英才計画（2024 年度化学科）
育成と研究の報告書

課題：化学反应の定量的構造活性相関モデリング

受講生：張一超（浙江省杭州第二高等学校）

指導教授：洪鑫（浙江大学）

2024 年 11 月

目录

育成概要

受講生紹介

張一超 Zhang Yichao

浙江省杭州第二高等学校在籍

2024 年 1 月：中学生英才計画（全国科学技術革新人材育成プログラム）化学科受講生に選抜され（全省で僅か 13 名）、浙江大学化学科の洪鑫教授に師事しました。

2024 年 7 月：浙江大学竺可楨学院「英才計画」サマーキャンプに参加し、「優秀キャンパー」に選ばれました（化学科で計 2 名）。

2024 年 11 月 3 日：浙江大学化学科主催の課題報告会に参加し、化学グループ代表として浙江大学英才計画 5 学科合同の課題答弁会に出席しました。

2024 年 11 月 29 日：浙江省受講生代表（化学科で計 2 名）として、中国科学技術大学で開催された中学生英才計画化学分野フォーラム活動に参加しました。

課題概要

テーマ

化学反応の定量的構造活性相関モデリング

特徴

情報学･統計学との学際融合を基盤とし、機械学習（ML）を化学研究に応用。

研究分野

化学における学際的融合応用の研究
機械学習の化学領域への適用

プロジェクト内容

１．文献研究プロジェクト「データ駆動型モデルによる化学反応の定量的予測」
２．実践応用プロジェクト「エファビレンジン（Efavirenz）合成プロセスにおける不斉アミド化反応に対する塩基性触媒の種類が反応に及ぼす影響の多変量線形回帰モデルの構築と分析」
３．文献研究プロジェクト「シンボリック回帰（Symbolic Regression）のメカニズムと化学分野における応用」
４．実践応用プロジェクト「ニッケル触媒による C−H アルキル化反応における HASPO プレリガンドのエナンチオ選択性モデル」

育成された能力

１．学際的学習能力：複数分野にわたる知識の統合能力及び応用実践能力
２．化学総合的スキル：学術文献の読解・分析能力、浙江大学における先端化学の講座、並びに基礎実験を通じて培われた実践的運用能力
３．情報化学関連技術：Python/SPSS/SISSO を活用した機械学習（ML）及びシンボリック回帰の実装・応用能力
４．統計モデリング能力：QSAR/QSPR（定量的構造活性相関／定量的構造特性相関）に基づく線形・非線形モデル（主成分回帰、段階的回帰法等を含む）の構築、評価及び検証能力

Ⅰ．文献研究プロジェクト

データ駆動型モデルによる化学反応の定量的予測

浙江省杭州第二高等学校 張一超

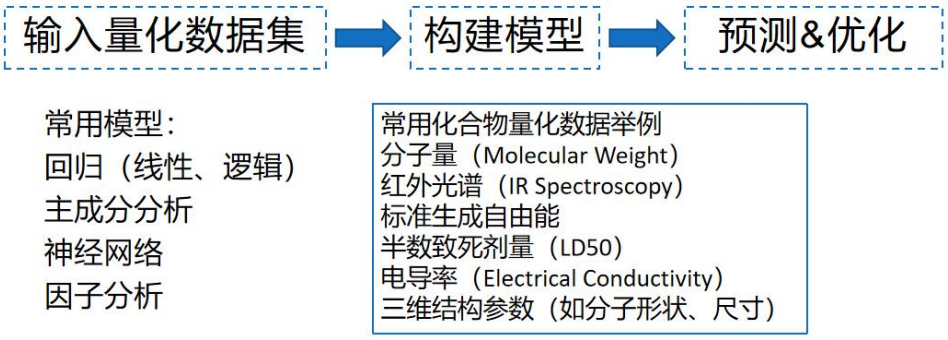
一、序論

データ駆動型モデルは、化学反応の定量予測において重要な役割を果たしており、特に不斉触媒の開発、反応機構の解明、新規分子材料の設計などの分野でその応用価値を示している。これらのモデルは通常、定量構造活性相関（QSAR）および定量構造性質相関（QSPR）の原理に基づいて構築される。

QSAR（Quantitative Structure-Activity Relationship）：化学物質の生物学的活性や反応性を予測するための統計モデリング手法。分子構造（原子数、官能基、立体化学など）と活性の相関を分析し、医薬品設計や環境リスク評価に応用される。

QSPR（Quantitative Structure-Property Relationship）：物理化学的性質（溶解度、沸点、反応速度など）と分子構造の関係をモデル化。実験データの統合により理論の発展を促進し、物質の深層理解に寄与する。

二、データ駆動型モデルの化学応用



図に示すように、データ駆動型モデルは以下の分野で活用される：

医薬品開発：大量の化学データを解析し、薬物毒性や薬効モデルを構築。開発プロセスの加速と安全性向上を実現。

不斉触媒の最適化：反応選択性に影響する特徴量を同定し、立体選択性の高い触媒設計を支援。

グリーンケミストリー：合成効率と環境負荷の定量化により、持続可能な反応条件を提案。

配体設計：対映体選択性を制御する配体構造の調整を指導。不斉合成の効率化に貢献。

物理化学的機構の解明：既存理論の改良を通じ、反応メカニズムの理解を深化させる。

三、関連研究の事例

Nature Reviews Chemistry（Reid & Sigman, 2018）不斉小分子触媒の発見における定量予測手法を比較し、データ駆動型モデルの有効性を実証。

Accounts of Chemical Research（Crawford et al., 2021）データ科学と物理有機化学の融合により、触媒設計の効率化と選択性向上を達成した事例を報告。

四、結論

データ駆動型モデルは、アルゴリズムの進化に伴い、化学研究の情報化・精密化をさらに推進する。医薬品・触媒・合成計画における応用だけでなく、基礎理論の発展にも新たな視座を提供する不可欠なツールである。

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

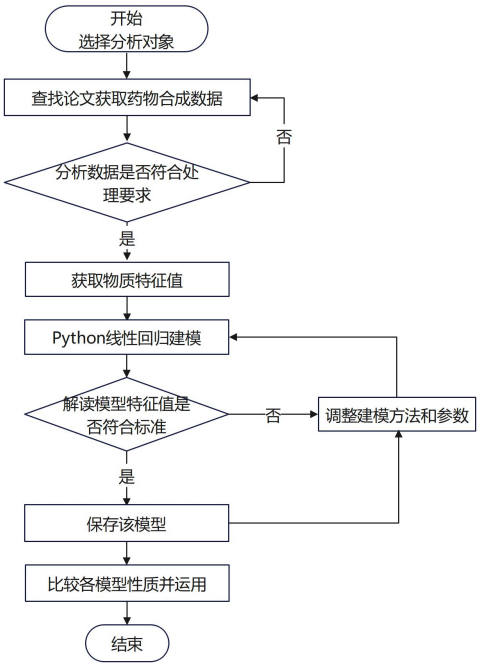
6.Williams, W. L., Zeng, L., Gensch, T., Sigman, M. S., Doyle, A. G., & Anslyn, E. V. (2021). The Evolution of Data-Driven Modeling in Organic Chemistry. ACS Central Science, 7(7), 1622–1637. https://doi.org/10.1021/acscentsci.1c00535

Ⅱ．実践応用プロジェクト

エファビレンジン（Efavirenz）合成プロセスにおける不斉アミド化反応に対する塩基性触媒の種類が反応に及ぼす影響の多変量線形回帰モデルの構築と分析

浙江省杭州第二高等学校 張一超

研究アプローチ



一、エファビレンジン

エファビレンジン（Efavirenz）は抗レトロウイルス薬の一種で、ヒト免疫不全ウイルス（HIV）感染症の治療に用いられる。非ヌクレオシド系レトロウイルス逆転写酵素阻害薬に分類され、HIV-1 逆転写酵素に非競合的に結合することで、ウイルス複製過程における DNA 合成を阻害し、体内でのウイルス増殖を抑制する [18][24][25]。

二、エファビレンジンの不斉合成法

「抗エイズ薬エファビレンジン（Efavirenz）の合成プロセス研究」[18] の論文に記載された当該薬剤の合成方案に基づき、不斉反応によるエファビレンジンの合成経路を確立した。

１．出発原料：2 - トリフルオロアセチル - 4 - クロロアニリンを出発原料として使用する。

２．グリニャール試薬の調製：クロロブタンと金属マグネシウムを反応させ、n - ブチルマグネシウムクロリドのグリニャール試薬を調製する。

３．不斉付加：水素化ナトリウム、塩化亜鉛、トリフルオロエタノール及びキラル配位子 (1R,2S)-1 - フェニル - 2-(1 - ピロリジニル)-1 - プロパノール（Lig A）の存在下で、シクロプロピルエチニルマグネシウムクロリドと反応させ、亜鉛原子を中心とする配位化合物を形成した後、出発原料と不斉付加反応を行い、キー中間体である (S)-1-(2 - アミノ - 5 - クロロフェニル)-1 - トリフルオロメチル - 3 - シクロプロピル - 2 - プロピン - 1 - オールを生成する。

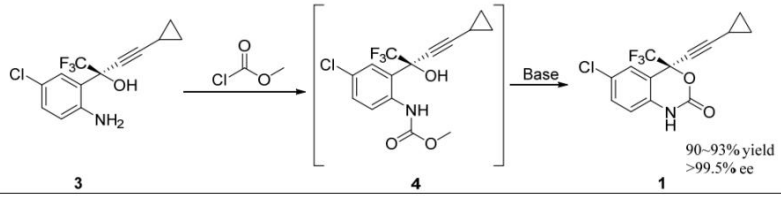
４．アミド化縮合：上記で得られたキラル中間体とクロロギ酸メチルを反応させ、中間体アミド化合物を形成する。

５．環化反応：カリウム tert - ブトキシドを触媒として用い、環化反応を経て目的生成物であるエファビレンジンを生成する。

６．精製：その後の洗浄、脱水、脱色及び結晶化工程を経て、最終的に高収率・高純度のエファビレンジン製品を得る。

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。



エファビレンジンの不斉合成法については参考文献 [11][18] を参照。

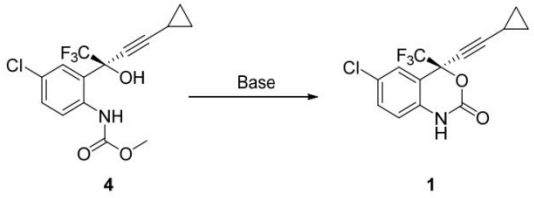
化合物 3 とクロロギ酸メチルを反応させた後、中間体アミド化合物 4 を生成し、脱水処理を行った上で、塩基性試薬の触媒下で環化反応を行い、目的生成物 1 であるエファビレンジンを得る。

三、アミド化縮合反応プロセス

アミド化縮合の反応プロセスについては参考文献 [18][21][24] を参照。

１．主要反応

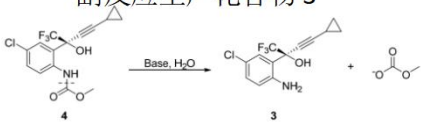
主要反应



4　　1

2．副反応による化合物 3 の生成

副反应生产化合物 3



4　　3

四、異なる塩基性触媒が反応に及ぼす影響

表：異なる塩基性触媒が反応に及ぼす影響[18]

番号	触媒	転化率①/%	化合物 3/%	選択性②/%
1	無	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	Na_2CO_3	98.1	1.47	96.5
5	KOH	98.5	1.15	97.1
6	$NaOCH_3$	98.5	0.45	98.0
7	t-BuOK	98.6	0.28	98.3

①化合物 4 の転化率；②エファビレンジンの選択性。

備考：化合物 3 を 210g（0.73mol）仕込み、減圧還流分水が終了した後（詳細な実験手順は 3.2 節を参照）、溶液を 7 等分し、各実験グループに対応する塩基性触媒 1.2g（50% NaOH 溶液は 2.4g）をそれぞれ添加し、45～50℃に昇温して 4 時間反応させた。

五、塩基性触媒の性質に関する定量データの取得

RDKit データベースを通じ、異なる塩基性触媒の多種類の定量性質データを取得した。具体的には、分子量（Molecular Weight）、トポロジカル極性表面積（Topological Polar Surface Area, TPSA）、回転可能結合数（NumRotatableBonds）、重原子数（HeavyAtomCount）、ヘテロ原子数（NumHeteroatoms）、CSP3 炭素原子比率（FractionCSP3）、脂溶性（LogP）を含む。

主要コード（一部コード省略）

```
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+]"
}
```

上記コードを通じ、Python により異なる塩基性触媒の性質データを取得できる。反応グループ 2 と反応グループ 3 における NaOH の濃度が一致しないため、添加した触媒の質量と濃度をいずれも独立変数として考慮した。

触媒	質量 g	濃度	分子量 g/mol	トポロジカル極性表面積Å²	回転可能結合数	重原子数	ヘテロ原子数	CSP3 炭素原子比率	脂溶性 LogP
無	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

六、多変量線形回帰モデルの構築

Python を用い、それぞれ「転化率」「化合物 3」「選択性」を従属変数とし、 塩基性触媒の物質定量データを独立変数として、多変量線形回帰分析を行った。

モデル構築の主要コード（一部コード省略）

```
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、多変量線形回帰の結果

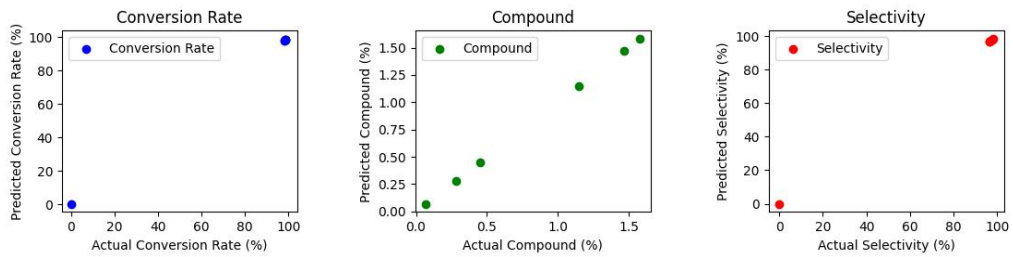
１．転化率回帰結果（Conversion Rate Regression Results）

```
                            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.
==============================================================================
```

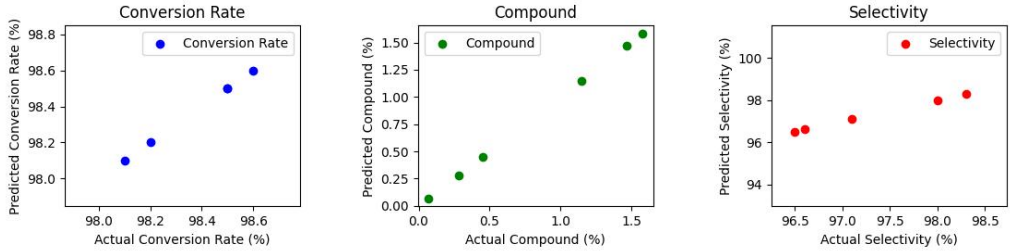
２．化合物回帰結果（Compound Regression Results）

```
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
```

Picture 1-1



Picture 1-2



図の説明：Picture1-1 においては、一部のサンプルが極めて近接した位置に存在するため、より大きな尺度を選択して全体図を作成した。Picture1-2 においては、サンプルが密集した領域を拡大表示し、多変量線形回帰の高い適合性をより明確に示している。

多変量線形回帰モデル式

１．転化率（ConversionRate）モデル

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）モデル

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．選択性（Selectivity）モデル

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、多変量線形回帰の分析

OLS（最小二乗法）回帰結果によれば、R 二乗値（R-squared）は 1.000 であり、これはモデルのデータへの適合度が極めて高く、従属変数の分散の 100％を説明できることを意味する。これは非常に高い適合度であり、モデルが従属変数の変動を良好に説明できるこ

とを示している。これは従属変数の変動を予測･説明する上で非常に意義のあることである。

ただし、警告や外れ値が発生したものの、回帰結果から依然として一定の結論を導き出すことができる。例えば、R 二乗値が 1 に極めて近いことから、モデルのデータへの適合度が高いことが確認できる。これは本モデルが従属変数（転化率：ConversionRate、化合物：Compound、選択性：Selectivity）の変動を良好に説明できることを示しており、積極的な側面である [7][8][9]。

サンプルサイズが小さいことから、係数の有意性及び信頼区間については慎重に解釈する必要がある。同時に、R 二乗値が 1 に極めて近いことから、過学習（オーバーフィッティング）が生じている可能性を考慮する必要がある。過学習が発生すると、新しいデータに対するモデルの予測能力が低下する可能性がある。

このような場合には、サンプル数を増やすことでモデルの頑健性（ロバスト性）を高めるべきである。

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.

Ⅲ. 文献研究プロジェクト

シンボリック回帰（Symbolic Regression）のメカニズムと化学分野における応用

浙江省杭州第二高等学校 張一超

一、シンボリック回帰の基本概念と特長

シンボリック回帰（Symbolic Regression, SR）は強力な機械学習手法の一つで、データから数学的表現式を推測し、データ背後に隠れた潜在的法則を明らかにすることができる。伝統的な数値回帰手法とは異なり、シンボリック回帰は予測モデルを提供するだけでなく、解釈性も備えており、データ間の関係を理解する上で役立つ。

シンボリック回帰は、データから数学モデルを自動的に発見する手法である。可能な数学的表現式の空間を探索し、データに最も適合する表現式を見つけ出す。これらの表現式は線形であっても非線形であってもよく、加減乗除、指数関数、対数関数、三角関数などを含む。

二、従来の回帰手法との比較

シンボリック回帰と伝統的回帰手法の主な違いは、伝統的回帰手法が主に入力変数と出力変数の間の数値関係の構築に注力するのに対し、シンボリック回帰は数値関係の探索だけでなく、数学的関係の探索を通じてより深い洞察とより高い解釈性を提供できる点にある。

三、主な実装手法とアルゴリズム

シンボリック回帰の主な実装手段には、遺伝的プログラミング（Genetic Programming, GP）、ニューラルネットワーク、勾配ブースティング決定木（Gradient Boosting Decision Tree, GBDT）などがある。これらの手法はそれぞれ長所を持ち、異なる問題やデータ型に対して異なる機能を発揮する。

シンボリック回帰モデルの構築ツールは多種多様である。SISSO（Sure Independence Screening and Sparsity Operator）は圧縮センシングに基づくアルゴリズムで、大量の候補記述子の中から最適な低次元記述子を識別するために特化して設計されている。gplearn は Python ベースの遺伝的プログラミングライブラリで、シンボリック回帰をサポートするだけでなく、分類や特徴構築などの機能も提供しており、多機能な機械学習ツールとして活用されている。PySR はオープンソースのシンボリック回帰ツールで、遺伝的プログラミングを利用してデータ中の数学的関係を明らかにし、特に高い解釈性を持つモデルが

必要なシーンに適している。

四、化学分野における応用事例

シンボリック回帰技術は化学分野において広く応用されている。材料化学では、材料の物理的・化学的性質を予測するために使用され、これにより新規材料の開発や既存材料の性能向上を指導する。化学合成経路の最適化においては、シンボリック回帰が異なる合成ステップにおける反応条件と生成物データを分析し、副反応を削減し、生成物の収率と純度を向上させるための改善策を提案することができる。同時に、それは配体選択、標的薬物設計などの分野でも役割を発揮することができる。

また、上海大学の欧陽潤海教授が開発した SISSO アルゴリズムは、高い操作性を備えているだけでなく、複数の化学分野において重要な成果を上げている。関連情報によれば、SISSO 法はペロブスカイト材料、トポロジカル絶縁体、触媒材料、超伝導体、二次元材料、ポリマーなどのモデル構築と新規材料の予測に使用されている。

Reference

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).

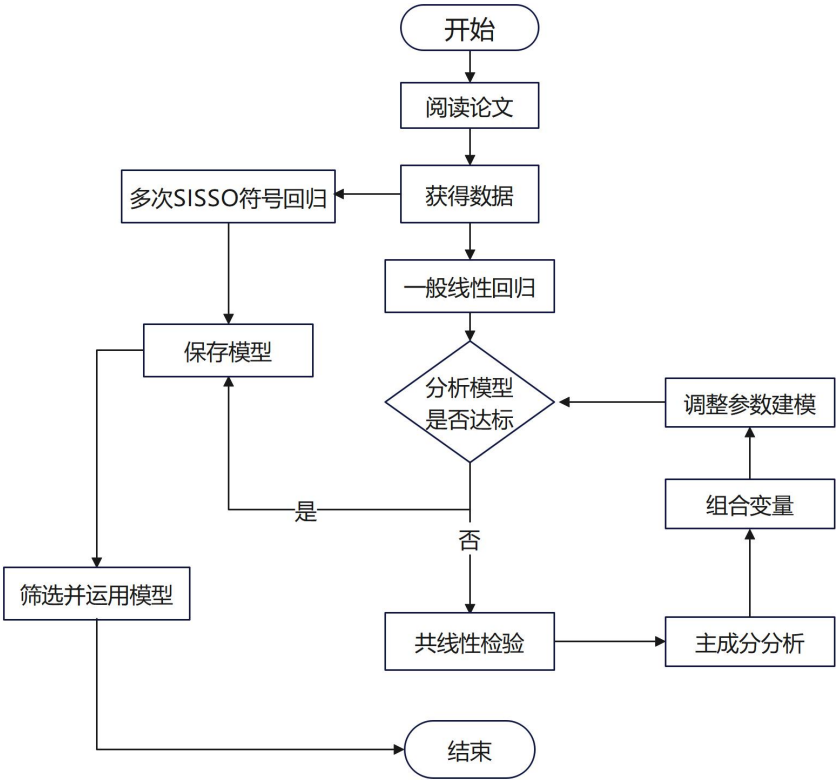
8. Shen, Y., Borowski, J. E., Hardy, M. A., Sarpong, R., Doyle, A. G., & Cernak, T. (2021). Automation and computer-assisted planning for chemical synthesis. Nature Reviews Methods Primers, 1, 23. https://doi.org/10.1038/s43586-021-00022-5

Ⅳ. 実践応用プロジェクト

ニッケル催化によるC–Hアルキル化反応におけるHASPOプレリガンドのエナンチオ選択性モデル

浙江省杭州第二高等学校 張一超

研究アプローチ



プロジェクトの説明：

本プロジェクトは、Zi-Jing Zhang, Matthias M. Simon, Shuang Yu, Shu-Wen Li, Xinran Chen, Silvia Cattani, Xin Hong, and Lutz Ackermann らによる論文 "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) のデータに基づいて研究を実施したものであり、ここに記して謝意を表します。

また、洪鑫（Xin Hong）教授と俞塽（Shuang Yu）氏から提供いただいた支援に対し、心より感謝申し上げます。

なお、データの一部は分量が多いため、本報告書では主要な部分および一部抜粋を示します。詳細な内容とデータは添付ファイルをご参照ください。

序論：論文内容の振り返り

本論文は、キラルなヘテロ原子置換二次ホスフィンオキシド（HASPO）リガンドを補助した Ni−Al 二金属触媒を用いる、革新的なニッケル触媒法を紹介している。この手法は C−N 軸不斉化合物の効率的な構築を目的とした C−H アルキル化反応を実現する。研究チームは反応条件の最適化を通じて、高いエナンチオ選択性を示す HASPO リガンドをスクリーニングし、反応の収率と立体選択性を大幅に向上させた。本論文では、さまざまな N-アリール置換ベンズイミダゾールやオレフィンを含む多様な基質に対する本手法の普遍性も実証されている。実験と密度汎関数理論（DFT）計算により、著者らはリガンド間水素移動や還元的脱離段階を含む反応メカニズムを解明した。さらに、多変量線形回帰（MVLR）解析を利用して、HASPO リガンドの構造とエナンチオ選択性の間の関係を研究し、将来のリガンド設計と最適化に対する理論的基盤を提供した。本研究は、コスト効率が高く毒性の低い触媒法を提供するだけでなく、生物活性や応用潜力を持つ軸不斉化合物の合成への新たな道を開くものである。

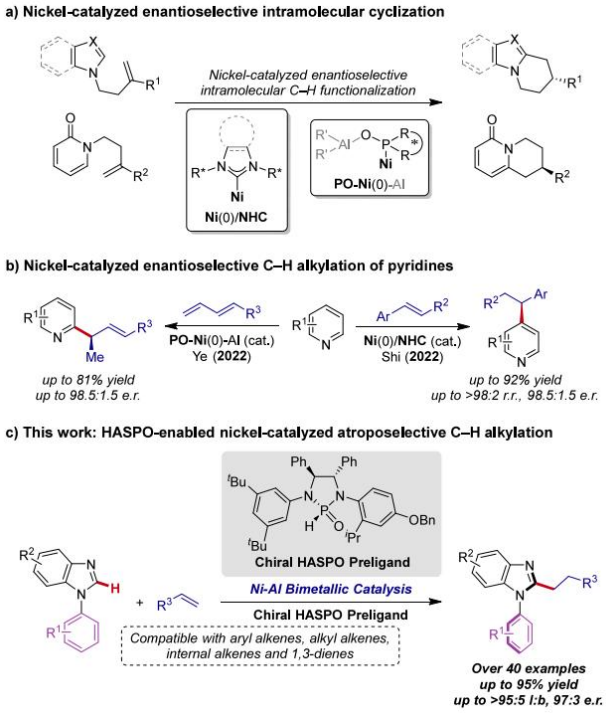


図 4 論文に記載された合成プロセスの概要図

HASPO リガンドは、ニッケル触媒によるアトロポ選択的 C−H アルキル化反応において中心的な役割を果たしており、その具体的な作用は以下の側面に現れている：
キラル源としての役割: HASPO リガンドは必要なキラル環境を提供し、反応の高いエナンチオ選択性を保証する。

- 触媒特性の調整: ニッケル触媒と錯体を形成することにより、配位作用を通じて触媒の電子状態および立体化学的特性を調整し、触媒活性と選択性に影響を与える。

- 立体化学的制御: HASPO リガンドの立体化学と空間構造は、反応の立体選択性に対して決定的な役割を果たし、その非 C2 対称性の特性が立体選択性制御において特に重要である。
- 基質/中間体の認識と安定化: HASPO リガンドの構造的特徴は、基質または中間体の認識と安定化に寄与し、化学選択性と位置選択性を向上させる。
- 触媒サイクルへの参加: 触媒サイクルにおいて、HASPO リガンドは触媒の初期形成、中間体の安定化、および最終生成物の放出に関与する。
- 効率と安定性の向上: HASPO リガンドの構造を最適化することにより、触媒の効率と安定性を高め、反応収率を向上させることができる。
- 反応条件の温和化: 適切な HASPO リガンドは、反応を温和な条件下で進行させ、副反応を減少させ、目的生成物の選択性を高めることを可能にする。

HASPO リガンドは、この高立体選択性反応を実現するための重要な構成要素であり、その多面的な作用を通じて、潜在的な応用価値を持つ軸不斉化合物の合成への効果的な道筋を提供する。

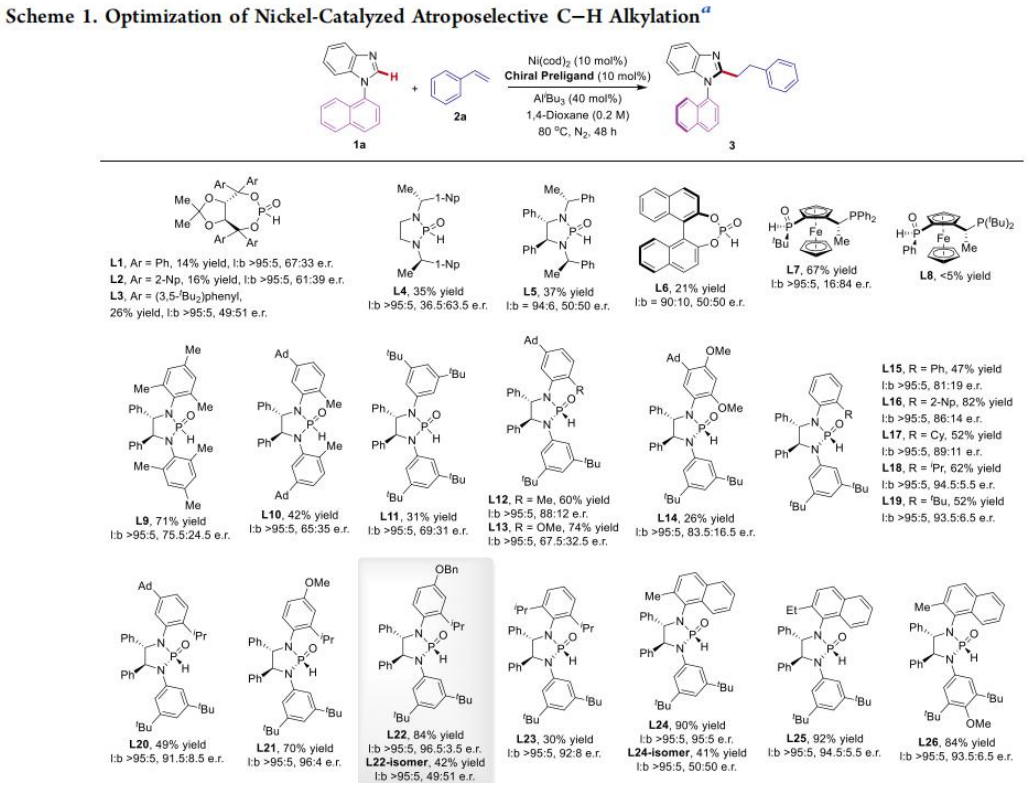


図 5 論文から抜粋したリガンドと関連データ

https://github.com/zju-ys/Nickel-MLR/blob/main/PhysOrg.csv を通じて論文内の各種リガンドの実験データを取得：

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

図 6 取得したリガンド反応データ

Part 1: SPSS に基づく線形回帰

第一步：一般線形回帰モデル

データに対して一般線形回帰モデリングを実施したところ、分析結果から強い多重共線性の存在が示唆された。

数式モデル：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

第二步：データ検定

データの多重共線性とピアソン相関の検定を実施し、主成分の選択または変数の統合による多重共線性がモデル精度に及ぼす影響の低減を図る。

多重共線性の検定

系数[a]

模型		未标准化系数 B	未标准化系数 标准错误	标准化系数 Beta	t	显著性	共线性统计 容差	共线性统计 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

ピアソン相関の検定

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	-.135	-.052	.229	.389	.343	-.015	.322	.046	-.222	.312	.416	.003	.509*	.310	.147	-.232	-.554**	-.296
	显著性（双尾）		.580	.818	.305	.074	.275	.948	.143	.841	.329	.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	.124	.256	-.208	.925**	.182	.509*	.051	-.106	-.057	-.557**
	显著性（双尾）	.580		.328	.166	.756	.272	.045	.005	.536	.584	.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	.005	-.326	-.378	-.267	-.534*	.499*	.261	.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	.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	-.330	.048	.109	-.194	.111	.179	-.235	-.070	-.225	.057	-.196	.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_9_10	皮尔逊相关性	.389	.070	.219	.148	1	.604*	.026	.279	.004	-.390	.522*	.584**	.285	.490*	.561**	.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	22
L_9_11	皮尔逊相关性	.243	.245	.005	.113	.604*	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	.073	.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**	.433*	.440*	.636**	-.377	.092	-.133	-.610**
	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	-.250	-.475*	-.105
	显著性（双尾）	.143	.005	.083	.833	.209	.045	.349		.410	.665	.655	.304	.020	.025	.398	.751	.262	.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：ステップワイズ回帰モデル

データに対してステップワイズ回帰モデリングを実施し、主要なキー変数を選択することで、多重共線性がモデル精度に及ぼす影響を低減する。

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：組合せ変数回帰モデル（主成分分析）

データに対し組合せ変数回帰モデルによる予備分析を実施し、以下の変数組合せを検討する：

B1_8_10 と B1_9_11：これらの変数間の相関係数は-0.125 であり、高い値ではないが、統計的に有意（P 値=0.580、有意ではない）である。さらに、これらは他の複数の変数とも高い相関を示しており、さらなる検討が必要である可能性がある。

BV_10 と BV_11：これらの変数間の相関係数は 0.431 であり、これは中程度の正の相関であり、統計的に有意（P 値=0.045、有意）である。

q_1 と q_2：これらの変数間の相関係数は-0.466 であり、これは強い負の相関であり、統計的に有意（P 値=0.029、有意）である。

q_3 と q_8：これらの変数間の相関係数は 0.644 であり、これは非常に強い正の相関であり、統計的に有意（P 値=0.001、有意）である。

d_1_2 と d_2_3：これらの変数間の相関係数は 0.811 であり、これは極めて強い正の相関であり、統計的に有意（P 値=0.000、有意）である。

第五步：一次組み合わせ変数回帰モデル

データに対して一次の組み合わせ変数回帰モデリングを実施した。

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

回帰モデル：

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：三次組み合わせ変数回帰モデル

データに対して三次の組み合わせ変数回帰モデリングを実施した。

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

回帰モデル：

ddg = -16.368 + 0.52 × PC1^3

ddg = -19.911 + 0.734 × PC1^3 - 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：多次組み合わせ変数回帰モデル

データに対して多次の組み合わせ変数回帰モデリングを実施した。

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

回帰モデル：

ddg = 261.178 + 5.432 × PC1^3 - 275.192 × PC2^3 - 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)				

第八部：モデルの整理と選択

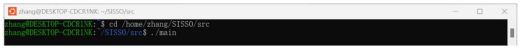
得られたモデルを整理･比較検討し、要件を満たす最適なモデルを選択して実用化した。

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: SISSO に基づくシンボリック回帰

第一步: システム設定とデータ前処理

SISSO ガイドライン及び関連指示に従い、Linux 環境下での SISSO のインストールとデバッグを実施した。次に、サンプル及びガイドの指示に基づき、csv データファイルを SISSO 要件の train.dat ファイル形式に変換した。最後に、標準設定に従い SISSO.in ファイルを設定した。



示例：在Ubuntu中启动SISSO


train.dat


SISSO.in 应当预先存入src文件夹的配置文件与数据文件

第二步: パラメータ調整

SISSO ガイドライン及び関連指示に従い、SISSO.in ファイル内の各種パラメータを設定した。以下にパラメータ項目の説明と解説を示す：

ptype: ターゲットプロパティのタイプ。1 は連続属性を持つ回帰、2 は分類属性を持つ分類を意味する。
ntask: タスク数。ntask = 1 は単一タスクの通常の機械学習、ntask > 1 はマルチタスク学習（MTL）を意味する。
task_weighting: (回帰のみ) マルチタスク回帰におけるタスクの重み付け。1 は重み付けなし（全てのタスクがデータセットサイズに関わらず同等に扱われる）、2 は各タスクがそのデータ量と全タスクの総データ量に対する比率に基づいて重み付けされる。
scmt: (回帰のみ) .true.に設定すると、符号制約付きマルチタスク回帰を呼び出す。
desc_dim: 記述子またはモデルの次元。
nsample: train.dat 内のサンプル数。単一タスク回帰では、入力は単一の整数。マルチタスク回帰では、入力は各タスクのサンプル数を表すカンマ区切りの整数。単一タスク分類では、各クラスのデータ数を示す括弧付きの整数が一つ。マルチタスク分類では、各タスクを定義するカンマ区切りの括弧が複数あり、各括弧内の整数がそのクラスのデータ数を定義する。
restart: ジョブの再開または継続。0 はジョブを最初から開始、1 はジョブを継続（前回のジョブの進捗情報は CONTINUE ファイルに保存されている）。
nsf: train.dat で提供されるスカラー特徴量の数。「スカラー特徴量」はデータセット内の各特徴量が一列を占めることを意味する。
ops: 特徴量構築のための数学演算子。ユーザーはリストから演算子をカスタマイズ可能。例: {+, -, *, /, exp, exp-, ^-1, ^2, ^3, sqrt, cbrt, log, |-|, scd, ^6, sin, cos}
fcomplexity: 特徴量の複雑さ。特徴量内の演算子数として定義される。fcomplexity=0 は特徴空間に入力変数のみが存在することを意味し、fcomplexity=3 は特徴空間内の全ての特徴量の複雑さが 3 を超えないことを意味する。

funit: train.dat ファイル内の特徴量が同じ単位を持つタイプを表す。例:
funit=(1:5)(6:9)(11:11) は、1～5 番目の特徴量が同じ単位、6～9 番目の特徴量が別の単位、10 番目の特徴量は無次元、11 番目の特徴量が異なる単位を持つことを意味する。
fmax_min と fmax_max: それぞれ、特徴量データ中の絶対値の最大値の閾値。fmax_min より小さい場合はゼロ特徴量として破棄され、fmax_max より大きい場合は無限大特徴量として破棄される。
nf_sis: SIS 部分空間のサイズ。SISSO-nD 計算では、n 個の SIS 部分空間が存在する。
method_so: スパースオペレータの方法。L0 ノルム最小化スパース化手法、または回帰専用の L1L0 手法。
nl1l0: (回帰のみ) LASSO で選択される特徴量の数。次の L0 で使用される。
fit_intercept: (回帰のみ) 線形モデルが非ゼロ/ゼロの切片をフィットするかどうか。
metric: (回帰のみ) 回帰におけるモデル選択に使用される指標。RMSE（二乗平均平方根誤差）または MaxAE（最大絶対誤差）を指定可能。
nmodels: 出力するランキング上位のモデル数。
isconvex: (分類のみ) 各データドメインが凸形状または非凸形状に制限できるかどうか。
bwidth: (分類のみ) ドメイン外に非常に近いデータを含めるための、各ドメインの境界許容度。

第三步: パラメータ設定

SISSO ガイドライン及び関連指示に従い、SISSO.in ファイル内の各種パラメータを設定した。一例として、ある設定例を示す。

```
! 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: データ解釈

Ubuntu 内で SISSO を実行後、結果は SISSO.out に保存される。SISSO.out のレポートを解読することにより、回帰結果を取得する。以下は、ある実行時の結果概要である:

```
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
```

第五步: モデル整理

パラメータを繰り返し調整し、データを統合した結果、以下の各モデルを取得した。

符号約束回归（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/

研修記録

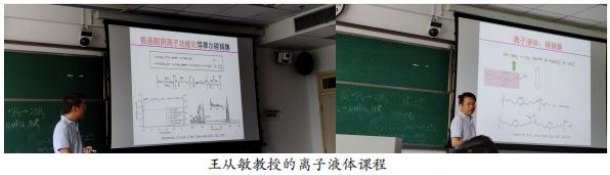
前沿講座

「英才計画」に参加して以来、累計でほぼ 10 回にわたり浙江大学で週末化学講座に参加し、数回の対面実験活動にも参加しました。この期間中、2 名の中国科学院院士による対面講義に恵まれました。

2024 年夏季には、浙江大学で開催された「英才計画」サマーキャンプに約 1 週間参加し、化学及びその他分野を含む総合講義と報告を十数回受講しました。

これらの前沿学術報告は豊富で生き生きとした化学知識をもたらし、様々な分野や専門方向の教授方々による多様な共有内容は、その都度、我々に大きな収穫をもたらしました。

図 7 院士講義



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

図 8 数回にわたる前沿講座

視野の拡大

浙江大学の何巧紅（He Qiaohong）老師の指導の下、浙江大学化学系の実験老師に引率され、化学系の各実験室と科学研究プラットフォームを見学し、基礎実験技能の学習と実践を展開しました

同時に、各教授方々に引率され、実験室に足を運び学習と理解を重ねました。

図 9 実験室での実践学習

研究室における育成

浙江大学の洪鑫（Hong Xin）教授に師事し、洪教授とはオンラインとオフラインの複数の方法を通じて学習を進めました。

洪教授の指導と助力により、学習と研究を進め、課題の各項目を順調に完了させることができました。

洪教授の指導のもと、洪教授の研究室に属する大学院生の俞塽（Yu Shuang）氏をはじめとする学生の助力を得て、各課題の研究を展開しました。

夏季休暇期間中には数回、対面で実験室を訪れて学習に参加し、オンライン方式を通じて深い学びを維持し、複数のプロジェクトを期限通りに完了させました。

ログ記録

各学習活動参加後、ログを記録しシステムに提出しました。現在までに 13 編を提出しています。

浙江大学「英才計画」夏季活動記録

浙江省杭州第二中学校 張一超（化学）

7 月 8 日から 12 日にかけて、我々は浙江大学紫金港キャンパスで「英才計画」夏季活動に参加しました。この活動は多彩で非常に意義深いものでした。我々は 5 日間の学習生活を通じて、前沿知識を探求し、科学研究の魅力を感じ取り、各自の研究課題に対する理解を一層深めました。

在石虎山机器人研究基地看到的机器狗表演

夏季活動は主に、知識講義、大学先修課（大学の単位認定を受ける高度な授業）、微专题研究（マイクロプロジェクト研究）、前沿基地の見学などの集団活動を含んでいました。同時に、指導教官と交流するための豊富な時間も提供されました。浙江大学竺可楨学院の主催により、各分野の教授方の熱意溢れる共有を通じて、我々は異なる学問の前沿動向を理解し、科学研究の革新的な方法を感じ取ることができました。例えば、葉高翔（Ye Gaoxiang）教授の講演では、哲学体系と科学発展が相互に促進し合う重要性を学びました。連日続いた大学先修課では、陳錦輝（Chen Jinhui）教授が微積分などの高等数学知識を初歩的に紹介してくれました。唐建軍（Tang Jianjun）教授の生態学をテーマとした講義では、生態研究の面白さを感じました。多数の専門家や学者が我々にもたらした前沿知識の饗宴は、科学知識を探求する意欲に満ちたものとしました。

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

同時に、各学科ごとに分かれて行われた微专题研究では、各学科が豊富な活動を企画しました。私が参加した化学グループを例にとると、王従敏（Wang Congmin）教授は、イオン液体が環境保護において重要であることを説明し、化学分野が二酸化炭素排出ピークアウト・カーボンニュートラルの実現に向けた取り組みにおいて重要であることを指摘しました。季鵬飛（Ji Pengfei）教授は、生物学と化学を結合し、生物触媒の手法を通じて材料研究を展開することについて話されました。王林軍（Wang Linjun）教授の共有では、我々は異なる化学の世界を見せられ、化学と他分野との融合による革新的な領域が力強く発展していることを認識しました。3 回の微专题研究活動終了後、我々はグループに分かれて微专题展示活動を行い、化学グループの学生は各自の課題と研究を紹介し、無機触媒から有機医薬品まで、実験化学から計算化学まで、我々は化学の各分野に対する理解を深め、相互に交流する能力も養うことができました。

付録

英才計画（全国科学技術革新人材育成プログラム）に関する基本情報

「中学生科技创新后备人才培养计划」（通称「英才計画」）は、中国科学技術協会と中華人民共和国教育部が 2013 年から共同で実施している人材育成プログラムであり、数学、物理学、化学、生物学、コンピューター科学の基礎学科に特長と潜在能力を持つ優秀な高校生を選抜し、著名な科学者の指導のもとで、週末や休暇を利用した研究活動を通じてその科学的素養と創新能力を育成することを目的としています。

2024 年現在、全国 58 の大学が参加し、約 1800 名の中学生が培養を受けており、大学進学後は基礎学科分野で深造する者や「基礎学科抜尖学生培養計画」に進む者も多く、高い教育効果を上げています。

関連公式リンク

英才計画の公式リンク
https://zxsycjh-kp.cast.org.cn/front/home
浙江大学 教員個人ホームページ
https://person.zju.edu.cn/hxchem
浙江大学化学系 洪鑫研究室（研究グループ）紹介
https://www.x-mol.com/groups/HongGroup

「英才計画」化学グループ参加生徒の集合写真

後列左から：
お二人目：浙江省科学技術協会の郭叶铭先生
六人目：浙江大学竺可楨学院副学院長の路欣先生
七人目：中国科学院院士の郭子建教授
八人目：中国科学院院士の麻生明教授
十二人目：浙江大学化学系の洪鑫教授（私の指導教授）
前列右から二人目が私です。

「英才計画」では、大変多くのことを学び、貴重な経験をさせていただきました。科学研究の道を志すようになったのも、この「英才計画」がきっかけです。先生方から賜りましたご指導、ご支援、ご教訓に、心から感謝申し上げます。