

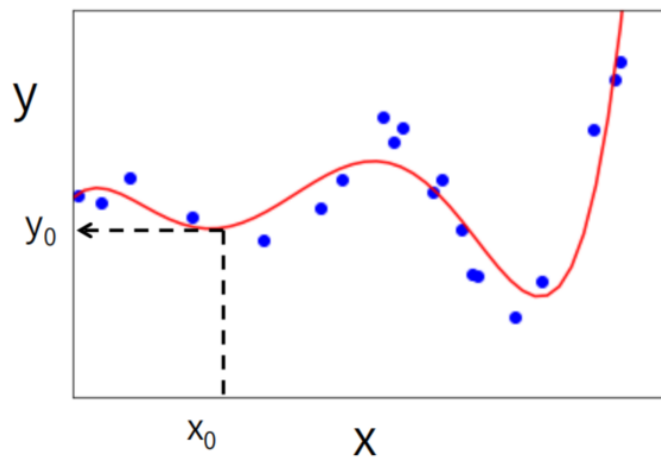
机器学习加速催化问题研究

机器学习加速催化研究





■ 机器学习模型





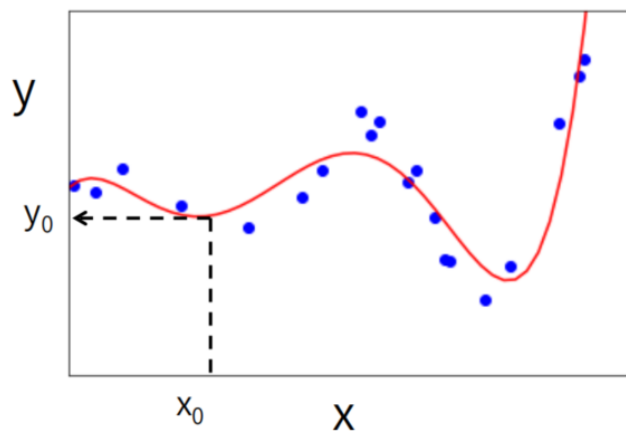
导论与环境配置

机器学习模型





■ 机器学习预测



通过函数拟合初始数据
进行插值



机器学习的价值就在于通过已有的数据
对未知的数据进行预测

导论与环境配置

■ Python环境安装及代码运行环境配置

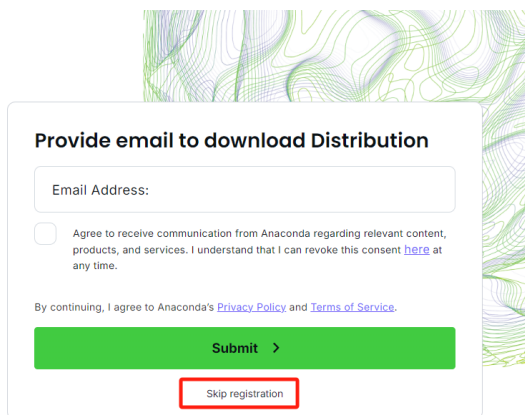
Anaconda官网: www.anaconda.com

Distribution

Free Download*

Register to get everything you need to get started on your workstation including Cloud Notebooks, Navigator, AI Assistant, Learning and more.

- ✓ Easily search and install thousands of data science, machine learning, and AI packages
- ✓ Manage packages and environments from a desktop application or work from the command line
- ✓ Deploy across hardware and software platforms
- ✓ Distribution installation on Windows, MacOS, or Linux



Provide email to download Distribution

Email Address:

☐ Agree to receive communication from Anaconda regarding relevant content, products, and services. I understand that I can revoke this consent [here](#) at any time.

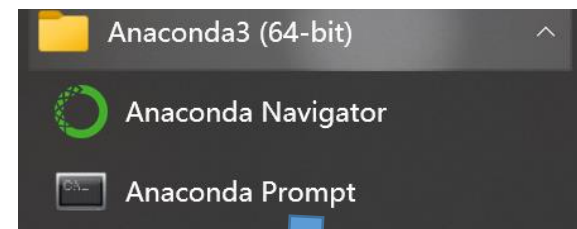
By continuing, I agree to Anaconda's [Privacy Policy](#) and [Terms of Service](#).



Windows

Python 3.12

64-Bit Graphical Installer (912.3M)



```
(base) C:\Users\ZJ>conda -V  
conda 24.1.2
```

```
C:\ProgramData\anaconda3  
C:\ProgramData\anaconda3\Scripts  
C:\ProgramData\anaconda3\Library\mingw-w64\bin  
C:\ProgramData\anaconda3\Library\bin
```

环境变量配置

机器学习开发环境准备

第一步：新建虚拟环境

```
(base) C:\Users\ZJ>conda create -n catalysis_ml python=3.9
```

虚拟环境创建完成

```
done
#
# To activate this environment, use
#
#     $ conda activate catalysis_ml
#
# To deactivate an active environment, use
#
#     $ conda deactivate
```

第二步：必备机器学习库的安装

激活环境

```
(base) C:\Users\ZJ>conda activate catalysis_ml
(catalysis_ml) C:\Users\ZJ>
```

机器学习库的安装

1. pip install 对应库名

2. conda install 对应库名

```
(catalysis_ml) C:\Users\ZJ>conda install numpy pandas scikit-learn matplotlib seaborn
```

```
conda config --add channels https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/main/
conda config --add channels https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/free/
conda config --add channels https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud/conda-forge/
conda config --set show_channel_urls yes
```

添加清华源

```
conda install shap xgboost
```

■ 相关机器学习库

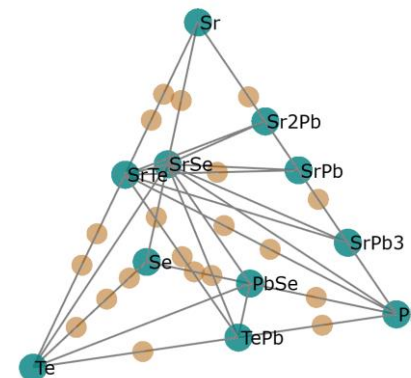
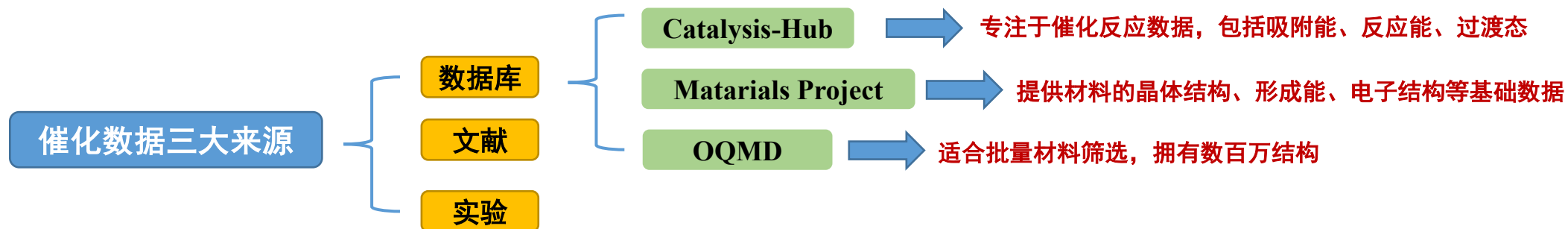
1 NumPy → NumPy  → 网址: numpy.org → 用于科学计算的python库

2 pandas →  → 网址: pandas.pydata.org → 用于进行大量数据处理分析的python库

3 Matplotlib →  → 网址: matplotlib.org → 数据可视化python库

4 SciKit-Learn →  → 网址: scikit-learn.org → 基于python语言的机器学习包

■ 数据库篇：Catalysis-Hub、Materials Project、OQMD三大数据库



■ Catalysis-Hub接口调用示例

导入python库

```
import requests
import pandas as pd
from tqdm import tqdm
```

API接口调用

```
# Catalysis-Hub GraphQL API
url = "https://api.catalysis-hub.org/graphql"

# 简单查询: 获取 5 条 reactionEnergy 数据
query = """
{
  reactions(first: 10) {
    edges {
      node {
        chemicalComposition
        reactionEnergy
        reactants
        products
      }
    }
  }
}
"""
```

数据获取

```
# 发送请求
response = requests.post(url, json={"query": query})
data = response.json()

# 解析 reactionEnergy, 并加 tqdm 显示进度
edges = data["data"]["reactions"]["edges"]
records = []

for edge in tqdm(edges, desc="解析 reactions 数据"):
    node = edge["node"]
    records.append({
        "composition": node["chemicalComposition"],
        "reaction_energy (eV)": node["reactionEnergy"],
        "reactants": node["reactants"],
        "products": node["products"]
    })

# 转为 DataFrame 并显示
df = pd.DataFrame(records)
print(df)
```

■ Catalysis-Hub接口调用示例

结果展示

```
解析 reactions 数据: 100% | 10/10 [00:00<?, ?it/s]
composition reaction_energy (eV) reactants
0 HH- MoS2 -0.220 {"CH0star": 1}
1 HH- MoSe2 -0.530 {"CH0star": 1}
2 HH- MoS2 -0.690 {"star": 1, "hfH2gas": 1}
3 HH- Ir-MoS2 0.770 {"star": 1, "COgas": 1}
4 HH- MoSe2 0.590 {"star": 1, "COgas": 1}
5 HH- MoS2 0.240 {"star": 1, "COgas": 1}
6 HH- Re-MoS2 0.156 {"star": 1, "hfH2gas": 1}
7 HH- Ir-MoS2 0.101 {"star": 1, "hfH2gas": 1}
8 HH- Co-MoS2 -0.226 {"star": 1, "hfH2gas": 1}
9 HH- MoS2 -0.320 {"CH0star": 1}

products
0 {"COgas": 1, "Hstar": 1}
1 {"COgas": 1, "Hstar": 1}
2 {"Hstar": 1}
3 {"COstar": 1}
4 {"COstar": 1}
5 {"COstar": 1}
6 {"Hstar": 1}
7 {"Hstar": 1}
8 {"Hstar": 1}
9 {"COgas": 1, "Hstar": 1}
```

相关说明

1. reactionEnergy为可执行数值性质，可用于直接分析或绘图
2. chemicalComposition、reactants、products为了说明这个反应涉及哪些物质
3. first:5可改成更大数值或添加filter筛选条件，批量获取更多数据



文献到数据：催化数据构建

■ Catalysis-Hub接口调用示例

数据结构

composition	reaction_energy (eV)	reactants	products
HH- MoS2	-0.22	{"CHOstar": 1}	{"COgas": 1, "Hstar": 1}
HH- MoSe2	-0.53	{"CHOstar": 1}	{"COgas": 1, "Hstar": 1}
HH- MoS2	-0.69	{"star": 1, "hfH2gas": 1}	{"Hstar": 1}
HH- Ir-MoS2	0.77	{"star": 1, "COgas": 1}	{"COstar": 1}
HH- MoSe2	0.59	{"star": 1, "COgas": 1}	{"COstar": 1}

1

composition

表示反应体系的组成或
表面催化剂

例如：HH-MoS₂就是表示六方晶系的结构类型

HH-Ir-MoS₂表示在MoS₂上掺杂了Ir的体系

2

Reaction_energy

表示反应的能量变化



文献到数据：催化数据构建

■ Catalysis-Hub接口调用示例

数据结构

composition	reaction_energy (eV)	reactants	products
HH- MoS2	-0.22	{"CHOstar": 1}	{"COgas": 1, "Hstar": 1}
HH- MoSe2	-0.53	{"CHOstar": 1}	{"COgas": 1, "Hstar": 1}
HH- MoS2	-0.69	{"star": 1, "hfH2gas": 1}	{"Hstar": 1}
HH- Ir-MoS2	0.77	{"star": 1, "COgas": 1}	{"COstar": 1}

3 reactants

字典形式表示反应物及其计数

0.59 {"star": 1, "COgas": 1} {"star": 1, "hfH2gas": 1}

表示表面吸附了1个CHO*中间体

表示表面有一个空位 (*), 还有一个H₂分子

“star”：通常表示活性位点

“hfH2gas”、“COgas”：等表示气相分子

4 products

字典形式表示反应产物及其数量

0.77 {"COgas": 1, "Hstar": 1} {"Hstar": 1}

表示生成了1个CO气体和1个H*吸附在表面

生成一个吸附H



文献到数据：催化数据构建

■ Catalysis-Hub接口调用示例

数据结构

composition	reaction_energy (eV)	reactants	products
HH- MoS2	-0.22	{"CHOstar": 1}	{"COgas": 1, "Hstar": 1}
HH- MoSe2	-0.53	{"CHOstar": 1}	{"COgas": 1, "Hstar": 1}
HH- MoS2	-0.69	{"star": 1, "hfH2gas": 1}	{"Hstar": 1}
HH- Ir-MoS2	0.77	{"star": 1, "COgas": 1}	{"COstar": 1}
HH- MoSe2	0.59	{"star": 1, "COgas": 1}	{"COstar": 1}

reactants/products 用字典表示，便于后续处理，例如可以计算H*吸附能，或者做统计分析

可以在这个基础上筛选，比如只看涉及“Hstar”的反应来研究氢吸附能

Materials Project接口调用示例

```
pip install mp_api
```

1. 硅结构数据

```
# 硅结构数据
from mp_api.client import MPRester

with MPRester("ROMy6FvBhpXZjk1kkTBIWDJCsZkuuSnk") as mpr:
    docs = mpr.materials.summary.search(material_ids=["mp-149"], fields=["structure"])
    structure = docs[0].structure
    # -- Shortcut for a single Materials Project ID:
    structure = mpr.get_structure_by_material_id("mp-149")
```



```
Structure Summary
Lattice
  abc : 3.8492784033699095 3.8492794116013456 3.849278
angles : 60.00001213094421 60.00000346645984 60.0000109754579
volume : 40.32952684741405
  A : np.float64(3.333573) np.float64(0.0) np.float64(1.924639)
  B : np.float64(1.111191) np.float64(3.142924) np.float64(1.924639)
  C : np.float64(0.0) np.float64(0.0) np.float64(3.849278)
  pbc : True True True
PeriodicSite: Si (3.889, 2.75, 6.736) [0.875, 0.875, 0.875]
PeriodicSite: Si (0.5556, 0.3929, 0.9623) [0.125, 0.125, 0.125]
```

2. 查找包含介电数据的条目的所有材料项目ID

```
from mp_api.client import MPRester
from emmet.core.summary import HasProps

with MPRester("ROMy6FvBhpXZjk1kkTBIWDJCsZkuuSnk") as mpr:
    docs = mpr.materials.summary.search(
        has_props = [HasProps.dielectric], fields=["material_id"]
    )
    mpids = [doc.material_id for doc in docs]

✓ 1.6s

Retrieving SummaryDoc documents: 100%|██████████| 7327/7327 [00:01<00:00, 6265.85it/s]
```



```
[MPID(mp-570778),
 MPID(mp-632291),
 MPID(mp-850274),
 MPID(mp-542605),
 MPID(mp-14),
 MPID(mp-19),
 MPID(mp-555760),
 MPID(mp-567313),
 MPID(mp-990424),
 MPID(mp-160),
 MPID(mp-611219),
 MPID(mp-12103),
```

■ Materials Project接口调用示例

```
pip install mp_api
```

3. 所有至少包含Si和O的材料的化学式

```
from mp_api.client import MPRester

with MPRester("ROMy6FvBhpXZjk1kkTBIWDJCsZkuuSnk") as mpr:
    docs = mpr.materials.summary.search(
        elements=["Si", "O"], fields=["material_id", "band_gap", "formula_pretty"]
    )
    mpid_formula_dict = {
        doc.material_id: doc.formula_pretty for doc in docs
    }
```

```
MPID(mp-600001): 'SiO2',
MPID(mp-561291): 'SiO2',
...
MPID(mp-1198966): 'SiAg2(S2O7)3',
MPID(mp-1211039): 'MnSi(OF)6',
MPID(mp-23731): 'CuSi(HO2)2',
MPID(mp-756818): 'LiSiBiO4',
...}
```

Materials Project接口调用示例

```
pip install mp_api
```

4. Stable materials (on the GGA/GGA+U hull) with large band gaps ($> 3\text{eV}$)

```
from mp_api.client import MPRester

with MPRester("ROMy6FvBhpXZjk1kkTBIWDJCsZkuuSnk") as mpr:
    docs = mpr.materials.summary.search(
        band_gap=(3, None), is_stable=True, fields=["material_id"]
    )
    stable_mpids = [doc.material_id for doc in docs]

    ## -- Alternative directly using energy above hull:
    docs = mpr.materials.summary.search(
        band_gap=(3, None), energy_above_hull=(0, 0), fields=["material_id"]
    )
    stable_mpids = [doc.material_id for doc in docs]
```

✓ 6.3s

```
Retrieving SummaryDoc documents: 100%|██████████| 6106/6106 [00:02<00:00, 2703.54it/s]
Retrieving SummaryDoc documents: 100%|██████████| 6106/6106 [00:02<00:00, 2741.48it/s]
```

Materials Project接口调用示例

```
pip install mp_api
```

5. 硅的能带结构

```
from mp_api.client import MPRester
from emmet.core.electronic_structure import BSPathType

with MPRester("ROMy6FvBhpXZjk1kkTBIWDJCsZkuuSnk") as mpr:
    # -- line-mode, Setyawan-Curtarolo (default):
    bs_sc = mpr.get_bandstructure_by_material_id("mp-149")

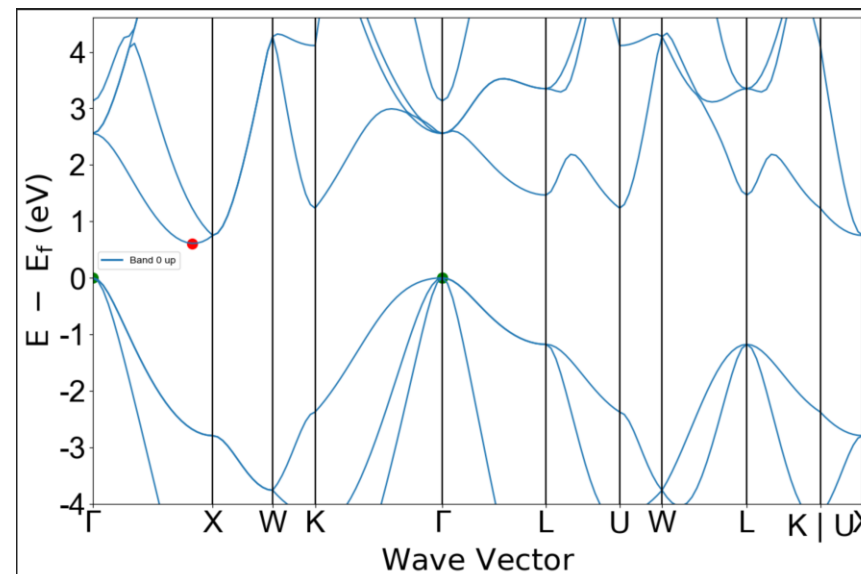
    # -- line-mode, Hinuma et al.:
    bs_hin = mpr.get_bandstructure_by_material_id("mp-149", path_type=BSPathType.hinuma)

    # -- line-mode, Latimer-Munro:
    bs_hin = mpr.get_bandstructure_by_material_id("mp-149", path_type=BSPathType.latimer_munro)

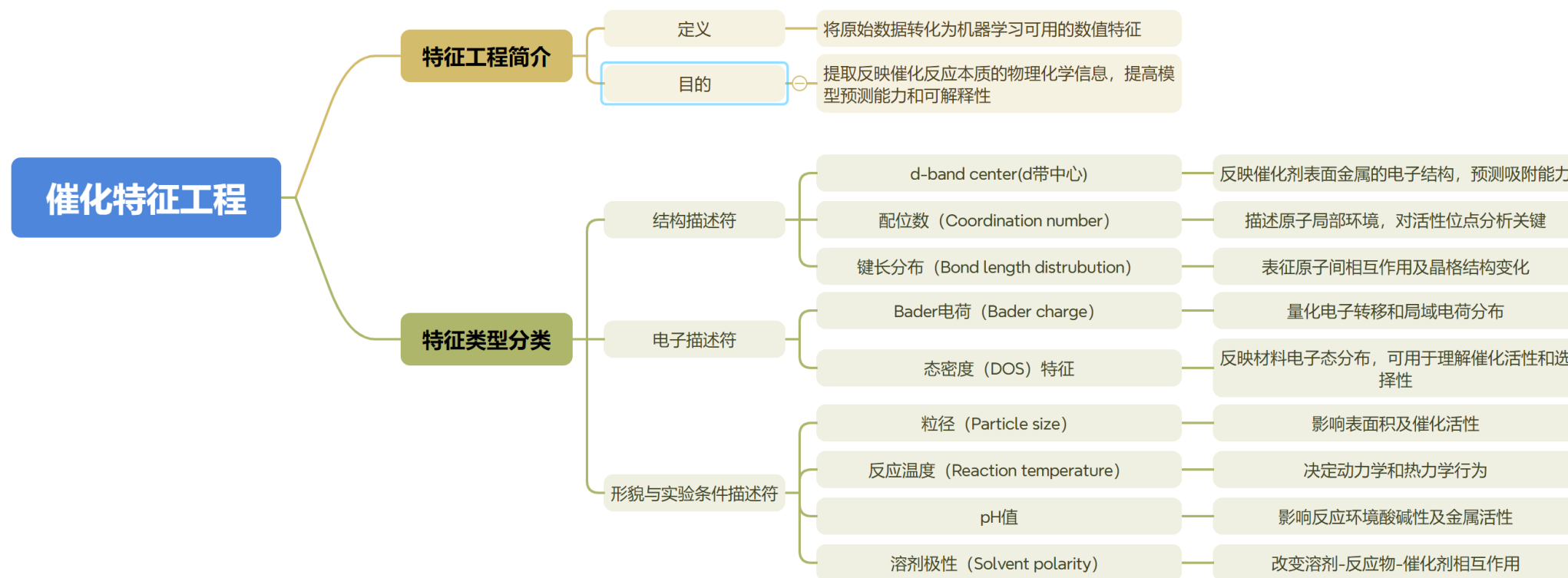
    # -- uniform:
    bs_uniform = mpr.get_bandstructure_by_material_id("mp-149", line_mode=False)
```

✓ 24.4s

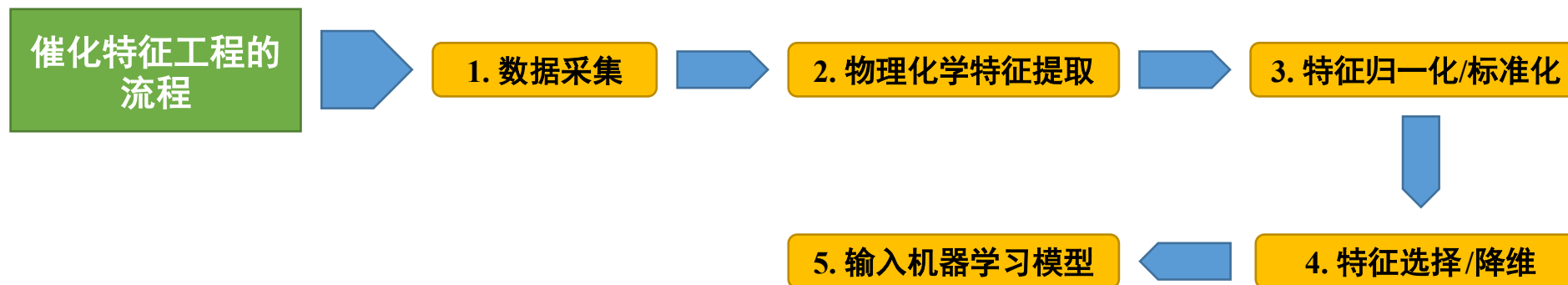
```
Retrieving ElectronicStructureDoc documents: 100%|██████████| 1/1 [00:00<?, ?it/s]
Retrieving ElectronicStructureDoc documents: 100%|██████████| 1/1 [00:00<?, ?it/s]
Retrieving ElectronicStructureDoc documents: 100%|██████████| 1/1 [00:00<?, ?it/s]
Retrieving ElectronicStructureDoc documents: 100%|██████████| 1/1 [00:00<?, ?it/s]
```



■ 催化领域特征工程



■ 催化领域特征工程

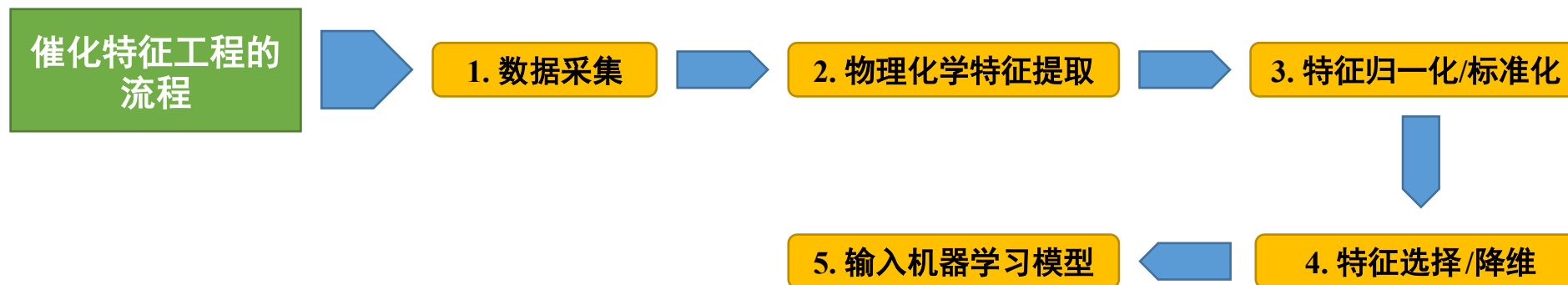


总结：1 特征工程是催化机器学习建模的核心环节；

2 结构、电子、实验条件描述符互为补充；

3 精准描述催化体系特征，有助于提高预测准确性和可解释性

■ 催化领域特征工程



总结：1 特征工程是催化机器学习建模的核心环节；

2 结构、电子、实验条件描述符互为补充；

3 精准描述催化体系特征，有助于提高预测准确性和可解释性



■ 实操案例：预测析氢反应（HER）活性

The Journal of Physical Chemistry C > Vol 124/Issue 36 > Article

Cite Share Jump to Expand

C: SURFACES, INTERFACES, POROUS MATERIALS, AND CATALYSIS | August 12, 2020

Descriptors for Hydrogen Evolution on Single Atom Catalysts in Nitrogen-Doped Graphene

Victor Fung, Guoxiang Hu, Zili Wu, and De-en Jiang*

Table S1. Properties and descriptors of transition-metal single-atom catalysts in nitrogen-doped graphene (see footnotes for definitions of each descriptor).

Size	atom	G ₀ (eV)	roov (Å)	rd (Å)	E _f (eV/atom)	IE	EN	ε _d (eV)	n	q (e)	d _{occ}	d _{unocc}	
Periodic Graphene	Sc	-0.190	1.700	0.310	-8.282	6.561	1.360	1.064	3	1.640	0.85	4.17	
	Ti	-0.500	1.600	0.280	-8.326	6.827	1.540	0.998	4	1.404	0.83	3.58	
	V	-0.270	1.530	0.260	-7.494	6.746	1.630	1.316	5	1.341	1.22	2.57	
	Cr	0.310	1.390	0.250	-6.604	6.766	1.660	0.944	6	1.263	2.36	0.98	
	Mn	0.430	1.390	0.230	-6.520	7.434	1.550	-0.513	7	1.279	1.33	2.46	
	Fe	0.250	1.320	0.220	-7.360	7.902	1.830	-0.846	8	1.080	2.98	1.23	
	Co	0.150	1.260	0.210	-7.562	7.880	1.880	-0.573	9	0.887	4.80	1.30	
	Ni	1.590	1.240	0.195	-7.699	7.639	1.910	-1.750	10	0.844	1.91	0.58	
	Cu	1.680	1.320	0.185	-5.097	7.726	1.900	-2.748	11	0.928	0.47	0.45	
	Zn	0.760	1.220	0.175	-3.174	9.394	1.650	-3.581	12	1.159	0.52	0.01	
	Zr	-0.900	1.750	0.540	-8.071	6.634	1.330	1.749	4	1.788	0.16	1.15	
	Nb	-0.760	1.640	0.510	-8.042	6.758	1.600	1.132	5	1.646	0.35	0.69	
	Mo	-0.910	1.540	0.490	-5.661	7.092	2.160	0.820	6	1.310	0.99	2.06	
	Tc	-0.460	1.470	0.455	-6.988	7.275	1.900	0.006	7	1.264	2.84	1.37	
	Ru	-0.560	1.420	0.450	-7.135	7.361	2.220	-0.191	8	0.951	3.54	1.46	
	Rh	-0.170	1.420	0.420	-7.579	7.459	2.280	-0.056	9	0.619	4.47	0.46	
	Pd	1.840	1.390	0.400	-5.704	8.337	2.200	-2.046	10	0.703	1.01	0.09	
	Ag	0.730	1.450	0.385	-1.862	7.576	1.930	-4.184	11	0.853	0.09	0.57	
	Cd	-0.060	1.440	0.370	-1.185	8.994	1.690	-2.385	12	0.984	0.15	0.02	
	Hf	-1.040	1.750	0.630	-9.420	6.824	1.300	1.392	4	1.589	0.36	2.45	
	Ta	-1.080	1.700	0.605	-8.892	7.887	1.500	0.841	5	1.645	0.73	2.46	
	W	-1.450	1.620	0.590	-7.101	7.980	2.360	0.096	6	1.558	1.91	2.01	
	Re	-0.910	1.510	0.565	-6.599	7.877	1.900	-0.055	7	1.379	1.42	1.93	
	Os	-0.730	1.440	0.543	-7.784	8.706	2.200	-0.262	8	1.179	3.59	0.91	
	Ir	-0.300	1.410	0.526	-8.361	9.121	2.200	-0.135	9	0.720	4.42	0.42	
	Pt	1.490	1.360	0.510	-7.677	9.017	2.280	-1.487	10	0.715	2.33	0.14	
	Au	2.240	1.360	0.488	-2.953	9.225	2.540	-3.348	11	0.918	0.02	0.57	
	Large nanographene	Ti	-0.498	1.600	0.280	-8.232	6.827	1.540	0.688	4	1.345	1.12	3.17
		V	-0.270	1.530	0.260	-7.494	6.746	1.630	0.890	5	1.275	1.92	2.12
		Cr	0.351	1.390	0.250	-6.809	6.766	1.660	0.826	6	1.212	2.43	0.37

实操训练的初始数据集

来源于支持信息

■ 实操案例：预测析氢反应（HER）活性

atom	G_H (eV)	rcov (Å)	rd (Å)	E_f (eV/atom)	IE	EN	ϵ_d (eV)	n	q (e)	d_{occ}	d_{unocc}
Sc	-0.190	1.700	0.310	-8.282	6.561	1.360	1.064	3	1.640	0.85	4.17
Ti	-0.500	1.600	0.280	-8.326	6.827	1.540	0.998	4	1.404	0.83	3.58
V	-0.970	1.520	0.260	-7.404	6.746	1.620	1.216	5	1.244	1.22	2.67

表格特征值讲解

这些参数可以分为三类：

1. 几何性质：rcov, rd
2. 热力学/电子性质： E_f , IE, EN, q, n
3. 电子结构/催化性质： ϵ_d , G_H , d_{occ} , d_{unocc}

G_H (Hydrogen adsorption free energy, 氢吸附自由能)	描述氢原子在催化表面上吸附的自由能变化 (ΔG_H)	意义：评价催化剂活性的关键指标
rcov (Covalent radius, 共价半径)	原子在共价键结合中表现出来的‘半径’	意义：对催化活性位点的几何结构和配位环境有影响
rd (Zunger radius of atomic d-orbital, d轨道半径)	描述过渡金属d轨道的有效半径	意义：与催化活性和电子结构密切相关
E_f (Formation energy, 形成能)	单原子位点（或掺杂位点）在基底上形成的能量变化	
IE (Ionization Energy, 电离能)	从原子中移去一个电子所需的能量	意义：电离能高的原子更难失去电子，对催化位点电子结构有影响
EN (Electronegativity, 电负性)	原子吸引电子的能力	意义：决定原子对电子云的偏移，影响吸附物与催化位点的相互作用强度
ϵ_d (d-band center, d带中心)	d轨道电子态密度 (DOS) 的加权平均能量	意义：d带中心越高，金属表面对吸附物的结合越强
n (Number of valence electrons, 价电子数)	原子最外层可参与化学反应的电子数	意义：直接影响原子配位化学，吸附能力及催化反应机理
q (Bader charge, Bader电荷)	原子在体系中的净电荷	意义：可以判断原子位点的电子供体/受体性质
d_{occ} (Number of occupied d states from 0 to -2 eV, 已占据 d 态数)	在费米能级附近 (0 到 -2 eV) 被占据的 d 轨道电子数。	意义：影响吸附物与金属的 σ/π 作用
d_{unocc} (Number of unoccupied d states from 0 to 2 eV, 未占据 d 态数)	费米能级附近 (0 到 2 eV) 未被占据的 d 态数	意义：对催化反应中的电子转移能力有直接影响。

■ 实操案例：预测析氢反应（HER）活性

1. **数据加载与探索** - 了解数据集的基本信息
2. **数据预处理** - 分离特征和目标变量
3. **模型训练** - 使用线性回归算法
4. **模型评估** - 多种评估指标和可视化
5. **交叉验证** - 5折交叉验证和留一法演示
6. **结果分析** - 全面的性能总结和可视化

 HER活性预测.ipynb

详细机器学习
代码实现流程