, another review article published in Accounts of Chemical Research (Crawford, J. M., Kingston, C., Toste, F. D., & Sigman, M. S. (2021). Data Science Meets Physical Organic Chemistry.) discusses the application of data-driven models in designing novel catalysts, highlighting their role in improving reaction efficiency and selectivity, with a specific focus on their application in chiral synthesis.

Data-driven models have become an indispensable tool in chemical research. They have not only demonstrated tremendous potential in fields such as drug design, catalyst discovery, and synthetic route optimization but also provided new perspectives in physical chemistry research. With the continuous advancement of algorithms, data-driven models will play an even more important role in chemical research and applications, constantly advancing the informatization development of chemistry.

References

1.Kulik, H. J., & Sigman, M. S. (2021). Advancing Discovery in Chemistry with Artificial Intelligence: From Reaction Outcomes to New Materials and Catalysts. Accounts of Chemical Research, 54(5), 2335–2336. https://doi.org/10.1021/acs.accounts.1c00232
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.Robinson, S. G., & Sigman, M. S. (2020). Integrating Electrochemical and Statistical Analysis Tools for Molecular Design and Mechanistic Understanding. Accounts of Chemical Research, 53(2), 289–299. https://doi.org/10.1021/acs.accounts.9b00527
4.江辰，尤田耙，等. (2006). 手性领域的定量构效关系研究. 中国科学技术大学.
5.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
6.Williams, W. L., Zeng, L., Gensch, T., Sigman, M. S., Doyle, A. G., & Anslyn, E. V. (2021). The Evolution of Data-Driven Modeling in Organic Chemistry. ACS Central Science, 7(7), 1622–1637. https://doi.org/10.1021/acscentsci.1c00535

Practical Application Project 2

Construction and Analysis of a Multiple Linear Regression Model for the Impact of Basic Catalyst Types on the Asymmetric Amidation Reaction in the Synthesis Process of Efavirenz

I. Project Operation Approach

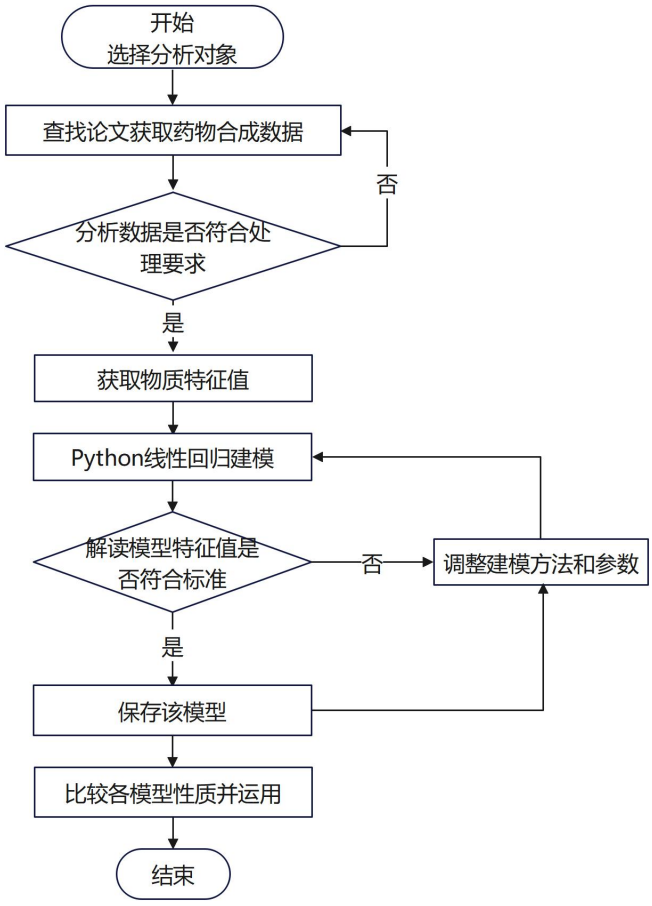


Figure 2 Schematic Diagram of the Project Operation Approach

II. Project Report

I. Efavirenz

Efavirenz is an antiretroviral drug used for the treatment of human immunodeficiency virus (HIV) infection. As a non-nucleoside reverse transcriptase inhibitor (NNRTI), it binds non-competitively to HIV-1 reverse transcriptase, inhibiting DNA synthesis during viral replication and thereby slowing down viral proliferation in the body.[18][24][25]

II. Asymmetric Synthesis Method of Efavirenz

Based on the synthetic scheme design of the drug described in the article Study on the Synthetic Process of Anti-AIDS Drug Efavirenz[18], a pathway for the asymmetric synthesis of efavirenz has been established.

1. Starting Material: 2-trifluoroacetyl-4-chloroaniline is used as the starting material.

2. Preparation of Grignard Reagent: n-butylmagnesium chloride Grignard reagent is prepared by the reaction of chlorobutane with metallic magnesium.

3. Asymmetric Addition: Under the action of sodium hydride, zinc chloride, trifluoroethanol, and the chiral ligand (1R,2S)-1-phenyl-2-(1-pyrrolidinyl)-1-propanol (Lig A), a reaction is carried out with cyclopropylethynylmagnesium chloride to form a zinc-centered coordination compound. Subsequent asymmetric addition with the starting material yields the key intermediate (S)-1-(2-amino-5-chlorophenyl)-1-trifluoromethyl-3-cyclopropylprop-2-yn-1-ol.

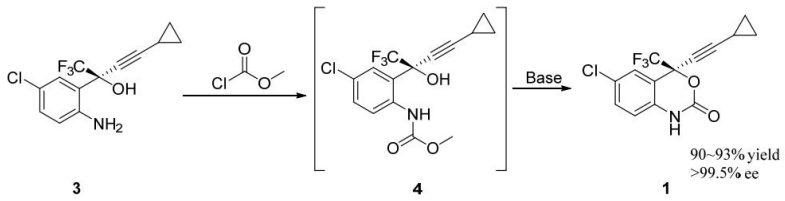
4. Amidation Condensation: The chiral intermediate obtained above reacts with methyl chloroformate to form an intermediate amide compound.

5. Cyclization Reaction: The target product efavirenz is generated via cyclization reaction catalyzed by potassium tert-butoxide.

6. Purification: Through subsequent steps of washing, dehydration, decolorization, and crystallization, high-yield and high-purity efavirenz is finally obtained.

Asymmetric Synthesis Method of Efavirenz[11][18]

After Compound 3 reacts with methyl chloroformate, it generates intermediate amide Compound 4. Following dehydration treatment, Compound 4 undergoes cyclization under the catalysis of a basic reagent to form the target product 1 (Efavirenz).

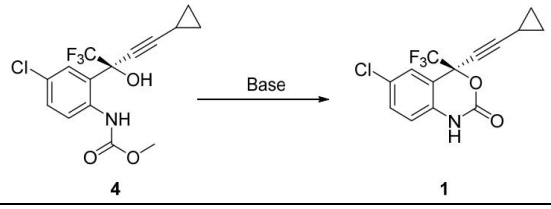


3 4 1

III. Process of the Amidation Condensation Reaction

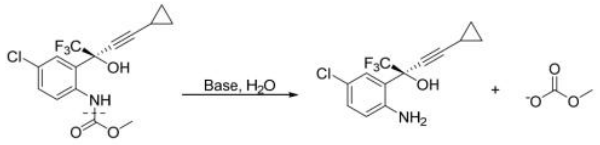
The Reaction Process of Amidation Condensation[18][21][24]

Main Reactions



4 1

The side reaction produces Compound 3.



4 3

IV. The Impact of Different Basic Catalysts on the Reaction

Table: Effects of Different Basic Catalysts on the Reaction[18]

No.	Catalyst	Conversion Rate① /%	Compound 3 /%	Selectivity② /%
1	无	0.1	0.07	0
2	NaOH	98.2	1.58	96.6
3	50% NaOH 溶液	96.6	5.62	90.2
4	Na2CO3	98.1	1.47	96.5
5	KOH	98.5	1.15	97.1
6	NaOCH3	98.5	0.45	98.0
7	t-BuOK	98.6	0.28	98.3

① Conversion rate of Compound 4; ② Selectivity of Efavirenz.

Note: 210 g (0.73 mol) of Compound 3 was charged as the feedstock. After the completion of vacuum reflux water separation (see Section 3.3.2 for detailed experimental procedures), the solution was evenly divided into 7 portions. To each portion, 1.2 g of the corresponding basic catalyst for each experimental group was added (2.4 g of 50% NaOH solution was added for the relevant group). The reaction mixture was heated to 45~50℃ and reacted for 4 hours.

V. Acquisition of Quantitative Data on the Properties of Basic Catalysts

Through the RDKit database, various quantitative property data of different basic catalysts were obtained, including Molecular Weight, Topological Polar Surface Area

(TPSA), Number of Rotatable Bonds (NumRotatableBonds), Heavy Atom Count (HeavyAtomCount), Number of Heteroatoms (NumHeteroatoms), Fraction of sp^3-Hybridized Carbon Atoms (FractionCSP3), and Lipophilicity (LogP).

Main Code (Partial Code Omitted)

```
import pandas as pd
from rdkit import Chem
from rdkit.Chem import Descriptors
from rdkit.Chem.Crippen import MolLogP
compounds_smiles = {
    "NaOH": " [Na+].[OH-]",
    "Na2CO3": "[Na+].[Na+].[O-]C(=O)O",
    "KOH": "[K+].[OH-]",
    "NaOCH3": "[Na+].[O-]C",
    "t-BuOK": "CC(C)(C)[O-].[K+]"
}
```

Through the above code, the property data of different basic catalysts can be obtained using Python. Since the concentrations of NaOH in Experimental Group 2 and Experimental Group 3 are inconsistent, both the mass and concentration of the added catalyst are set as independent variables for consideration.

Catalyst	**Mass (g)**	**Concentration**	**Molecular Weight (g/mol)**	**Topological Polar Surface Area (Å²)**	**Number of Rotatable Bonds**	**Heavy Atom Count**	**Number of Heteroatoms**	**Fraction of sp^3-Hybridized Carbon Atoms**	**Lipophilicity (LogP)**
None	0	0	0	0	0	0	0	0	0
NaOH	1.2	1	39.997	30	0	2	2	0	-3.1728
NaOH	2.4	0.5	39.997	30	0	2	2	0	-3.1728
Na_2CO_3	1.2	1	106.996	60.36	0	6	5	0	-7.1043
KOH	1.2	1	56.105	30	0	2	2	0	-3.1728
$NaOCH_3$	1.2	1	54.024	23.06	0	3	2	1	-4.0195
t-BuOK	1.2	1	112.213	23.06	0	6	2	1	-2.8508

VI. Construction of the Multiple Linear Regression Model

Using Python, multiple linear regression analysis was performed with "Conversion Rate", "Compound 3" (content), and "Selectivity" as the dependent variables, and the quantitative property data of basic catalysts as the independent variables, respectively.

Main Code for Model Construction (Partial Code Omitted)

```
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
data = {
    'Compound': ['无', 'NaOH', 'Na2CO3', 'KOH', 'NaOCH3', 't-BuOK'],
    'MolecularWeight': [0, 39.997, 106.996, 56.105, 54.024, 112.213],
    'TPSA': [0, 30, 60.36, 30, 23.06, 23.06],
    'NumRotatableBonds': [0, 0, 0, 0, 0, 0],
    'HeavyAtomCount': [0, 2, 6, 2, 3, 6],
    'NumHeteroatoms': [0, 2, 5, 2, 2, 2],
    'FractionCSP3': [0, 0, 0, 1, 1, 1],
    'LogP': [0, -3.1728, -7.1043, -3.1728, -4.0195, -2.8508],
    'ConversionRate': [0.1, 98.2, 98.1, 98.5, 98.5, 98.6],
    'Compound': [0.07, 1.58, 1.47, 1.15, 0.45, 0.28],
    'Selectivity': [0, 96.6, 96.5, 97.1, 98, 98.3]
}
df = pd.DataFrame(data)
X = df[['MolecularWeight', 'TPSA', 'NumRotatableBonds', 'HeavyAtomCount', 'NumHeteroatoms', 'FractionCSP3', 'LogP']]
y_conversion = df['ConversionRate']
y_compound = df['Compound']
y_selectivity = df['Selectivity']
X = sm.add_constant(X)
model_conversion = sm.OLS(y_conversion, X).fit()
model_compound = sm.OLS(y_compound, X).fit()
model_selectivity = sm.OLS(y_selectivity, X).fit()
print('Conversion Rate Regression Results:')
print(model_conversion.summary())
print('\nCompound Regression Results:')
print(model_compound.summary())
print('\nSelectivity Regression Results:')
print(model_selectivity.summary())
```

VII. Results of Multiple Linear Regression

Conversion Rate Regression Results

OLS Regression Results

Dep. Variable:	ConversionRate	R-squared:	1.000
Model:	OLS	Adj. R-squared:	nan
Method:	Least Squares	F-statistic:	nan
Date:	Tue, 26 Mar 2024	Prob (F-statistic):	nan
Time:	20:34:23	Log-Likelihood:	167.07
No. Observations:	6	AIC:	-322.1
Df Residuals:	0	BIC:	-323.4
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1000	inf	0	nan	nan	nan
MolecularWeight	-1.5825	inf	-0	nan	nan	nan
TPSA	10.4653	inf	0	nan	nan	nan
NumRotatableBonds	-5.285e-14	inf	-0	nan	nan	nan
HeavyAtomCount	42.8935	inf	0	nan	nan	nan
NumHeteroatoms	-168.7190	inf	-0	nan	nan	nan
FractionCSP3	25.7903	inf	0	nan	nan	nan
LogP	-31.2300	inf	-0	nan	nan	nan

Omnibus:	nan	Durbin-Watson:	2.728
Prob(Omnibus):	nan	Jarque-Bera (JB):	1.035
Skew:	1.006	Prob(JB):	0.596
Kurtosis:	3.305	Cond. No.	834.

Compound Regression Results

OLS Regression Results

Dep. Variable:	Compound	R-squared:	1.000
Model:	OLS	Adj. R-squared:	nan
Method:	Least Squares	F-statistic:	nan
Date:	Tue, 26 Mar 2024	Prob (F-statistic):	nan
Time:	20:34:24	Log-Likelihood:	193.55
No. Observations:	6	AIC:	-375.1
Df Residuals:	0	BIC:	-376.4
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0700	inf	0	nan	nan	nan
MolecularWeight	-0.0289	inf	-0	nan	nan	nan
TPSA	0.1827	inf	0	nan	nan	nan
NumRotatableBonds	-6.321e-16	inf	-0	nan	nan	nan
HeavyAtomCount	0.5051	inf	0	nan	nan	nan
NumHeteroatoms	-1.9183	inf	-0	nan	nan	nan
FractionCSP3	0.0355	inf	0	nan	nan	nan
LogP	-0.0033	inf	-0	nan	nan	nan

Omnibus:	nan	Durbin-Watson:	2.286
Prob(Omnibus):	nan	Jarque-Bera (JB):	1.220
Skew:	-1.103	Prob(JB):	0.543
Kurtosis:	3.107	Cond. No.	834.

Selectivity Regression Results

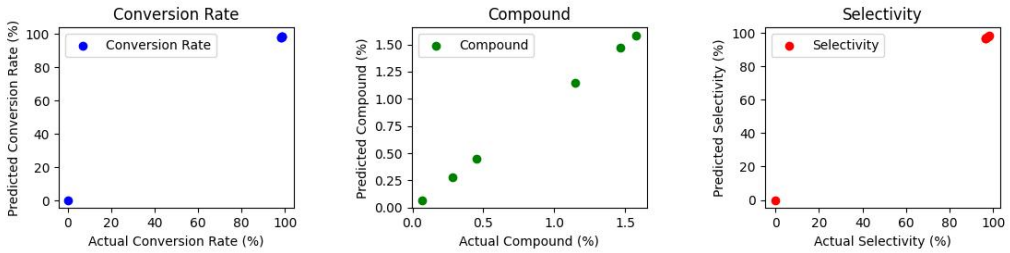
OLS Regression Results

Dep. Variable:	Selectivity	R-squared:	1.000
Model:	OLS	Adj. R-squared:	nan
Method:	Least Squares	F-statistic:	nan
Date:	Tue, 26 Mar 2024	Prob (F-statistic):	nan
Time:	20:34:24	Log-Likelihood:	167.62
No. Observations:	6	AIC:	-323.2
Df Residuals:	0	BIC:	-324.5
Df Model:	5		
Covariance Type:	nonrobust		

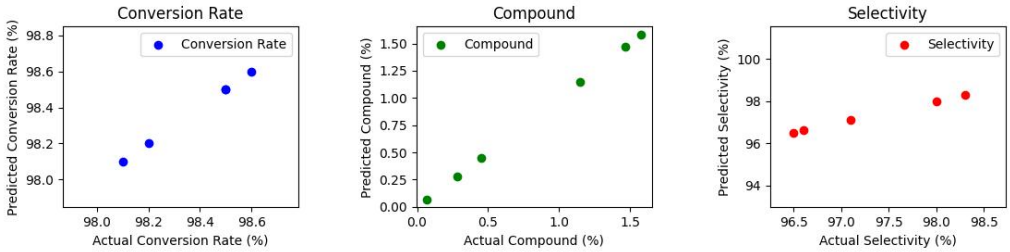
	coef	std err	t	P>|t|	[0.025	0.975]
const	2.915e-14	inf	0	nan	nan	nan
MolecularWeight	-1.5635	inf	-0	nan	nan	nan
TPSA	10.2981	inf	0	nan	nan	nan
NumRotatableBonds	-5.236e-14	inf	-0	nan	nan	nan
HeavyAtomCount	42.6172	inf	0	nan	nan	nan
NumHeteroatoms	-167.1690	inf	-0	nan	nan	nan
FractionCSP3	25.6843	inf	0	nan	nan	nan
LogP	-31.2956	inf	-0	nan	nan	nan

Omnibus:	nan	Durbin-Watson:	2.925
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.034
Skew:	-0.039	Prob(JB):	0.983
Kurtosis:	2.640	Cond. No.	834.

Picture 1-1



Picture 1-2



Note: In Picture 1-1, since some samples are extremely close to each other in position, a larger scale is adopted to draw the global graph. In Picture 1-2, the sample-dense area is enlarged to more clearly demonstrate the high fitting degree of the multiple linear regression.

Multiple Linear Regression Model Equations

1. Conversion Rate Model: 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
2. Compound 3 Model: 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
3. Selectivity Model: 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

VIII. Analysis of Multiple Linear Regression

According to the ordinary least squares (OLS) regression results, the R-squared value is 1.000, which indicates that the model fits the data extremely well and explains 100% of the variance in the dependent variables. This represents an exceptionally high degree of fit, demonstrating that the model can effectively account for changes in the dependent variables. This is highly meaningful for predicting and explaining

variations in the dependent variables. However, despite the presence of warnings and outliers, we can still draw certain conclusions from the regression results. For instance, the R-squared value is very close to 1, signifying an excellent fit of the model to the data. This indicates that our model can well explain the changes in the dependent variables (ConversionRate, Compound 3, Selectivity), which is a positive outcome.[7][8][9]

Due to the small sample size, we need to interpret the significance of the coefficients and confidence intervals with caution. Meanwhile, since the R-squared value is already very close to 1, it may be necessary to consider the possibility of overfitting. Overfitting can lead to a decrease in the model’s predictive power when applied to new data.

In this case, the sample size should be increased to enhance the robustness of the model.

References

[1] 蔡玉磊,田磊,程俊.一种新型不对称合成依非韦伦的方法[J].安徽化工,2022,48(05):44-47+51.

[2] 杨尧.依非韦伦关键中间体的合成工艺研究[D].武汉工程大学,2022.DOI:10.27727/d.cnki.gwhxc.2022.000313.

[3] 李灿,张方方,周毅博等.依非韦伦中间体的不对称合成[J].武汉工程大学学报,2020,42(05):496-500.DOI:10.19843/j.cnki.cn42-1779/tq.201909028.

[4] 王瑜.抗艾滋病药物依非韦伦（Efavirenz）的合成工艺研究[D].浙江工业大学,2019.

[5] 胡争朋.依非韦伦关键中间体的合成研究[D].武汉工程大学,2018.

[6] 胡争朋,吴广文,熊奇等.依非韦伦关键中间体的合成[J].中国医药工业杂志,2018,49(01):49-52.DOI:10.16522/j.cnki.cjph.2018.01.005.

[7] 李运丽.依非韦伦的合成工艺改进[D].郑州大学,2016.

[8] 翟洪.依非韦伦及喹啉衍生物的合成[D].安徽中医药大学,2013.

[9] 江辰.手性领域的定量构效关系研究[D].中国科学技术大学,2006.

[10] 萝卜. Python 实战多元线性回归模型，附带原理+代码. Retrieved from https://blog.csdn.net/csdnsevenn/article/details/107888173

[11] Landrum.RDKit: Open-source cheminformatics. Release 2014.03.1[J].2010.

[12] RDKit: "RDKit: Open-source cheminformatics. n.d. https://www.rdkit.org. Accessed 14 Aug. 2024."

[13] pandas: " McKinney, Wes. "Data Structures for Statistical Computing in Python." Proceedings of the 9th Python in Science Conference, 2010, pp. 51-56."

[14] statsmodels: "Seabold, Skipper, and Josef Perktold. "Statsmodels: Econometric and Statistical Modeling with Python." Proceedings of the 9th Python in Science Conference, 2010, pp. 91-96."

[15] matplotlib: "Hunter, John D. "Matplotlib: A 2D Graphics Environment." Computing in Science & Engineering, vol. 9, no. 3, 2007, pp. 90-95."

Literature Study Project 3

Mechanism of Symbolic Regression (SR) and Its Applications in the Chemical Field

Symbolic Regression (SR) is a powerful machine learning method that can infer mathematical expressions from data and reveal the underlying laws behind it. Unlike traditional numerical regression methods, symbolic regression not only provides predictive models but also offers interpretability, helping us understand the relationships between data.

SR is an approach for automatically discovering mathematical models from data. It searches the space of possible mathematical expressions to find the one that best fits the data. These expressions can be linear or nonlinear, including basic arithmetic operations (addition, subtraction, multiplication, division), exponentials, logarithms, trigonometric functions, and more.

The main difference between symbolic regression and traditional regression methods lies in their core focus: traditional regression methods primarily aim to establish numerical relationships between input and output variables. In contrast, symbolic regression not only seeks numerical correlations but also provides deeper insights and higher interpretability by identifying mathematical relationships.

The primary implementation approaches of symbolic regression include Genetic Programming (GP), neural networks, and Gradient Boosting Decision Trees (GBDT). Each of these methods has its own advantages and exhibits distinct functionalities across different problem types and data characteristics.

There are various tools available for constructing symbolic regression models:

- SISSO (Sure Independence Screening and Sparsity Operator) is a compressed sensing-based algorithm specifically designed to identify optimal low-dimensional descriptors from a large set of candidate descriptors.
- gplearn is a Python-based genetic programming library that supports not only symbolic regression but also classification and feature construction, making it a versatile machine learning tool.
- PySR is an open-source symbolic regression tool that leverages genetic programming to uncover mathematical relationships in data, and it is particularly suitable for scenarios requiring highly interpretable models.

Symbolic regression technology has wide-ranging applications in the chemical field:

- In materials chemistry, it is used to predict the physical and chemical properties of materials, thereby guiding the development of new materials and the improvement of existing material performance.
- In the optimization of chemical synthesis routes, symbolic regression can analyze

reaction conditions and product data across different synthesis steps, proposing improvement measures to reduce side reactions and enhance product yield and purity.

- Additionally, it can play a role in ligand selection, targeted drug design, and other related areas.

Notably, the SISSO algorithm developed by Professor Ouyang Runhai from Shanghai University is not only highly operable but has also achieved significant results in multiple chemical subfields. According to relevant reports, the SISSO method has been applied to model construction and new material prediction in perovskite materials, topological insulators, catalytic materials, superconductors, two-dimensional materials, polymers, and more.

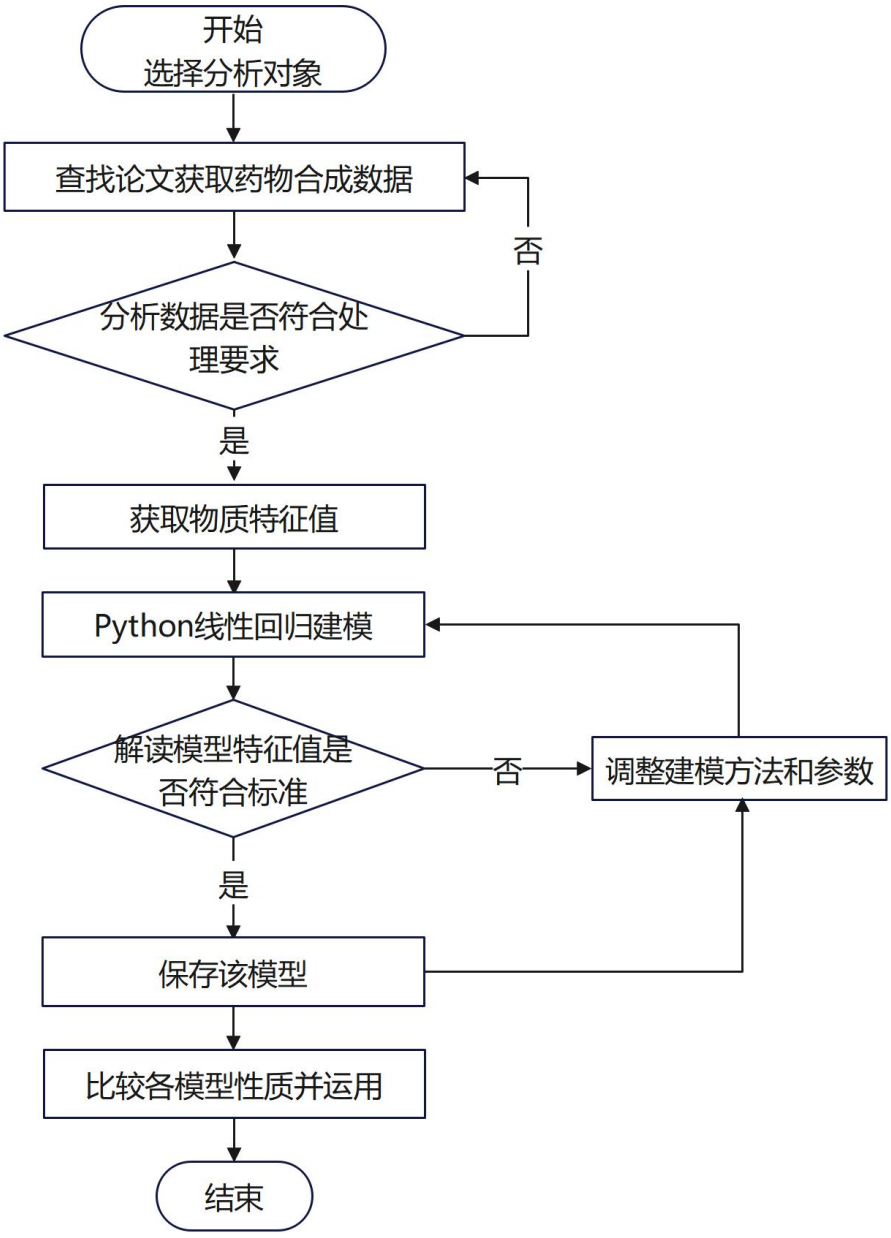
References

1. Purcell, T. A. R., Schäffler, M., & Ghiringhelli, L. M. (2023, May 3). Recent advances in the SISSO method and their implementation in the SISSO++ code (Version 1). arXiv:2305.01242. Retrieved from https://arxiv.org/pdf/2305.01242
2. 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
3. Ouyang, R. (2023, Sep 12). SISSO. SISSO.3.3, July, 2023. https://rouyang2017.github.io/SISSO/
4. Makke, N., & Chawla, S. (2024). Explaining Scientific Discoveries with Symbolic Regression: A Survey. Artificial Intelligence Review. https://doi.org/10.1007/s10462-023-10622-0
5. Lou, S., Liu, C., Chen, Y., & Mo, F. (2024). Empowering Machines to Think Like Chemists: Unveiling Molecular Structure-Polarity Relationships via Hierarchical Symbolic Regression. arXiv:2401.13904.
6. Poli, R. (2008). A Field Guide to Genetic Programming.
7.R. Ouyang, E. Ahmetcik, C. Carbogno, M. Scheffler, and L. M. Ghiringhelli, J. Phys. Mater., in press, https://doi.org/10.1088/2515-7639/ab077b (2019).
8. Shen, Y., Borowski, J. E., Hardy, M. A., Sarpong, R., Doyle, A. G., & Cernak, T. (2021). Automation and computer-assisted planning for chemical synthesis. Nature Reviews Methods Primers, 1, 23. https://doi.org/10.1038/s43586-021-00022-5

Practical Application Project 4

Enantioselectivity Model of HASPO Preligands in Nickel-Catalyzed C−H Alkylation Reactions

I. Project Operation Approach



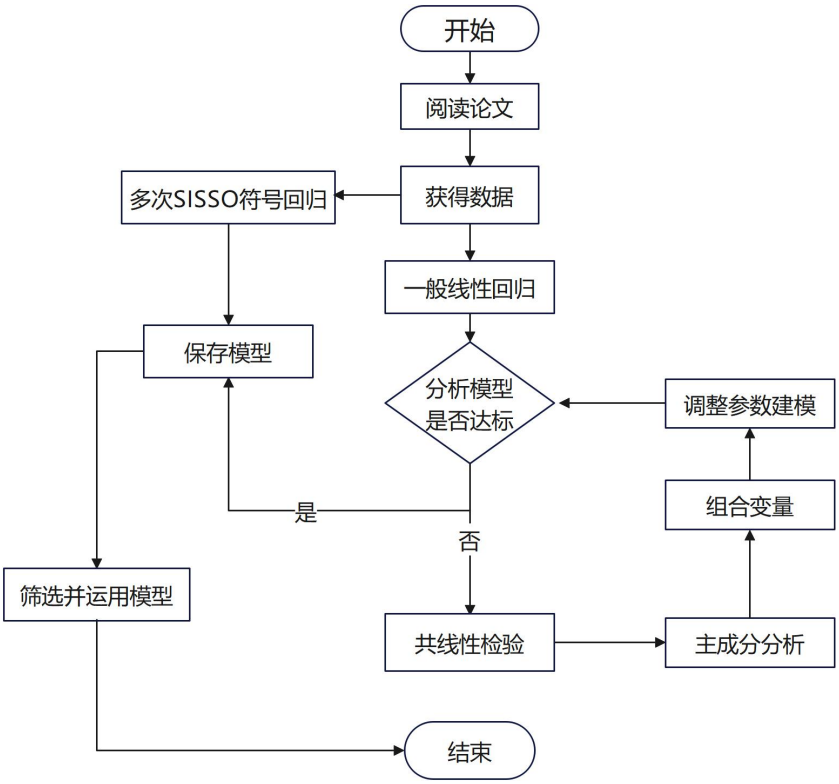


Figure 3 Schematic Diagram of the Project's Operational Framework

II. Project Report

Note

This project is conducted based on the data from the paper: 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. I hereby express my gratitude to Professor Xin Hong and Senior Shuang Yu for their valuable assistance.
Additionally, due to the extensive volume of some data, this report only presents the main parts and partial excerpts. For detailed content and complete data, please refer to the attached documents.

Introduction: Literature Review

This paper introduces an innovative nickel-catalyzed method for the efficient construction of C−N axially chiral compounds via C−H alkylation, enabled by a chiral heteroatom-substituted secondary phosphine oxide (HASPO) ligand-assisted Ni − Al bimetallic catalyst. The research team optimized the reaction conditions and screened HASPO ligands with high enantioselectivity, significantly improving the reaction yield and stereoselectivity. The paper also demonstrates the versatility of this method across a broad range of substrates, including various N-aryl-substituted benzimidazoles and diverse alkenes.

Through experiments and density functional theory (DFT) calculations, the authors elucidated the reaction mechanism, including ligand-to-ligand hydrogen transfer (LLHT) and reductive elimination steps. Furthermore, multivariate linear regression (MVLR) analysis was employed to investigate the relationship between the structure of HASPO ligands and enantioselectivity, providing a theoretical basis for future ligand design and optimization.

This study not only offers a cost-effective and low-toxicity catalytic approach but also opens up new avenues for the synthesis of axially chiral compounds with biological activity and application potential.

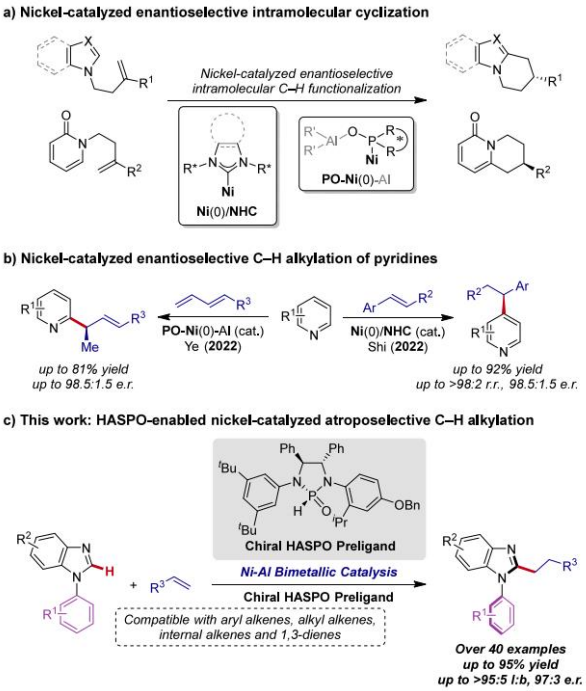


Figure 4 Schematic Diagram of the Synthetic Process in the Paper

HASPO ligands play a central role in nickel-catalyzed atroposelective C − H alkylation reactions, with their specific functions reflected in the following aspects: as chiral sources, HASPO ligands provide the necessary chiral environment to ensure high enantioselectivity of the reaction; by forming complexes with nickel catalysts, HASPO ligands regulate the electronic and stereochemical properties of the catalysts through coordination interactions, influencing catalytic activity and selectivity; the stereochemistry and spatial structure of HASPO ligands play a decisive role in the stereocontrol of the reaction, and their non-C2 symmetry feature is particularly crucial in stereoselectivity control; the structural characteristics of HASPO ligands facilitate the recognition and stabilization of substrates or intermediates, improving chemoselectivity and regioselectivity; in the catalytic cycle, HASPO ligands are involved in the initial formation of the catalyst, the stabilization of intermediates, and the release of the final product; by optimizing the structure of HASPO ligands, the efficiency and stability of the catalyst can be enhanced, thereby improving the reaction yield; suitable HASPO ligands enable the reaction to proceed under mild conditions, reducing side reactions and increasing the selectivity of the target product. HASPO ligands are key components for achieving this highly stereoselective reaction, and through their multifaceted roles, they provide an effective approach for the synthesis of axially chiral compounds with potential application value.

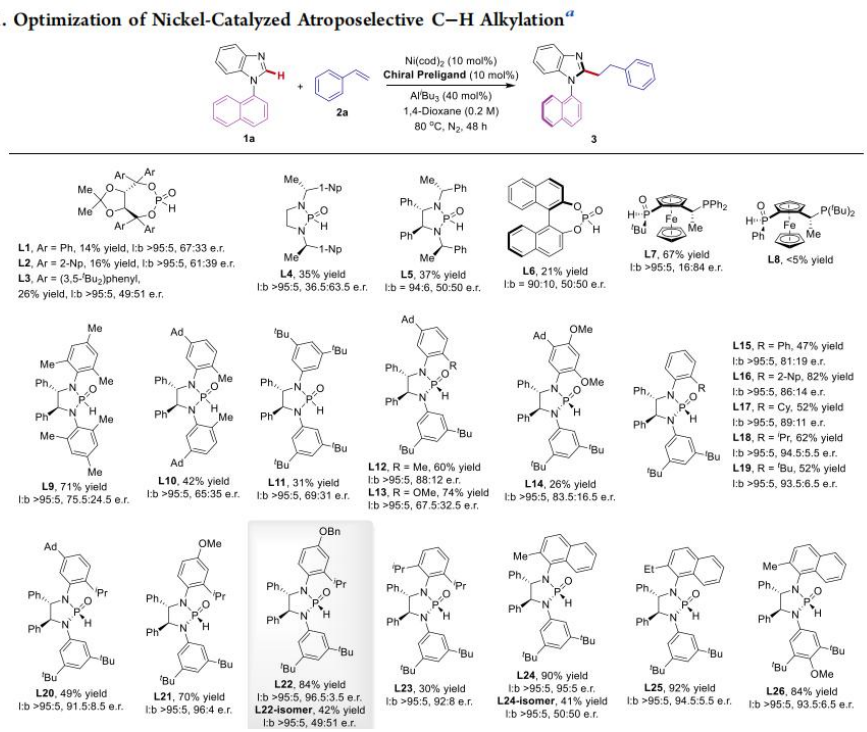


Figure 5 Excerpts of Ligands and Related Data from the Paper

Experimental data of various ligands in the paper were obtained from https://github.com/zju-ys/Nickel-MLR/blob/main/PhysOrg.csv.

ddg	B1_8_10	B5_8_10	L_8_10	BV_10	B1_9_11	B5_9_11	L_9_11	BV_11	q_1	q_2	q_3	q_8	q_9	q_10	q_11	d_1_2	d_2_3	d_9_11	d_10_12
−0. 388946	1. 7	7. 4296553	5. 7629159	0. 6819092	2. 0329402	7. 4718725	5. 6430446	0. 6788751	−1. 039	2. 00608	−0. 08846	−0. 86861	−0. 87723	−0. 03173	−0. 02345	1. 4924619	1. 4250983	1. 4693375	1. 4650537
0	1. 7632793	6. 1784451	5. 4374385	0. 7185869	2. 9432972	6. 1317296	5. 5465926	0. 7213532	−1. 05405	2. 00334	−0. 0853	−0. 86786	−0. 8706	−0. 02687	−0. 02759	1. 4968031	1. 4185372	1. 4679952	1. 4727254
0. 7905382	1. 8819142	4. 4815647	7. 8585811	0. 8145034	2. 1075711	4. 460685	7. 8547358	0. 8230875	−1. 03787	2. 01448	−0. 08186	−0. 86954	−0. 87554	0. 14501	0. 14983	1. 4915714	1. 4180347	1. 4312338	1. 4303502
0. 4348216	2. 5111604	7. 8376077	8. 8028728	0. 7709187	2. 4209684	9. 4963049	8. 9224935	0. 7956695	−1. 02937	1. 99388	−0. 06051	−0. 82913	−0. 8555	0. 19396	0. 18414	1. 4915528	1. 4079593	1. 4224277	1. 422483
0. 5620147	3. 2412684	5. 8051073	7. 8846047	0. 7281611	3. 2388224	5. 7928513	7. 4517964	0. 720037	−1. 03607	1. 98974	−0. 06491	−0. 81513	−0. 81882	0. 20575	0. 20125	1. 4931489	1. 4168374	1. 4116038	1. 4154465
1. 3995102	2. 4984086	7. 8258199	8. 8165076	0. 7572738	3. 3128314	5. 7835747	7. 8605654	0. 7277651	−1. 03586	1. 99049	−0. 06464	−0. 82563	−0. 82039	0. 19595	0. 19816	1. 4934706	1. 4091216	1. 4267729	1. 4134303
0. 5133854	2. 2717998	7. 3149553	9. 1400932	0. 7485848	3. 2308771	5. 7893561	7. 865743	0. 7277593	−1. 05277	1. 98111	−0. 04809	−0. 8092	−0. 81854	0. 16537	0. 20092	1. 496008	1. 4068717	1. 4105079	1. 4118137
1. 1389541	2. 2609738	7. 441784	9. 045662	0. 7501048	3. 2263198	5. 8066355	7. 8833555	0. 7279165	−1. 05289	1. 98168	−0. 05022	−0. 81405	−0. 81648	0. 1397	0. 20235	1. 4959901	1. 4075183	1. 4149952	1. 4106513
1. 018507	2. 0824919	7. 128398	6. 8250532	0. 7893974	3. 2330722	5. 661711	7. 8756272	0. 7373975	−1. 04888	2. 00165	−0. 06535	−0. 85305	−0. 82186	0. 17462	0. 19581	1. 4957895	1. 4170231	1. 4268874	1. 4151883
1. 2750845	2. 1107794	8. 6971519	6. 8162737	0. 7739354	3. 2466584	5. 651684	7. 88062	0. 7358484	−1. 03479	1. 99543	−0. 06296	−0. 85406	−0. 82574	0. 17603	0. 19525	1. 4917458	1. 4169393	1. 4232815	1. 4163381
1. 4685651	2. 1332227	7. 5315859	6. 815889	0. 7824439	3. 2356962	5. 7948018	7. 8736751	0. 7289939	−1. 04455	1. 99868	−0. 06398	−0. 86722	−0. 82185	0. 15617	0. 19919	1. 495208	1. 4164254	1. 4287729	1. 4140751
1. 9975603	2. 1297207	5. 6186439	6. 823591	0. 7882152	3. 2020011	5. 8277351	7. 9000536	0. 7304964	−1. 02866	2. 01086	−0. 07993	−0. 86059	−0. 82647	0. 15205	0. 20242	1. 4885243	1. 4211465	1. 4338672	1. 4151735
1. 8727467	2. 1914012	5. 7975667	6. 8176449	0. 8240135	3. 2390039	5. 6678726	7. 8793435	0. 7418235	−1. 03504	1. 99238	−0. 06593	−0. 87125	−0. 82482	0. 1483	0. 19578	1. 4930473	1. 4148129	1. 4406147	1. 4118463
1. 6691265	3. 2726079	7. 7457548	8. 9433897	0. 7957103	3. 3088212	7. 6400555	7. 8801526	0. 7328142	−1. 04623	1. 99781	−0. 06236	−0. 86882	−0. 82179	0. 16255	0. 19683	1. 4961726	1. 4158523	1. 4277233	1. 4130305
2. 2323084	2. 1387408	5. 6209923	8. 9316077	0. 7881511	3. 2380642	5. 7008396	7. 8996072	0. 7343692	−1. 03157	2. 01219	−0. 08111	−0. 8614	−0. 82624	0. 12192	0. 20039	1. 4893049	1. 4213668	1. 4339912	1. 4149083
2. 3297516	2. 2643829	7. 0989974	8. 9233781	0. 7881103	3. 2034481	5. 8286505	7. 9008642	0. 730048	−1. 02988	2. 0113	−0. 08134	−0. 86185	−0. 8258	0. 12344	0. 20321	1. 4889788	1. 4213234	1. 4340694	1. 4145396
1. 7155379	3. 3063354	5. 6213675	6. 8190651	0. 8306525	3. 2478413	5. 6557607	7. 8763247	0. 7420099	−1. 04479	2. 01345	−0. 07278	−0. 87657	−0. 82737	0. 15065	0. 19355	1. 4935753	1. 4183455	1. 4334635	1. 4157463
2. 0682142	1. 7685765	5. 7550974	6. 8301454	0. 8177122	3. 2077383	5. 8327083	7. 9049784	0. 7314864	−1. 01958	2. 00444	−0. 08075	−0. 87054	−0. 82707	0. 17739	0. 20253	1. 4878347	1. 4198945	1. 4306815	1. 4151466
1. 9975603	2. 4895425	5. 7739051	6. 8189946	0. 8353697	3. 2079092	5. 8333755	7. 9054085	0. 7303741	−1. 02053	2. 00726	−0. 08357	−0. 87697	−0. 82613	0. 17833	0. 20296	1. 4881244	1. 4193596	1. 4314699	1. 4147093
1. 8727467	1. 7	5. 8270676	6. 8312933	0. 7997985	3. 4219744	5. 7996223	8. 5643067	0. 7423651	−1. 04506	2. 01078	−0. 06828	−0. 87822	−0. 829	0. 16451	0. 18313	1. 4940802	1. 417577	1. 4289446	1. 4160723
−0. 0281	3. 2559046	5. 7847135	7. 8641613	0. 7364948	2. 133421	7. 2479317	8. 9168286	0. 8013476	−1. 03029	1. 99937	−0. 07248	−0. 82109	−0. 86797	0. 20344	0. 13611	1. 4923631	1. 4159182	1. 4092765	1. 4247119
0	3. 2431543	5. 662814	7. 8843125	0. 7462029	1. 9140127	5. 7370919	6. 8741757	0. 8106539	−1. 03681	1. 99826	−0. 06581	−0. 82177	−0. 8773	0. 19907	0. 17183	1. 4936515	1. 4158795	1. 4107916	1. 4289269

Figure 6 Acquired Ligand Reaction Data

Part 1: Linear Regression Based on SPSS

Step 1: General Linear Regression Model

A general linear regression model was constructed for the data, and the analyzed feedback indicated the presence of strong multicollinearity.

ddg=251.488−0.167 · B1810+1.610 · B1911+0.031 · B5810+0.172 · B5911−0.070 · L810−0.898 · L911−8.745 · BV10+15.826 · BV11+39.023 · q1+42.782 · q2−24.417 · q3

模型摘要

模型	R	R 方	调整后 R 方	标准估算的错误
1	.998a	.996	.959	.168397174645

a. 预测变量:(常量), d_10_12, B5_8_10, q_8, q_1, B1_8_10, BV_11, B5_9_11, L_8_10, d_2_3, B1_9_11, q_2, L_9_11, d_9_11, d_1_2, BV_10, q_10, q_3, q_11, q_9

系数a

模型		未标准化系数		标准化系数	t	显著性	共线性统计	
		B	标准错误	Beta			容差	VIF
1	(常量)	251.488	314.929		.799	.508		
	B1_8_10	-.167	.236	-.110	-.709	.552	.081	12.367
	B1_9_11	1.610	1.223	.951	1.317	.318	.004	265.574
	B5_8_10	.031	.089	.041	.355	.757	.148	6.756
	B5_9_11	.172	.177	.211	.972	.433	.042	23.980
	L_8_10	-.070	.109	-.094	-.639	.588	.090	11.136
	L_9_11	-.898	.976	-.878	-.921	.454	.002	463.809
	BV_10	-8.745	9.448	-.414	-.926	.452	.010	101.816
	BV_11	15.826	29.786	.646	.531	.648	.001	752.232
	q_1	39.023	89.506	.462	.436	.705	.002	572.100
	q_2	42.782	85.809	.507	.499	.668	.002	526.849
	q_3	-24.417	67.697	-.327	-.361	.753	.002	419.139
	q_8	-15.354	16.048	-.442	-.957	.440	.009	108.632
	q_9	-13.469	95.484	-.358	-.141	.901	.000	3283.040
	q_10	-10.235	6.607	-.770	-1.549	.261	.008	125.944
	q_11	-11.148	29.503	-.887	-.378	.742	.000	2807.339
	d_1_2	25.175	173.902	.084	.145	.898	.006	172.846
	d_2_3	-93.684	79.805	-.539	-1.174	.361	.009	107.470
	d_9_11	-15.933	27.297	-.297	-.584	.618	.008	132.270
	d_10_12	-143.022	69.740	-2.802	-2.051	.177	.001	951.288

a. 因变量：ddg

Step 2: Data Validation

Multicollinearity and Pearson correlation tests were performed on the data to select principal components or integrate variables, thereby mitigating the impact of multicollinearity on model accuracy.

Multicollinearity Test

系数[a]

模型		未标准化系数		标准化系数			共线性统计	
		B	标准错误	Beta	t	显著性	容差	VIF
1	（常量）	251.488	314.929		.799	.508		
	B1_8_10	-.167	.236	-.110	-.709	.552	.081	12.367
	B1_9_11	1.610	1.223	.951	1.317	.318	.004	265.574
	B5_8_10	.031	.089	.041	.355	.757	.148	6.756
	B5_9_11	.172	.177	.211	.972	.433	.042	23.980
	L_8_10	-.070	.109	-.094	-.639	.588	.090	11.136
	L_9_11	-.898	.976	-.878	-.921	.454	.002	463.809
	BV_10	-8.745	9.448	-.414	-.926	.452	.010	101.816
	BV_11	15.826	29.786	.646	.531	.648	.001	752.232
	q_1	39.023	89.506	.462	.436	.705	.002	572.100
	q_2	42.782	85.809	.507	.499	.668	.002	526.849
	q_3	-24.417	67.697	-.327	-.361	.753	.002	419.139
	q_8	-15.354	16.048	-.442	-.957	.440	.009	108.632
	q_9	-13.469	95.484	-.358	-.141	.901	.000	3283.040
	q_10	-10.235	6.607	-.770	-1.549	.261	.008	125.944
	q_11	-11.148	29.503	-.887	-.378	.742	.000	2807.339
	d_1_2	25.175	173.902	.084	.145	.898	.006	172.846
	d_2_3	-93.684	79.805	-.539	-1.174	.361	.009	107.470
	d_9_11	-15.933	27.297	-.297	-.584	.618	.008	132.270
	d_10_12	-143.022	69.740	-2.802	-2.051	.177	.001	951.288

a. 因变量：ddg

Pearson Correlation Test

皮尔逊相关性检验

		B1_8_10	B1_9_11	B5_8_10	B5_9_11	L_9_11	BV_10	BV_11	q_1	q_2	q_3	q_8	q_9	q_10	q_11	d_1_2	d_2_3	d_9_11	d_10_12
B1_8_10	皮尔逊相关性	1	-.125	-.052	.229	.343	-.015	.322	.046	-.222	.312	.416	.023	.509*	.310	.147	-.232	-.554**	-.296
	显著性（双尾）		.580	.818	.305	.275	.948	.143	.841	.320	.157	.054	.885	.016	.160	.514	.300	.007	.182
	个案数	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22
B1_9_11	皮尔逊相关性	-.125	1	.219	-.306	.070	.245	.431*	-.581**	-.139	-.134	.256	-.208	.929**	.182	.509*	.061	-.108	-.057	-.597**
	显著性（双尾）	.580		.328	.166	.756	.272	.045	.005	.536	.554	.251	.354	.000	.418	.015	.823	.634	.802	.007
	个案数	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22
B5_8_10	皮尔逊相关性	-.052	.219	1	.431*	.219	.003	-.326	-.378	-.267	-.534*	.499*	.351	.290	-.032	.027	.340	-.402	-.043	-.073
	显著性（双尾）	.818	.328		.045	.328	.981	.138	.083	.230	.011	.018	.240	.190	.888	.905	.122	.054	.849	.747
	个案数	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22
B5_9_11	皮尔逊相关性	.229	-.306	.431*	1	.148	.113	-.330	.048	.109	-.194	.111	.179	-.295	-.070	-.235	.057	-.198	.058	.214
	显著性（双尾）	.305	.166	.045		.510	.615	.134	.833	.628	.388	.624	.426	.251	.758	.292	.801	.376	.799	.339
	个案数	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22
L_8_10	皮尔逊相关性	.389	.070	.219	.148	1	.504*	.026	.279	.004	-.393	.522*	.584**	.285	.490*	.581**	.017	-.579**	-.640**	-.547**
	显著性（双尾）	.074	.756	.328	.510		.017	.908	.209	.986	.070	.013	.004	.199	.021	.008	.939	.005	.001	.008
	个案数	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22
L_9_11	皮尔逊相关性	.343	.245	.005	.113	.504*	1	.549**	.432*	.309	-.071	.403	.158	.422	.802**	.777**	-.239	-.391	-.685**	-.802**
	显著性（双尾）	.275	.272	.981	.615	.017		.008	.045	.162	.755	.063	.483	.051	.000	.000	.284	.072	.000	.000
	个案数	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22
BV_10	皮尔逊相关性	-.015	-.431*	-.326	-.330	.026	.549**	1	.210	.326	.396	-.077	-.548**	.423*	.440*	.626**	-.377	.092	-.133	-.619**
	显著性（双尾）	.948	.045	.138	.134	.908	.008		.349	.139	.068	.735	.008	.044	.041	.001	.083	.685	.557	.003
	个案数	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22
BV_11	皮尔逊相关性	.322	-.581**	-.378	.048	.279	.432*	.210	1	.185	.098	.101	.230	-.494*	.476*	.190	-.072	-.290	-.475*	-.105
	显著性（双尾）	.143	.005	.083	.833	.309	.045	.349		.410	.665	.655	.304	.020	.025	.398	.751	.362	.025	.643
	个案数	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22	22

Step 3: Stepwise Regression Model

Stepwise regression modeling was performed on the data to select key variables, thereby mitigating the impact of multicollinearity on model accuracy.

模型 1：ddg=-11.826+16.765· BV10
模型 2：ddg=-10.858+12.397· BV10+0.812· B1911
模型 3：ddg=136.367+9.025· BV10+0.956· B1911-97.166· d12

模型误差

模型摘要

模型	R	R 方	调整后 R 方	标准估算的错误
1	.793a	.629	.611	.517553989011
2	.904b	.817	.797	.373513621323
3	.949c	.901	.885	.281257819922

a. 预测变量：(常量)，BV_10
b. 预测变量：(常量)，BV_10，B1_9_11
c. 预测变量：(常量)，BV_10，B1_9_11，d_1_2

Step 4: Combined Variable Regression Model (Principal Component Analysis)

A preliminary analysis of the combined variable regression model was performed on the data to evaluate variable combinations:

B1_8_10 and B1_9_11: The correlation coefficient between these two variables is -0.125. Although not very high, the result is statistically insignificant (P-value = 0.580), and both variables exhibit high correlations with multiple other variables, requiring further investigation.
BV_10 and BV_11: The correlation coefficient between these two variables is 0.431, indicating a moderate correlation, which is statistically significant (P-value = 0.045).
q_1 and q_2: The correlation coefficient between them is -0.466, representing a strong negative correlation that is statistically significant (P-value = 0.029).
q_3 and q_8: The correlation coefficient between them is 0.644, indicating a very strong positive correlation that is statistically significant (P-value = 0.001).
d_1_2 and d_2_3: The correlation coefficient between them is 0.811, representing an extremely strong positive correlation that is statistically significant (P-value = 0.000).

Step 5: Primary Combined Variable Regression Model

A primary combined variable regression model was constructed for the data.
Principal Component 1 (PC1):PC1 = 0.861 * BV10 + 0.929 * q1 + 0.859 * q2 + 0.883 * q3 + 0.928 * q8 + 0.923 * d12 + 0.906 * d23
Principal Component 2 (PC2):PC2 = 0.561 * BV11
ddg = −61.986 + 23.144 · PC1 − 27.823 · PC2

模型摘要

模型	R	R 方	调整后 R 方	标准估算的错误
1	.845[a]	.715	.685	.465888163804

a. 预测变量：(常量)，PC2，PC1

Step 6: Tertiary Combined Variable Regression Model

A tertiary combined variable regression model was constructed for the data.
Principal Component 1 (PC1):PC1 = 0.861 * BV10 + 0.929 * q1 + 0.859 * q2 + 0.883 * q3 + 0.928 * q8 + 0.923 * d12 + 0.906 * d23
Principal Component 2 (PC2):PC2 = 0.561 * BV11
ddg = -16.368 + 0.52 × PC1³ddg = −19.911 + 0.734 × PC1³ − 50.056 ×PC2

模型摘要

模型	R	R 方	调整后 R 方	标准估算的错误
1	.625[a]	.390	.360	.663690353415
2	.848[b]	.719	.689	.462282978876

a. 预测变量：(常量)，sanPC1

b. 预测变量：(常量)，sanPC1，sanPC2

Step 7: Multiple Combined Variable Regression Model

Multiple combined variable regression modeling was performed on the data.
Principal Component 1 (PC1):PC1 = 0.861 * BV10 + 0.929 * q1 + 0.859 * q2 + 0.883 * q3 + 0.928 * q8 + 0.923 * d12 + 0.906 * d23
Principal Component 2 (PC2):PC2 = 0.561 * BV11
ddg = 261.178 + 5.432 × PC1³ − 275.192 × PC2³ − 294.506 × PC1^(1/2) + 252.284 × PC2^(1/2)

模型摘要				
模型	R	R 方	调整后 R 方	标准估算的错误
1	.853[a]	.728	.664	.480615869992
2	.000[b]	.000	.000	.829464803808
a. 预测变量：(常量), kaiPC2, sanPC1, sanPC2, kaiPC1				
b. 预测变量：(常量)				

Step 8: Model Collation and Selection

The data were collated and analyzed for model selection, resulting in the identification of a model that meets the requirements, which is then put into application.

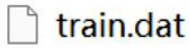
一般线性回归（SPSS）					
模型	特征值项	方法	R方	残差均方	F
ddg=251.488−0.167· B1810+1.610· B1911+0.031· B5810+0.172· B5911−0.070· L810−0.898· L911−8.745· BV10+15.826· BV11+39.023· q1+42.782· q2−24.417· q3		输入	0.959	0.028	26.711
ddg=−11.826+16.765· BV10		步进	0.611	0.268	33.939
ddg=−10.858+12.397· BV10+0.812· B1911		步进	0.797	0.140	42.281
ddg=136.367+9.025· BV10+0.956· B1911−97.166· d12		步进	0.885	0.079	54.881
ddg=−61.986+23.144· PC1−27.823· PC2	主成分1 (PC1) PC1 = 0.861 * BV10 + 0.929 * q1 + 0.859 * q2 + 0.883 * q3 + 0.928 * q8 + 0.923 * d12 + 0.906 * d23	一次主成分	0.685	/	/
ddg = −19.911 + 0.734 * (0.861 * BV10 + 0.929 * q1 + 0.859 * q2 + 0.883 * q3 + 0.928 * q8 + 0.923 * d12 + 0.906 * d23)**3 − 50.056 * (0.561 * BV11)**3		三次	0.689	/	/
$ddg=-16.368+0.52\times PC1^{3}$		三次	0.360	/	/
$ddg=261.178+5.432\times PC1^{3}-275.192\times PC2^{3}-294.506\times PC1^{1/2}+252.284\times PC2^{1/2}$	主成分2 (PC2): PC2 = 0.561 * BV11	多次	0.664	/	/

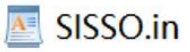
Part 2: Symbolic Regression Based on SISSO

Step 1: System Configuration and Data Preprocessing

In accordance with the SISSO Guide standards and relevant instructions, the installation and debugging of SISSO were performed in a Linux environment.
Following the examples and Guide instructions, the CSV data file format was converted into the train.dat file that meets SISSO’s requirements.
Additionally, the SISSO.in file was configured according to the standard specifications.

示例：在Ubuntu中启动SISSO





应当预先存入src文件夹的配置文件与数据文件

Step 2: Parameter Tuning

In accordance with the SISSO Guide standards and relevant instructions, various parameters in the SISSO.in file were configured. The following is an explanation and interpretation of each parameter item:

- *ptype: Type of the target property. 1 denotes regression with continuous properties; 2 denotes classification with categorical properties.*
- *ntask: Number of tasks. ntask = 1 indicates conventional machine learning for a single task; ntask > 1 indicates multi-task learning (MTL).*
- *task_weighting: (Regression-only) Task weighting in multi-task learning regression. 1 means no weighting (all tasks are treated equally regardless of their dataset size); 2 means each task is weighted by the ratio of its data volume to the total data volume of all tasks.*
- *scmt: (Regression-only) If set to .true., signed-constrained multi-task regression is invoked.*
- *desc_dim: Dimension of descriptors or models.*
- *nsample: Number of samples in train.dat. For single-task regression, the input is a single integer. For multi-task regression, the input is multiple integers separated by commas, representing the number of samples for each task. For single-task classification, there will be only one integer enclosed in parentheses, indicating the number of data points for each category. For multi-task classification, there will be multiple parentheses separated by commas, each defining a task, and the integers inside specify the number of data points for the corresponding category.*
- *restart: Restart or continue the job. 0 means starting the job from scratch; 1 means continuing the job (progress information from the last job is stored in the CONTINUE file).*

- *nsf: Number of scalar features provided in train.dat. "Scalar features" means each feature in the dataset occupies one column.*
- *ops: Mathematical operators used for feature construction. Users can customize operators from the list, such as {+, -, *, /, exp, exp-, ^-1, ^2, ^3, sqrt, cbrt, log, |-|, scd, ^6, sin, cos}.*
- *fcomplexity: Feature complexity defined as the number of operators in a feature. fcomplexity=0 means only input variables exist in the feature space, while fcomplexity=3 means the complexity of all features in the feature space does not exceed 3.*
- *funit: Indicates the type of features in the train.dat file that share the same unit. For example, funit=(1:5)(6:9)(11:11) means features 1 to 5 have the same unit, features 6 to 9 have another unit, feature 10 is dimensionless, and feature 11 has a different unit.*
- *fmax_min and fmax_max: Thresholds for the maximum absolute value in the feature data. Features with values less than fmax_min are regarded as zero features and discarded; those greater than fmax_max are regarded as infinite features and discarded.*
- *nf_sis: Size of the SIS-subspace. For SISSO-nD calculations, there will be n SIS-subspaces.*
- *method_so: Method of the sparsity operator, which can be the L0-norm minimization sparsification method or the L1L0 method (regression-only).*
- *nl1l0: (Regression-only) Number of features selected using LASSO for subsequent L0 optimization.*
- *fit_intercept: (Regression-only) Whether the linear model fits a non-zero/zero intercept.*
- *metric: (Regression-only) Metric used for model selection in regression, which can be RMSE (Root Mean Square Error) or MaxAE (Maximum Absolute Error).*
- *nmodels: Number of top-ranked models to output.*
- *isconvex: (Classification-only) Whether each data domain can be constrained to be convex or non-convex.*
- *bwidth: (Classification-only) Boundary tolerance for each domain to include data very close but outside the domain.*

Step 3: Parameter Configuration

In accordance with the SISSO Guide standards and relevant instructions, various parameters in the SISSO.in file were configured. The following is an example of a primary configuration:

```
! SISSO Control Parameters
ptype = 1          ! Regression
ntask = 1          ! Single task
task_weighting = 1  ! No weighting for single task
scmt = .false.      ! Not sign-constrained multi-task learning
desc_dim = 5       ! Dimension of the descriptor (set based on your
requirement)
nsample = 22       ! Number of samples (from the train.dat.txt)
restart = 0         ! Start from scratch
nsf = 22           ! Number of scalar features (one for each column
excluding the first)
ops =  '(+)(-)(*)(/)(log)'  ! Mathematical operators for feature
construction
fcomplexity = 3    ! Feature complexity (set based on your model complexity
requirement)
funit = (2:22)     ! If all features are dimensionless, leave it empty
fmax_min = 1e-3    ! Features with absolute values smaller than this are
discarded
fmax_max = 1e9     ! Features with absolute values larger than this are
discarded
nf_sis = 300       ! Size of the SIS-subspaces (can be adjusted based on
computational resources)
method_so = 'L0'  ! Sparsity method, 'L0' for regression
nl1l0 = 100        ! Number of features selected by LASSO (if method_so is
'L1L0')
fit_intercept = .true.  ! Fit a nonzero intercept
metric = 'RMSE'    ! Model selection metric for regression
nmodels = 7        ! Number of top models to output
isconvex = .false.! Not applicable for regression
bwidth = 0.1       ! Boundary width for classification (not used in
regression)
```

Step 4: Data Interpretation

After running SISSO in the Ubuntu environment, the results are saved in the SISSO.out file. Regression results were obtained by interpreting the report in SISSO.out. Below is a summary of the results from one specific run:

```
Dimension:      1
-------------------
Feature Construction (FC) starts ...
Population Standard Deviation (SD) of the task 001:       0.81046
Total number of features in the space phi00:                    22
Total number of features in the space phi01:                   688
Total number of features in the space phi02:                542038
Size of the SIS-selected subspace from phi02:               300
Time (second) used for this FC:                    0.33

Descriptor Identification (DI) starts ...
Total number of SIS-selected features from all dimensions:              300

   1D descriptor:
       d001 = ((q_3*q_11)*(q_8+BV_11))       feature_ID:000001

   1D model(y=sum(ci*di)+c0):
       coeff.(ci)_task001:         0.9161889536E+03
                  c0_task001:         0.3565908600E-01
       RMSE,MaxAE_task001:         0.2585603429E+00       0.6108099105E+00

       RMSE and MaxAE of the model:      0.258560    0.610810
--------------------------------------------------------------------------------
Time (second) used for this DI:                    0.00
```

Step 5: Model Collation

After multiple rounds of parameter tuning and continuous data integration, the following models were obtained:

符号约束回归（SISSO）											
No	描述符维度DD	coeff.(ci)	特征值项	c0	稀释模式SM	空间层级FS	选择性子空间SIS	最多操作符MF	操作符	RMSE	MaxAE
1	1	916.1889536	d001 = ((q_3*q_11)*(q_8+BV_11))	0.035659086	L0	2	100-100-100	3	(+)(-)(*)	0.258560343	0.610809911
	2	-4.800899871	d001 = ((q_10-q_2)*(q_11-q_8))	6.391823734						0.161254495	0.358187173
		-7.9199913	d002 = ((q_3*B1_9_11)+(d_2_3*d_10_12))								
	3	203.6234498	d001 = ((q_3*B1_9_11)*(d_10_12-d_1_2))	7.095999257						0.139705011	0.301084836
		5.440928931	d002 = ((BV_11-q_10)*(q_11+d_2_3))								
		-12.83892716	d003 = ((d_2_3-q_1)-(q_3+d_9_11))								
2	1	与1-1组相同			L0	2	500-500-500	3	(+)(-)(*)	/	/
	2	-16.71618794	d001 = ((q_3*B1_9_11)*(q_11*B1_9_11))	-5.415589707						0.158993916	0.350418208
		-10.00633031	d002 = ((q_8+q_10)*(q_3+BV_11))								
	3	-24.0686122	d001 = ((q_3*B1_9_11)*(q_11*B1_9_11))	16.92556002						0.126080266	0.233273538
		22.01468794	d002 = ((d_1_2-q_2)-(q_9+d_2_3))								
		9.013415961	d003 = ((BV_11-q_10)-(q_8-q_9))								
3	1	8.391555126	d001 = ((d_9_11*(q_11+d_9_11))-(q_3*B1_9_11))	-19.76751669	L0	3	100-100-100-100	4	(+)(-)(*)	0.21293301	0.549436095
	2	8.303188962	d001 = ((d_9_11*(q_11+d_9_11))-(q_3*B1_9_11))	-18.9984024						0.152809484	0.327464131
		50.31508678	d002 = ((q_3+q_11)*((d_1_2-d_2_3)-q_10))								
	3	6.910015417	d001 = ((d_9_11*(q_11+d_9_11))-(q_3*B1_9_11))	-18.10994559						0.113640519	0.259986489
		0.203555406	d002 = ((L_8_10+L_9_11)*(q_8*(q_8+q_10)))								
		-2.819536326	d003 = ((L_8_10-L_9_11)*(q_3*(L_8_10-L_9_11)))								
	4	7.546557957	d001 = ((d_9_11*(q_11+d_9_11))-(q_3*B1_9_11))	4.689449251						0.093447747	0.202795725
		51.31391725	d002 = ((q_3+q_11)*((d_1_2-d_2_3)-q_10))								
		-0.205932076	d003 = ((L_8_10-L_9_11)*((L_9_11-L_8_10)-q_11))								
		7.023904378	d004 = ((BV_10*(q_1-q_8))-(q_2*d_1_2))								
4	1	916.1889536	d001 = ((q_3*q_11)*(q_8+BV_11))	0.035659086	L1L0_100	2	100-100-100	3	(+)(-)(*)	0.258560343	0.610809911
	2	-4.800899871	d001 = ((q_10-q_2)*(q_11-q_8))	6.391823734						0.161254495	0.358187173
		-7.9199913	d002 = ((q_3*B1_9_11)+(d_2_3*d_10_12))								
	3	7.807636163	d001 = ((d_9_11-q_8)+(q_11+d_9_11))	-16.23073477						0.141941942	0.305469742
		-7.428358192	d002 = ((q_3*B1_9_11)+(d_1_2*d_10_12))								
		0.210381111	d003 = ((BV_11-q_10)*(L_8_10+B1_9_11))								

References

1.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.
2.Landrum.RDKit: Open-source cheminformatics. Release 2014.03.1[J].2010.
3.RDKit: "RDKit: Open-source cheminformatics. n.d. https://www.rdkit.org. Accessed 14 Aug. 2024."
4.pandas: " McKinney, Wes. "Data Structures for Statistical Computing in Python." Proceedings of the 9th Python in Science Conference, 2010, pp. 51-56."
5.statsmodels: "Seabold, Skipper, and Josef Perktold. "Statsmodels: Econometric and Statistical Modeling with Python." Proceedings of the 9th Python in Science Conference, 2010, pp. 91-96."
6.matplotlib: "Hunter, John D. "Matplotlib: A 2D Graphics Environment." Computing in Science & Engineering, vol. 9, no. 3, 2007, pp. 90-95."
7.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
8.Ouyang, R. (2023, Sep 12). SISSO. SISSO.3.3, July, 2023. https://rouyang2017.github.io/SISSO/

Training Records

Training Status

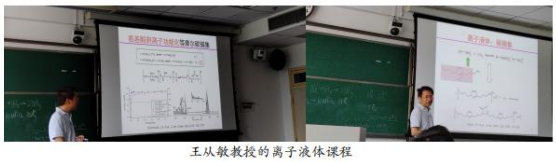
Cutting-Edge Lectures

Since participating in the Youth Talent Program, I have attended nearly ten weekend chemistry lectures and several offline experimental activities at Zhejiang University. During this period, two academicians of the Chinese Academy of Sciences delivered face-to-face lectures for us.

In the summer of 2024, I participated in the Summer Camp of the Youth Talent Program at Zhejiang University, which lasted for nearly a week. During the camp, I attended more than ten comprehensive lectures and reports on chemistry and other disciplines.

These cutting-edge interdisciplinary reports have provided rich and vivid chemical knowledge. Professors from different fields and research directions shared diverse content, and each lecture and learning experience has brought me substantial gains.

Figure 7 On-Site Scene of the Academician's Lecture



王从敏教授的离子液体课程

季鹏飞教授的生物催化课程

王林军教授的智能化学讲座

Figure 8 Multiple Cutting-Edge Lectures

Expanding Horizons

Under the guidance of Professor He Qiaohong from Zhejiang University and led by experimental teachers from the Department of Chemistry, Zhejiang University, I have visited various laboratories and research platforms of the department multiple times to carry out basic experimental skill learning and practice. Meanwhile, I have entered laboratories for learning and understanding on numerous occasions under the leadership of various professors.

Figure 9 On-Site Learning and Practice in the Laboratory

Intra-Group Training

I am supervised by Professor Hong Xin from Zhejiang University and have conducted learning through various online and offline methods with Professor Hong.
With Professor Hong’ s guidance and assistance, I have carried out studies and research, successfully completing all items of the research project.
Under Professor Hong’ s supervision and with the help of graduate students in his group, such as Yu Shuang, I have conducted research on various research topics.
During the summer vacation, I participated in on-site learning at the laboratory several times, maintained in-depth learning through online means, and completed multiple projects on schedule.

Keeping Learning Journals

After each learning session, I recorded my experiences in a journal and submitted them to the system. To date, 13 journals have been submitted.

Summer Activity Record of the Elite Talent Program at Zhejiang University

From July 8 to 12, we participated in the summer activities of the Elite Talent Program at the Zijingang Campus of Zhejiang University. The program's summer activities were rich and varied, with profound significance. Through five days of study and life, we explored cutting-edge knowledge, experienced the charm of scientific research, and gained a deeper understanding of our respective research topics.

在石虎山机器人研究基地看到的机器狗表演

The summer activities mainly included group events such as knowledge lectures, university elective courses, micro-topic research, and visits to cutting-edge bases, while also providing ample time for us to communicate with our supervisors. Organized by the Chu Kochen Honors College of Zhejiang University, we learned about the frontier developments in various disciplines and felt the innovative methods of scientific research through the enthusiastic sharing of professors from different fields. For example, in Professor Ye Gaoxiang's report, we understood the importance of the mutual promotion between philosophical systems and scientific development; in the consecutive days of university advanced placement courses, Professor Chen Jinhui guided us to initially grasp higher mathematics knowledge such as calculus; in Professor Tang Jianjun's ecology-themed lecture, we experienced the interesting aspects of ecological research. Many experts and scholars brought us a feast of cutting-edge knowledge, igniting our passion for exploring scientific truths.

张克俊教授的智能产品设计讲座

陈锦辉教授的大中数学衔接课

Meanwhile, in the micro-topic research carried out in groups by different disciplines, various rich activities were organized. Taking the chemistry group I participated in as an example, Professor Wang Congmin introduced the importance of ionic liquids for environmental protection and pointed out the role of chemistry in the "dual carbon" initiative; Professor Ji Pengfei conducted materials research through biocatalysis, combining biology with chemistry; from Professor Wang Linjun ’ s sharing, we witnessed a diverse chemical world and recognized the vigorous development of innovative fields integrating chemistry with multiple disciplines. After three micro-topic research activities, we held group presentations. Students in the chemistry group introduced their respective research topics, ranging from inorganic catalysis to organic pharmaceuticals, and from experimental chemistry to computational chemistry. These presentations not only deepened our understanding of various chemical fields but also fostered our ability to communicate with one another.

浙江大学紫金港校区美丽的校园环境与住宿环境

Appendix

Basic Information about the Youth Talent Program (National Program for Cultivating Reserve Talents in Science, Technology and Innovation for High School Students)

The "Program for Cultivating Reserve Talents in Science and Technology Innovation for High School Students" (commonly known as the "Youth Talent Program") is a talent development initiative jointly implemented by the China Association for Science and Technology (CAST) and the Ministry of Education of the People's Republic of China (MOE) since 2013. Its purpose is to select outstanding high school students with strengths and potential in the basic disciplines of mathematics, physics, chemistry, biology, and computer science, and foster their scientific literacy and innovative capabilities through research activities conducted on weekends and vacations under the guidance of renowned scientists.

As of 2024, 58 universities nationwide have participated in the program, nurturing approximately 1,800 high school students. After entering university, many of these students pursue advanced studies in basic discipline fields or join the "Program for Cultivating Top Students in Basic Disciplines," achieving remarkable educational outcomes.

Related Official Links

Official Link of the Youth Talent Program
https://zxsycjh-kp.cast.org.cn/front/home
Personal Homepage of the Faculty Member at Zhejiang University
https://person.zju.edu.cn/hxchem
Introduction to the Hong Xin Research Group, Department of Chemistry, Zhejiang University
https://www.x-mol.com/groups/HongGroup

Group Photo of Participating Students in the Chemistry Group of the Youth Talent Program

From left to right in the back row:
2nd person: Mr. Guo Yeming from the Zhejiang Association for Science and Technology
6th person: Professor Lu Xin, Vice Dean of the Chu Kochen Honors College, Zhejiang University
7th person: Professor Guo Zijian, Academician of the Chinese Academy of Sciences (CAS)
8th person: Professor Ma Shengming, Academician of the Chinese Academy of Sciences (CAS)
12th person: Professor Hong Xin from the Department of Chemistry, Zhejiang University (my supervisor)
I am the 2nd person from the right in the front row.