机器学习

在材料设计中的应用

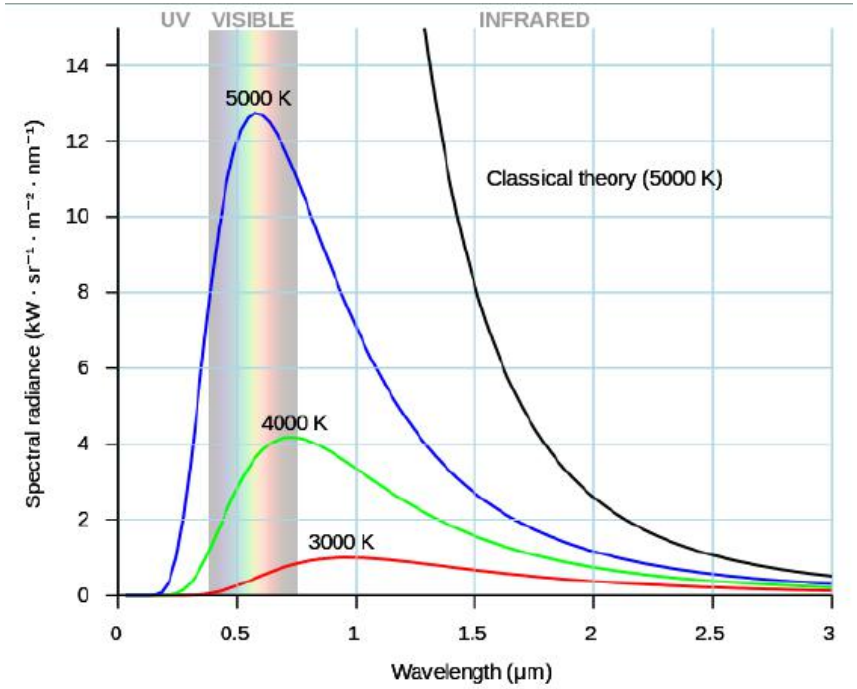
The Story of Light Bulb

6000+ materials have been tested

The Story of Light Bulb

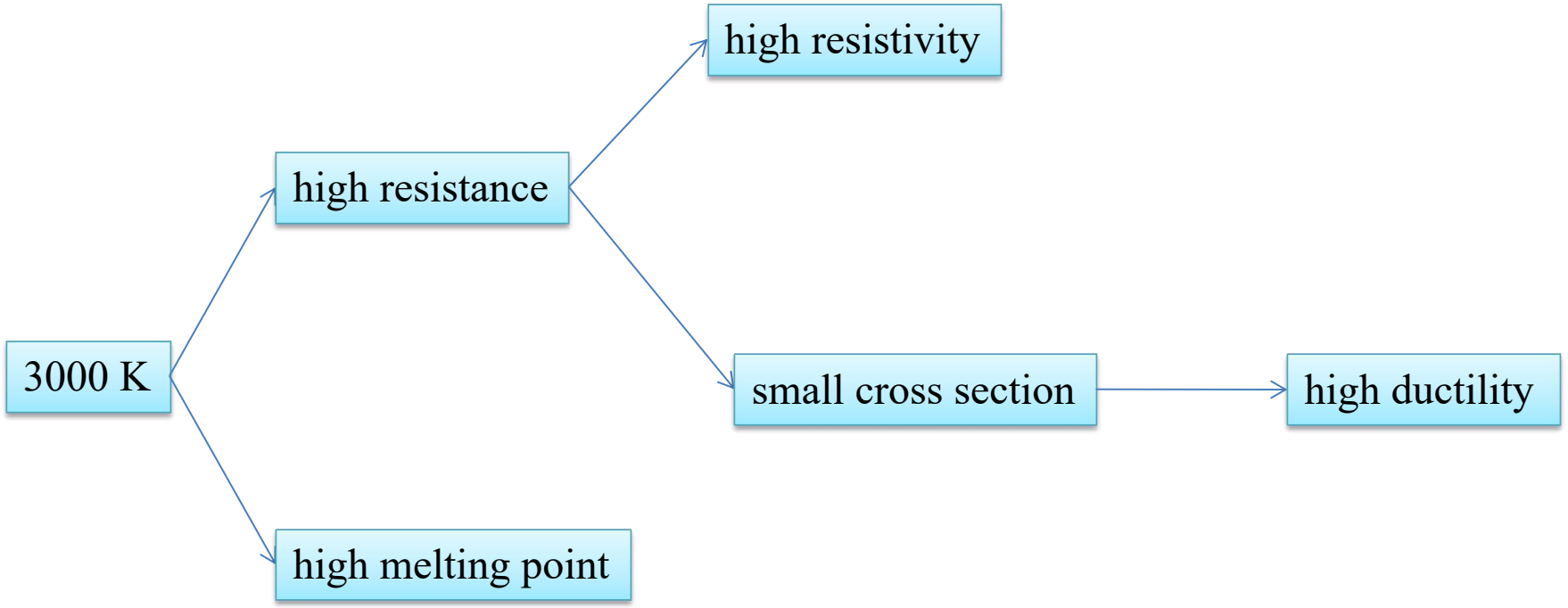
Among numerous experiments, Edison's team finally discovered that carbonized bamboo filament could last over 1,200 hours.

Physics: black body radiation

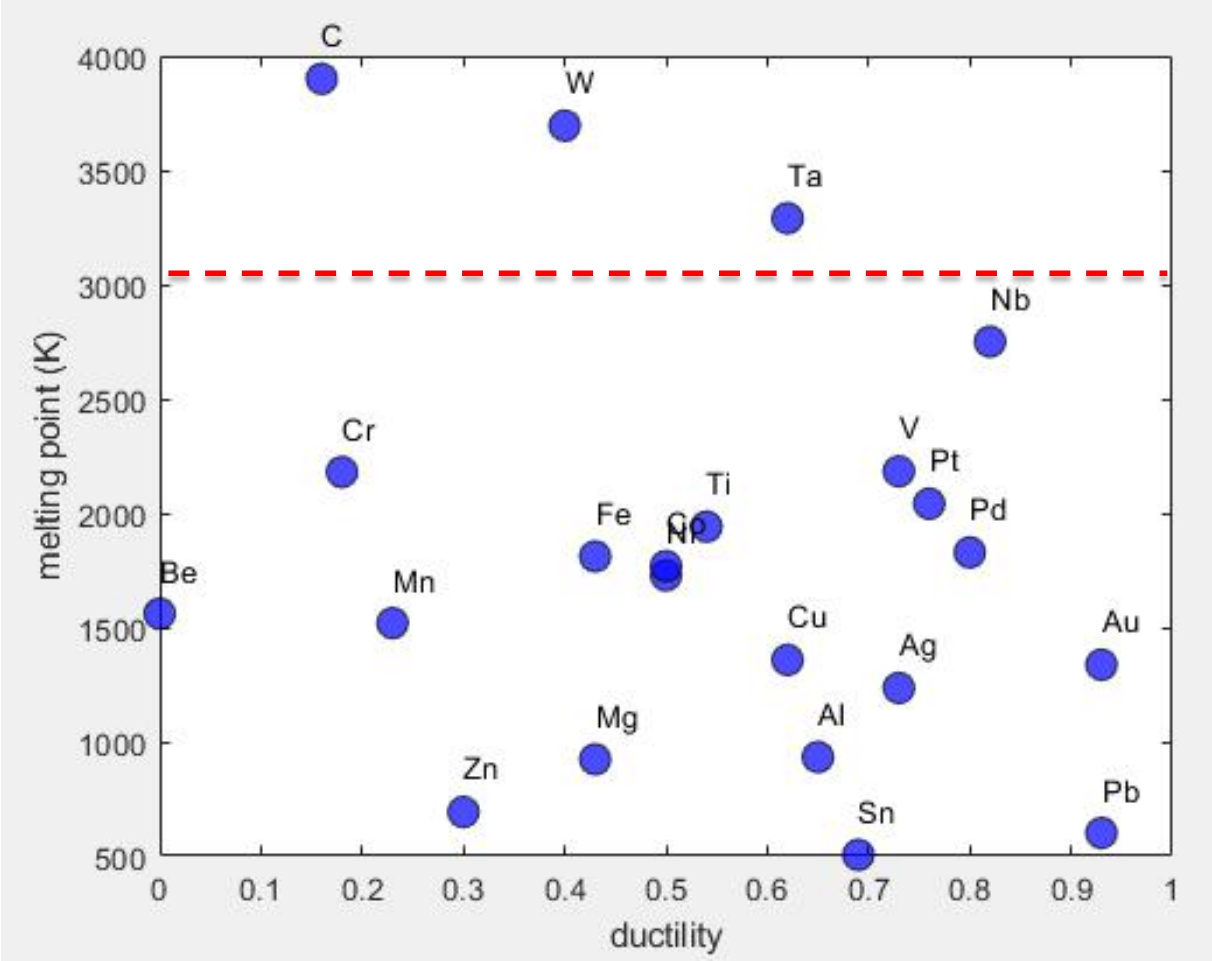


$$B(\nu, T) = \frac{2h\nu^3}{c^2} \frac{1}{e^{\frac{h\nu}{k_B T}} - 1}$$

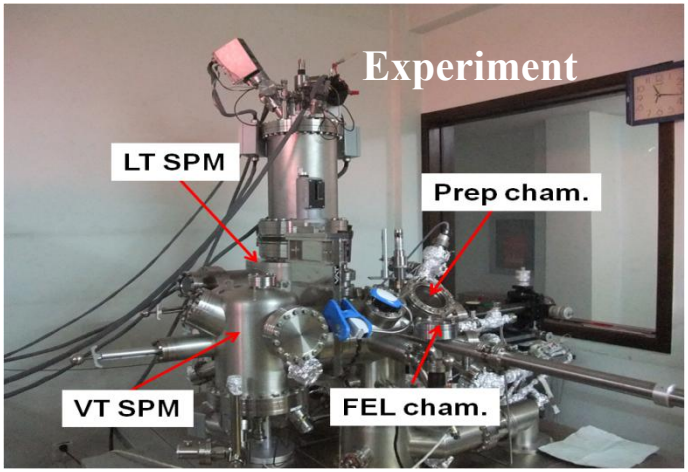
Melting point and ductility

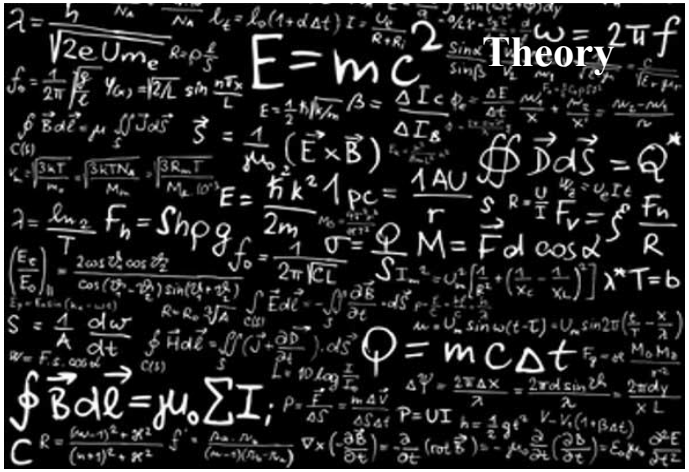


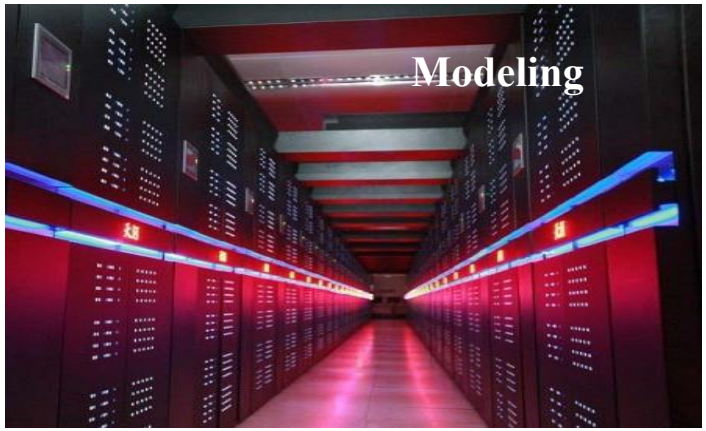
Melting point and ductility



Material Informatics







Material Informatics=Materials Science + Data Science

Advantages

Although the data-driven approach is internally related to other approaches, it has unique features and advantages:

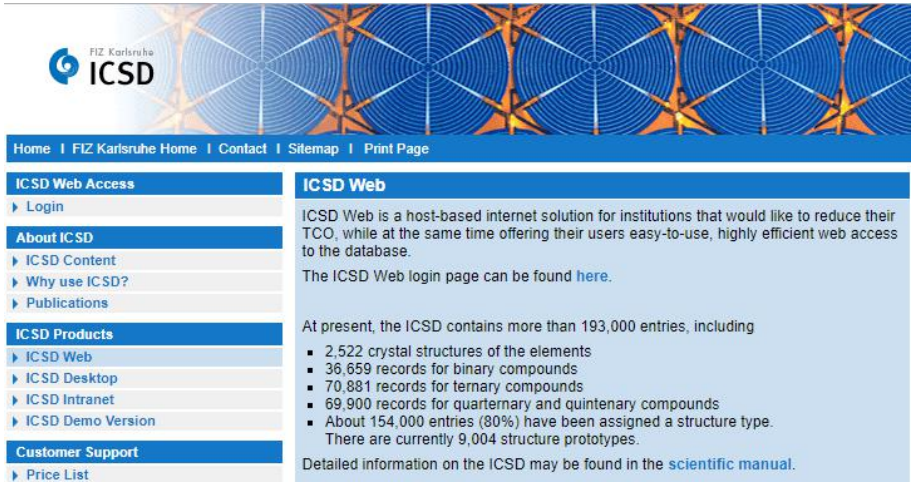
- **Enlarge the search space**
- **Have a global vision**
- **Find hidden patterns**

Enlarge the search space

Human: Thomas Edison did thousands of experiments before finding an acceptable materials for the filament of electric bulb

MI: About 200,000 experimentally realized materials are documented in Inorganic Crystal Structure Database

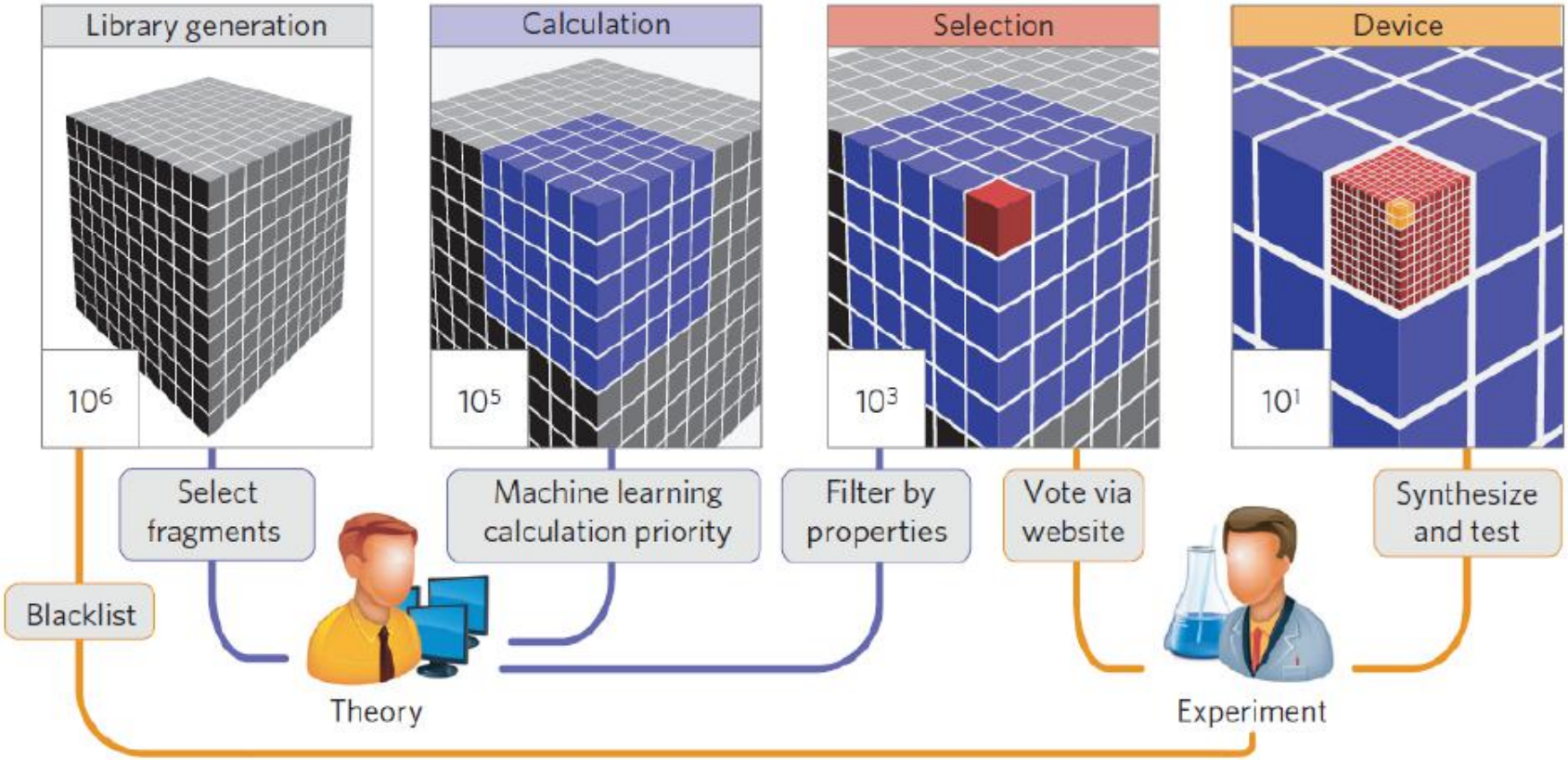
10^3



10^5

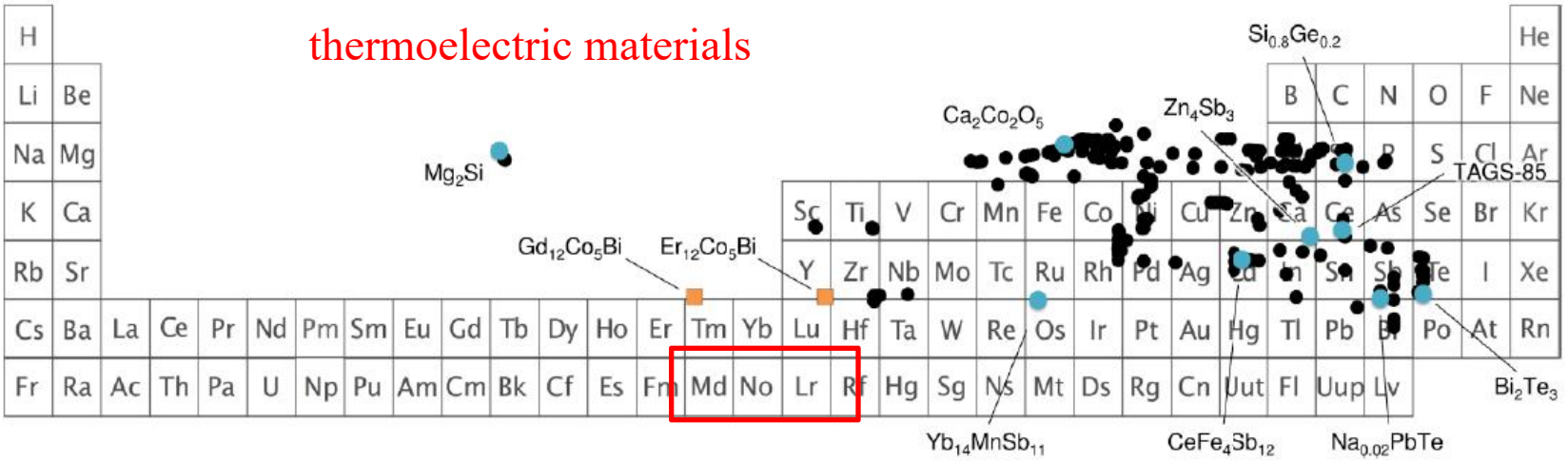
OLED molecule design

Nature Mater. 15, 1120 (2016)



1.6 million molecules → **400,000 calculations** → **900 candidates** → **4 molecules validated**

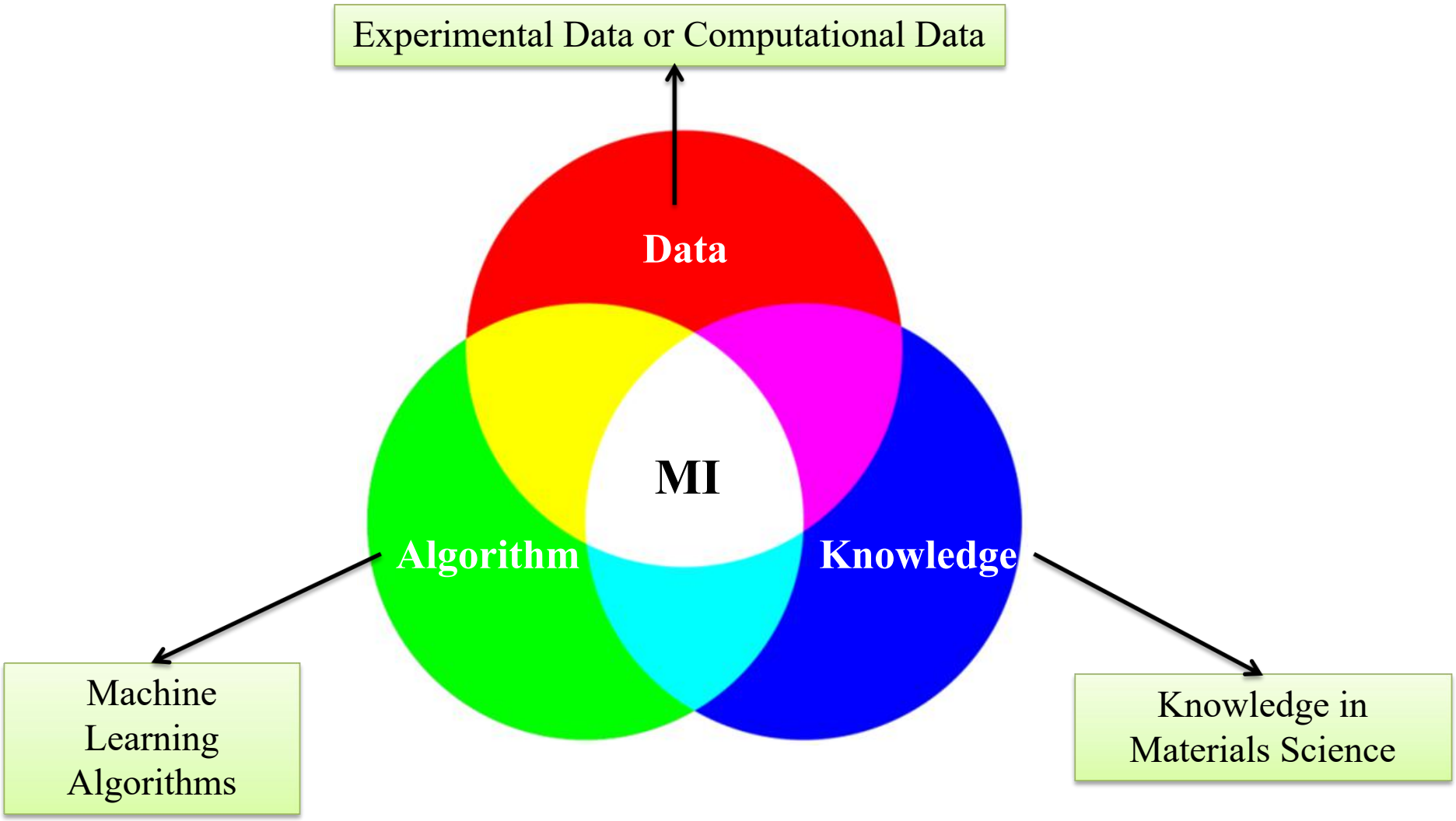
Have a global vision



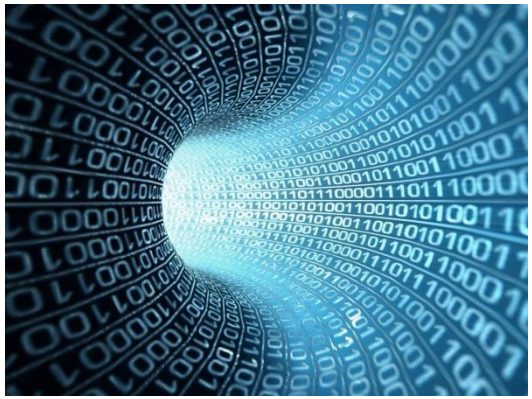
Human: Researchers tend to improve properties of known materials (nano-structuring, doping etc) or make substitution in known structures.

MI: Recommendation engines have no bias toward any material class, hence more likely to identify new materials.

Three key components

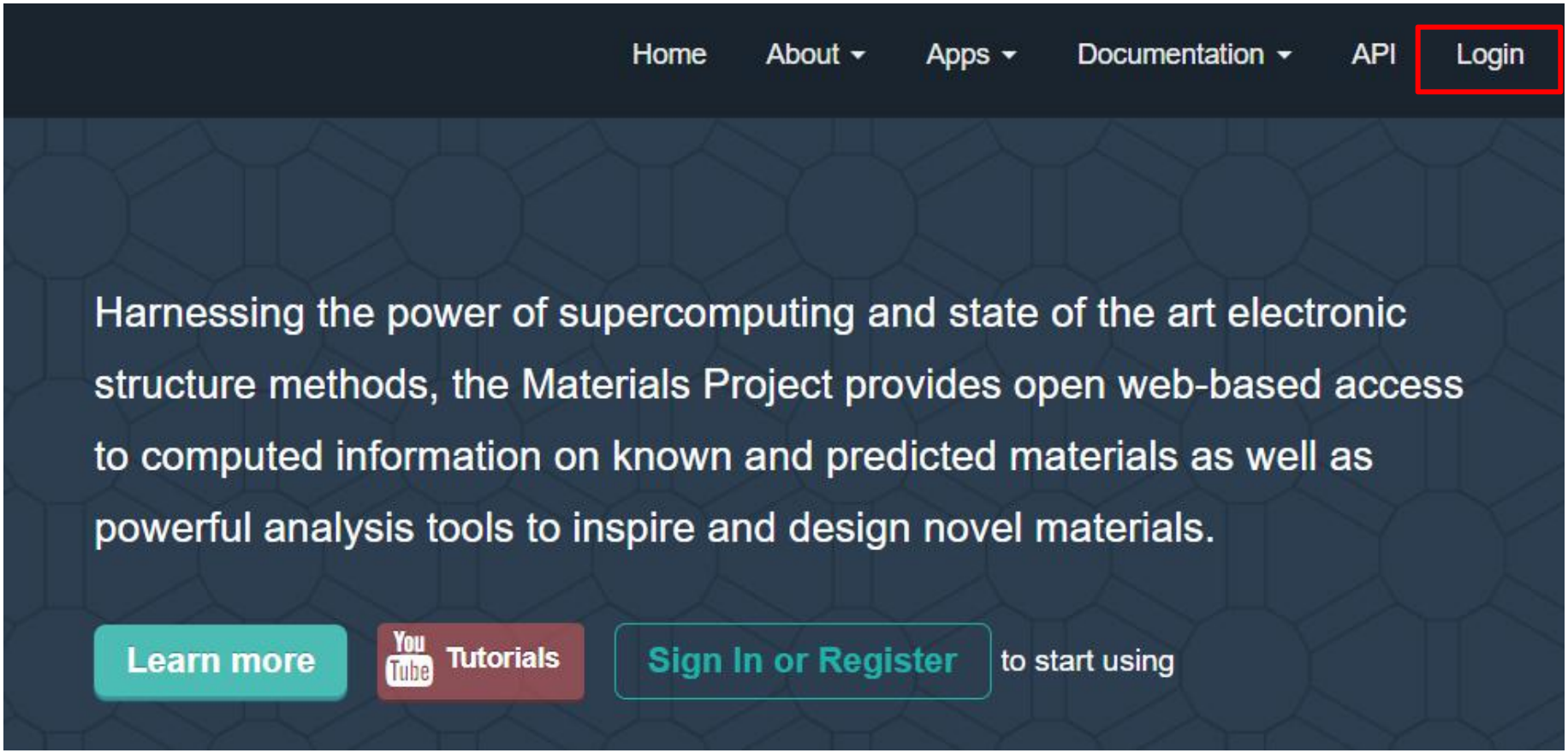


Data Resources



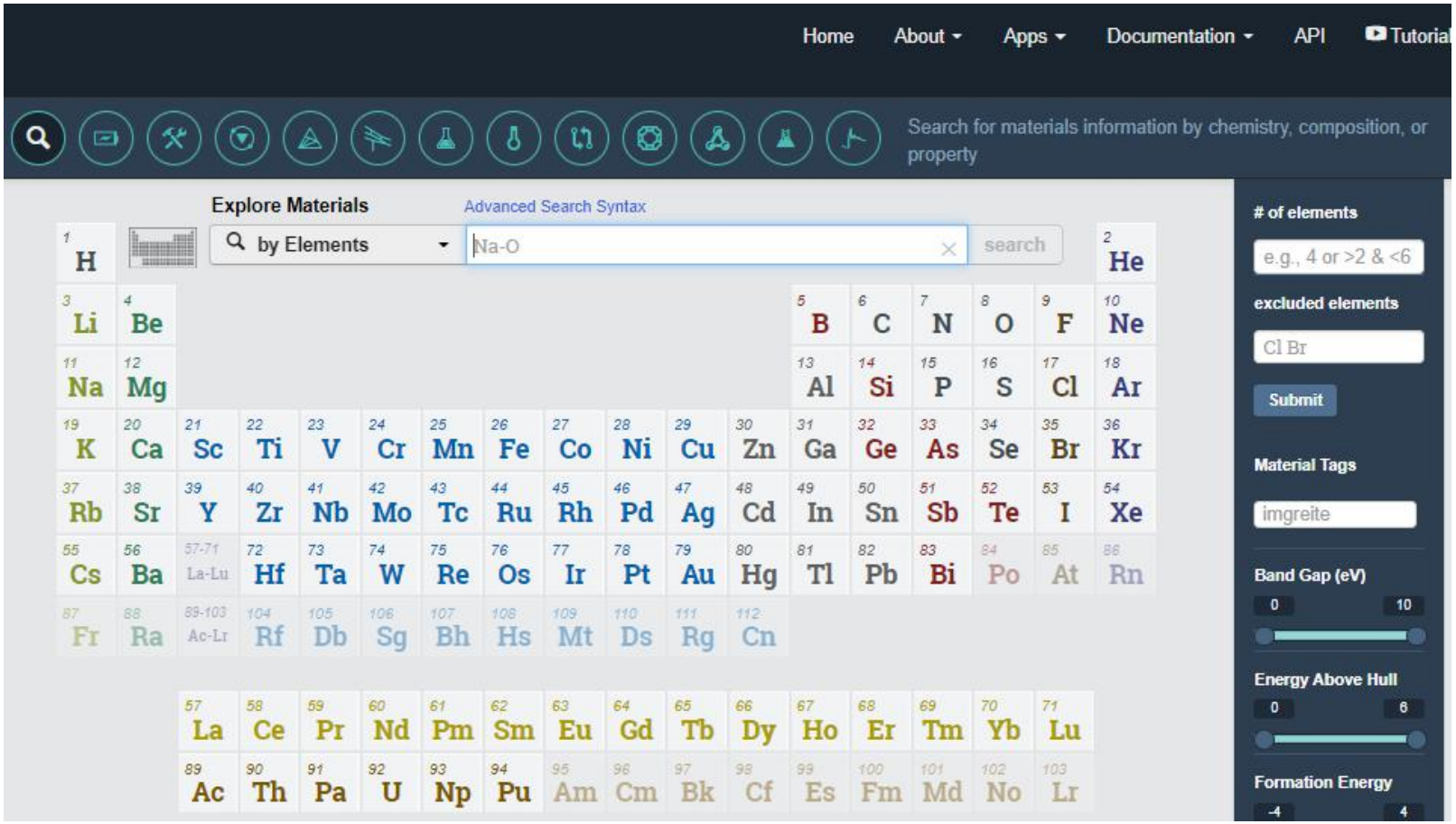
- Experimental data (e.g., x-ray diffraction)
- Computational data (e.g., VASP, Qchem)
- Database
  - **Materials Project**
  - **COD**
  - **ICSD**
  - AFLOWLIB
  - OQMD
  - Mat Navi
  - NoMaD
  - MMDB
  - CATAPP
  - MAGNDATA

Materials Project



注册一个Materials Project账号并登陆 https://www.materialsproject.org/

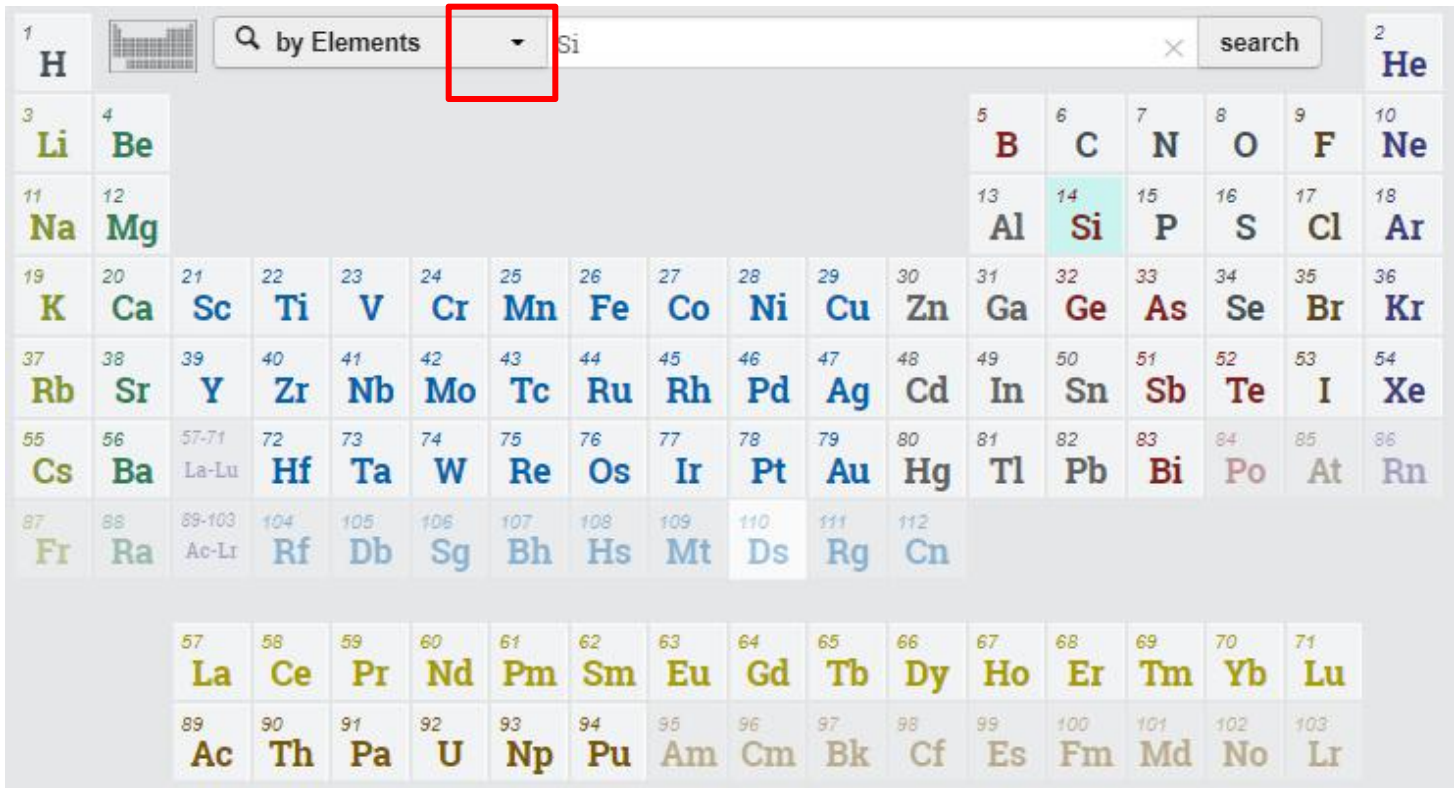
Materials Project



材料领域的机器学习

如何获取数据--推荐的几个数据库

搜索得到Si的结构、能带和态密度DOS



Search Materials

Materials Id	Formula	Spacegroup	Formation Energy (eV)	E Above Hull (eV)	Band Gap (eV)	Nsites	Density (gm/cc)	Volume
mp-149	Si	$Fd\bar{3}m$	0	0	0.612	2	2.281	40.888
mp-165	Si	$P6_3/mmc$	0.011					

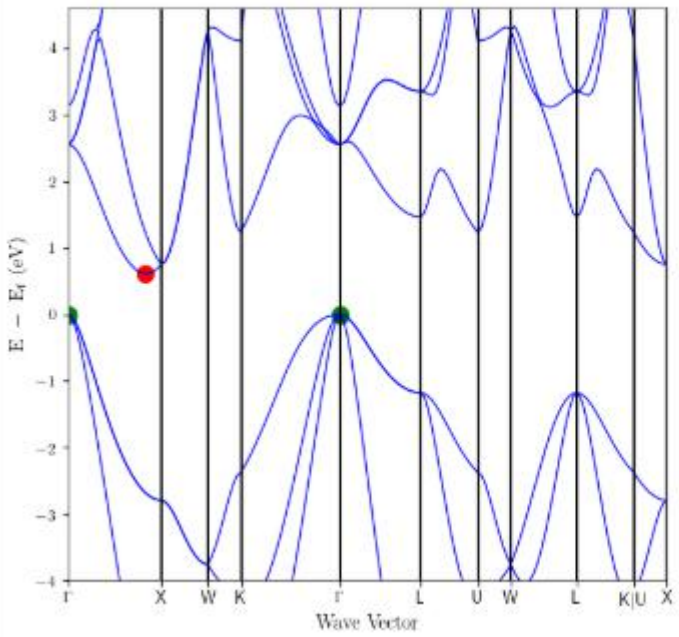
MATERIAL Si | ID: mp-149 | DOI: 10.17188/1190959

Electronic Structure | Phonon Dispersion | X-Ray Diffraction | X-Ray Absorption | Substrates | Calculation Summary | Provenance/Citation

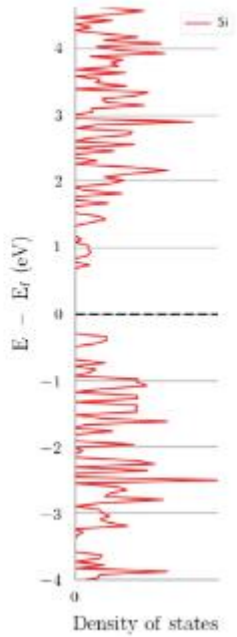
HM:P 1
a=5.469Å
b=5.469Å
c=5.469Å
α=90.000°
β=90.000°
γ=90.000°

Structure Type: Conventional Standard | Primitive | Refined | CIF

Band Structure



Density of States



Band Gap				
0.612 eV	4	3.037	61.423	

材料领域的机器学习—练习

任务：

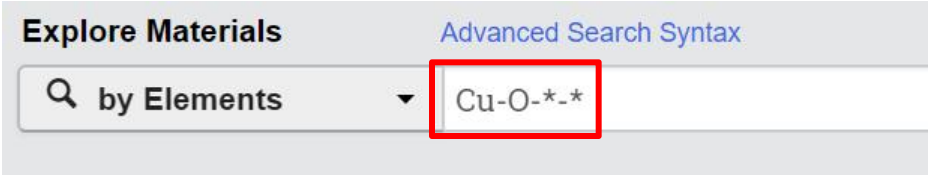
mp-66和mp-48哪个是金刚石哪个是石墨，哪种更稳定？

石墨:mp-48, 0.000 eV formation energy/atom, $P6_3/mmc$
金刚石:mp-66, 0.134 eV formation energy/atom, Fd-3m

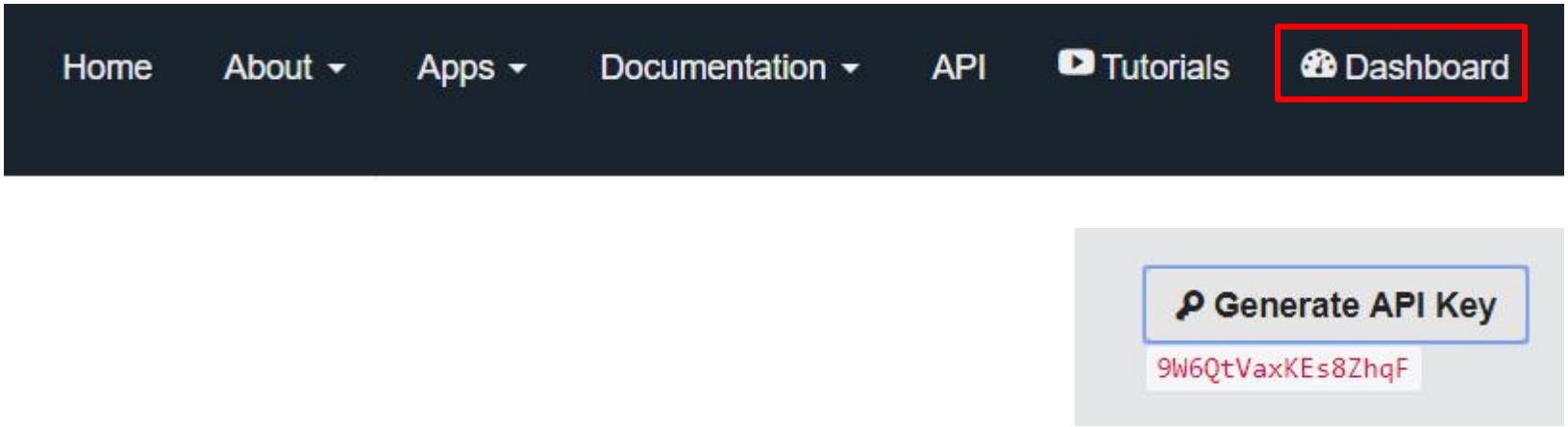
比较MoS_2和$MoSe_2$的带隙？哪个更宽？

MoS_2:mp-1434　Eg=1.204 eV
$MoSe_2$:mp-1027692　Eg=1.138 eV

搜索含Cu, O的四元素化合物



材料领域的机器学习


Home
About
Apps
Documentation
API
Tutorials
Dashboard
Generate API Key
9W6QtVaxKEs8ZhqF

材料领域的机器学习

查看材料信息：
https://www.materialsproject.org/rest/v2/materials/mp-24972/vasp/?API_KEY=ue01GUMslBL9JRIH

查看材料能带信息：
https://www.materialsproject.org/rest/v2/materials/mp-24972/vasp/bandstructure?API_KEY=ue01GUMslBL9JRIH

查看材料态密度信息：
https://www.materialsproject.org/rest/v2/materials/mp-24972/vasp/dos?API_KEY=ue01GUMslBL9JRIH

查看材料结构信息：
https://www.materialsproject.org/rest/v2/materials/mp-24972/vasp/structure?API_KEY=ue01GUMslBL9JRIH

材料领域的机器学习

elasticity
Mechanical properties in the elastic limit. Contains the full elastic tensor as well as derived properties, e.g. Poisson ratio and bulk (K) and shear (G) moduli. Consult our hierarchical documentation for the particular names of sub-keys.
piezo
Piezoelectric properties. Contains a tensor and derived properties. Again, consult our repository for the names of sub-keys.
diel
Dielectric properties. Contains a tensor (one each for total and electronic contribution) and derived properties, e.g. dielectric constant, refractive index, and recognized potential for ferroelectricity.

Calculation parameters

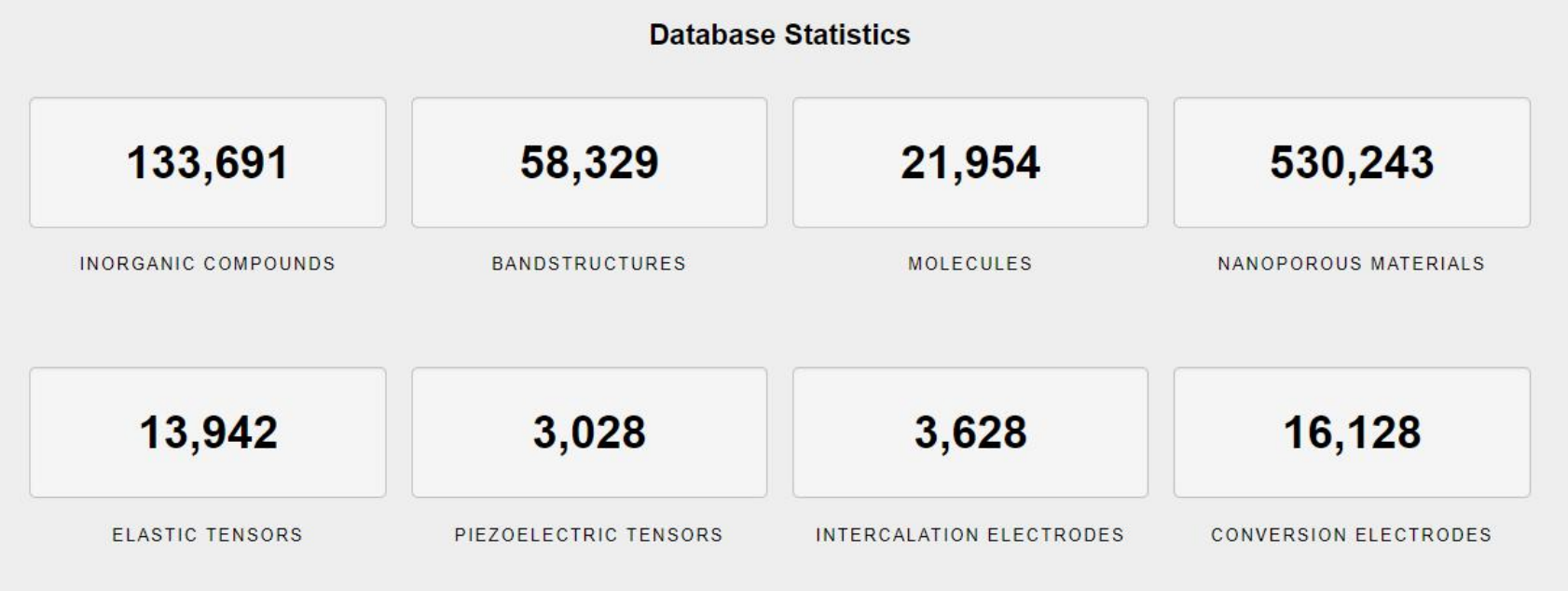
is_hubbard
A boolean indicating whether the structure was calculated using the Hubbard U extension to DFT
hubbards
An array of Hubbard U values, where applicable.
is_compatible
Whether this calculation is considered compatible under the GGA/GGA+U mixing scheme.

Electronic structure

band_gap
The calculated band gap
dos
The calculated density of states in the pymatgen json representation
bandstructure
The calculated "line mode" band structure (along selected symmetry lines -- aka "branches", e.g. Γ to Z -- in the Brillouin zone) in the pymatgen json representation
bandstructure_uniform
The calculated uniform band structure in the pymatgen json representation

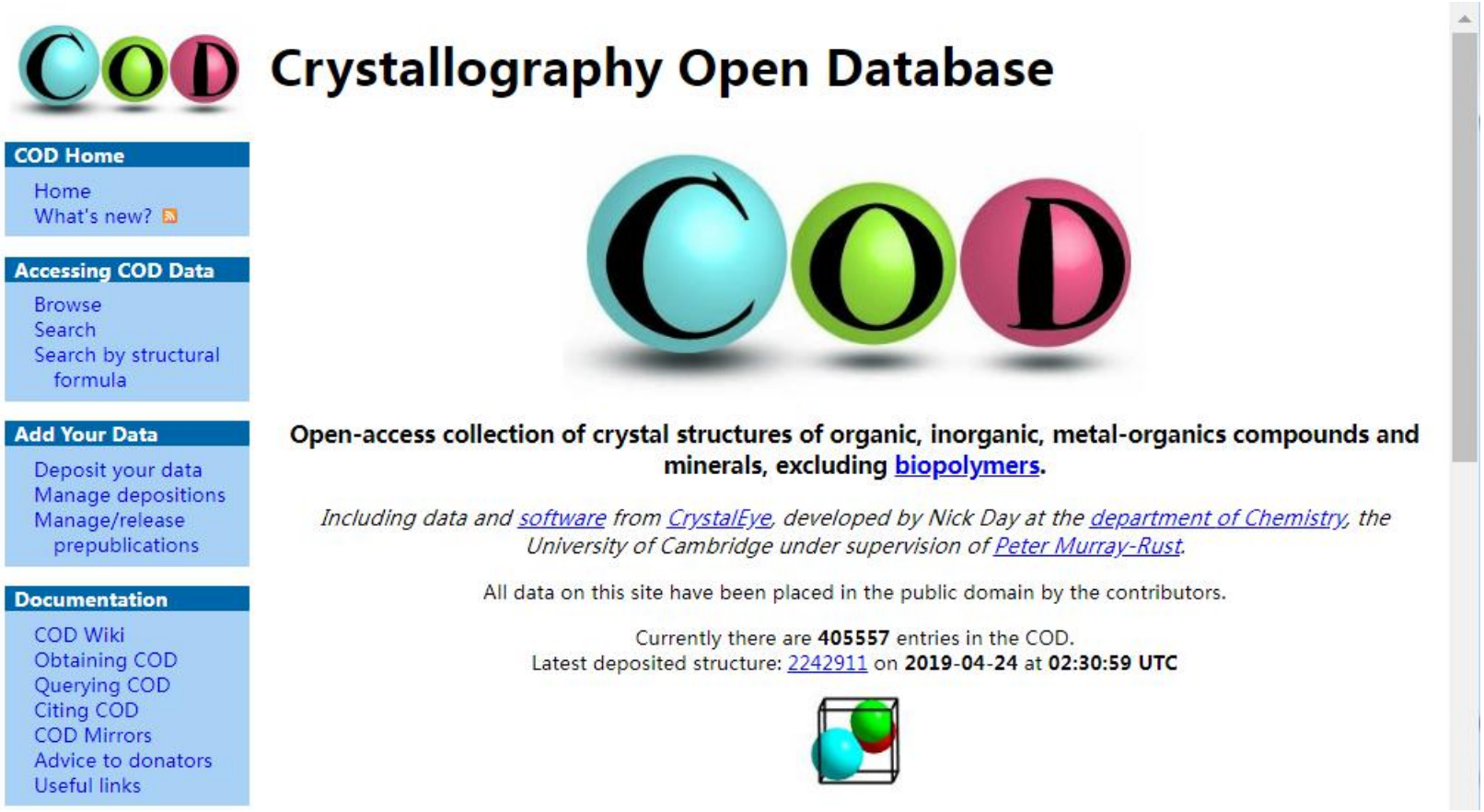
https://materialsproject.org/docs/api

材料领域的机器学习



材料基本信息数据已经整理到mp_index.xlsx文件，供大家参考

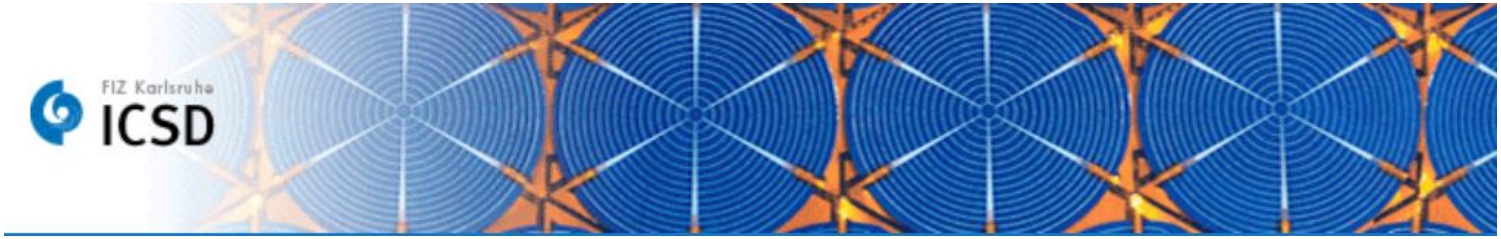
材料领域的机器学习

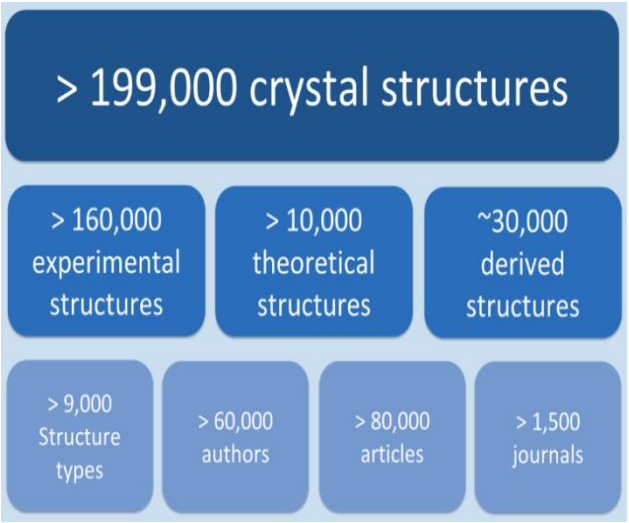


Crystallograph Open Database拥有约40万种晶体结构。
优点：材料数量庞大、提供多种下载方式、免费、API完备、提供材料的文献出处
缺点：缺乏材料的非结构信息

https://www.crystallography.net/cod/

材料领域的机器学习





ICSD拥有接近20万种无机晶体结构及相关的电子结构信息。

优点：晶体结构主要来自实验数据；学术界认可度比较高；搜索方式多样。

缺点：**昂贵，无永久license；**只有结构数据，没有物理性质。

http://www2.fiz-karlsruhe.de/icsd_home.html

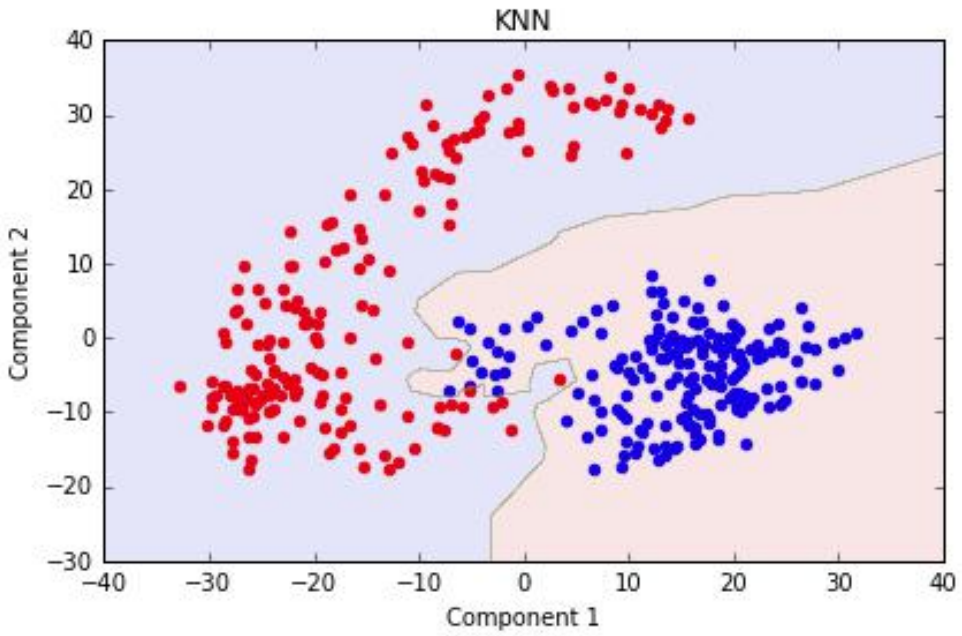
材料领域的机器学习

常用的分类算法

缩写	全称	线性
KNN	K-nearest neighbors	nonlinear
LogReg	Logistic regression	linear
LDA	Linear discriminant analysis	linear
DT	Decision tree	nonlinear
SVM	Supporting vector machine	Linear or nonlinear

材料领域的机器学习

常用的分类算法--KNN



根据最近邻的K个点的类别来判断“新来的”点归为哪一类。

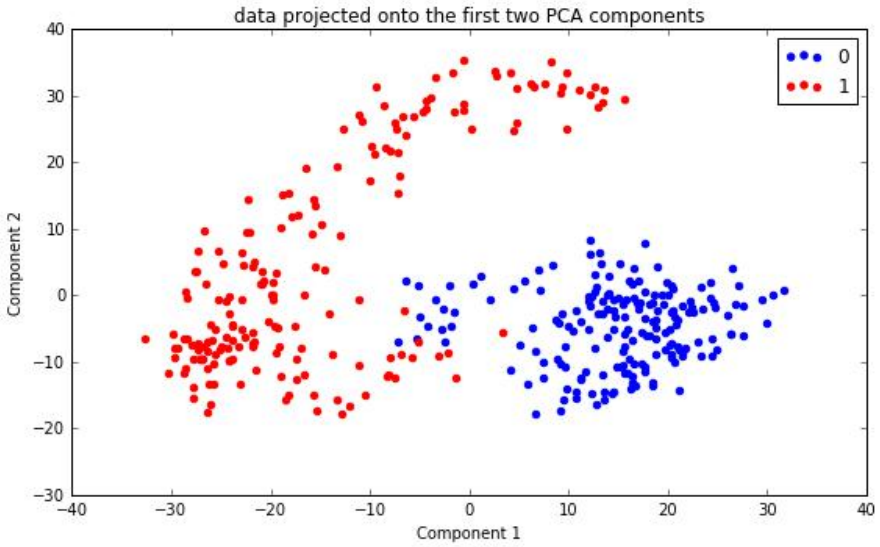
如对于某个新点，K=5时，其最近邻的5个点里3红2蓝，那么归为红点。

特点：易于理解和实现；点数量多时计算距离耗时；距离的定义；K如何选取

KNN: 近朱者赤，近墨者黑!

材料领域的机器学习

常用的分类算法--LogReg



LogReg: 名为逻辑回归，其实是分类算法

根据点的分量的线性组合结果，来给出每个点被判为红（蓝）点的概率。如果模型对训练集里点的概率判得不准，则施加惩罚，迫使模型继续调整。

$$\max_{\{\beta_0,\beta_1,\beta_2\}} \ell$$

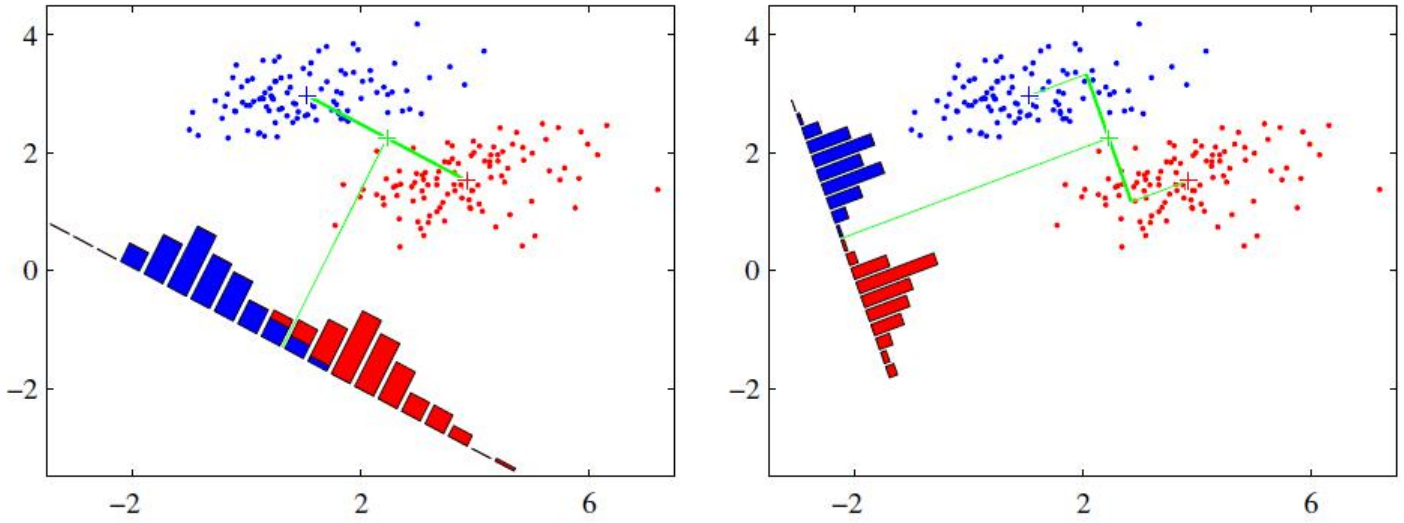
$$\ell \equiv \prod_{y_i=1} p(x_i) \prod_{y_j=0} (1 - p(x_j))$$

$$p(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}$$

材料领域的机器学习

常用的分类算法—线性判别分析LDA

LDA将数据在低维度上进行投影，投影后希望每一种类别数据的投影点尽可能的接近，而不同类别的数据的类别中心之间的距离尽可能的大。



LDA可以用于分类也可以用于降维

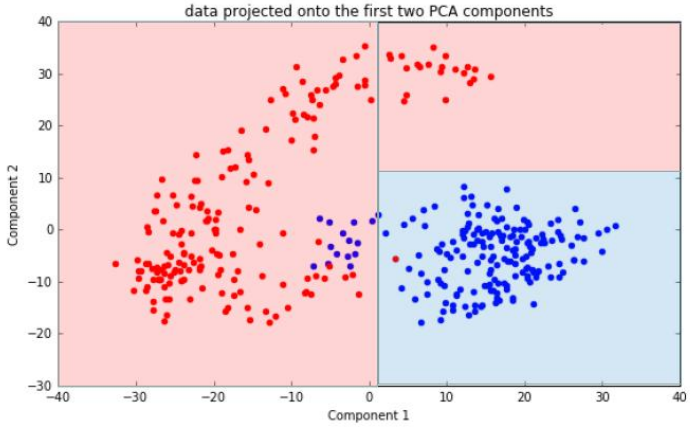
材料领域的机器学习

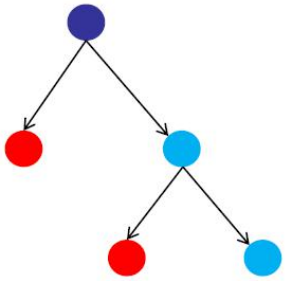
常用的分类算法—决策树

Entropy $S \equiv -\sum_{k=1}^{K} p_k \ln p_k \qquad \sum_{k=1}^{K} p_k = 1$

Entropy gain $\Delta S = S - \left(\frac{N_1}{N_1 + N_2} S_1 + \frac{N_2}{N_1 + N_2} S_2 \right)$

$\max \Delta S$





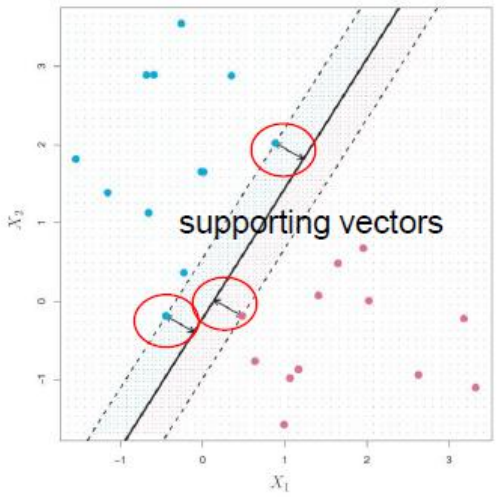
决策树有点像一系列if()语句，通过特征分割来使得信息增益最大。

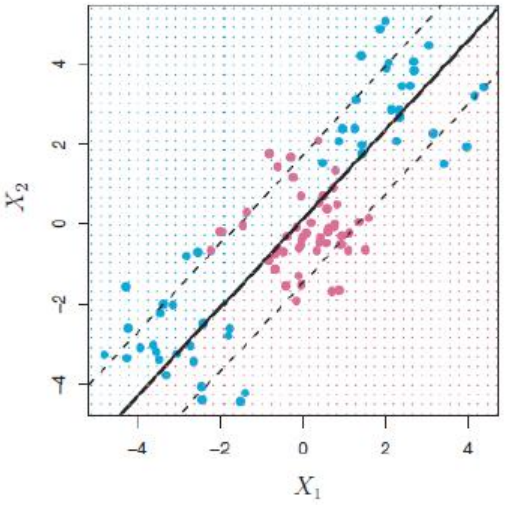
信息熵有这样的特点：各状态的概率越接近，熵越大，概率越悬殊，熵越小。那么如果分割出的集合里分布越悬殊，熵越小，分割前和分割后的信息增量（“熵增”）越大。

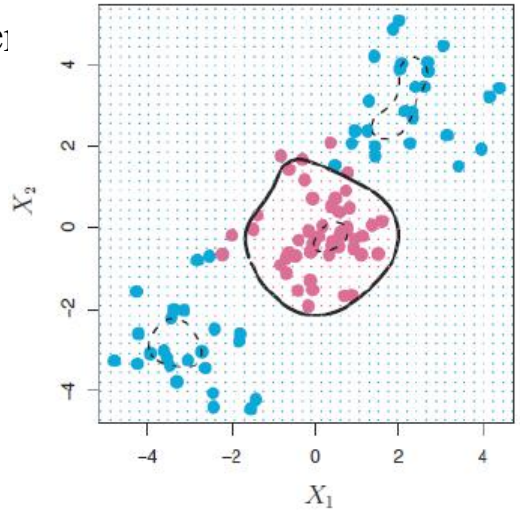
材料领域的机器学习

常用的分类算法—SVM

SVM采用超平面(hyper plane)来划分数据点，使得离超平面最近的点尽可能远（最大间隔）。



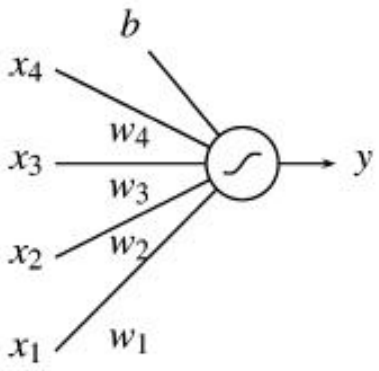


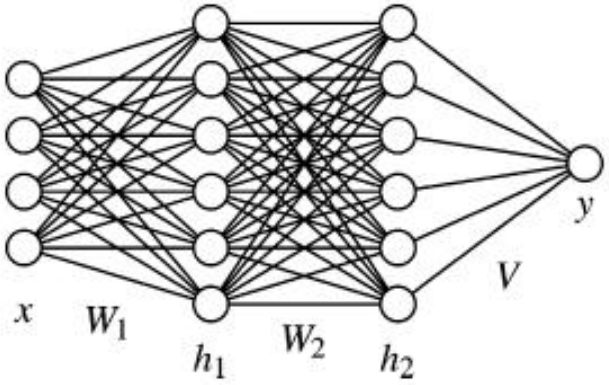


材料领域的机器学习

常用的分类算法—神经网络

神经网络是一种多层结构的模型，每一层由多个特征组成，每个特征与前一层的若干特征相连，即线性组合及非线性激活。





http://playground.tensorflow.org

材料领域的机器学习

模型选择---交叉验证

为了更准确地评估模型的有效性，可以采用交叉验证(cross-validation)的方法。

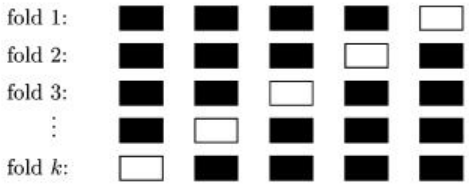
K-交叉验证从训练集里拆成K份，每次选一份用于验证评估，其余用于训练，综合K次评估结果得到模型的成绩。

1. Split data into k sets of roughly the same size, e.g. into $k = 5$ splits.
2. Use $(k - 1)$ splits for training and model selection. Then test the generated model on the remaining hold-out or test split.

training

3. Repeat step 2. k-times, i.e. until each subset has been once used for testing.

validation

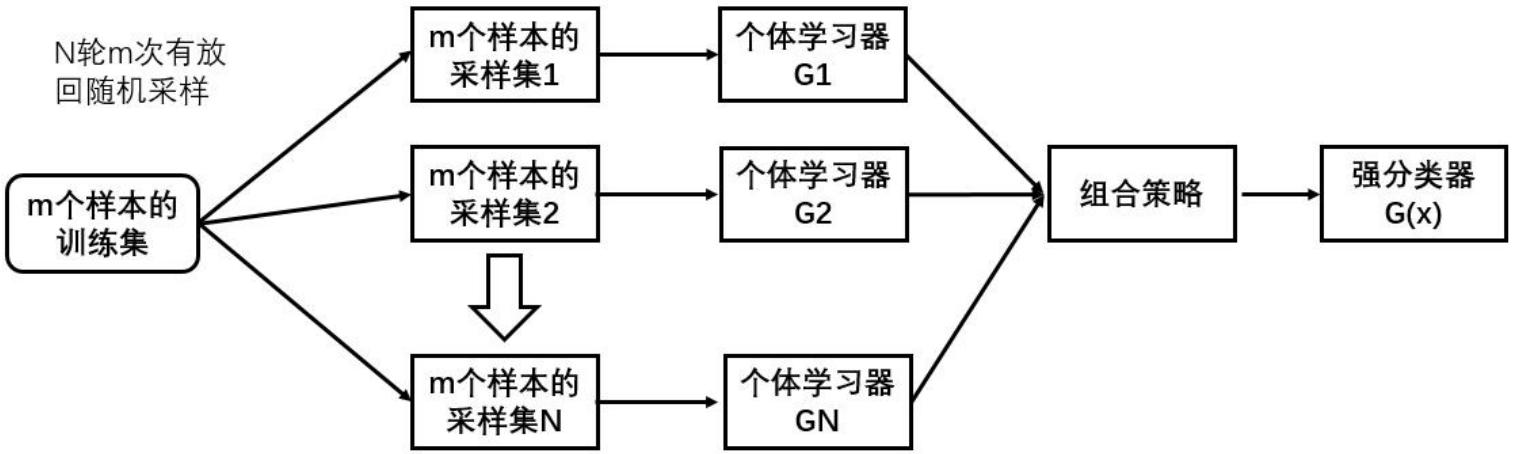


4. Combine the k estimates of the prediction error into one cross-validation error.

材料领域的机器学习

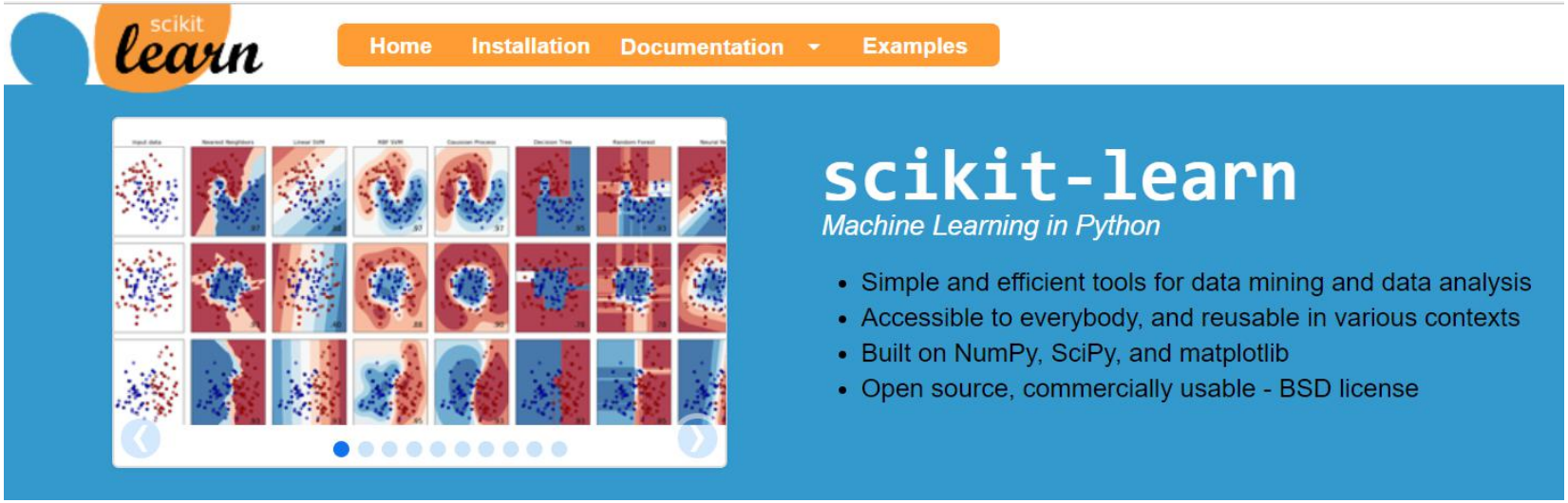
模型选择---bagging

通过捆绑组合多个模型（学习器）的方式，提高整体的鲁棒性（Robustness），准确程度。除了bagging外还有boost等方法。



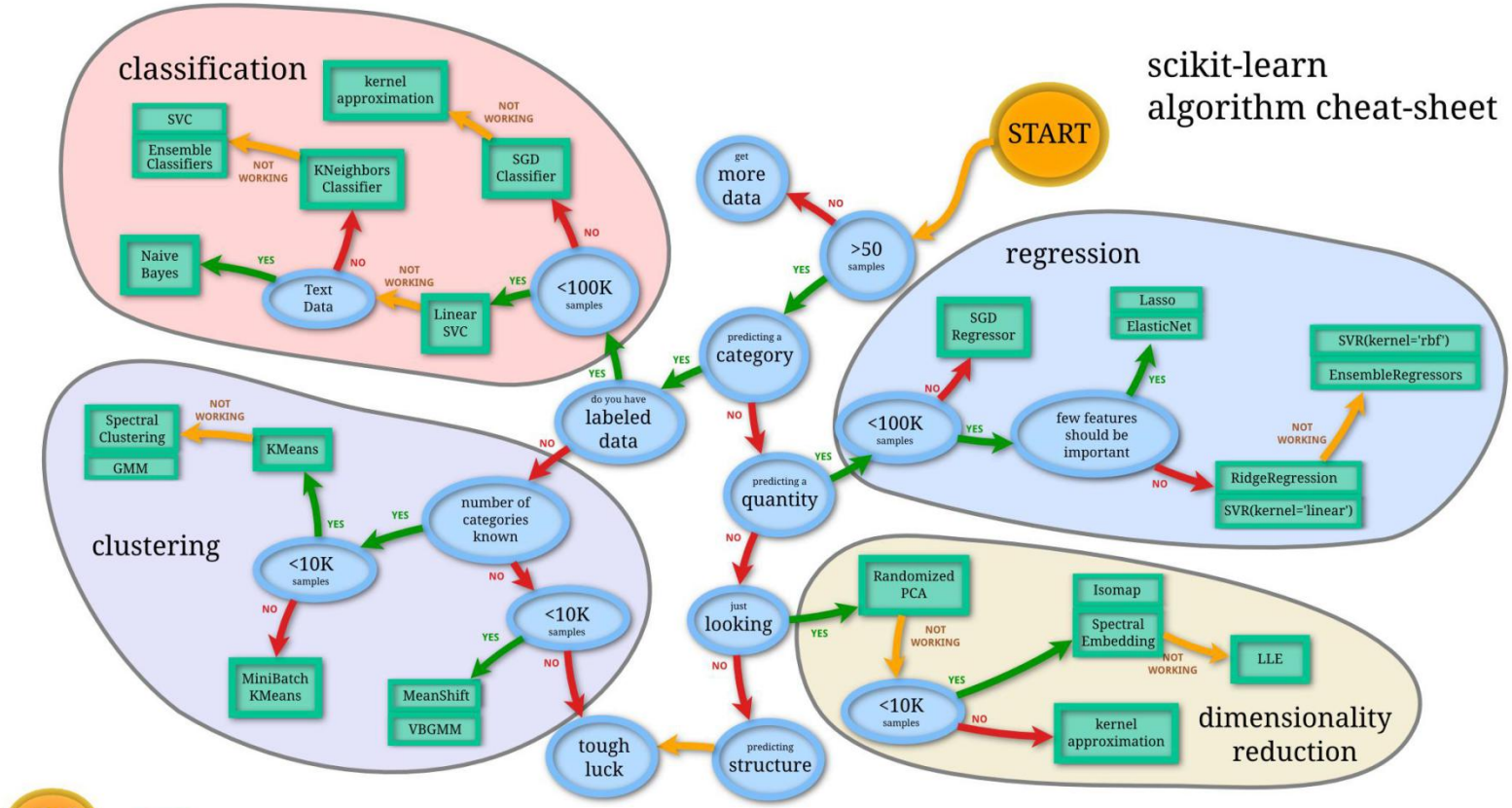
Python里常用的机器学习、材料学相关库

经典的机器学习库scikit-learn

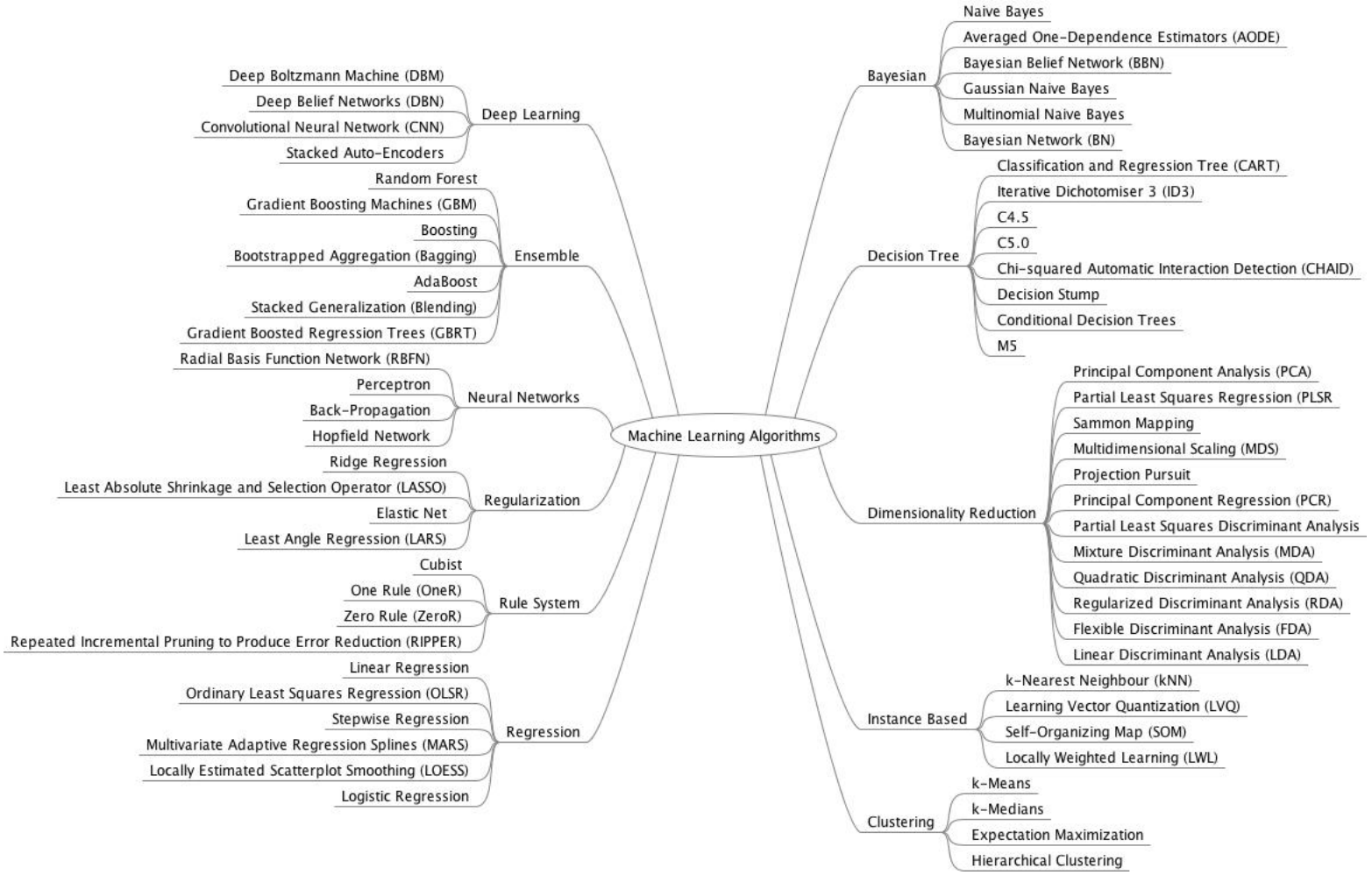


支持分类、 回归、聚类 、降维、模型选择、数据预处理等多类型任务。封装程度很高，易于使用。

Machine learning algorithms

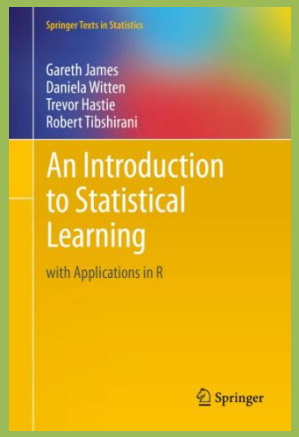


Machine learning algorithms



Data analysis tools

Statistical Learning

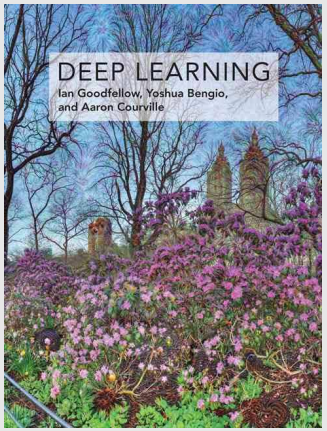


scikit learn

10^2 to 10^3 data points

physics interpretation

Deep Learning



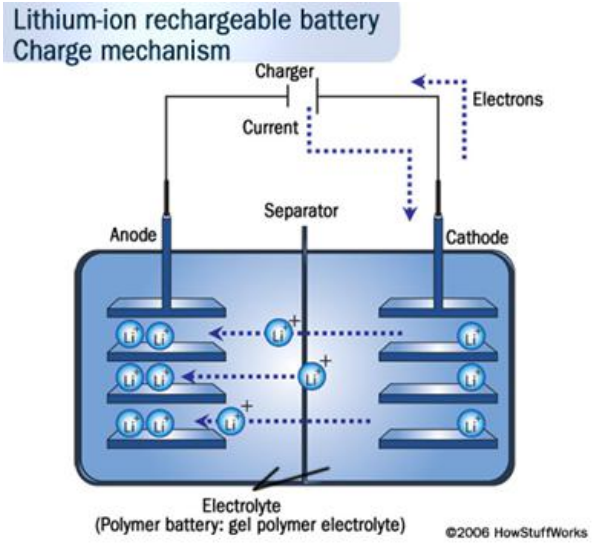
TensorFlow

$> 10^4$ data points

large model compacity

固态锂离子导体材料的虚拟筛选

固态锂离子导体材料



大多数锂电池里，电解质是液态的，即锂离子盐溶解在有机溶剂里。

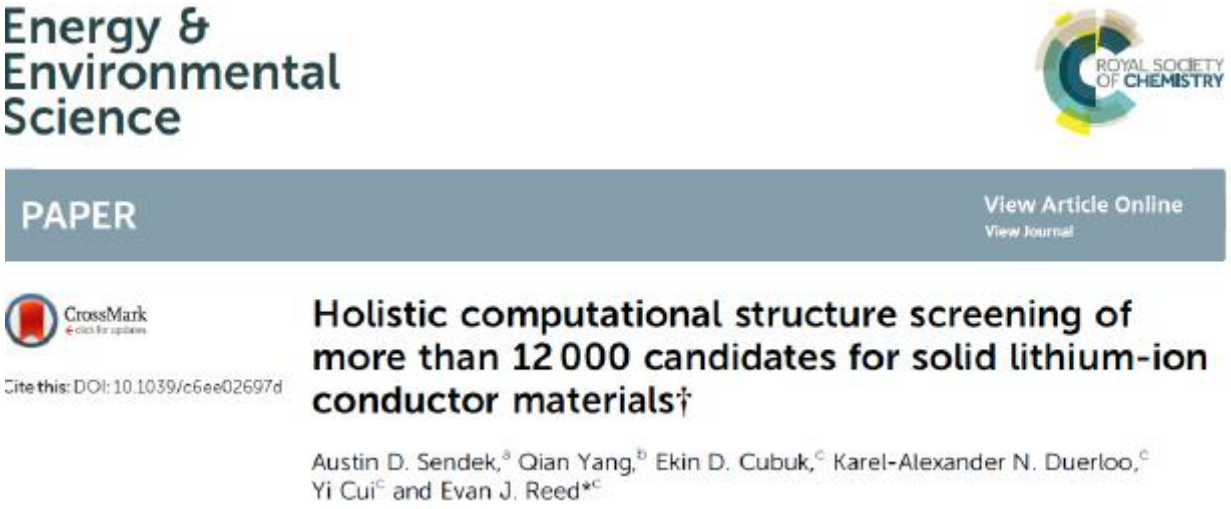
优点：廉价，高的锂离子电导率

缺点：安全问题如机械穿刺引起的短路，发热导致压力增大等

如何寻找固态的锂离子导体材料？
Guess-and-check? 花费了数十年，仅得到了数十种

Knowledge

- **The size of training set is relatively small in materials science**
- **Human knowledge can help design effective descriptors**



Energy & Environmental Science

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

CrossMark

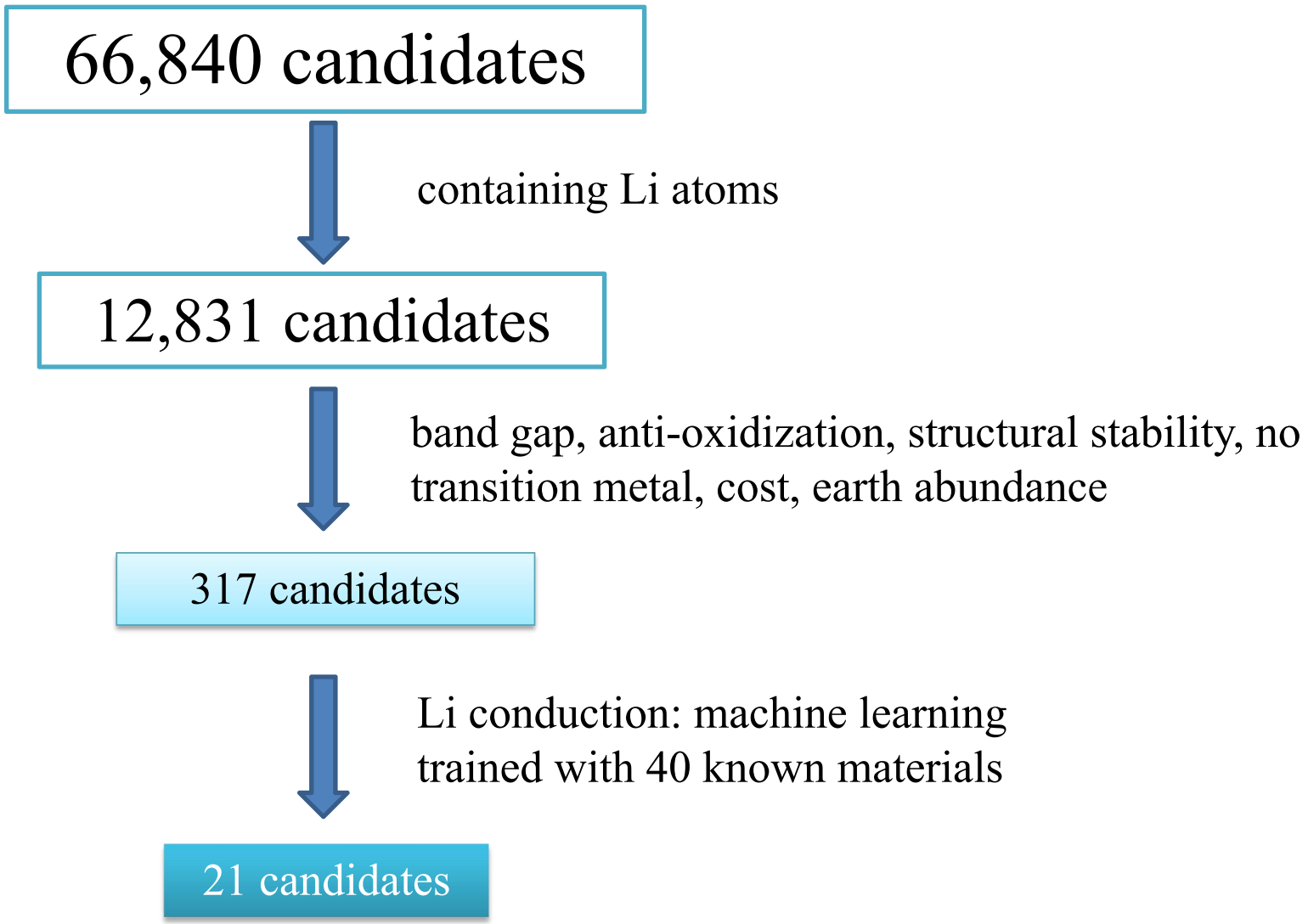
Cite this: DOI: 10.1039/c6ee02697d

Holistic computational structure screening of more than 12 000 candidates for solid lithium-ion conductor materials†

Austin D. Sendek,[a] Qian Yang,[b] Ekin D. Cubuk,[c] Karel-Alexander N. Duerloo,[c] Yi Cui[c] and Evan J. Reed*[c]

Material Project里含有约12000种含锂的晶体结构，很可能存在“沉睡的”优秀目标材料。 如何找到它们？

Solid lithium-ion conductors



Screening

Requirement	Reasoning	Filter
Electronic conductivity	$\sigma = (\mu_e + \mu_h) q \sqrt{N_C N_V} e^{-E_g/2kT}$	$E_g > 1eV$
Structure stability	Energy above hull: The decomposition energy of this material into the set of most stable materials	$E_{hull} = 0$
Stability vs. oxidation at cathode	$Li_x A_a B_b C_c \xrightarrow{\Delta G_f} xLi^0 + aA^0 + bB^0 + cC^0$	$\tilde{V}_{OX} \equiv \Delta G_f / xF$ $\tilde{V}_{OX} > 4V$
Stability vs. reduction at anode	Transition metals have many stable oxidation states	No transition metal

固态锂离子导体材料的虚拟筛选

机器学习描述符筛选---训练集的准备

从文献中搜集了40种有离子电导率的含Li材料，其中高离子电导率的11种。

晶体结构从ICSD种下载。

$$\tilde{\sigma} = \begin{cases} 1, \sigma \geq 10^{-4}\ \mathrm{S\ cm^{-1}} \\ 0, \sigma < 10^{-4}\ \mathrm{S\ cm^{-1}} \end{cases}$$

Composition	RT bulk ionic conductivity (S cm^{-1})	Ionic conductivity ref.	Structure ref.
$LiLa(TiO_3)_2$	1×10^{-3}	17	29
$Li_{9.81}Sn_{0.81}P_{2.19}S_{12}$	5.5×10^{-3}	30	30
$Li_{10}Ge(PS_6)_2$	1.4×10^{-2}	7	31
$Li_{10.35}Si_{1.35}P_{1.65}S_{12}$	6.5×10^{-3}	30	30
$Li_{14}ZnGe_4O_{16}(2)$	1.0×10^{-6}	32	33 and 34
$Li_2Ca(NH)_2$	6.4×10^{-6}	35	36
$Li_2Ge_7O_{15}$	5.0×10^{-6}	37	38
Li_2NH	2.5×10^{-4}	35	39
Li_2S	1.0×10^{-13}	40	41
$Li_{13.6}Si_{2.8}S_{1.2}O_{16}$	6.0×10^{-7}	42	43
$Li_{14}Ge_2V_2O_{16}$	7.0×10^{-5}	44	45

固态锂离子导体材料的虚拟筛选

机器学习描述符筛选---20个Li相关描述符

	Feature	Pearson correlation coefficient
1	Volume per atom[a]	0.20
2	Standard deviation in Li neighbour count	0.22
3	Standard deviation in Li bond ionicity	−0.04
4	Li bond ionicity[a]	−0.18
5	Li neighbour count[a]	−0.19
6	*Li–Li bonds per Li*[a]	*0.06*
7	*Bond ionicity of sublattice*[a]	*−0.28*
8	Sublattice neighbour count[a]	−0.13
9	*Anion framework coordination*[a]	*−0.06*
10	Minimum anion–anion separation distance[a] (Å)	0.09
11	Volume per anion (Å^3)	−0.01
12	*Minimum Li–anion separation distance*[a] *(Å)*	*0.20*
13	*Minimum Li–Li separation distance*[a] *(Å)*	*−0.10*
14	Electronegativity of sublattice[a]	−0.16
15	Packing fraction of full crystal	0.16
16	Packing fraction of sublattice	0.19
17	Straight-line path width[a] (Å)	0.07
18	Straight-line path electronegativity[a]	−0.29
19	Ratio of features (4) and (7)	−0.03
20	Ratio of features (5) and (8)	−0.18
	Constant term	—

20个描述符与离子电导率的相关系数都很低。单独使用是不合适的。

解决方案---把低相关度的描述符组合成高相关度特征！

$$P_{\mathrm{LR}}\left(\tilde{\sigma}^{(i)}=1\right)=\left[1+\exp\left(-\sum_{j=0}^{n}\theta_j x_j^{(i)}\right)\right]^{-1}$$

PLR > 50%, 则把材料判为1类(高离子电导率）
PLR < 50%, 则判为0类。

Features design

Features design relies on professional knowledge to identify the most relevant factors

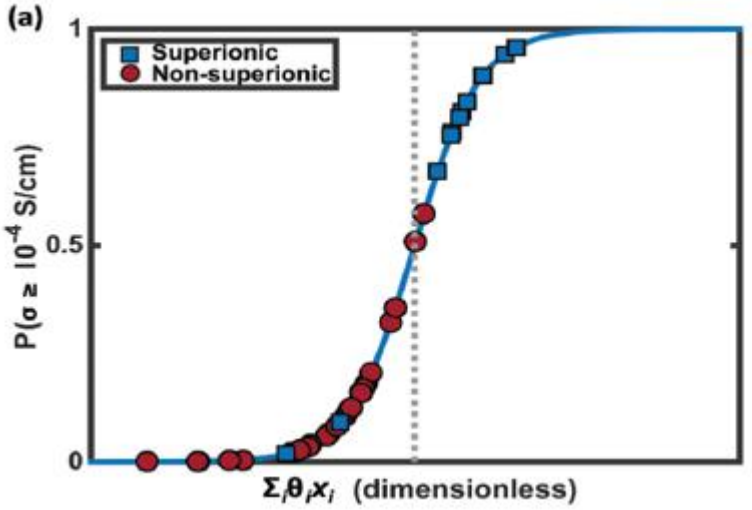
Features	Definition
LLB	The average number of Li neighbors for each Li
SBI	The average sublattice bond iconicity
AFC	The average anion-anion coordination number in the anion framework
LASD	The average shortest Li-anion distance in angstroms
LSSD	The average shortest Li-Li distance in angstroms

Ionic conductivity screening

$P(x) > 0.5$ good ionic conductor

$P(x) < 0.5$ bad ionic conductor

$$P(x) \equiv \frac{\exp(\sum_i \theta_i x_i)}{1 + \exp(\sum_i \theta_i x_i)}$$



$$\sum_i \theta_i x_i = 0.184 \times LLB - 4.009 \times SBI - 0.467 \times AFC$$
$$+8.699 \times LASD - 2.17 \times LLSD - 6.564$$

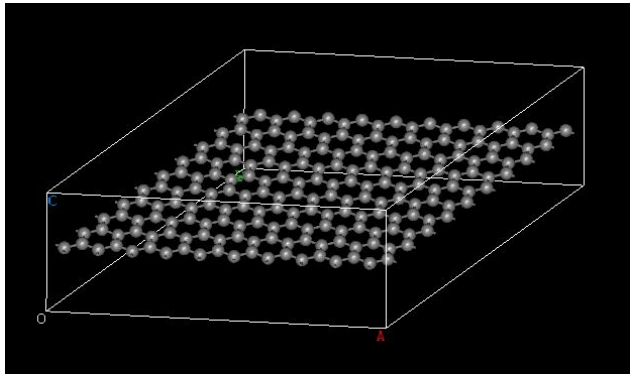
Logistic regression

训练结果

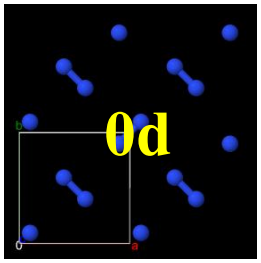
MPID	Chemical formula
mp-554076	$BaLiBS_3$
mp-532413	$Li_5B_7S_{13}$
mp-569782[a]	$Sr_2LiCBr_3N_2$
mp-558219	$SrLi(BS_2)_3$
mp-15797	$LiErSe_2$
mp-29410	$Li_2B_2S_5$
mp-676361	Li_3ErCl_6
mp-643069[a]	Li_2HIO
mp-19896	Li_2GePbS_4
mp-7744[a]	$LiSO_3F$
mp-22905[b]	$LiCl$
mp-34477	$LiSmS_2$
mp-676109	Li_3InCl_6
mp-559238	$CsLi_2BS_3$
mp-866665[a]	$LiMgB_3(H_9N)_2$
mp-8751	$RbLiS$
mp-15789	$LiDyS_2$
mp-15790	$LiHoS_2$
mp-15791	$LiErS_2$
mp-561095[a]	$LiHo_3Ge_2(O_4F)_2$
mp-8430	$KLiS$

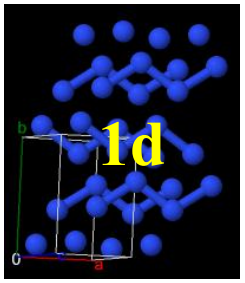
二维铁磁材料的虚拟筛选

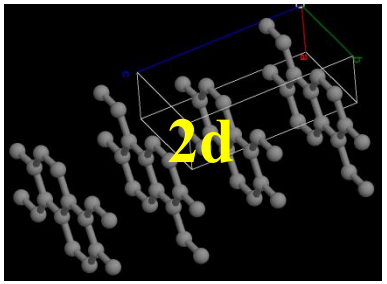
二维材料

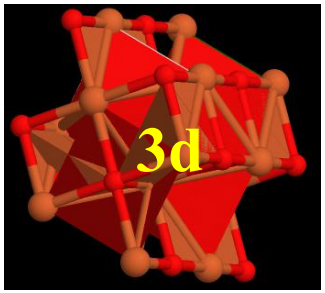


如何用尽可能低的成本从材料数据库里把二维材料找出来？









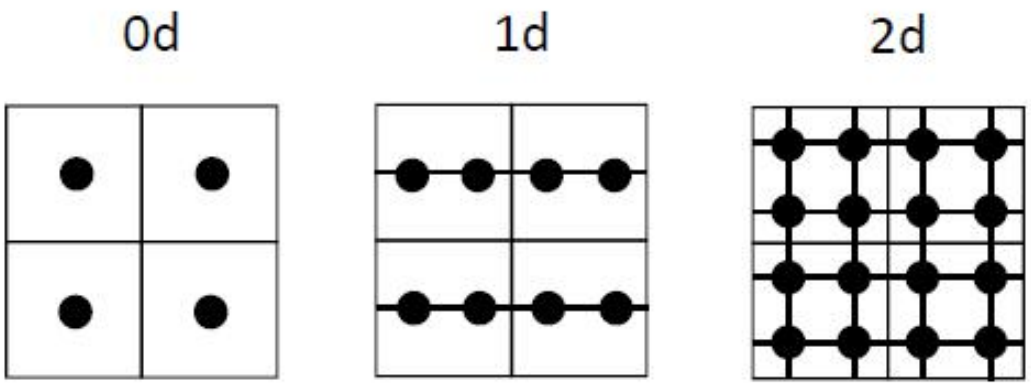
二维铁磁材料的虚拟筛选

二维材料—维度描述符

扩晶胞，计算连接的原子数量，
对比值求2的对数。

$$2^d = \frac{N_{2\times2\times2}}{N_{1\times1\times1}}.$$

$$D_{12} < R_1 + R_2 + \delta R,$$



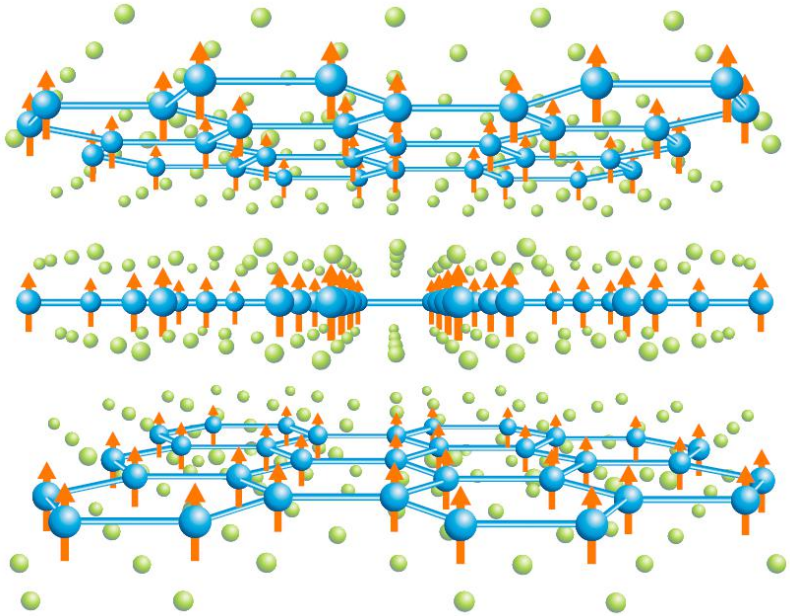
二维铁磁材料的虚拟筛选

dimensionality	dR = 0.5	dR = 1.0	dR = 1.5
0d	3,506	1,839	910
1d	1,257	867	577
2d	3,531	2,510	1,802
3d	59,188	62,266	64,194

通过简单的1个描述符，马上找到了数据库里全部2d材料。调控dR可以提高维度描述符的严格程度。

Two recent experiments

Ferromagnetism was found in 2d materials experimentally



LETTER

doi:10.1038/nature22060

Discovery of intrinsic ferromagnetism in two-dimensional van der Waals crystals

Cheng Gong[1]*, Lin Li[2]*, Zhenglu Li[3,4]*, Huiwen Ji[5], Alex Stern[2], Yang Xia[1], Ting Cao[3,4], Wei Bao[1], Chenzhe Wang[1], Yuan Wang[1,4], Z. Q. Qiu[3], R. J. Cava[5], Steven G. Louie[3,4], Jing Xia[2] & Xiang Zhang[1,4]

LETTER

doi:10.1038/nature22391

Layer-dependent ferromagnetism in a van der Waals crystal down to the monolayer limit

Bevin Huang[1]*, Genevieve Clark[2]*, Efrén Navarro-Moratalla[3]*, Dahlia R. Klein[3], Ran Cheng[4], Kyle L. Seyler[1], Ding Zhong[1], Emma Schmidgall[1], Michael A. McGuire[5], David H. Cobden[1], Wang Yao[6], Di Xiao[4], Pablo Jarillo-Herrero[3] & Xiaodong Xu[1,2]

UC-Berkeley group	$Cr_2Ge_2Te_6$
UW and MIT group	CrI_3

Pre-screening for 2d ferromagnetic materials

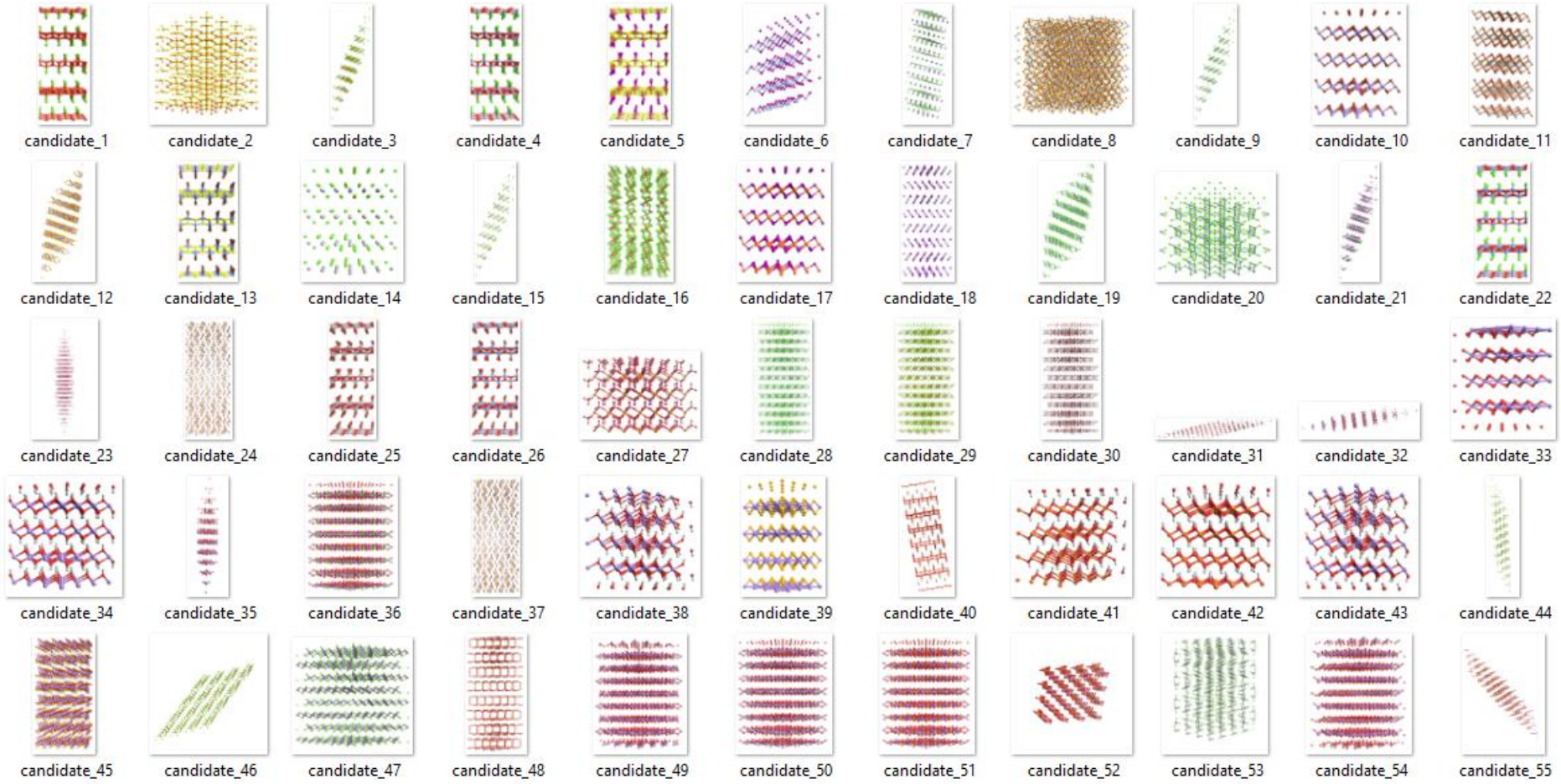
Goal: Search for materials with layered structure and ferromagnetism

Pre-screening criterion

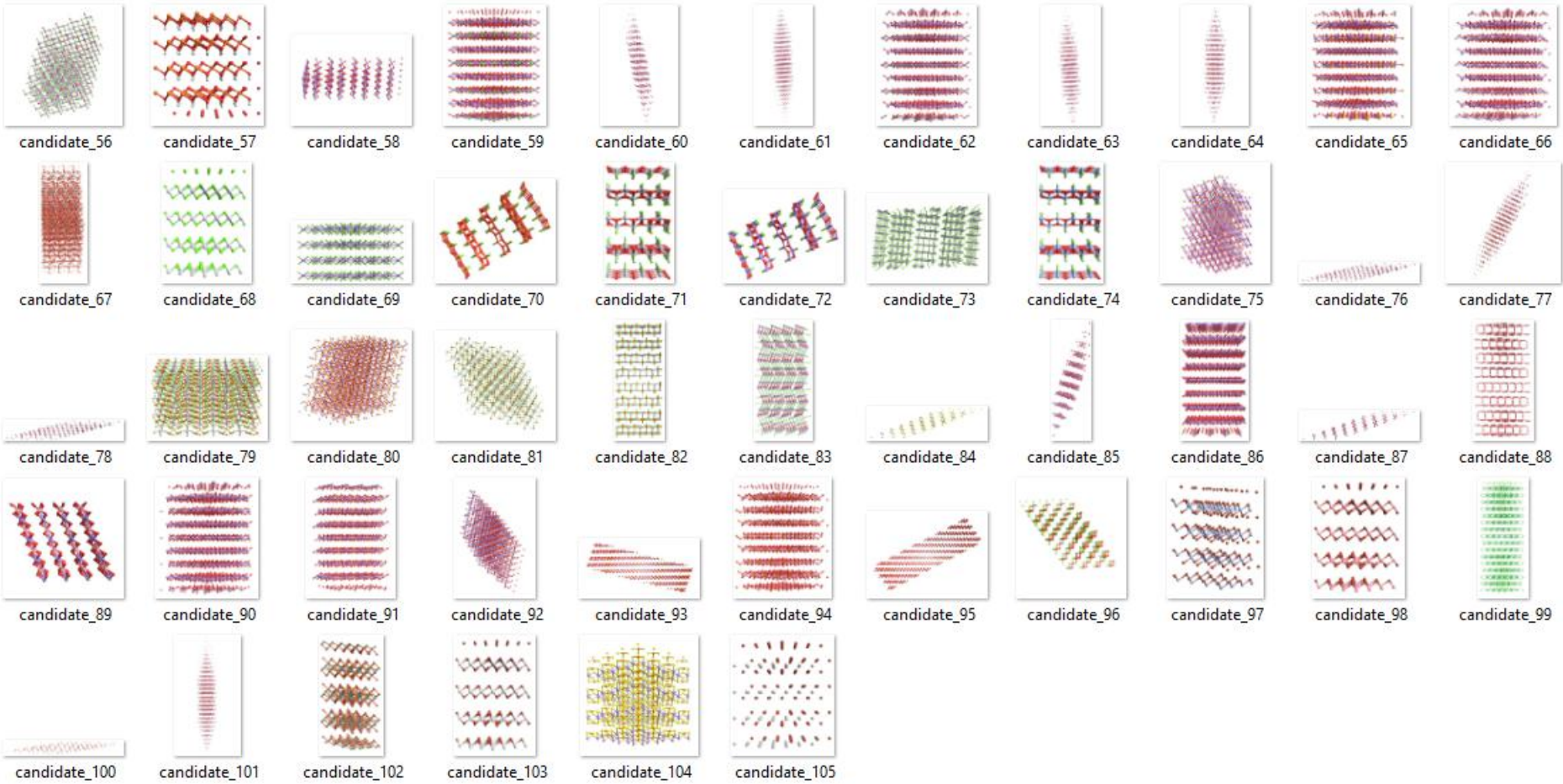
1) Material exited (use ICSD)
2) dimension = 2
3) packing factor < 0.354
4) magnetic moment per atom > 0.625 m_B
5) Energy above hull < 0.136 eV

materials	mp-id	quantity
diamond	mp-66	$\delta E = 0.136$
graphite	mp-568286	p = 0.143
MoS_2	mp-1434	p = 0.354
Fe	mp-13	$\mu = 2.332$
Co	mp-54	$\mu = 3.319$
Ni	mp-23	$\mu = 0.625$

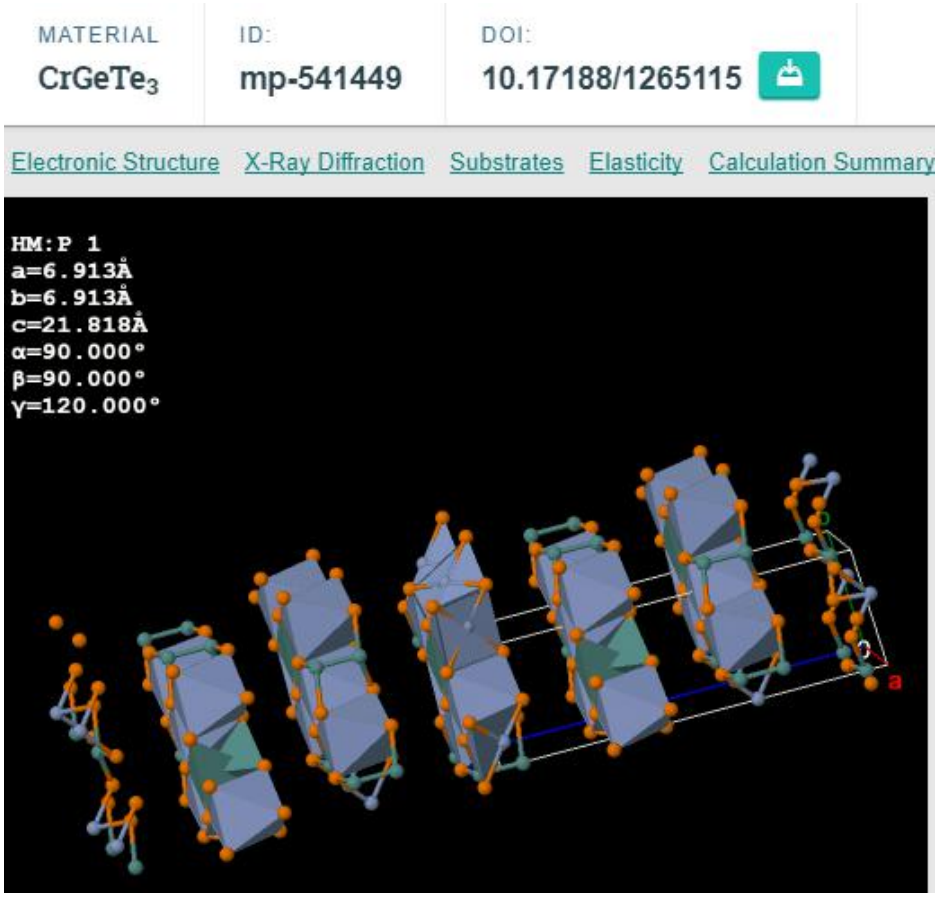
Results: 105 candidates out of 187,093 compounds (1-55)



Results: 105 candidates (56-105)



Rediscover experimental materials



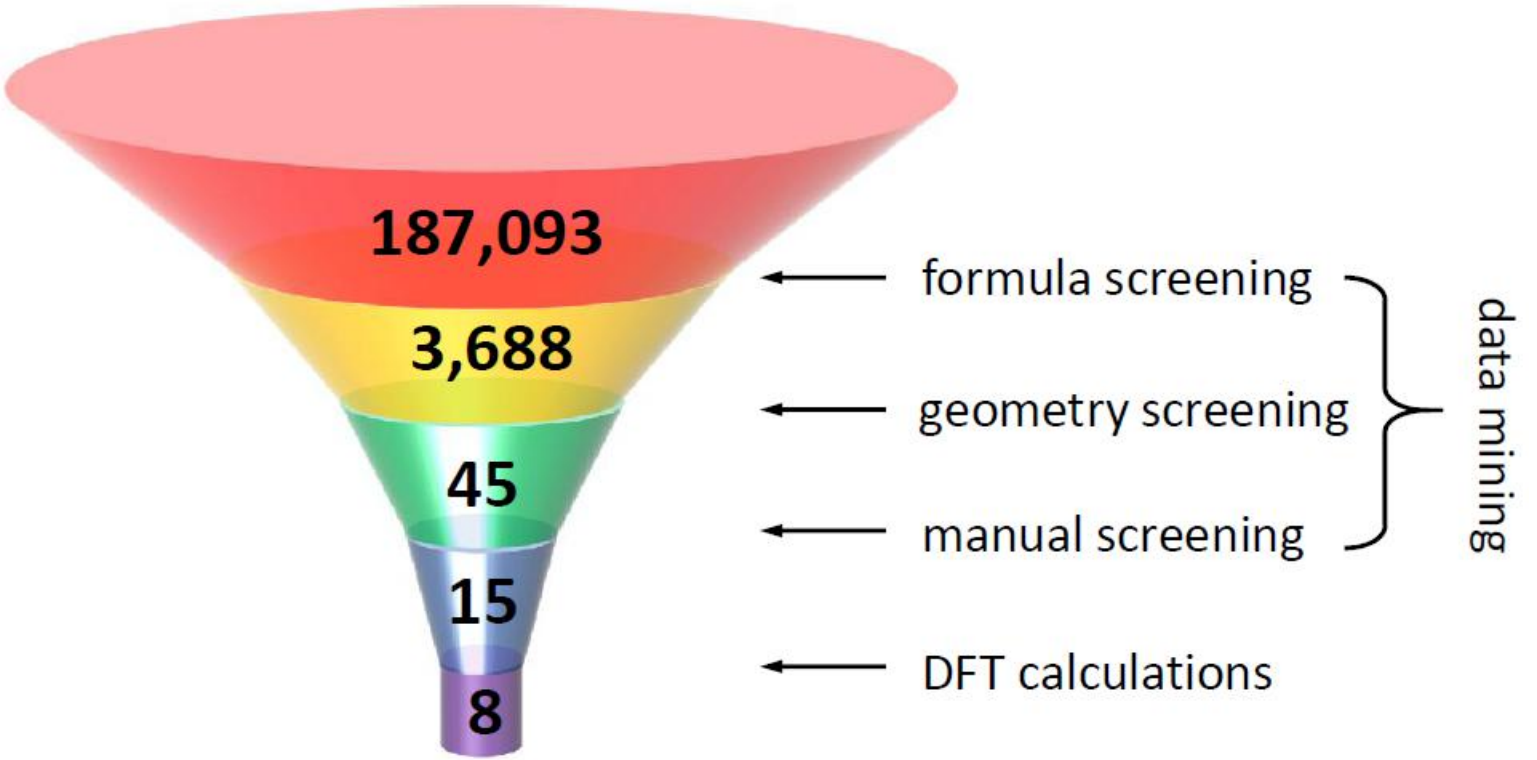
UC group	$Cr_2Ge_2Te_6$
UW and MIT group	CrI_3
mp-541449	$CrGeTe_3$
mp-27734	$CrBr_2$
mp-567504	$CrCl_3$

Add one piece of physics to the search criterion:

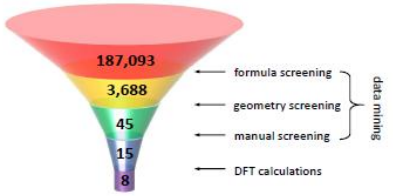
screening criteria

1) **Material exited (so use ICSD)**
2) **dimension = 2**
3) **packing factor < 0.354**
4) **magnetic moment per atom > 0.625 m_B**
5) **Energy above hull < 0.136 eV**
6) **Must have M-Z-M network: M = transition metal atom, Z=heavy atom (Z > 49). This ensures some super-exchange.**

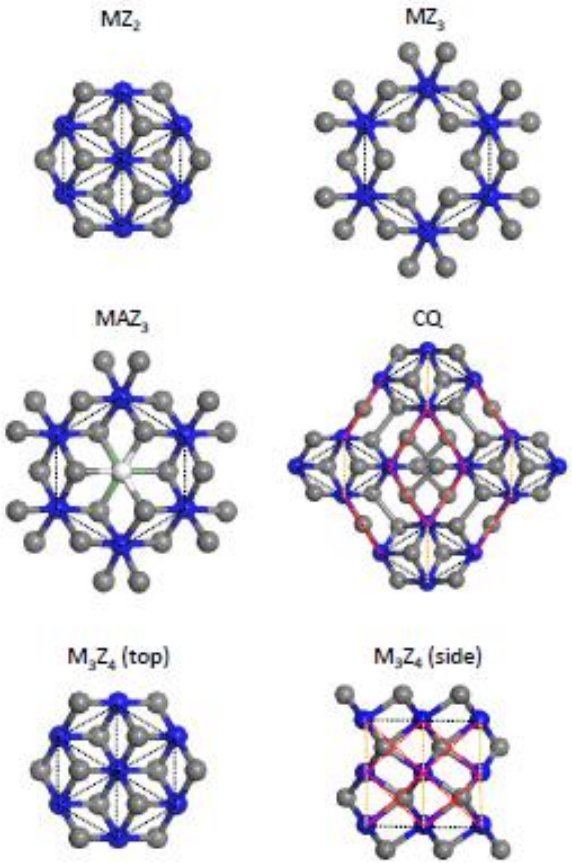
Mining the ICSD for 2D magnetic materials



Phys. Rev. Mater. 2, 081001(R) (2018)



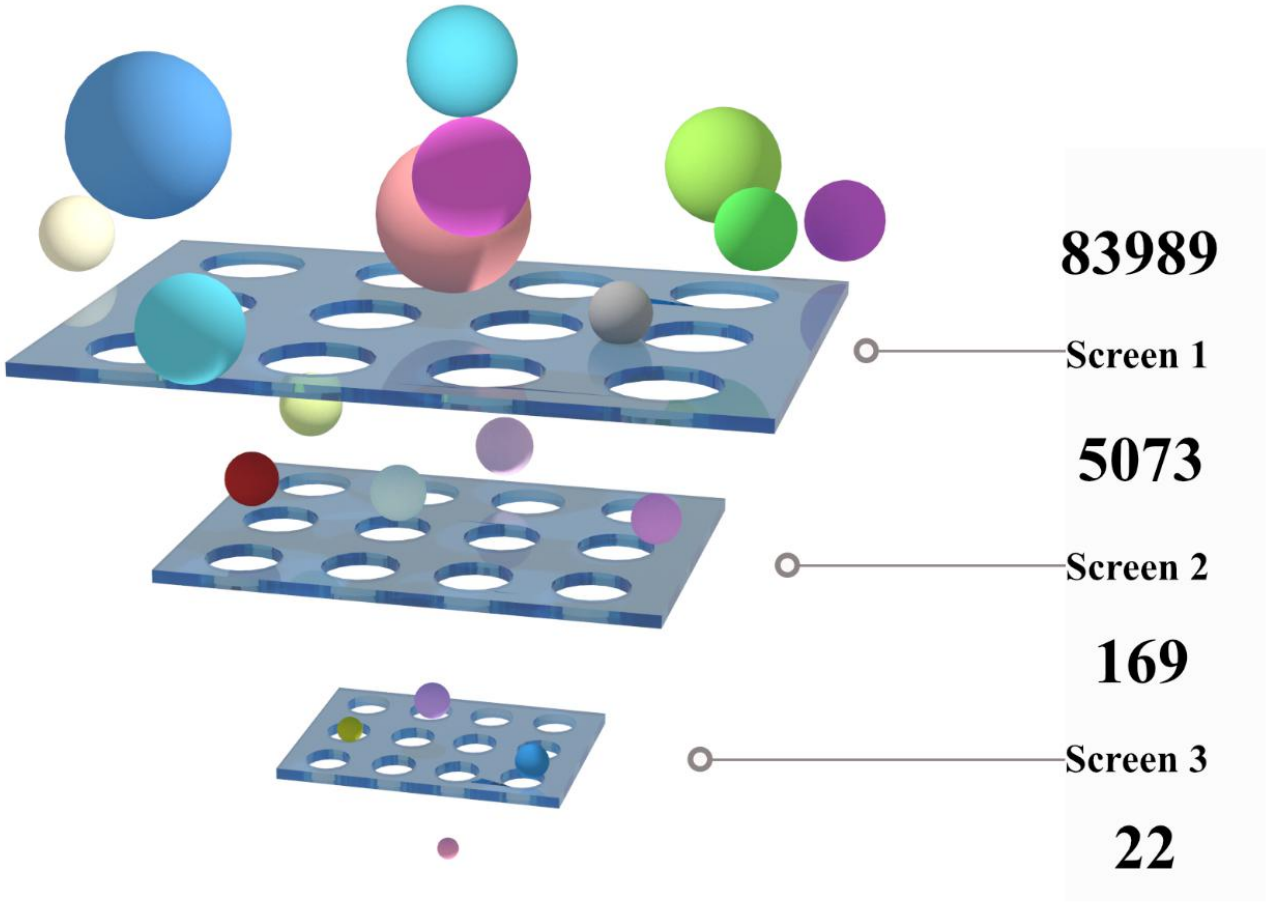
DFT and Monte Carlo data



#	ICSD	formula	structure	ordering	J(meV)	T_c(K)
1	52368	CoI_2	-	AFM	-	-
2	52369	FeI_2	-	AFM	-	-
3	33673	MnI_2	MZ_2	FM	0.3	15
4	22108	NiI_2	MZ_2	FM	1.5	63
5	246907	VI_2	-	AFM	-	-
6	603582	VTe_2	MZ_2	FM	3.0	128
7	626809	$CrSiTe_3$	MAZ_3	FM	12.2	214
8	252343	$CrGeTe_3$	MAZ_3	FM	17.8	314
9	251655	CrI_3	MZ_3	FM	9.1	161
10	180602	FeTe	-	AFM	-	-
11	66958	MnCaSn	-	AFM	-	-
12	4073	CrI_2	-	AFM	-	-
13	76730	NiTe	-	AFM	-	-
14	35266	$CrTe_3$	CQ	FM	5.0	71
					16.5	
					1.5	
15	626873	Cr_3Te_4	sMZ_2	FM	16.6	2057
					18.8	
					12.0	

Using the 2 experimental data, get a heuristic scaling factor of T_c : 0.2 ~ 0.4.

搜索光催化材料



(1) Allowed elements
(2) ICSD
(3) E_{hull}<=0.02
(4) 4<=K_{comp}<=6
(5) 0<E_{gap}<=3 eV
(6) Direct band gap

J. Phys. Chem. Lett. 2019, 10, 5211−5218

搜索光催化材料

Band edge positions:

$$E_{VB}^{0} = \chi(S) - E_e + \frac{1}{2}E_g$$

$$\chi(S) = \sqrt[N]{\chi_1^r \chi_2^s \cdots \chi_{n-1}^p \chi_n^q}$$

https://github.com/nevbac/photocatalytic-material-searching/tree/master/searching_engine