英才计划（2024 届化学）
培养报告

课题：**化学反应的定量构效关系建模**

学员：**张一超（浙江省杭州第二中学）**

导师：**洪鑫（浙江大学）**

2024.10.17

Contents 目录

中学生科技创新后备人才培养计划，简称中学生英才计划，是**中国科学技术协会和中华人民共和国教育部**自 2013 年开始共同组织实施的人才培养计划 。旨在发现一批具有学科特长、创新潜质的优秀中学生。

培养概况

学员介绍

张一超，就读于浙江省杭州第二中学；
2024 年 2 月入选中学生英才计划化学学科，师从浙江大学化学系洪鑫教授；
2024 年 7 月**参加浙江大学竺可桢学院“英才计划”夏令营被并评为“优秀营员”**；

课题简介

以课题“化学反应的定量构效关系建模”为核心，基于信息学、统计学的多学科结合基础，综合运用机器学习（ML）对化学学科进行学习和研究。

主要方向：**化学的多学科结合应用；机器学习在化学领域的应用；符号回归的化学应用（SISSO）**

课题包含 2 个实操应用项目和 2 个文献学习项目。分别是：
文献学习项目：数据驱动模型对化学反应的定量预测
实操应用项目：依非伟伦合成过程中不对称酰胺化反应的碱性催化剂种类对反应影响的多元线性回归模型构建与分析
文献学习项目：符号回归（Symbolic Regression）的机理与化学领域的应用
实操应用项目：镍催化的 C–H 烷基化反应中 HASPO 预配体对映体选择性模型

主要培养能力：
1. 多学科结合的广泛学习能力，培养跨学科知识应用能力，注重知识的多领域应用能力；
2. 化学学科的综合能力，掌握一定的文献阅读能力；并通过浙江大学组织的多次化学学科前沿讲座了解前沿化学动态；在浙江大学的学习过程中参加数次基础实验活动，培养一定的化学实验能力；
3. 信息化学的操作能力，学习机器学习（Medicine Learning）的基本原理，掌握 Python、SPSS、SISSO 等信息化学技术和符号回归（Symbolic Regression）等信息技术实践应用；
4. 统计学方面的建模能力，掌握 QSAR 和 QSPR 的基本操作思路，对各类线性和非线性模型的建模、检验、应用进行学习和应用，尝试应用主成分回归、逐步回归等线性建模，尝试应用基于 SISSO 的符号回归技术。

文献学习项目报告一

数据驱动模型对化学反应的定量预测

浙江省杭州第二中学 张一超

数据驱动模型在化学反应的定量预测中发挥着日渐明显的作用，尤其是在发现新型手性催化剂、理解反应机理和设计新分子材料等领域，体现了强大的应用价值。而这些模型通常基于定量构效关系（QSAR）和定量构性关系（QSPR）的原理构建。

QSAR（Quantitative Structure-Activity Relationship）是一种建模概念，用于预测化学物质的生物活性或反应活性。它通过统计方法关联分子的结构特征与它们的活性。QSAR 模型通常可用于药物设计、化学品合成工业、环境污染物的风险评估等领域。

QSPR（Quantitative Structure-Property Relationship）与 QSAR 类似，但关注的是化学物质的物理化学性质，如溶解度、沸点、反应速率等。QSPR 模型有助于理解分子结构与它们所表现的物理化学性质之间的关系，具有用实验数据的整合反哺理论的发展和完善的潜力，对更深层次认识物质有重要作用。

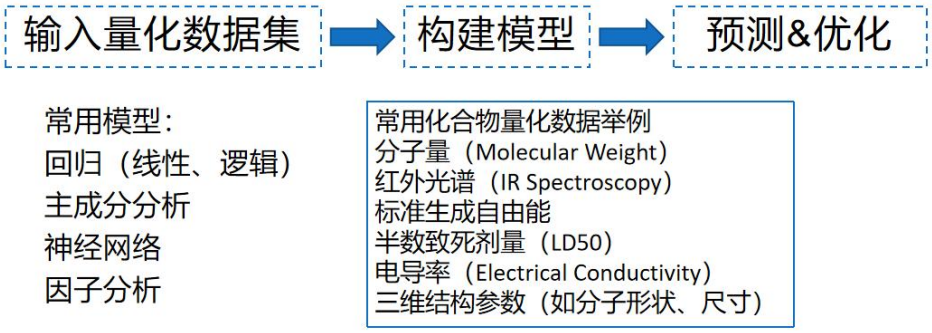


图 1 数据驱动模型建模的初级思路

数据驱动模型在化学领域发挥着至关重要的作用，它们通过分析大量的化学数据来预测和设计新药物，构建药物毒性或药性模型，从而加速药物发现过程并提高其安全性；在不对称反应催化剂的预测中，这些模型能够处理选择性与催化剂量化数据，识别影响反应选择性的关键特征点，进而优化催化剂设计，提高反应的立体选择性；数据驱动模型还能规划合成反应环境，通过输入合成效率与环境量化特征，计算出最高效率的环境条件，以实现绿色化学和可持续合成；在配体特征优化方面，模型可以指导配体结构的调整，优化对映体选择性，这对于手性药物的合成尤为重要；物理化学性质的机理研究也受益于数据驱动模型，它们能够通过模型对已有理论进行改良和完善，提供更深入的机理理解。

以一篇发表在 Nature Reviews Chemistry 的综述论文为例（Reid, J. P., & Sigman, M. S. (2018).）。这篇文章探讨了定量预测技术在小分子手性催化剂发现中的应用，评估了不同方法在预测催化剂性能方面的准确性和可靠性，充分证明了数据

驱动模型在化学领域具有的重要作用。

同样，另一篇发表在 Accounts of Chemical Research 的综述论文（Crawford, J. M., Kingston, C., Toste, F. D., & Sigman, M. S. (2021).）则探讨了数据驱动模型在设计新型催化剂的应用，具有提高反应的效率和选择性的作用，并且特别关注了数据驱动模型在手性合成方面的应用。

数据驱动模型已经成为化学研究中不可或缺的工具，它们不仅已经在药物设计、催化剂发现、合成路径优化等领域展现出巨大的潜力，也在物理化学研究方面提供了新的视角。随着算法的不断进步，数据驱动模型将在化学领域的研究和应用中将发挥更加重要的作用，不断完善化学的信息化发展。

参考文献：

[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

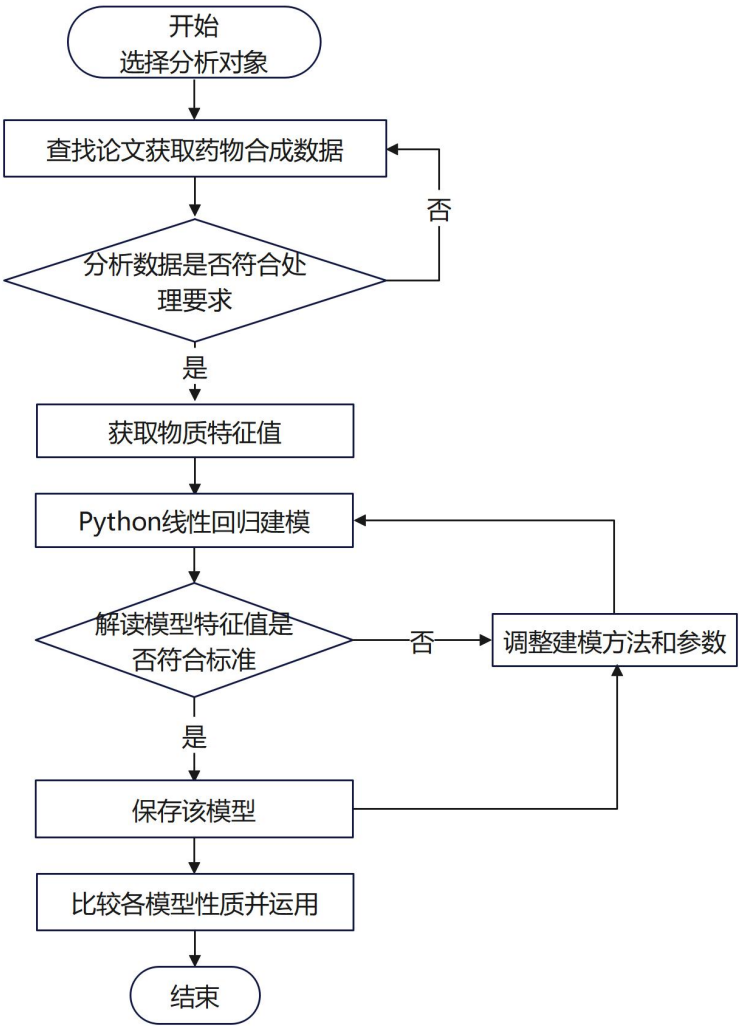
[6] Williams, W. L., Zeng, L., Gensch, T., Sigman, M. S., Doyle, A. G., & Anslyn, E. V. (2021). The Evolution of Data-Driven Modeling in Organic Chemistry. ACS Central Science, 7(7), 1622–1637. https://doi.org/10.1021/acscentsci.1c0053

实操应用项目报告二

依非伟伦合成过程中不对称酰胺化反应的碱性催化剂种类对反应影响的多元线性回归模型构建与分析

一、项目操作思路

图 2 项目操作思路示意图



二、项目报告

依非伟伦合成过程中不对称酰胺化反应的碱性催化剂种类对反应影响的多元线性回归模型构建与分析

浙江省杭州第二中学 张一超

一、依非韦伦

依非伟伦（Efavirenz）是一种抗逆转录病毒药物，用于治疗人类免疫缺陷病毒（HIV）感染。它属于非核苷类逆转录酶抑制剂，通过非竞争性地结合到 HIV-1 逆转录酶上，抑制病毒复制过程中的 DNA 合成，从而减缓病毒在体内的增殖。[18][24][25]

二、依非伟伦的不对称合成方法

根据《抗艾滋病药物依非韦伦（Efavirenz）的合成工艺研究》[18]一文中对该药物的合成方案设计，建立了一种提供不对称反应合成依非韦伦的路径。

1.起始原料：使用 2-三氟乙酰基-4-氯苯胺作为起始原料。

2.格氏试剂的制备：通过氯丁烷与金属镁的反应制备正丁基氯化镁格氏试剂。

3.不对称加成：在氢化钠、氯化锌、三氟乙醇和手性配体(1R,2S)-1-苯基-2-(1-吡咯烷基)-1-丙醇（Lig A）的作用下，与环丙乙炔基氯化镁进行反应，形成以锌原子为中心的配位化合物，然后与起始原料进行不对称加成反应，生成关键中间体(S)-1-(2-氨基-5-氯苯基)-1-三氟甲基-3-环丙基-2-丙炔-1-醇。

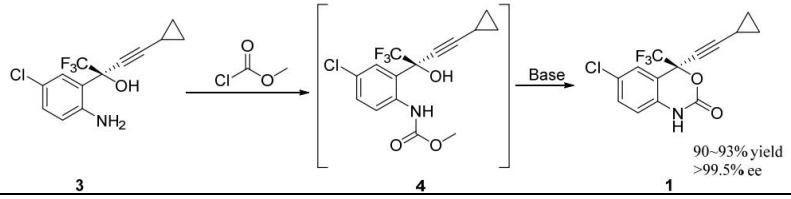
4.酰胺化缩合：将上述得到的手性中间体与氯甲酸甲酯反应，形成中间体酰胺化合物。

5.环合反应：在叔丁醇钾的催化下，通过环合反应生成目标产物依非韦伦。

6.精制：通过后续的洗涤、脱水、脱色和结晶步骤，最终得到高收率、高纯度的依非韦伦产品。

依非伟伦的不对称合成方法[11][18]

化合物 3 与氯甲酸甲酯反应后，生成中间体酰胺化合物 4，经脱水处理后，在碱性试剂的催化下环合生成目标产物 1 依非韦伦。

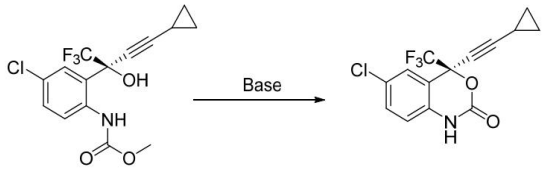


3 4 1

三、酰胺化缩合反应过程

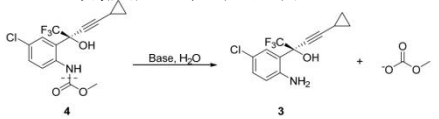
酰胺化缩合的反应过程[18][21][24]

主要反应



4 1

副反应生产化合物 3



4 3

四、不同碱性催化剂对反应的影响

表：不同碱性催化剂对反应的影响[18]

序号	催化剂	转化率①/%	化合物 3/%	选择性②/%
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化合物 4 的转化率；②依非韦伦的选择性。

备注：以 210 g（0.73 mol）化合物 3 投料，做至减压回流分水结束后（具体实验步骤见 3.3.2 节），将溶液平均分为 7 份，分别加入各实验组对应的碱性催化剂 1.2 g（50%NaOH 溶液 加入 2.4 g），升温至 45~50℃反应 4 小时。

五、碱性催化剂性质的量化数据获得

通过 RDKit 数据库，获得不同碱性催化剂的多种量化性质数据，包括分子质量 (Molecular Weight) 、拓扑极性表面积 (Topological Polar Surface Area, TPSA)、可旋转键数 (NumRotatableBonds)、重原子数 (HeavyAtomCount)、杂原子数 (NumHeteroatoms) 、CSP3 碳原子比例 (FractionCSP3)、脂溶性 (LogP)。

主要代码（部分代码省略）

```
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+]"
}
```

可通过 Python 得到不同碱性催化剂的性质数据。由于反应组 2 与反应组 3 中 NaOH 的浓度不一致，因此将加入的催化剂质量与浓度均设为自变量考虑。

催化剂	质量 g	浓度	分子质量 g/mol	拓扑极性表面积 Å²	可旋转键数	重原子数	杂原子数	CSP3 碳原子比例	脂溶性 LogP
无	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

Na2CO3	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
NaOCH3	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

六、多元线性回归模型的构建

通过 Python，分别以“转化率”“化合物 3”“选择性”作为因变量，碱性催化剂的物质量化数据作为自变量，进行多元线性回归分析。

主要构建模型代码（部分代码省略）

```
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())
```

七、多元线性回归结果

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

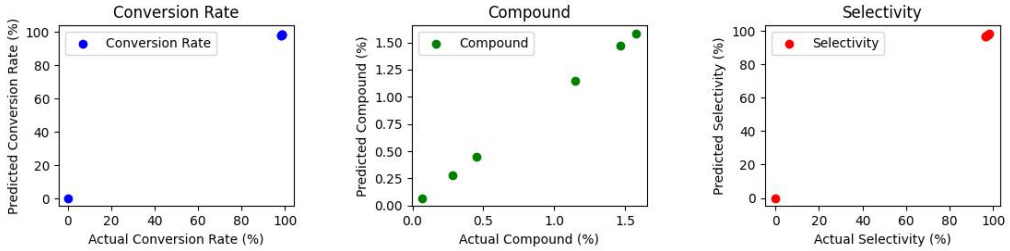
OLS Regression Results

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		

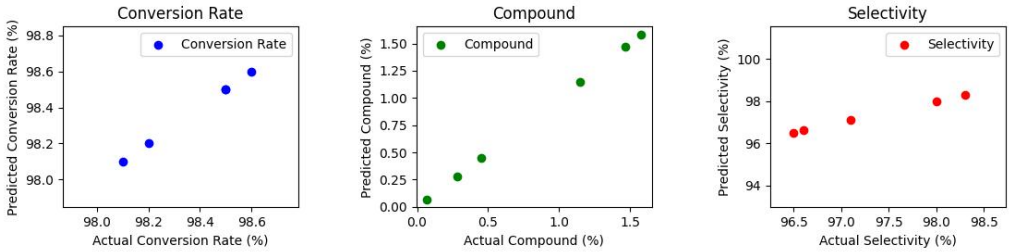
	coef	std err	t	P>|t|	[0.025	0.975]
const	2.915e-14	inf	0	nan	nan	nan
MolecularWeight	-1.5635	inf	-0	nan	nan	nan
TPSA	10.2981	inf	0	nan	nan	nan
NumRotatableBonds	-5.236e-14	inf	-0	nan	nan	nan
HeavyAtomCount	42.6172	inf	0	nan	nan	nan
NumHeteroatoms	-167.1690	inf	-0	nan	nan	nan
FractionCSP3	25.6843	inf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



说明：Picture1-1 中，由于部分样本量处于极其接近位置，故选取较大量度绘制全局图；在 Picture1-2 中，则放大样本密集区域，更明显地展现多元线性回归的较高拟合性。

多元线性回归模型式

1.转化率（ConversionRate）模型: 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.化合物（Compound）模型: 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.选择性（Selectivity）模型: 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

八、多元线性回归的分析

根据 OLS 回归结果，R-squared 值为 1.000，这意味着模型对数据的拟合非常好，解释了因变量方差的 100%。这是一个非常强的拟合度，表明模型可以很好地解释因变量的变化。这对于预测和解释因变量的变化非常有意义。然而，尽管出现了警告和异常值，但我们仍然可以从回归结果中得出一些结论。例如，R-squared 值非常接近 1，这意味着模型对数据拟合得非常好。这表明我们的模型能够很好地解释因变量（ConversionRate、Compound、Selectivity）的变化，这是一个积极的方面。[7][8][9]

由于样本量较小，我们需要谨慎解释系数的显著性和置信区间。同时，R-squared 值已经非常接近 1，因此可能需要考虑是否存在过拟合的情况。过拟合可能会导致模型在新数据上的预测能力下降。

在这种情况下，应当增加样本数量，增强模型稳健性。

参考文献

[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."

文献学习项目报告三

符号回归（Symbolic Regression）的机理和化学领域运用

浙江省杭州第二中学 张一超

符号回归（Symbolic Regression, SR）是一种强大的机器学习方法，它能够从数据中推断出数学表达式，揭示数据背后的潜在规律。与传统的数值回归方法不同，符号回归不仅能够提供预测模型，还能够提供解释性，帮助我们理解数据之间的关系。

符号回归是一种从数据中自动发现数学模型的方法。它通过搜索可能的数学表达式空间，找到最佳匹配数据的表达式。表达式的运算符号可以自主选择，具有良好的适应性，也便于调整。

符号回归与传统回归方法存在显著的优点。传统回归受制于统计方法的局限性，不如符号回归更具有解释性。但二者本质上都属于机器学习，也都具有机器学习的基本优势，对数据处理有重要作用。

符号回归的主要实现途径包括遗传编程（GP）、神经网络和梯度提升决策树（GBDT）。这些方法各有优势，在不同的问题和数据类型上具有不同的功能。

符号回归模型的构建工具有很多种。SISSO（Sure Independence Screening and Sparsity Operator）是一种基于压缩感知的算法，专门设计用于从大量的候选描述符中识别出最优的低维描述符。gplearn 是一个基于 Python 的遗传编程库，它不仅支持符号回归，还提供了分类和特征构造等功能，使得它成为一个多功能的机器学习工具。PySR 是一个开源的符号回归工具，它利用遗传编程来揭示数据中的数学关系，特别适用于需要高度解释性模型的场景。

符号回归技术在化学领域具有广泛的应用。在材料化学中，它被用来预测材料的物理和化学性质，从而指导新材料的开发和现有材料性能的改进。在化学合成路径优化方面，符号回归能够分析不同合成步骤中的反应条件和产物数据，提出改进措施，以减少副反应、提高产物收率和纯度。同时，其也在配体选择、靶向药物设计等方面可以发挥作用。

而由上海大学欧阳润海教授研发的 SISSO 算法，不仅具有较强的可操作性，而且在多个化学领域已经取得了重要的成果。根据相关介绍，SISSO 方法已被用于钙钛矿材料、拓扑绝缘体、催化材料、超导、二维材料、聚合物等的模型构建和新材料预测。

参考文献

[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

实操应用项目报告四

镍催化的 C-H 烷基化反应中 HASPO 预配体对映体选择性模型

一、项目操作思路

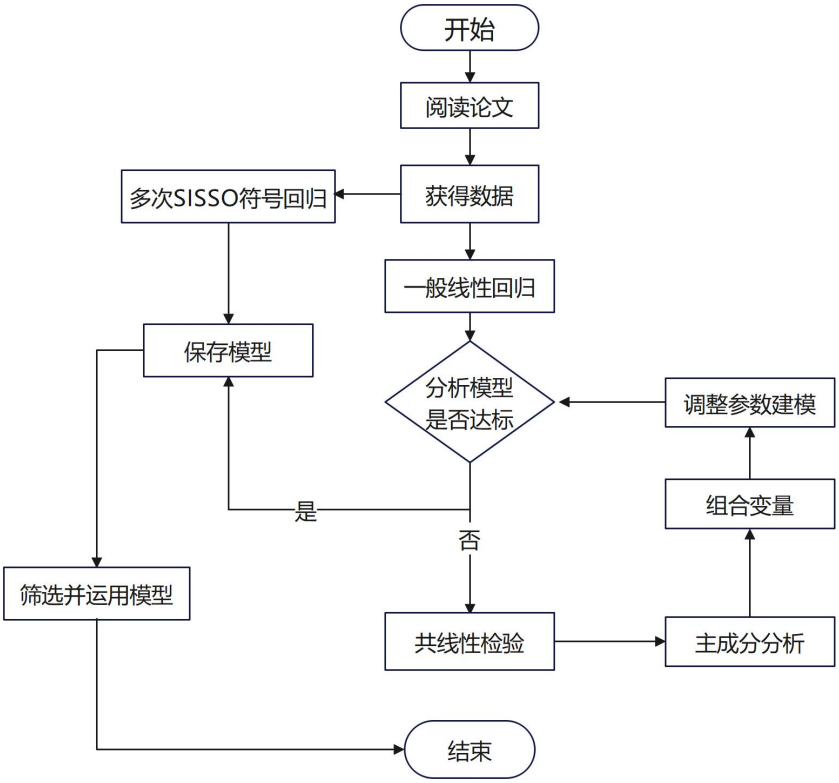


图 3 项目操作思路示意图

【本项目使用到的相关工具概述】

SPSS：https://www.ibm.com/spss

一款由 IBM 公司研发的，用于数据分析和统计学研究的软件；

SISSO：https://github.com/rouyang2017/SISSO

一款由上海大学欧阳润海教授研发的用于符号约束回归的工具。

二、项目报告

镍催化的 C−H 烷基化反应中 HASPO 预配体对映体选择性模型

浙江省杭州第二中学 张一超

说明：

本项目基于 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.论文的数据进行研究，特此向洪鑫教授和俞塽学长提供的帮助表示感谢。

同时，由于部分数据内容较多，本报告仅展示主要部分和部分摘录，具体内容和数据详见附件。

Introduction 论文内容回顾

本文介绍了一种创新的镍催化方法，通过使用手性杂原子取代的二级膦氧化物（HASPO）配体辅助的 Ni−Al 双金属催化剂，实现了 C−H 烷基化反应，用于高效构建 C−N 轴手性化合物。研究团队通过优化反应条件，筛选出具有高对映体选择性的 HASPO 配体，显著提升了反应的产率和立体选择性。论文还展示了该方法对多种底物的普适性，包括不同的 N-芳基取代的苯并咪唑和各种烯烃。通过实验和密度泛函理论（DFT）计算，作者揭示了反应机理，包括配体到配体的氢转移和还原消除步骤。此外，利用多变量线性回归（MVLR）分析，研究了 HASPO 配体的结构与对映体选择性之间的关系，为未来的配体设计和优化提供了理论依据。这项研究不仅提供了一种成本效益、低毒性的催化方法，而且为合成具有生物活性和应用潜力的轴手性化合物开辟了新的途径。

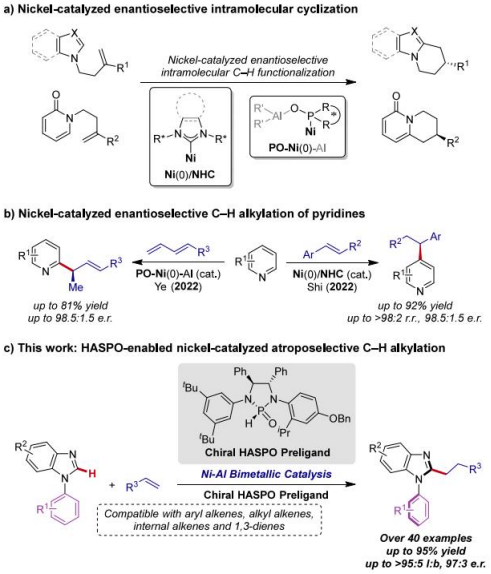


图 4 文中合成过程简要示意图

HASPO 配体在镍催化的 atroposelective C−H 烷基化反应中发挥着核心作用，其具体作用体现在以下几个方面：作为手性源，HASPO 配体提供必要的手性环境，确保了反应的高对映体选择性;通过与镍催化剂形成配合物，HASPO 配体通过配位作用调节催化剂的电子和立体化学性质，影响催化活性和选择性；HASPO 配体的立体化学和空间结构对反应的立体控制起着决定性作用，其非 C2 对称性特征在立体选择性控制中尤为关键；HASPO 配体的结构特征有助于识别和稳定底物或中间体，提高化学选择性和区域选择性；在催化循环中，HASPO 配体参与了催化剂的初始形成、中间体的稳定以及最终产物的释放；通过优化 HASPO 配体的结构，可以提高催化剂的效率和稳定性，从而提升反应产率；合适的 HASPO 配体可以使得反应在温和条件下进行，减少副反应，提高目标产物的选择性。HASPO 配体是实现这一高立体选择性反应的关键组分，通过其多方面的作用，为合成具有潜在应用价值的轴手性化合物提供了有效途径。

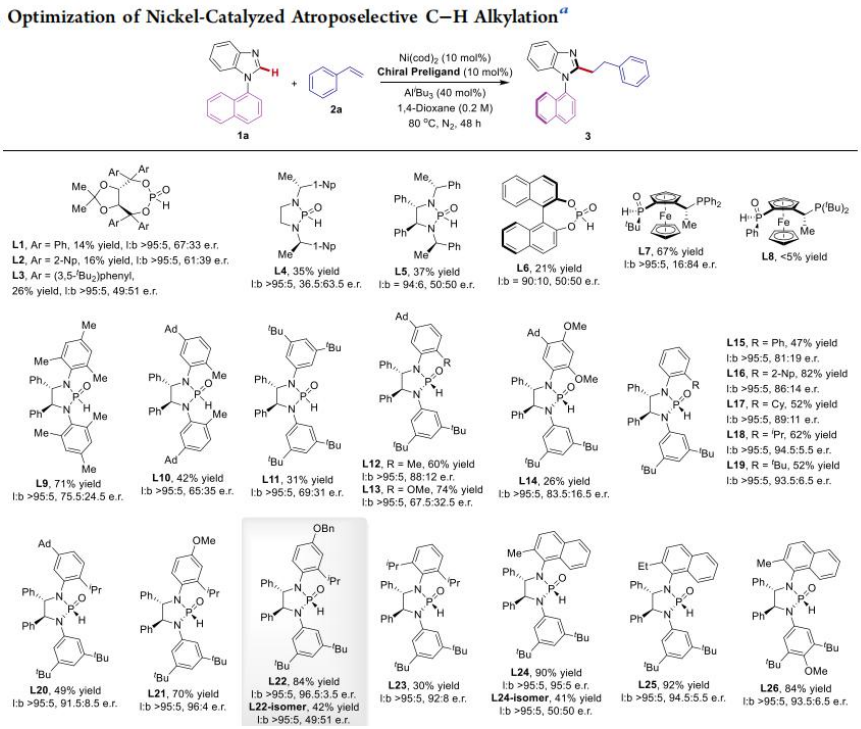


图 5 文中的配体与相关数据摘录

通过 https://github.com/zju-ys/Nickel-MLR/blob/main/PhysOrg.csv 获得文中各类配体的实验数据。

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

图 6 获得的配体反应数据

Part 1：基于 SPSS 的线性回归

第一步：一般线性回归模型

对数据进行一般回归线性建模，分析反馈数据认为其存在较强共线性。

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

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

第二步：数据检验

对数据进行多重共线性和皮尔逊检验，以达到选择主成分或整合变量的目的，从而削弱共线性对模型准确性的影响。

多重共线性检验

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

皮尔逊检验

皮尔逊相关性检验

		B1_8_10	B1_9_11	B5_8_10	B5_9_11	L_8_10	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	.389	.243	-.015	.322	.046	-.222	.312	.416	.033	.509*	.310	.147	-.232	-.554**	-.296
	显著性（双尾）		.580	.818	.305	.074	.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	22
B1_9_11	皮尔逊相关性	-.125	1	.219	-.306	.070	.245	.431*	-.581**	-.139	-.134	.256	-.208	.926**	.182	.509*	.051	-.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	22
B5_8_10	皮尔逊相关性	-.052	.219	1	.431*	.219	.003	-.226	-.378	-.267	-.534*	.499*	.251	.290	-.020	.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	.064	.849	.747
	个案数	22	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	-.336	.048	.109	-.194	.111	.179	-.255	-.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	22
L_8_10	皮尔逊相关性	.389	.070	.219	.148	1	.504*	.026	.379	.004	-.393	.322*	.584**	.285	.490*	.591**	.017	-.575**	-.640**	-.547**
	显著性（双尾）	.074	.756	.328	.510		.017	.908	.309	.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	22
L_9_11	皮尔逊相关性	.243	.245	.003	.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*	-.226	-.336	.026	.549**	1	.210	.326	.396	-.077	-.548**	.433*	.440*	.626**	-.377	.092	-.123	-.610**
	显著性（双尾）	.948	.045	.138	.134	.908	.008		.349	.139	.068	.735	.008	.044	.041	.001	.083	.685	.587	.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	.379	.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

第三步：逐步回归模型

对数据进行逐步回归建模，选择主要关键变量，从而削弱共线性对模型准确性的影响。

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

第四步：组合变量回归模型（主成分）

对数据进行组合变量回归建模预分析，研判变量组合：

B1_8_10 和 B1_9_11：这两个变量的相关系数为-0.125，虽然不是很高，但在统计上显著（P 值为 0.580，不显著），并且它们与多个其他变量都有较高的相关性，可能需要进一步考察。

BV_10 和 BV_11：这两个变量的相关系数为 0.431，这是一个中等程度的相关性，并且在统计上显著（P 值为 0.045，显著）。

q_1 和 q_2：它们之间的相关系数为-0.466，这是一个较强的负相关性，并且在统计上显著（P 值为 0.029，显著）。

q_3 和 q_8：它们之间的相关系数为 0.644，这是一个很强的正相关性，并且在统计上显著（P 值为 0.001，显著）。

d_1_2 和 d_2_3：它们之间的相关系数为 0.811，这是一个非常强的正相关性，并且在统计上显著（P 值为 0.000，显著）。

第五步：一次组合变量回归模型

对数据进行一次组合变量回归建模。

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*

主成分 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

第六步：三次组合变量回归模型

对数据进行三次组合变量回归建模。

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*

主成分 2 (PC2):

*PC2 = 0.561 * BV11*

ddg=-16.368+0.52×PC13

ddg=−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

第七步：多次组合变量回归模型

对数据进行多次组合变量回归建模。

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*

主成分 2 (PC2):

*PC2 = 0.561 * BV11*

ddg=261.178+5.432×PC13−275.192×PC23−294.506×PC11/2+252.284×PC21/2

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)				

第八步：整理和选择模型

对数据进行整理与选择，得到符合需求的模型并将其投入运用。

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×$PC1^3$		三次	0.360	/	/
ddg=261.178+5.432×$PC1^3$−275.192×$PC2^3$−294.506×$PC1^{1/2}$+252.284×$PC2^{1/2}$	主成分2 (PC2): PC2 = 0.561 * BV11	多次	0.664	/	/

Part 2：基于 SISSO 的符号回归

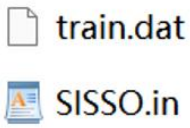
第一步：系统配置与数据预处理

根据 SISSO Guide 标准与相关指导，进行 Linux 环境下 SISSO 安装和调试。

再根据范例与 Guide 指导，将 csv 数据文件格式转换为符合 SISSO 要求的 train.dat 文件。

再根据标准配置 SISSO.in 文件。

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

第二步：调整参数

根据 SISSO Guide 标准与相关指导，配置 SISSO.in 文件中的各项参数。以下是参数项的说明和解释：

ptype: 目标属性的类型。1 代表具有连续属性的回归；2 代表具有分类属性的分类。

ntask: 任务的数量。ntask = 1 表示单个任务的常规机器学习；ntask > 1 表示多任务学习（MTL）。

task_weighting: (仅回归) 多任务学习回归中的任务加权。1 表示没有加权（所有任务无论其数据集大小都同等对待）；2 表示每个任务根据其数据量与所有任务的总数据量的比值进行加权。

scmt: (仅回归) 如果设置为 .true.，则调用有符号约束的多任务回归。

desc_dim: 描述符或模型的维度。

nsample: train.dat 中的样本数量。对于单任务回归，输入仅为一个整数。对于多任务回归，输入是以逗号分隔的多个整数，表示各个任务的样本数量。对于单任务分类，将只有一个圆括号内的整数，表示各个类别的数据数量。对于多任务分类，将有多个以逗号分隔的圆括号，每个括号定义一个任务，里面的整数定义该类别的数据数量。

restart: 重新开始或继续作业。0 表示从头开始作业，1 表示继续作业（上次作业的进度信息存储在 CONTINUE 文件中）。

nsf: train.dat 中提供的标量特征的数量。"标量特征"意味着数据集中的每个特征占一列。

*ops: 用于特征构建的数学运算符。用户可以从列表中自定义运算符，例如 {+, -, *, /, exp, exp-, ^-1, ^2, ^3, sqrt, cbrt, log, |-|, scd, ^6, sin, cos}。*

fcomplexity: 特征复杂性定义为特征中的运算符数量。fcomplexity=0 表示特征空间中只有输入变量，而 fcomplexity=3 表示特征空间中所有特征的复杂性不超过 3。

funit: 表示 train.dat 文件中特征具有相同单位的类型。例如，funit=(1:5)(6:9)(11:11) 表示第 1 到第 5 个特征具有相同的单位，第 6 到第 9 个特征具有另一种单位，第 10 个特征是无量纲的，第 11 个特征具有不同的单位。

fmax_min 和 fmax_max: 分别是特征数据中的最大绝对值的阈值，如果小于 fmax_min 则视为零特征并丢弃，如果大于 fmax_max 则视为无穷大特征并丢弃。

nf_sis: SIS-子空间的大小。对于 SISSO-nD 计算，将有 n 个 SIS-子空间。

method_so: 稀疏算子的方法，可以是 L0 范数最小化稀疏化方法，或者仅用于回归的 L1L0 方法。
nl1l0: (仅回归) 使用 LASSO 选择的特征数量，用于后续的 L0。
fit_intercept: (仅回归) 线性模型是否拟合非零/零截距。
metric: (仅回归) 回归中模型选择使用的度量，可以是 RMSE（均方根误差）或 MaxAE（最大绝对误差）。
nmodels: 输出排名最高的模型数量。
isconvex: (仅分类) 每个数据域是否可以被限制为凸形或非凸形。
bwidth: (仅分类) 每个域的边界容差，以包含非常接近但位于域外的数据。

第三步：配置参数

根据 SISSO Guide 标准与相关指导，配置 SISSO.in 文件中的各项参数。以一次配置为例：

```
! 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)
```

第四步：解读数据

在 Ubuntu 中，运行 SISSO 后会将结果保存在 SISSO.out 中。通过解读 SISSO.out 的报告，获得回归结果。以下是某一次运行的结果摘要：

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

第四步：整理模型

经过多次调整参数，将数据不断整合，获得以下各个模型：

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))								

参考文献：

[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/

三、项目数据

（1）实验源数据：

由浙江大学俞璐提供 https://github.com/zju-ys/Nickel-MLR

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	8.8250532	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

（2）线性回归模型（SPSS）

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）	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	PC1 = 0.861 * BV10 + 0.929 * q1 + 0.859 * q2 + 0.883 * q3 + 0.928 * q8 + 0.923 * d12 + 0.906 * d23	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	/	/

（3）符号回归模型（SISSO）

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))								

培养记录

培养情况

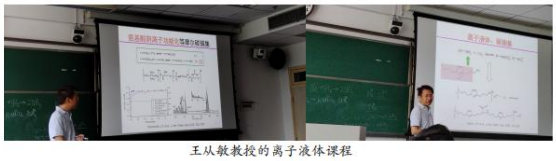
前沿讲座

自参加英才计划以来，累计在浙江大学参加周末化学讲座近十次，参加线下实验活动数次。期间，由 2 位中国科学院院士为我们带来面对面讲座。

2024 年暑期参加浙江大学英才计划夏令营，共近一周时间，期间参加化学和其他学科综合讲座与报告十余次。

前沿学科报告带来了丰富生动的化学知识，不同领域和方向的教授们带来不同的分享内容，每一次报告和学习都让我们收获颇丰。

图 7 院士讲座现场



王从敏教授的离子液体课程

李鹏飞教授的生物催化课程

王林军教授的智能化学讲座

图 8 多次前沿讲座

拓展视野

在浙江大学何巧红老师的指导下，由浙江大学化学系实验老师带领，多次参观浙江大学化学系各实验室和科研平台，开展基础实验技能学习实践。同时多次在各位教授的带领下进入实验室进行学习了解。

图 9 实验室现场学习实践

组内培养

师从浙江大学洪鑫教授，与洪教授通过线上线下多种方式展开学习。

在洪教授的指导和帮助下，进行学习和研究，并顺利完成课题各项目。

由洪教授指导，在洪教授组内研究生俞璞等同学的帮助下，开展各项课题的研究。

于暑假间线下到实验室参加数次学习，并通过线上方式保持深度学习，按时完成多个项目。

浙江大学英才计划暑期活动记录

浙江省杭州第二中学 张一超（化学）

7 月 8 日至 12 日，我们在浙江大学紫金港校区参加了英才计划暑期活动。英才计划暑期活动丰富多彩，意义非凡。我们通过五天的学习生活，探索了前沿知识，感受到科学研究的魅力，对各自的课题有了进一步的认识。

在石虎山机器人研究基地看到的机器狗表演

暑期活动主要包含知识讲座、大学选修课、微专题研究、参观前沿基地等集体活动，同时也提供了丰富的时间供我们与导师交流谢谢。在浙江大学竺可桢学院的组织下，我们通过各个领域的教授的热情分享，了解到不同学科的前沿动态，感受到科学研究的创新方法。比如在叶高翔教授的报告中，我们了解到哲学体系与科学发展相促进的重要性；在连续多天的大学先修课上，陈锦辉教授带领我们初步认识了微积分等高等数学知识；在唐建军教授的生态主题讲座中，我们感受到生态研究的趣味性。众多专家学者为我们带来一场前沿知识的盛宴，让我们对探索科学知识充满兴趣。

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

同时，在由各个学科分组进行的微专题研究中，不同学科组织了丰富的活动。以我

参加的化学组为例，王从敏教授为我们介绍了离子液体对环境保护的重要性，指出了化学学科在双碳行动中的重要性；季鹏飞教授通过生物与化学结合，通过生物催化的方式展开材料的研究；来自王林军教授的分享中，我们看到了不一样的化学世界，认识到化学与多学科结合的创新领域的蓬勃发展。在三次微专题研究活动结束后，我们分组进行了微专题展示活动，化学组的同学介绍了各自的课题与研究，从无机催化到有机药物，从实验化学到计算化学，我们对化学各领域有了更深的了解，也培养了我们相互交流的能力。

活动回顾与展望

自 2024 年 2 月正式加入英才计划以来，我顺利参加所有学习研究活动。这一年内，我参加了由浙江大学组织的各类前沿科学讲座、报告数十次，并且在夏令营活动中参加了大学课程先修课。

我在一次次活动中不断拓宽我的眼界，丰富知识面，不仅对化学学科的认识更加深入，也对其他科学学科进行了学习和拓展。

在这一年中，我最大的收获，是对当今多学科结合的研究大趋势有了更全面的理解，也通过自己的实践参与其中，感受多学科交叉发展对科学的重要意义。

我有幸跟随浙江大学化学系洪鑫教授进行学习和研究，在洪老师的帮助下，我一步步地开展项目。科学学习的过程中必然会有困难和挫折，但在师长的无私帮助下，通过坚持不懈的挑战，总可以有所收获。由于我没有很好的编程基础，在进行项目二的工作时我面临巨大的困难，但一步步学习和解读，最终完善代码的过程，带给我的不仅是学科能力的提升，也是精神意志的打磨。

在活动中我对化学的学科特征有了更深的认识：化学是一门尝试的学科，也是一门覆盖面极广的学科。从化学的发展历史上来看，一次次的试错和突破，为人类科学事业带来不断的进步。如果没有尝试，化学的基础理论研究便难以不断推进，也难以赋能生产技术。同时化学也是覆盖面极其广阔的学科，传统意义上的四大化学，即有机化学、无机化学、计算化学、物理化学的边界正在打破，越来越多综合的研究方法正在投入实践；而化学的研究方向则更为广阔，从生活中的一切化工产品，到新能源行业的创新发展，一切都离不开化学。随着时间的推移，化学将发挥更加重要的作用，也必然会激发更多层次的创新。

在这个人工智能技术取得蓬勃发展的时代，化学的深层次发展将进一步推进。如果是机器学习（ML）是人工智能（AI）技术的前一步，那我们也足以从中见识到技术的巨大作用。从最简单的机器学习模型，到如今神经网络技术的推广，不仅仅是化学学科发生了翻天覆地的变化，所有学科都在产生新一轮的变革。以 2024 年诺奖为例，蛋白质预测模型正是这样一种技术的应用。在数十年前可能需要多年工作才能得出的结果，如今或许只需几个小时。这不仅仅是技术的突破，也是创新能力的突破。或许未来简单化的思维将难以应对挑战，我们需要的不是单一的科学思维，而是全面综合的科学视角。多学科结合与智能技术，将为科学研究带来更加深远的变化。

参加英才计划，让我对未来有了更明晰的视野。我将带着英才计划带给我的新思维、新知识，不断挑战和超越自己，培养更为全面的能力，更加重视理论学习与实践的结合，继续进步。

致谢

在英才计划的培养期间，特别感谢浙江省科协的各位负责老师与浙江大学化学系的多位教授和老师，为我们组织和安排了一系列丰富多彩的活动；同时向浙江大学竺可桢学院组织英才计划夏令营的志愿者表示感谢。

我的导师，浙江大学化学系洪鑫教授，在我的学习和研究方面给予了丰富的指导和帮助，为我的项目指明方向。洪鑫教授组内的研究生学长提供了软件安装和数据方面的帮助。特此表示感谢。