Python与机器学习

——机器学习与高通量筛选

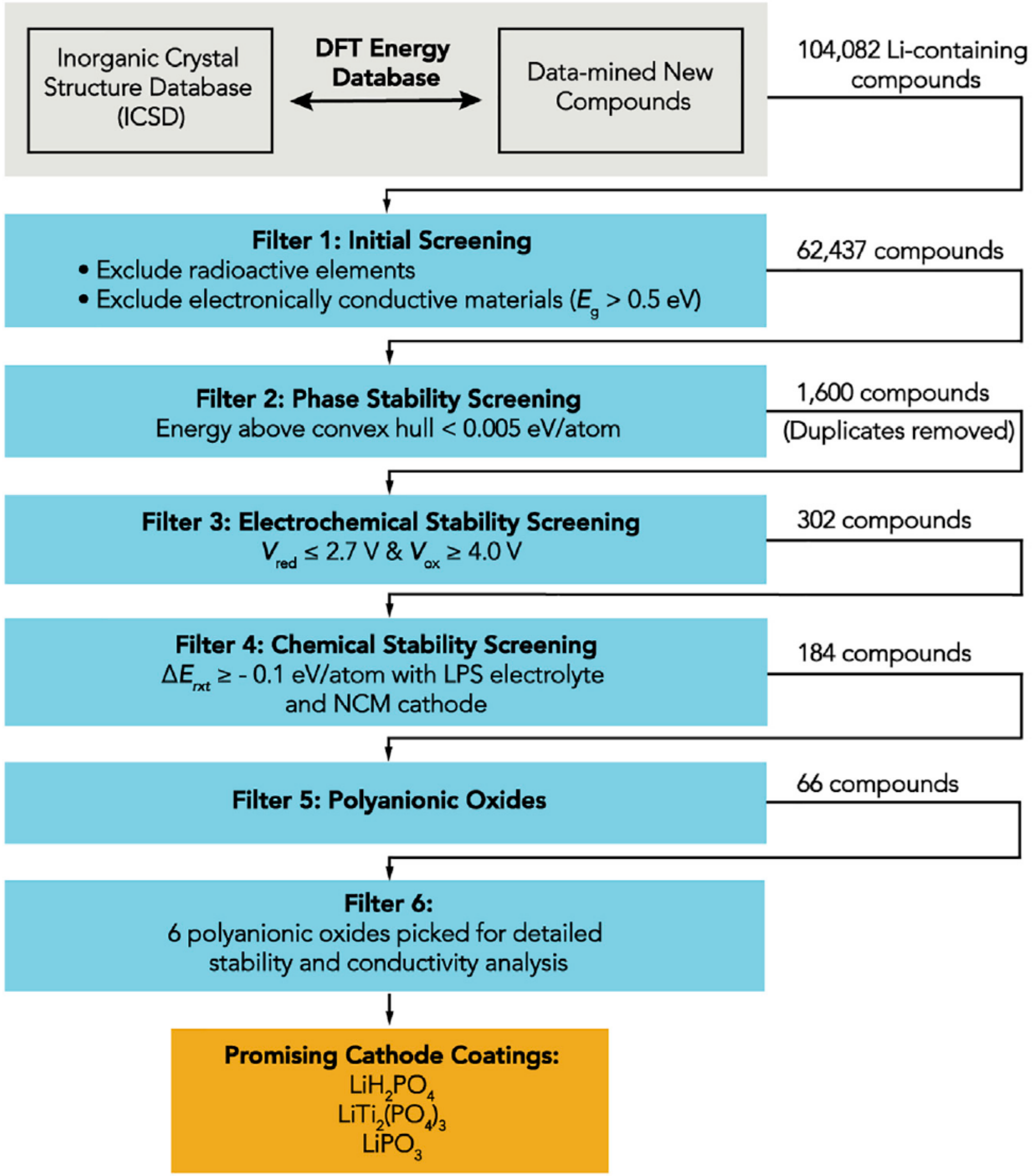
华算科技 黄老师

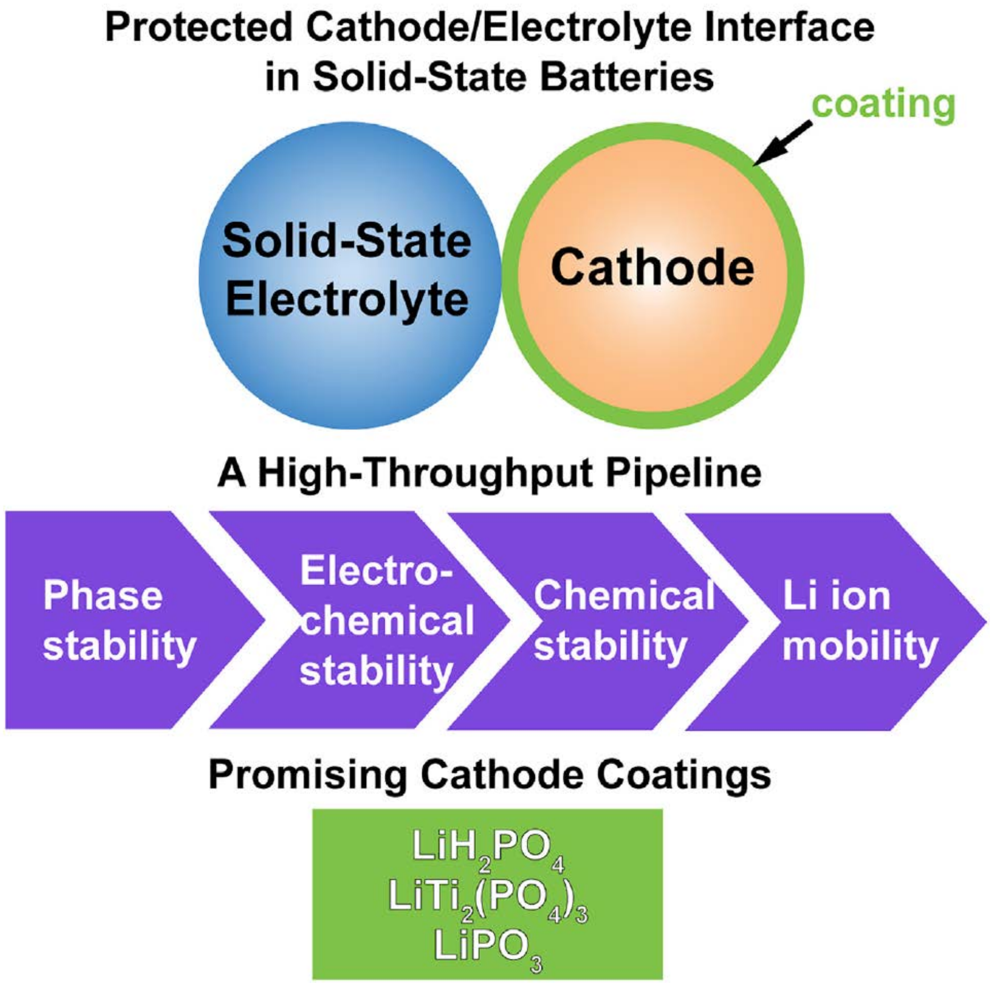
2022年2月24日

目录

目录

高通量筛选

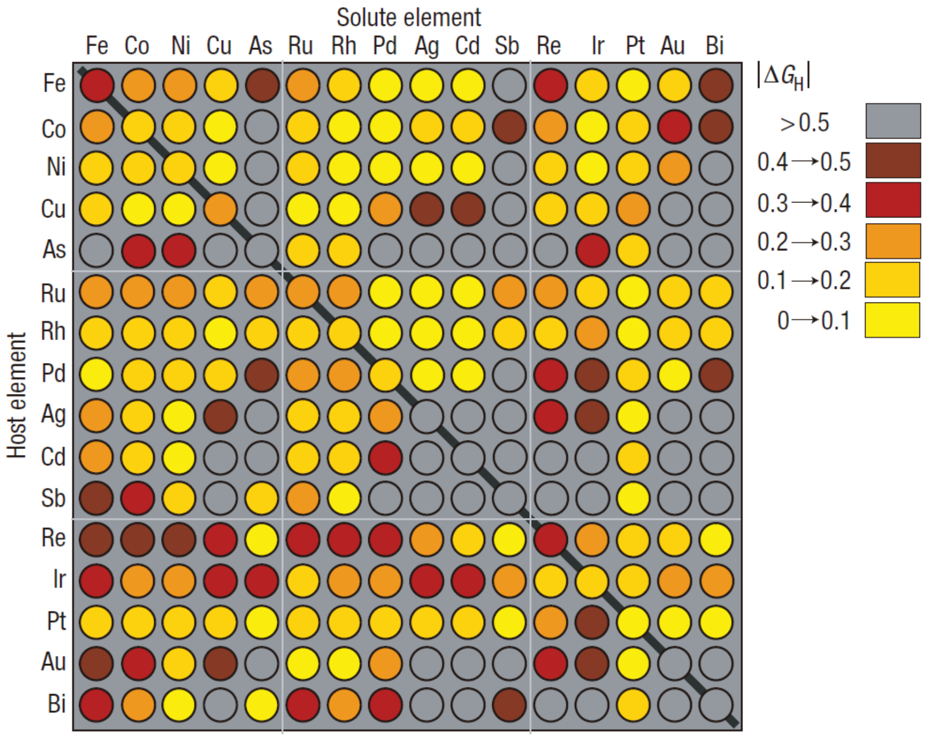


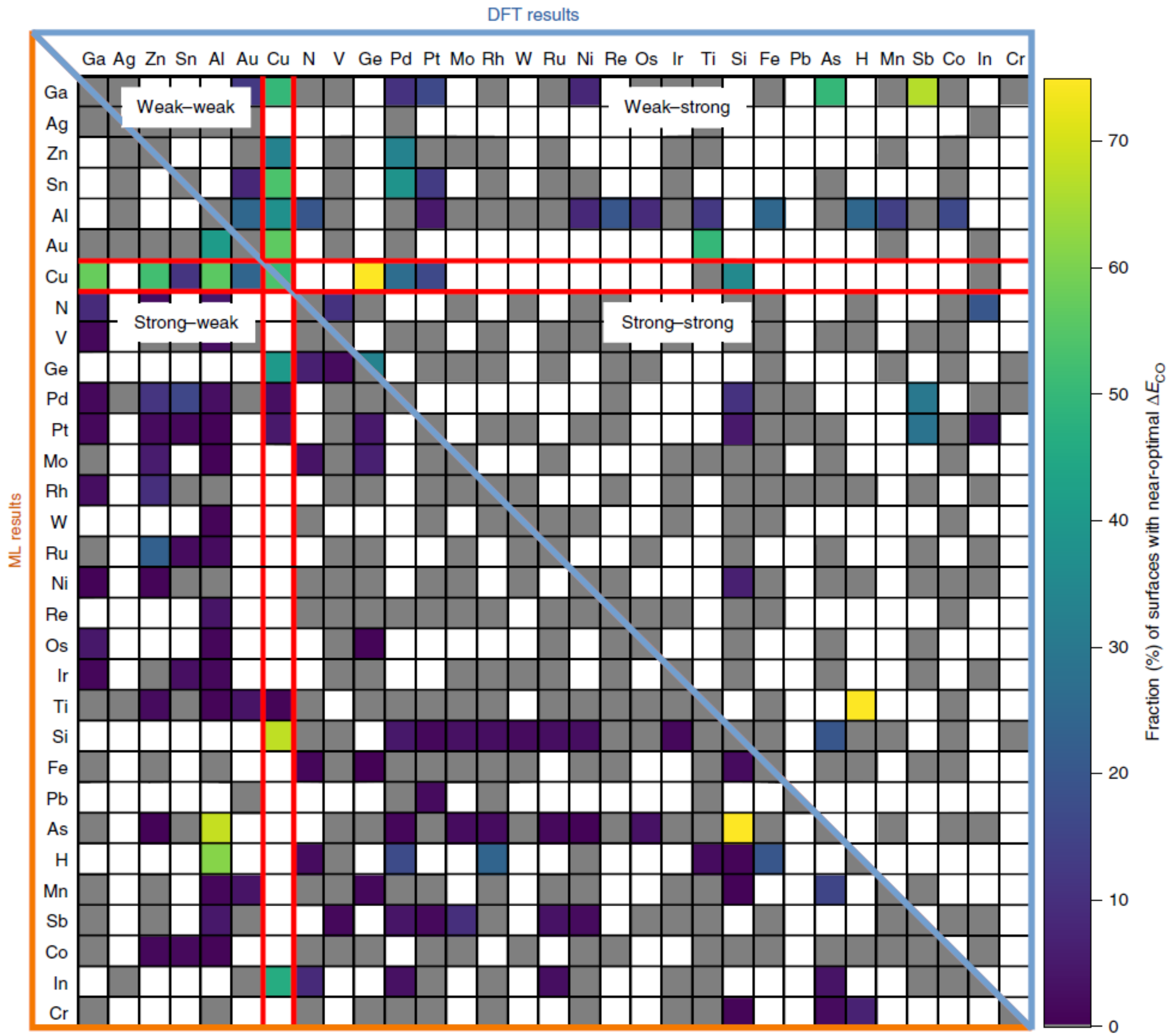


Y. Xiao, G. Ceder *et al.* *Joule*. **2019**, 3, 1-24.

高通量筛选

HER电极的筛选



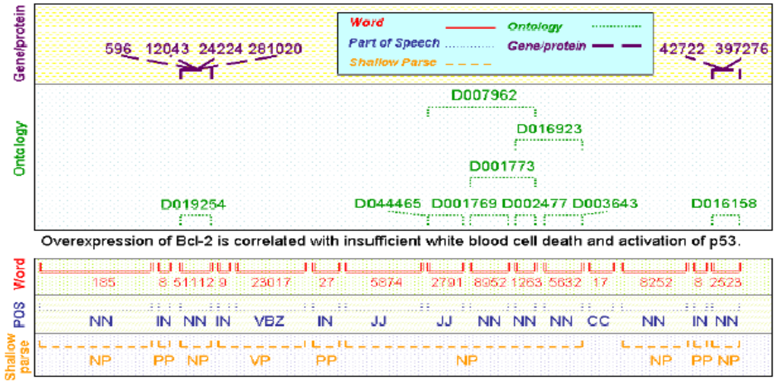


J. Greeley, J. K. Norskov *et al. Nat. Mater* **2006**, 5, 909.

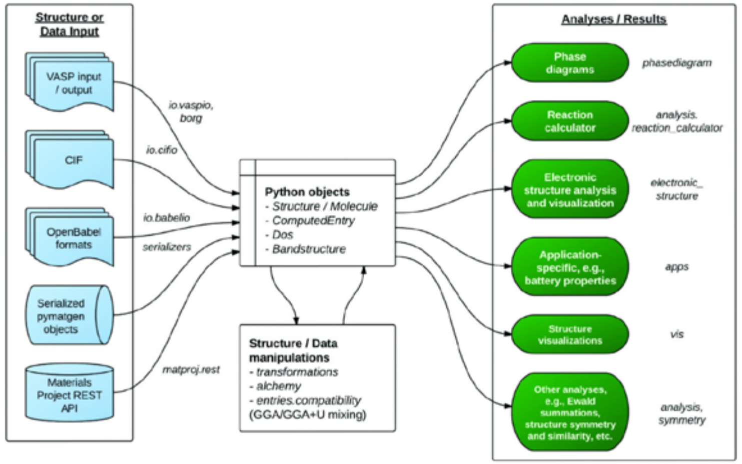
K. Tran and Z. W. Ulissi. *Nat. Cata* **2018**, 1, 696.

高通量数据来源

The Materials Project

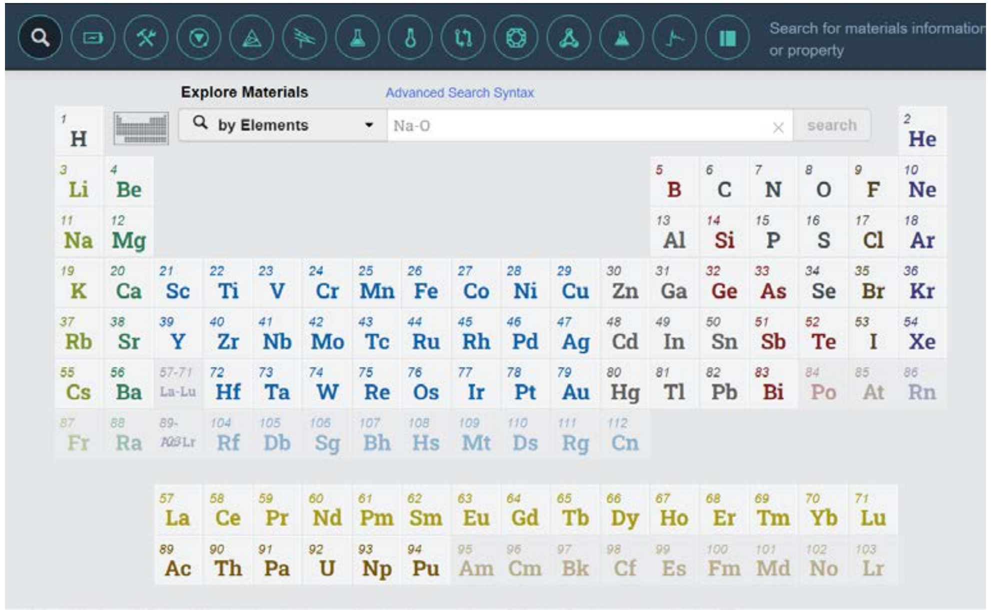


- 数据库
- DFT计算
- 文献数据挖掘
- 实验获取



目录

The Materials Project

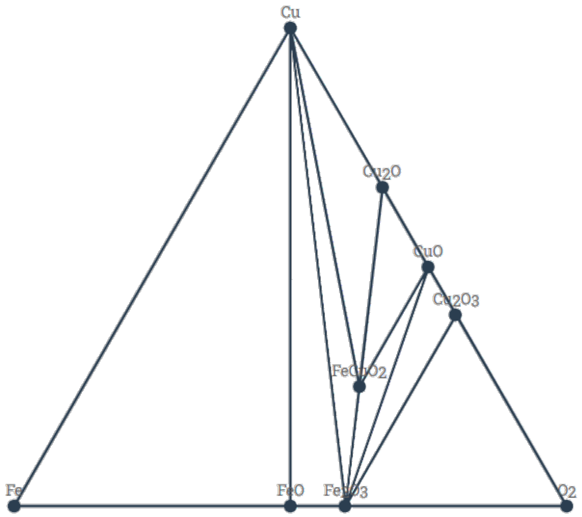


https://materialsproject.org

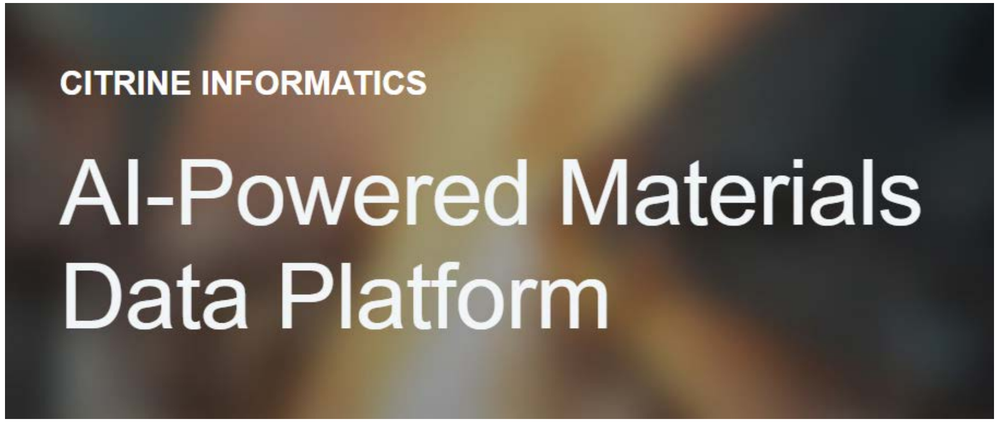
The Materials Project提供基于 Web 的开放式访问，可访问已知和预测材料的计算信息， 以及用于激发和设计新颖材料的强大分析工具。

$$MgO + Al_2O_3 \rightarrow MgAl_2O_4$$

$\Delta H_{calculated}$	$\Delta H_{experimental}$
-0.381 eV (-37 kJ mol^{-1})	-0.369 eV (-36 kJ mol^{-1})



Citrine Informatics



材料学数据库

https://citrination.com

大量的理论与实验数据

Names: alumina, alumina

Chemical Formula: Al_2O_3

Properties

Purity: 99.9 %

Elastic tensor

$$\begin{bmatrix} 497.5 & 162.7 & 115.5 & 22.5 & 0.0 & 0.0 \\ 162.7 & 497.5 & 115.5 & -22.5 & 0.0 & 0.0 \\ 115.5 & 115.5 & 503.3 & 0.0 & 0.0 & 0.0 \\ 22.5 & -22.5 & 0.0 & 147.4 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 147.4 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 22.5 & 167.4 \end{bmatrix} \text{GPa}$$

Data Type

EXPERIMENTAL

Methods

Resonant ultasound spectroscopy (RUS)

Elastic tensor

$$\begin{bmatrix} 495 & 171 & 130 & 20 & 0 & 0 \\ 171 & 495 & 130 & -20 & 0 & 0 \\ 130 & 130 & 486 & 0 & 0 & 0 \\ 20 & -20 & 0 & 148 & 0 & 0 \\ 0 & 0 & 0 & 0 & 148 & 0 \\ 0 & 0 & 0 & 0 & 20 & 162 \end{bmatrix} \text{GPa}$$

Data Type

COMPUTATIONAL

Methods

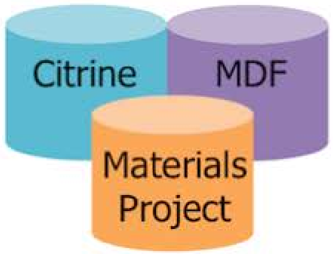
VASP 4.6 GGA/PBE

matminer

matminer

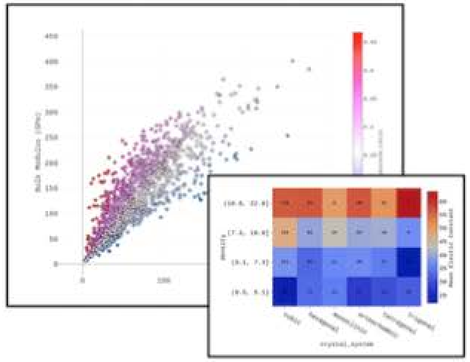
https://hackingmaterials.lbl.gov/matminer

数据检索 → 特征提取 → 可视化



Materials Databases

MATERIAL	FEATURES				PROPERTY
TiO_2 rutile	F_{11}	F_{12}	…	F_{1N}	gap = 3.0 eV
C diamond	F_{21}	F_{22}	…	F_{2N}	gap = 5.5 eV
…	…	…	…	…	…
PbTe rocksalt	F_{M1}	F_{M2}	…	F_{MN}	gap = 0.3 eV



scikit learn 机器学习 Keras

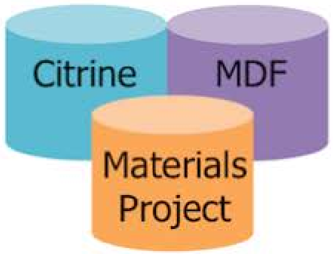
Ward, L., Jain, A., *et al. Comput. Mater. Sci.* **2018**, 152, 60-69.

matminer

matminer

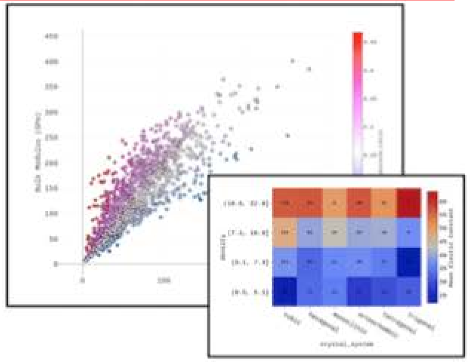
https://hackingmaterials.lbl.gov/matminer

数据检索 → 特征提取 → 可视化



Materials Databases

MATERIAL	FEATURES				PROPERTY
TiO_2 rutile	F_{11}	F_{12}	...	F_{1N}	gap = 3.0 eV
C diamond	F_{21}	F_{22}	...	F_{2N}	gap = 5.5 eV
...	...	...	...	...	...
PbTe rocksalt	F_{M1}	F_{M2}	...	F_{MN}	gap = 0.3 eV



↓

scikit learn — 机器学习 — Keras

Ward, L., Jain, A., *et al. Comput. Mater. Sci.* **2018**, 152, 60-69.

matminer简介

※ 可在线访问40多个现成的数据集

matminer.datasets

※ 从数据库中创建自己的数据集

matminer.data_retrieval

※ 将材料属性转换为描述符信息

matminer.featurizers

Table of Datasets

Find a table of all 42 datasets available in matminer here.

Name	Description	Entries
boltztrap_mp	Effective mass and thermoelectric properties of 8924 compounds in The Materials Project database that are calculated by the BoltzTraP software package run on the GGA-PBE or GGA+U density functional theory calculation results	8924
brgoch_superhard_training	2574 materials used for training regressors that predict shear and bulk modulus.	2574
castelli_perovskites	18,928 perovskites generated with ABX combinatorics, calculating gllbsc band gap and pbe structure, and also reporting absolute band edge positions and heat of formation.	18928

bandstructure

Features derived from a material’ s electronic bandstructure.

matminer. featurizers. bandstructure

Name	Description
BranchPointEnergy	Branch point energy and absolute band edge position.
BandFeaturizer	Featurizes a pymatgen band structure object.

base

Parent classes and meta-featurizers.

模块介绍

matminer 0.7.4 documentation » matminer (Materials Data Mining)

modules | index

Table of Contents

matminer

matminer is a Python library for data mining the properties of materials.

Matminer contains routines for:

- **one-line access to 40+ ready-made datasets** (`matminer.datasets`)
  - Spans various domains of materials data
  - Full list of datasets here: Table of Datasets
- **easily creating your own datasets from online repositories** (`matminer.data_retrieval`)
  - such as The Materials Project and Citrination, among others
- **transforming and featurizing complex materials attributes into numerical descriptors** (`matminer.featurizers`)
  - 70+ featurizers adapted from scientific publications
  - Feature generation routines for

modules:
介绍matminer中的所有模块、子模块以及模块中包含的方法

index:
以检索的形式列出了matminer中所有的方法

目录

数据载入

load_dataset()函数

用于导入matminer
集成的任一数据集

In [1]:
```python
from matminer.datasets import load_dataset
df = load_dataset('expt_gap')
```

In [2]:
```python
df
```

Out[2]:

	formula	gap expt
0	Hg0.7Cd0.3Te	0.35
1	CuBr	3.08
2	LuP	1.30
3	Cu3SbSe4	0.40
4	ZnO	3.44
...	...	...
6349	Tm2MgTl	0.00
6350	Nb5Ga4	0.00
6351	Tb2Sb5	0.00
6352	Lu2AlTc	0.00
6353	CeZnPO	0.00

6354 rows × 2 columns

数据载入

Submodules

matminer.datasets.convenience_loaders module

子模块介绍：convenience_loaders用于数据载入

matminer.datasets.convenience_loaders.**load_expt_gap**(*data_home=None, download_if_missing=True*)

Convenience function for loading the expt_gap dataset.me

方法介绍：导入对应数据集的方法

Args:

data_home (str, None): Where to look for and store the loaded dataset

download_if_missing (bool): Whether or not to download the dataset if it isn't on disk

参数介绍：默认参数可完成

Returns: (pd.DataFrame)

返回值类型：DataFrame

数据载入

使用convenience_loaders方法

In [3]:
```
from matminer.datasets.convenience_loaders import load_expt_gap

df = load_expt_gap()
df
```

Out[3]:

	formula	gap expt
0	Hg0.7Cd0.3Te	0.35
1	CuBr	3.08
2	LuP	1.30
3	Cu3SbSe4	0.40
4	ZnO	3.44
...	...	...
6349	Tm2MgTl	0.00
6350	Nb5Ga4	0.00
6351	Tb2Sb5	0.00
6352	Lu2AlTc	0.00
6353	CeZnPO	0.00

6354 rows × 2 columns

查看数据集

查看数据集中不同列的含义

expt_gap

Experimental band gap of 6354 inorganic semiconductors.

Number of entries: 6354

化学式

带隙

Column	Description
`formula`	chemical formula
`gap expt`	band gap (in eV) measured experimentally

Reference

https://pubs.acs.org/doi/suppl/10.1021/acs.jpclett.8b00124

查看数据集

查看单个数据　　df.loc[0]

```
In  [12]: df.loc[0]

Out[12]: formula      Hg0.7Cd0.3Te
         gap expt             0.35
         Name: 0, dtype: object
```

字段查找　　df.loc[df['formula']== 'CuBr']

```
In  [14]: df.loc[df['formula'] == 'CuBr']

Out[14]:
```

	formula	gap expt
1	CuBr	3.08
16	CuBr	2.94
455	CuBr	3.08
457	CuBr	2.91
1705	CuBr	3.07
2783	CuBr	2.90
2789	CuBr	3.02
3774	CuBr	2.99

练习

练习：导入并查看
elastic_tensor_2015
数据集

以Materials Project为例

In [25]:

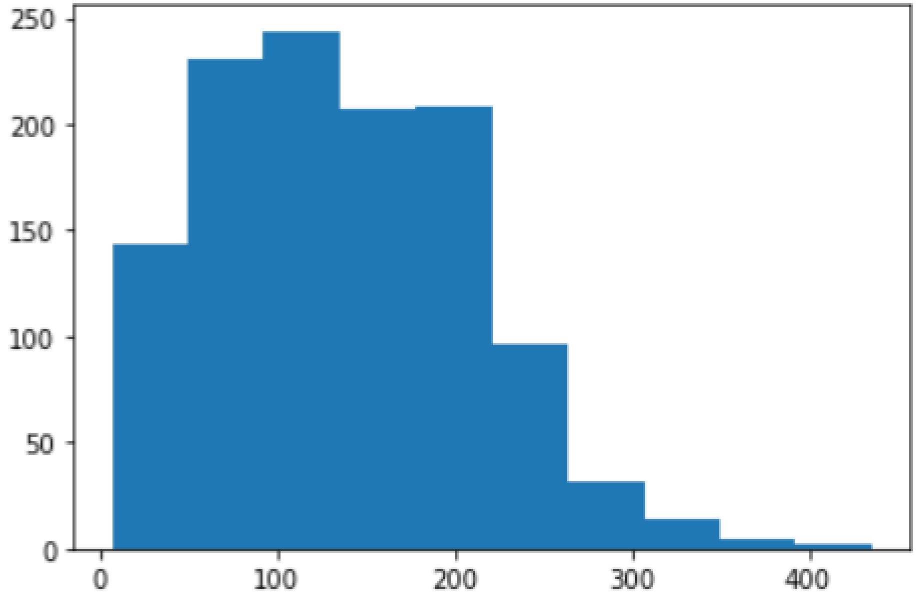
```
from matminer.data_retrieval.retrieve_MP import MPDataRetrieval
mpd = MPDataRetrieval(api_key="YOUR API KEY")
data = mpd.get_data('Fe2O3',['formula','band_gap'])
df_test = mpd.get_dataframe(criteria='Si-O',properties=['formula','band_gap'])
for d1 in data:
    print(d)
print(df_test)
for d2 in df_test:
    print(dt)
```

```
{'formula': {'Fe': 2.0, 'O': 3.0}, 'band_gap': 0.22019999999999995, 'material_id': 'mp-1244869'}
{'formula': {'Fe': 2.0, 'O': 3.0}, 'band_gap': 1.5673, 'material_id': 'mp-1456'}
{'formula': {'Fe': 2.0, 'O': 3.0}, 'band_gap': 1.4248000000000003, 'material_id': 'mp-715276'}
{'formula': {'Fe': 2.0, 'O': 3.0}, 'band_gap': 0.37980000000000014, 'material_id': 'mp-1245154'}
{'formula': {'Fe': 2.0, 'O': 3.0}, 'band_gap': 0.0, 'material_id': 'mp-1078361'}
{'formula': {'Fe': 2.0, 'O': 3.0}, 'band_gap': 1.1123, 'material_id': 'mp-1245078'}
{'formula': {'Fe': 2.0, 'O': 3.0}, 'band_gap': 0.0, 'material_id': 'mp-716814'}
{'formula': {'Fe': 2.0, 'O': 3.0}, 'band_gap': 1.3464999999999998, 'material_id': 'mp-715572'}
{'formula': {'Fe': 2.0, 'O': 3.0}, 'band_gap': 0.0, 'material_id': 'mp-1068212'}
{'formula': {'Fe': 2.0, 'O': 3.0}, 'band_gap': 1.4464000000000001, 'material_id': 'mp-510080'}
```

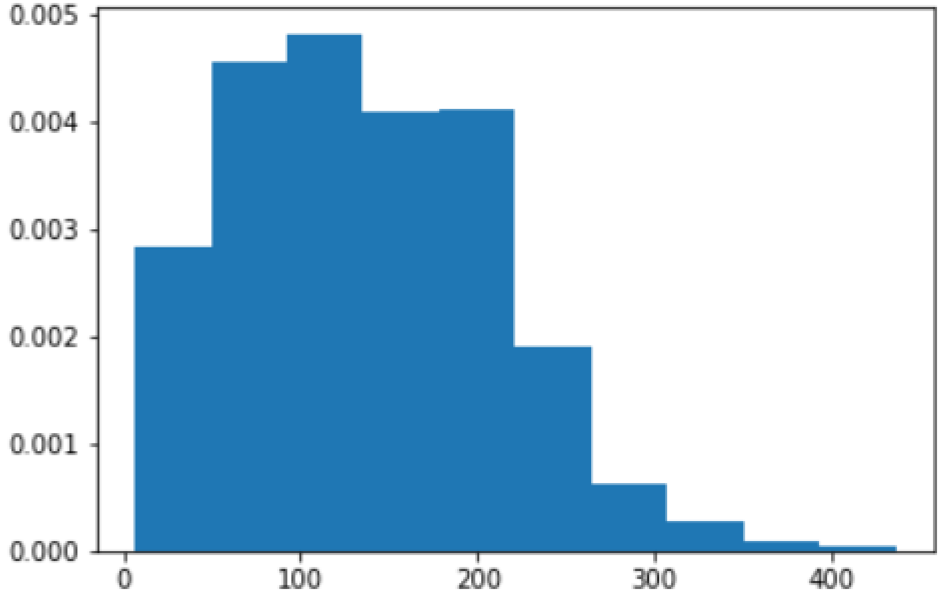
目录

数据可视化——分布图

In [7]:
```
import matplotlib.pyplot as plt
plt.hist(df_el['K_VRH'])
plt.show()
```

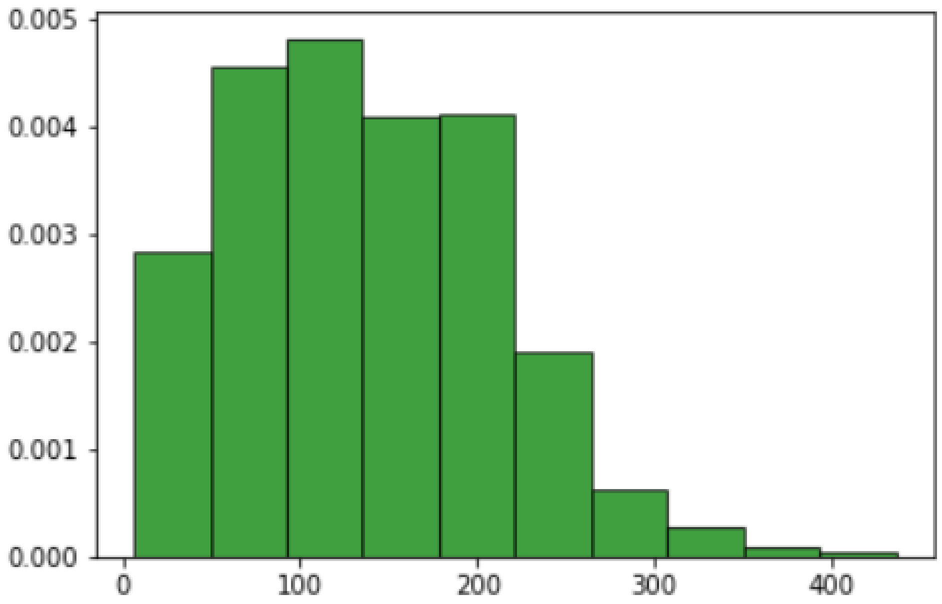


In [8]:
```
plt.hist(df_el['K_VRH'], density = True)
plt.show()
```



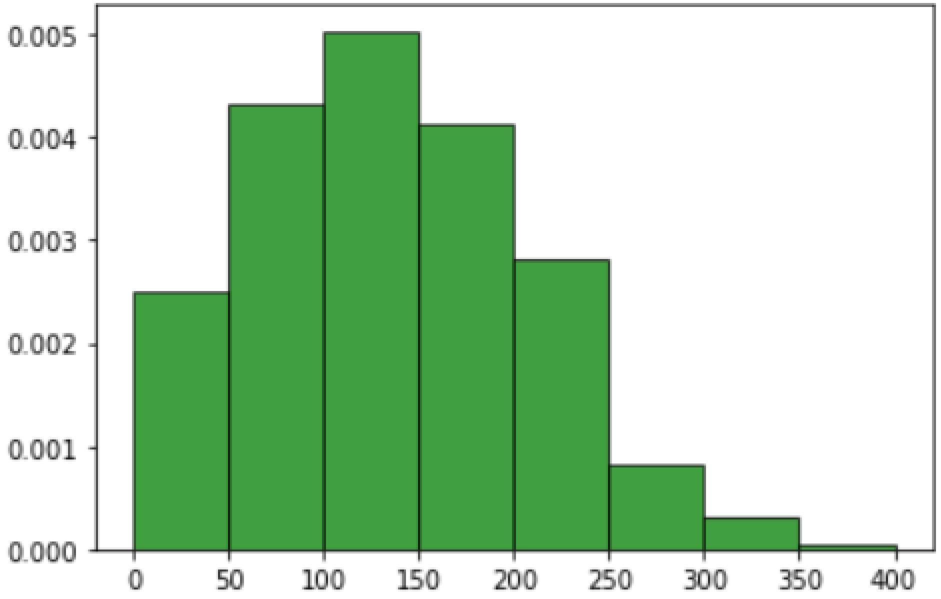
数据可视化——分布图

In [9]:
```
plt.hist(df_el['K_VRH'], density = True, color = 'g', edgecolor = 'k', alpha = 0.75)
plt.show()
```



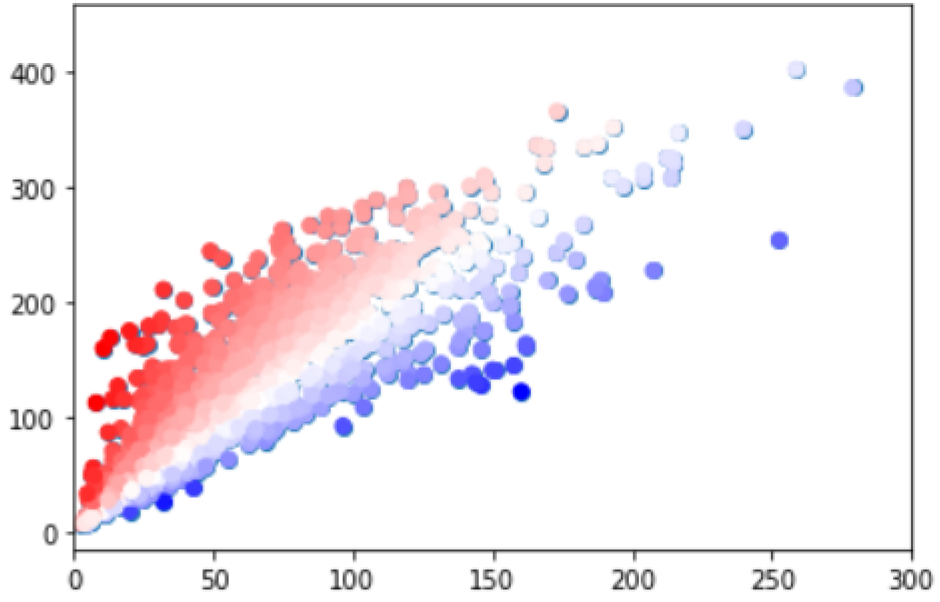
数据可视化——分布图

In [10]:
```
plt.hist(df_el['K_VRH'], density = True, color = 'g', edgecolor = 'k', alpha = 0.75,
         bins = [0, 50, 100, 150, 200, 250, 300, 350, 400])
plt.show()
```



数据可视化——弹性模量

In [11]:
```
import numpy as np
x = np.array(df_el['G_VRH'])
y = np.array(df_el['K_VRH'])
plt.scatter(x,y)
z = np.array(df_el['poisson_ratio'])
plt.xlim(0,300)
plt.scatter(x,y,c=z,cmap='bwr')
plt.show()
```



数据可视化——弹性模量

```
In  [12]: import plotly.express as px

fig = px.scatter(df_el,x = 'G_VRH',y = 'K_VRH',color= 'poisson_ratio',
                 color_continuous_scale = 'emrld',range_x=[0,300])
fig.show()
```

目录

数据导入

In [1]:
```
from matminer.datasets import load_dataset
df_mp = load_dataset('mp_nostruct_20181018')
```

In [2]:
```
df_mp
```

Out[2]:

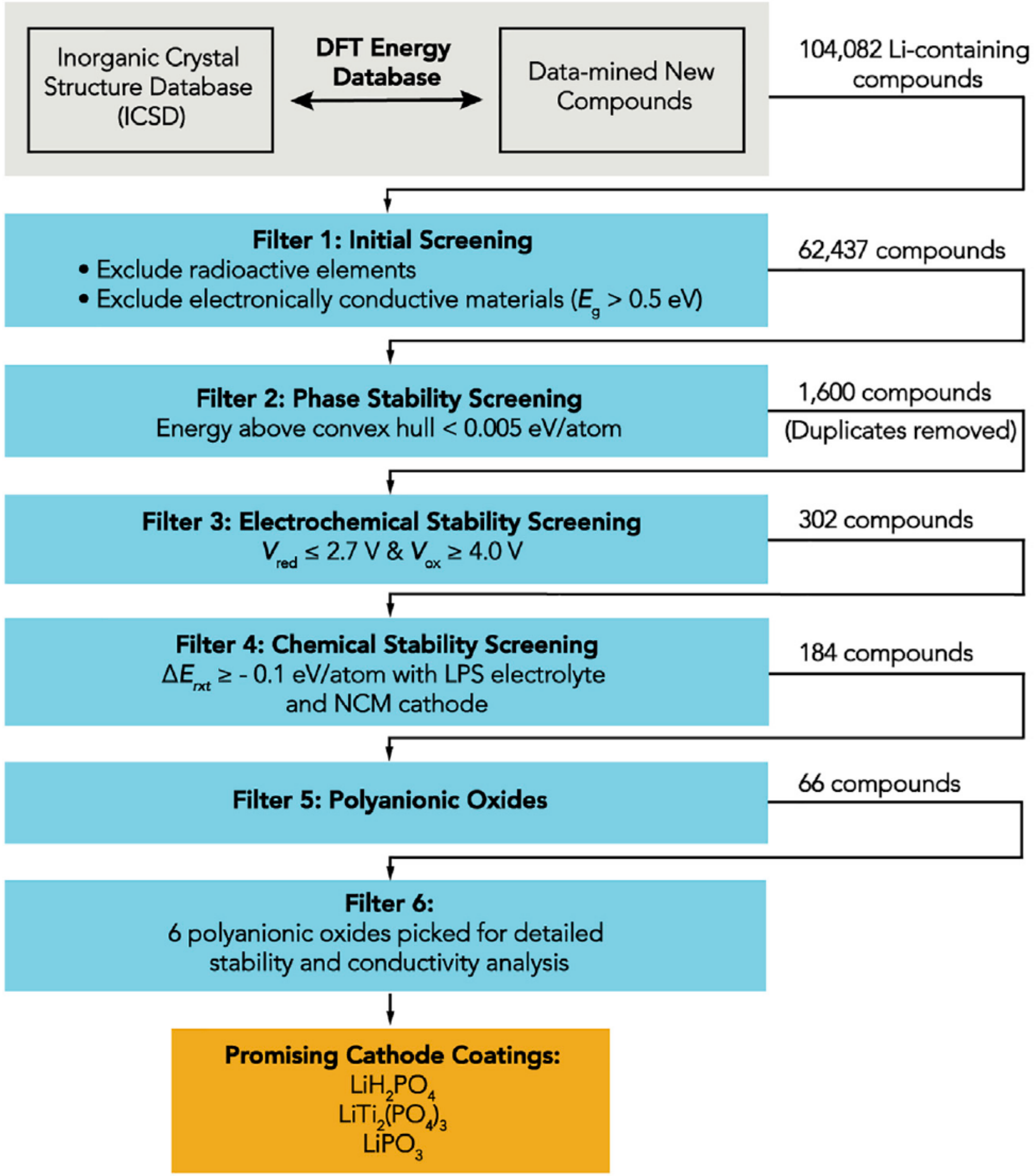
	mpid	formula	e_hull	gap pbe	mu_b	elastic anisotropy	bulk modulus	shear modulus	e_form
0	mp-85	In	0.003319	0.0000	2.700000e-05	1.044699	33.154748	4.904836	0.003319
1	mp-110	Mg	0.039182	0.0000	-1.360000e-05	-11.326659	35.636106	1.830272	0.039182
2	mp-20	Be	0.108143	0.0000	1.000000e-07	8.030000	124.000000	84.000000	0.108143
3	mp-8640	Hf	0.071216	0.0000	-2.050000e-05	0.881277	101.242732	44.836516	0.071216
4	mp-674158	P	3.509988	2.0113	3.000042e+00	10.884643	0.327165	-0.064038	3.509988
...	...	...	...	...	...	...	...	...	...
83984	mp-4446	Sr3(GaO3)2	0.000691	3.5262	0.000000e+00	NaN	NaN	NaN	-2.832238
83985	mp-3393	Sr3Al2O6	0.000000	4.2046	0.000000e+00	NaN	NaN	NaN	-3.358646
83986	mp-24696	MgSb2(H4O3)6	0.028109	3.2827	-4.338000e-04	NaN	NaN	NaN	-1.533338
83987	mp-23984	GaH18C3(N3F2)3	0.000000	4.9759	2.460000e-05	NaN	NaN	NaN	-1.066094
83988	mp-24554	AlH18C3(N3F2)3	0.000000	5.3705	5.312000e-04	NaN	NaN	NaN	-1.161128

83989 rows × 9 columns

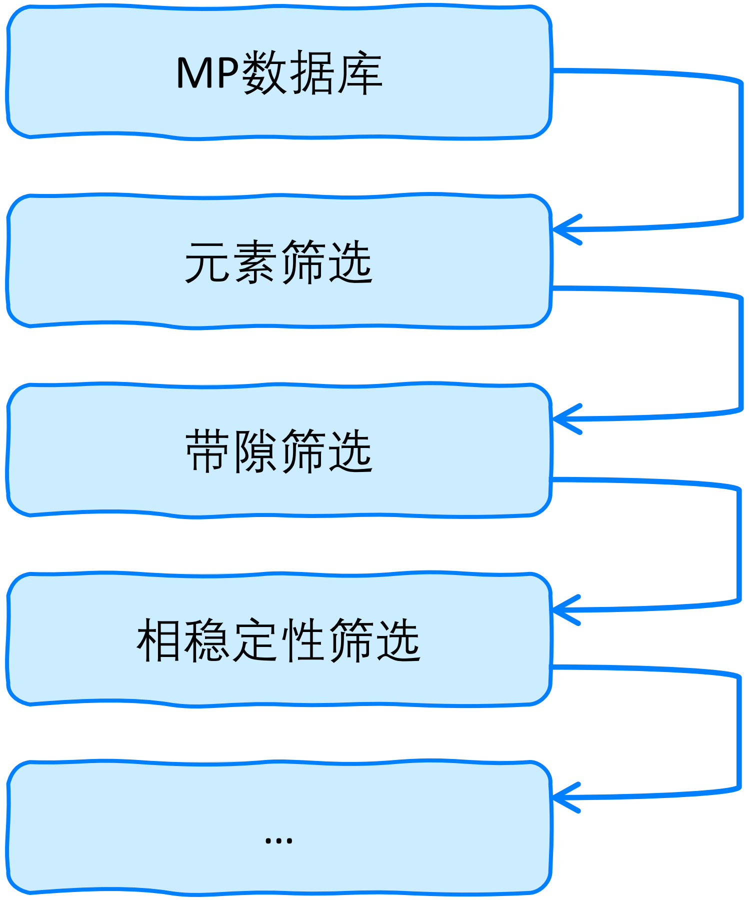
数据

结果统计

高通量筛选



Inorganic Crystal Structure Database (ICSD)
DFT Energy Database
Data-mined New Compounds
104,082 Li-containing compounds
Filter 1: Initial Screening
• Exclude radioactive elements
• Exclude electronically conductive materials (E_g > 0.5 eV)
62,437 compounds
Filter 2: Phase Stability Screening
Energy above convex hull < 0.005 eV/atom
1,600 compounds
(Duplicates removed)
Filter 3: Electrochemical Stability Screening
$V_{red} \leq 2.7$ V & $V_{ox} \geq 4.0$ V
302 compounds
Filter 4: Chemical Stability Screening
$\Delta E_{rxt} \geq -0.1$ eV/atom with LPS electrolyte and NCM cathode
184 compounds
Filter 5: Polyanionic Oxides
66 compounds
Filter 6:
6 polyanionic oxides picked for detailed stability and conductivity analysis
Promising Cathode Coatings:
LiH_2PO_4
$LiTi_2(PO_4)_3$
$LiPO_3$



MP数据库
元素筛选
带隙筛选
相稳定性筛选
…

预处理

```
In  [3]: num = df_mp.isna().sum()
         num

Out[3]: mpid                      0
        formula                   2
        e_hull                    0
        gap pbe                   0
        mu_b                      0
        elastic anisotropy    76313
        bulk modulus          76313
        shear modulus         76313
        e_form                    0
        dtype: int64
```

包含nan数据

预处理

筛去nan数据（83989 → 83987）

In [5]:

```
df_mp = df_mp.dropna(axis = 0, subset = ['formula'])
df_mp
```

Out[5]:

	mpid	formula	e_hull	gap pbe	mu_b	elastic anisotropy	bulk modulus	shear modulus	e_form
0	mp-85	In	0.003319	0.0000	2.700000e-05	1.044699	33.154748	4.904836	0.003319
1	mp-110	Mg	0.039182	0.0000	-1.360000e-05	-11.326659	35.636106	1.830272	0.039182
2	mp-20	Be	0.108143	0.0000	1.000000e-07	8.030000	124.000000	84.000000	0.108143
3	mp-8640	Hf	0.071216	0.0000	-2.050000e-05	0.881277	101.242732	44.836516	0.071216
4	mp-674158	P	3.509988	2.0113	3.000042e+00	10.884643	0.327165	-0.064038	3.509988
...	...	...	...	...	...	...	...	...	...
83984	mp-4446	Sr3(GaO3)2	0.000691	3.5262	0.000000e+00	NaN	NaN	NaN	-2.832238
83985	mp-3393	Sr3Al2O6	0.000000	4.2046	0.000000e+00	NaN	NaN	NaN	-3.358646
83986	mp-24696	MgSb2(H4O3)6	0.028109	3.2827	-4.338000e-04	NaN	NaN	NaN	-1.533338
83987	mp-23984	GaH18C3(N3F2)3	0.000000	4.9759	2.460000e-05	NaN	NaN	NaN	-1.066094
83988	mp-24554	AlH18C3(N3F2)3	0.000000	5.3705	5.312000e-04	NaN	NaN	NaN	-1.161128

83987 rows × 9 columns

筛选1

是否含Li（83987 → 13943）

In [6]:
```
df_mp_Li = df_mp.loc[df_mp['formula'].str.contains('Li')]
df_mp_Li
```

Out[6]:

	mpid	formula	e_hull	gap pbe	mu_b	elastic anisotropy	bulk modulus	shear modulus	e_form
15	mp-51	Li	0.002860	0.0000	0.000100	-4.976255	13.860513	15.128887	0.002860
29	mp-567337	Li	0.017123	0.0000	0.000390	1.890000	14.000000	7.000000	0.017123
143	mp-135	Li	0.000000	0.0000	0.000072	12.177018	14.012877	4.480159	0.000000
207	mp-2314	LiPb	0.000000	0.0000	0.000062	NaN	NaN	NaN	-0.273765
260	mp-934	LiTl	0.000000	0.0000	-0.000068	0.764436	31.438228	16.386834	-0.230930
...	...	...	...	...	...	...	...	...	...
83839	mp-601344	LiZr3H18N4F19	0.000000	6.0167	-0.033311	NaN	NaN	NaN	-2.311442
83862	mp-686484	LiCa9Mg(PO4)7	0.000000	4.7063	-0.000302	NaN	NaN	NaN	-3.286062
83946	mp-686230	Li20Nb19O60	0.055496	0.0000	6.630597	NaN	NaN	NaN	-2.777481
83968	mp-723059	Li3Nd2H6(N3O10)3	0.000000	3.5123	-0.007566	NaN	NaN	NaN	-1.506936
83970	mp-722330	Li3La2H6(N3O10)3	0.000000	3.5098	-0.006423	NaN	NaN	NaN	-1.526244

13943 rows × 9 columns

筛选2

是否是金属（813943 → 8990）

```
In [9]: df_mp_Li_gp = df_mp_Li.loc[df_mp_Li['gap pbe'] > 0.5]
        df_mp_Li_gp
```

Out[9]:

	mpid	formula	e_hull	gap pbe	mu_b	elastic anisotropy	bulk modulus	shear modulus	e_form
333	mp-23259	LiBr	0.025492	4.9234	0.000000e+00	0.290976	21.062752	15.948641	-1.547844
942	mp-23703	LiH	0.000000	2.9737	0.000000e+00	0.096891	36.063260	42.924750	-0.489313
1355	mp-1138	LiF	0.000000	8.7161	-9.000000e-07	0.158661	69.881504	50.943440	-3.180880
1447	mp-22899	LiI	0.036396	4.2306	0.000000e+00	0.057074	20.634770	12.967352	-1.199312
1451	mp-22905	LiCl	0.000000	6.2500	0.000000e+00	0.206676	31.939069	21.114162	-2.107280
...	...	...	...	...	...	...	...	...	...
83837	mp-699932	Ba3Li2Mo4P6(ClO14)2	0.002449	2.1138	1.599965e+01	NaN	NaN	NaN	-2.688131
83839	mp-601344	LiZr3H18N4F19	0.000000	6.0167	-3.331140e-02	NaN	NaN	NaN	-2.311442
83862	mp-686484	LiCa9Mg(PO4)7	0.000000	4.7063	-3.017000e-04	NaN	NaN	NaN	-3.286062
83968	mp-723059	Li3Nd2H6(N3O10)3	0.000000	3.5123	-7.565500e-03	NaN	NaN	NaN	-1.506936
83970	mp-722330	Li3La2H6(N3O10)3	0.000000	3.5098	-6.423500e-03	NaN	NaN	NaN	-1.526244

8990 rows × 9 columns

筛选3

是否稳定（8990 → 1482）

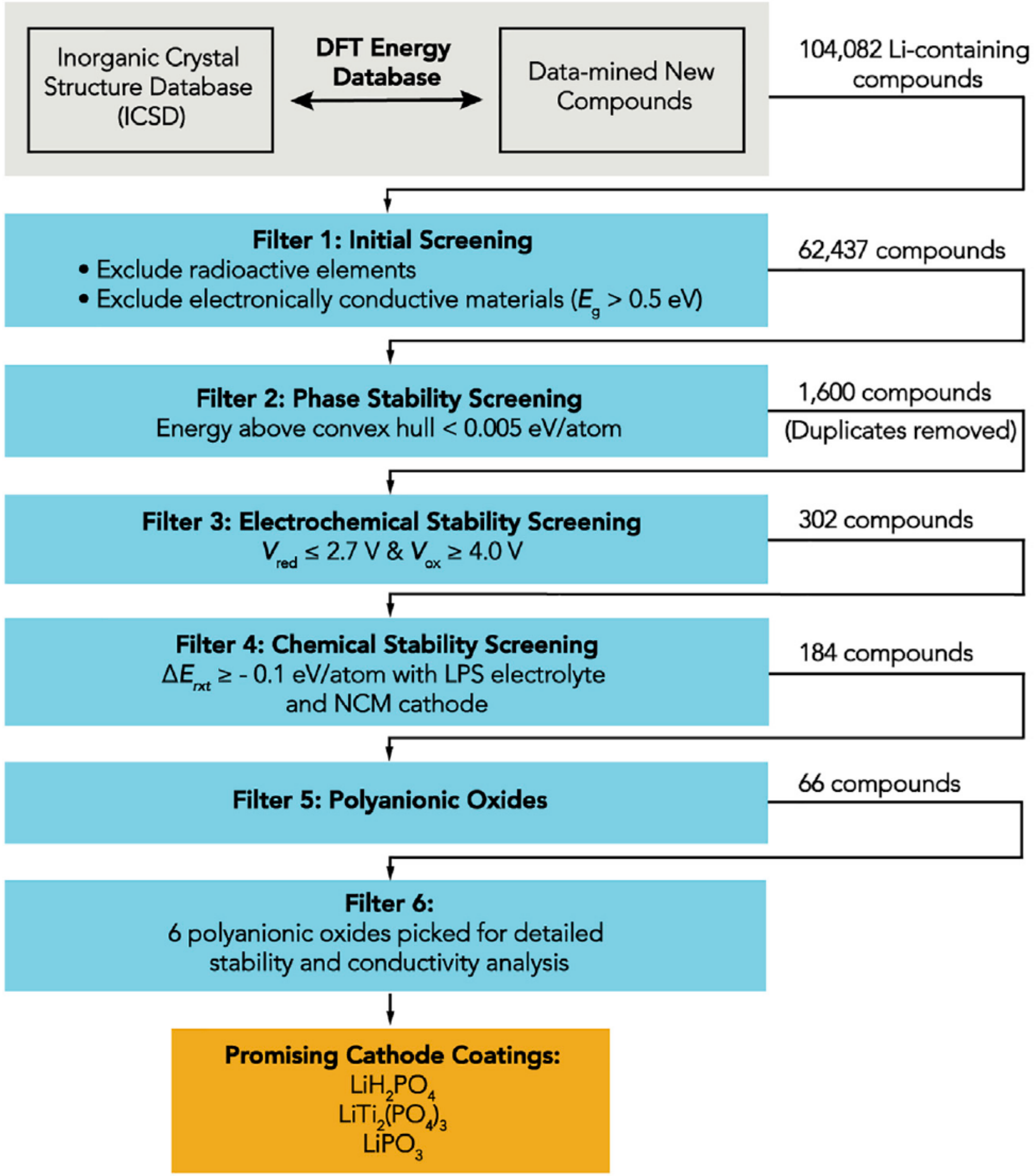
In [7]:
```python
df_mp_Li_gp_hull = df_mp_Li_gp.loc[df_mp_Li_gp['e_hull'] < 0.005]
df_mp_Li_gp_hull
```

Out[7]:

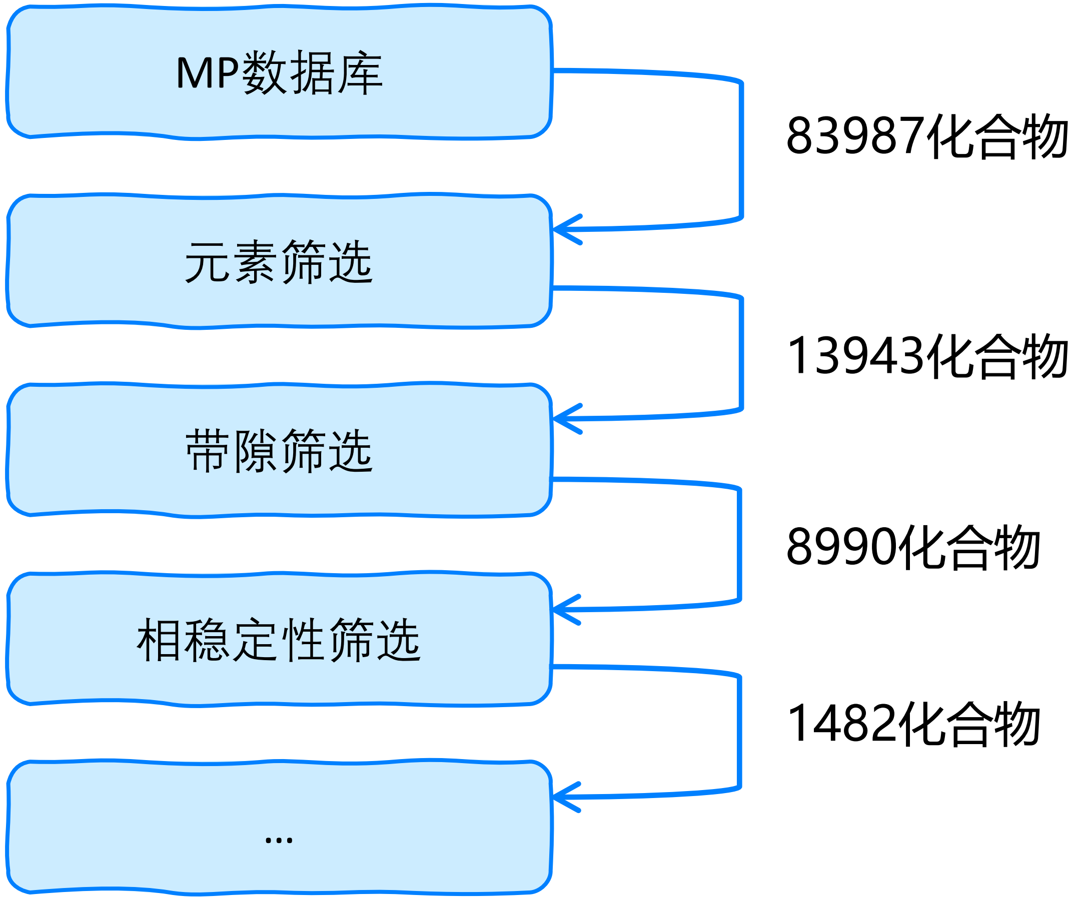
	mpid	formula	e_hull	gap pbe	mu_b	elastic anisotropy	bulk modulus	shear modulus	e_form
942	mp-23703	LiH	0.000000	2.9737	0.000000e+00	0.096891	36.063260	42.924750	-0.489313
1355	mp-1138	LiF	0.000000	8.7161	-9.000000e-07	0.158661	69.881504	50.943440	-3.180880
1451	mp-22905	LiCl	0.000000	6.2500	0.000000e+00	0.206676	31.939069	21.114162	-2.107280
1587	mp-7575	LiZnN	0.000000	0.5083	0.000000e+00	0.345628	115.754088	84.897980	-0.389165
1700	mp-9124	LiZnAs	0.000000	0.5475	1.642000e-04	0.071256	54.738221	40.055683	-0.519940
...	...	...	...	...	...	...	...	...	...
83837	mp-699932	Ba3Li2Mo4P6(ClO14)2	0.002449	2.1138	1.599965e+01	NaN	NaN	NaN	-2.688131
83839	mp-601344	LiZr3H18N4F19	0.000000	6.0167	-3.331140e-02	NaN	NaN	NaN	-2.311442
83862	mp-686484	LiCa9Mg(PO4)7	0.000000	4.7063	-3.017000e-04	NaN	NaN	NaN	-3.286062
83968	mp-723059	Li3Nd2H6(N3O10)3	0.000000	3.5123	-7.565500e-03	NaN	NaN	NaN	-1.506936
83970	mp-722330	Li3La2H6(N3O10)3	0.000000	3.5098	-6.423500e-03	NaN	NaN	NaN	-1.526244

1482 rows × 9 columns

高通量筛选



Inorganic Crystal Structure Database (ICSD)
DFT Energy Database
Data-mined New Compounds
104,082 Li-containing compounds
Filter 1: Initial Screening
• Exclude radioactive elements
• Exclude electronically conductive materials (E_g > 0.5 eV)
62,437 compounds
Filter 2: Phase Stability Screening
Energy above convex hull < 0.005 eV/atom
1,600 compounds
(Duplicates removed)
Filter 3: Electrochemical Stability Screening
$V_{red} \leq 2.7$ V & $V_{ox} \geq 4.0$ V
302 compounds
Filter 4: Chemical Stability Screening
$\Delta E_{rxt} \geq -0.1$ eV/atom with LPS electrolyte and NCM cathode
184 compounds
Filter 5: Polyanionic Oxides
66 compounds
Filter 6:
6 polyanionic oxides picked for detailed stability and conductivity analysis
Promising Cathode Coatings:
LiH_2PO_4
$LiTi_2(PO_4)_3$
$LiPO_3$

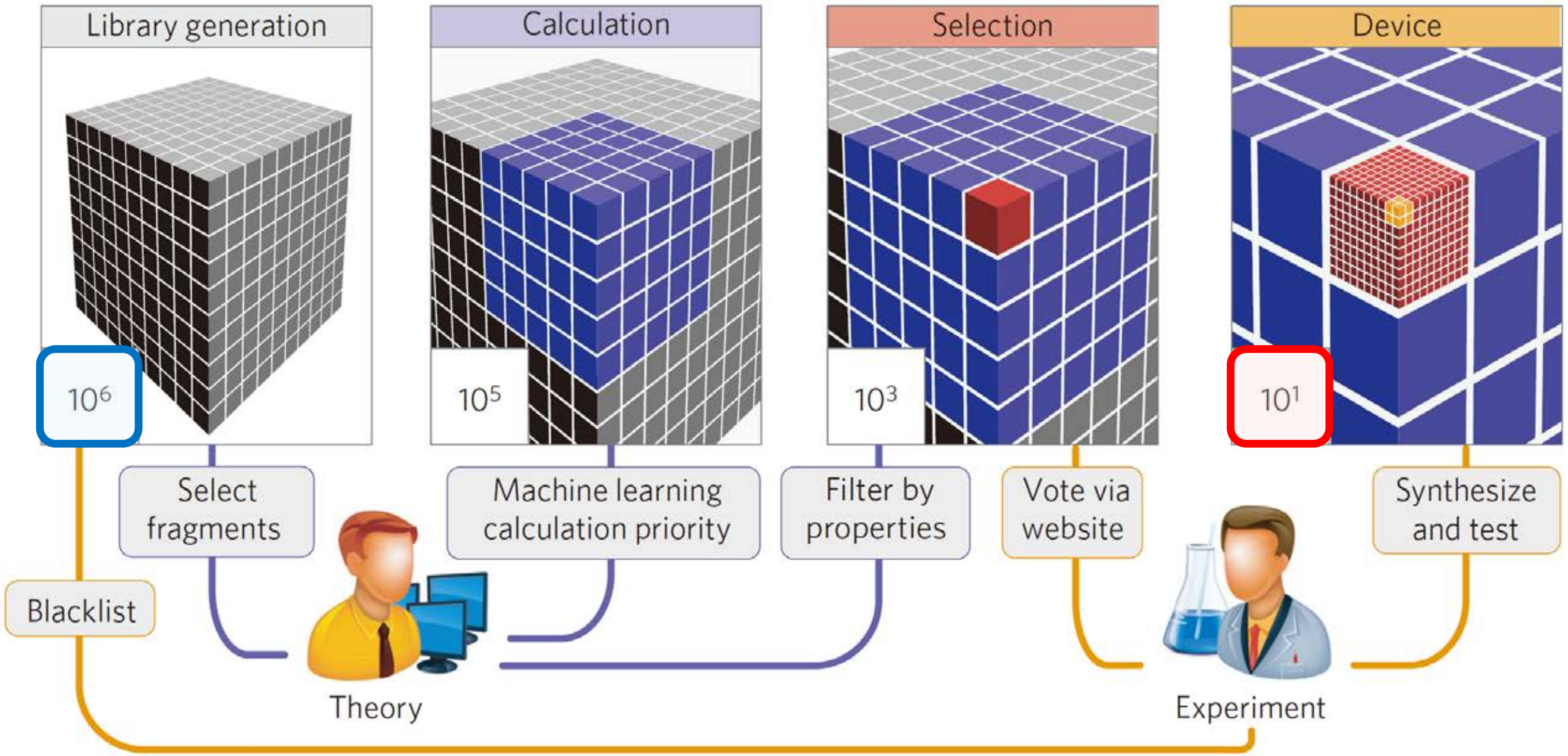


MP数据库
83987化合物
元素筛选
13943化合物
带隙筛选
8990化合物
相稳定性筛选
1482化合物
…

目录

高通量筛选中的机器学习

各节点的分类依据可借助机器学习方法进行判断



R. G. Bombarelli, A. A. Guzik *et al. Nat. Mater.* **2016**, 15, 1120-1127.

高通量筛选

TD-DFT吸收光谱近似为实验发射光谱

随机选择分子进行计算回归获得机器学习模型