機器學習於材料資訊的應用
Machine Learning on Material Informatics

陳南佑(NAN-YOW CHEN)

nanyow@narlabs.org.tw

楊安正(AN-CHENG YANG)

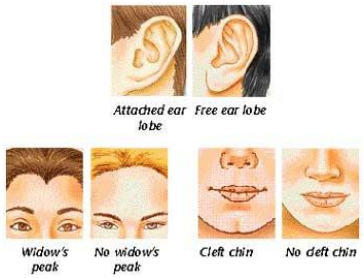
acyang@narlabs.org.tw

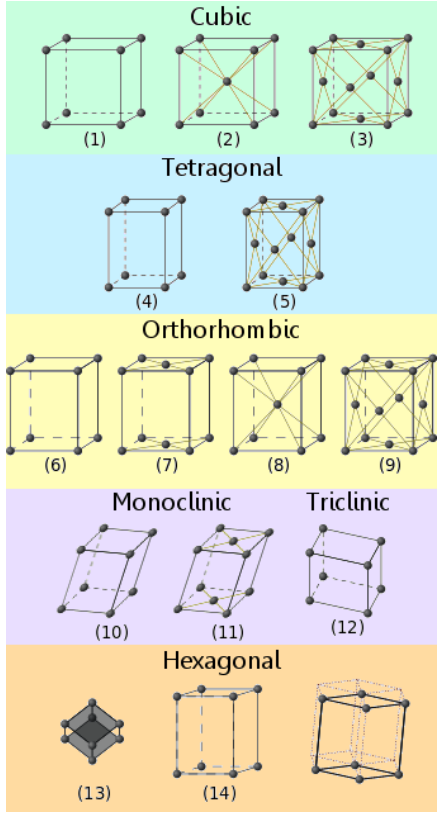
Data comes from Computational Materials Science

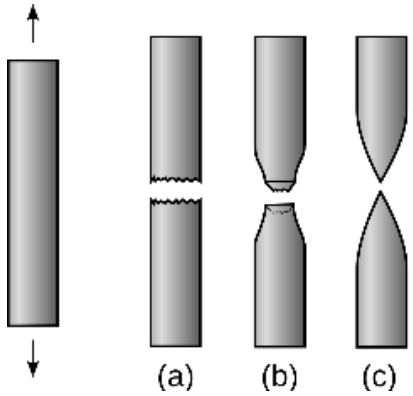
A new powerful approach to discover novel material

材料基因

Human Genome V.S. Material Genome



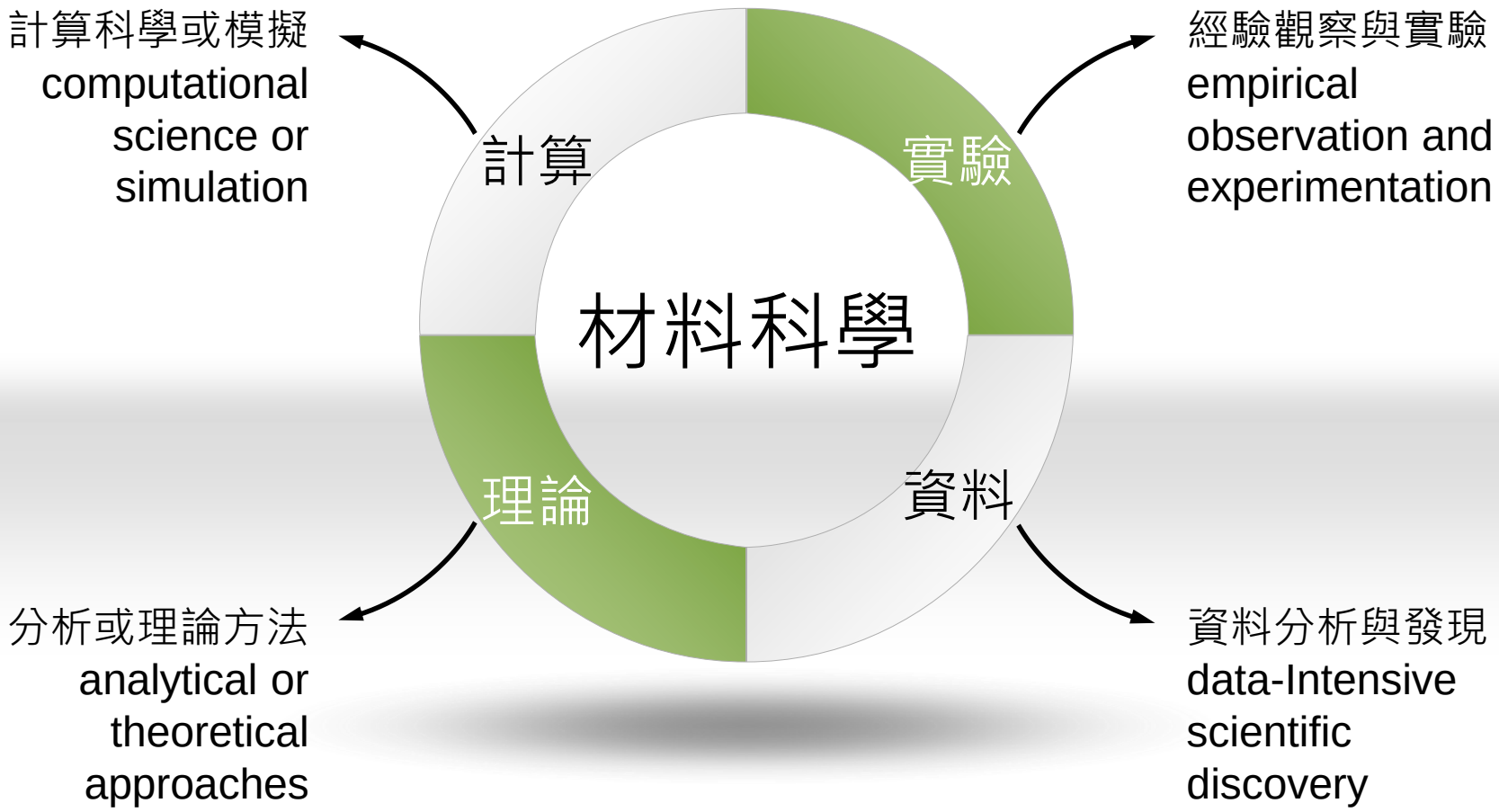




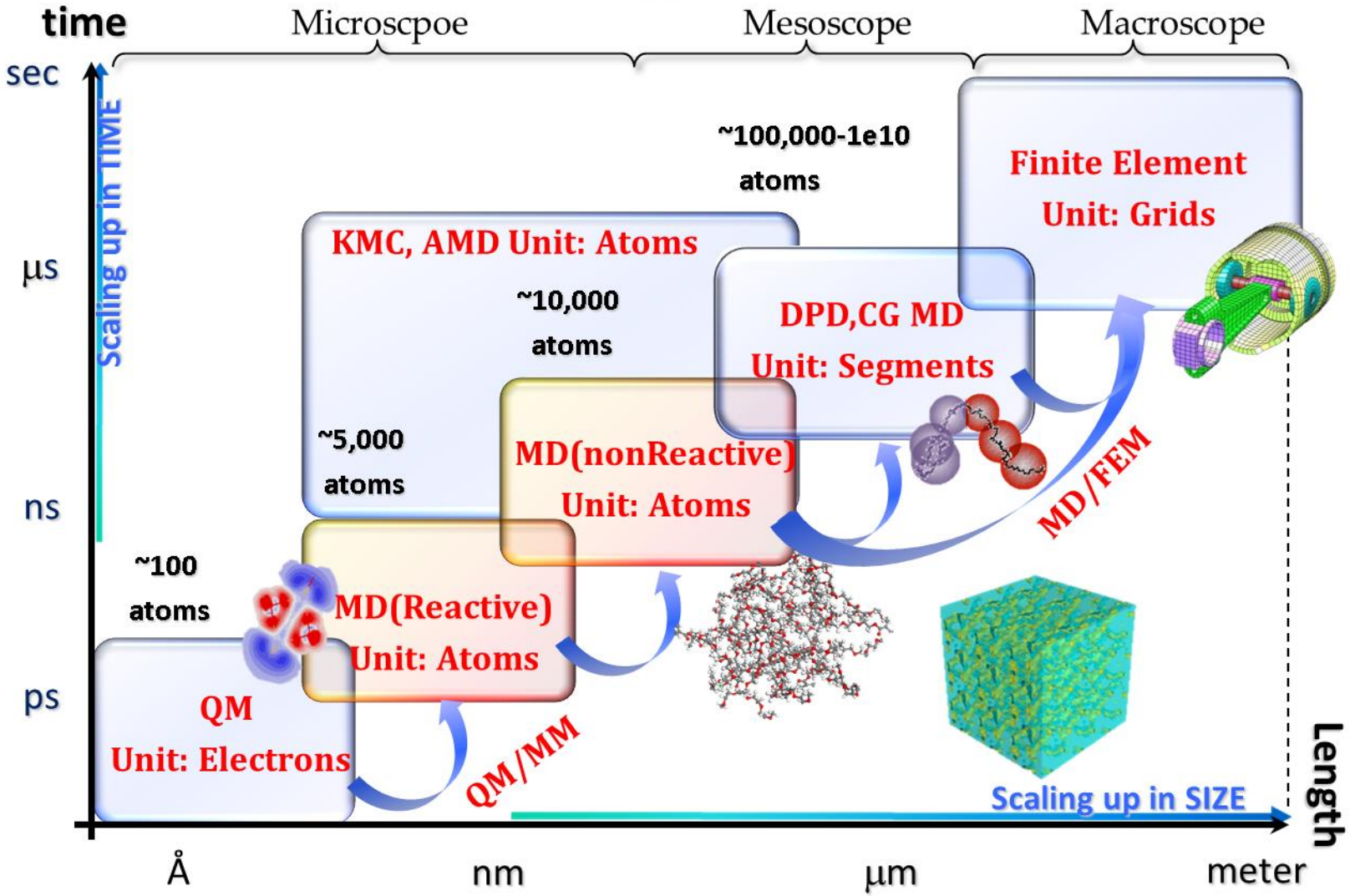
生物個體的外在表現叫作「表現型（phenotype）」，由內在的「基因型（genotype）」決定。改變表現型，首先要改變基因型。.

材料在巨觀尺度表現出來的材料特性是由微觀尺度下的原子組態所決定的。要改變材料特性，首先要改變微觀的原子組態。.

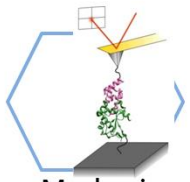
材料科學的分支


計算科學或模擬
computational
science or
simulation
計算
實驗
經驗觀察與實驗
empirical
observation and
experimentation
材料科學
理論
資料
分析或理論方法
analytical or
theoretical
approaches
資料分析與發現
data-Intensive
scientific
discovery

Multi-Scale Modeling for Materials Simulation

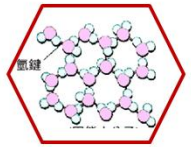


Properties you can get from calculation



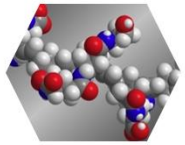
Mechanics

Elastic constant
Young's modus
Stress-strain
Hardness



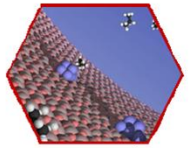
Thermodynamics

Glass transition temperature
Coefficient of Thermal expansion
Thermal conductivity(Green Kubo)
Phonon density
Solubility Parameters



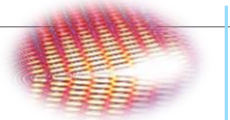
Structure

Radius distribution function
Bond length, angle, dihedral angle
Gyration Radius
Atomic shear strain



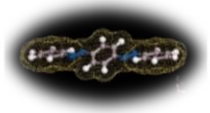
Dynamics

Diffusion coefficient
Viscosity (Green Kubo)
Auto-correlation function



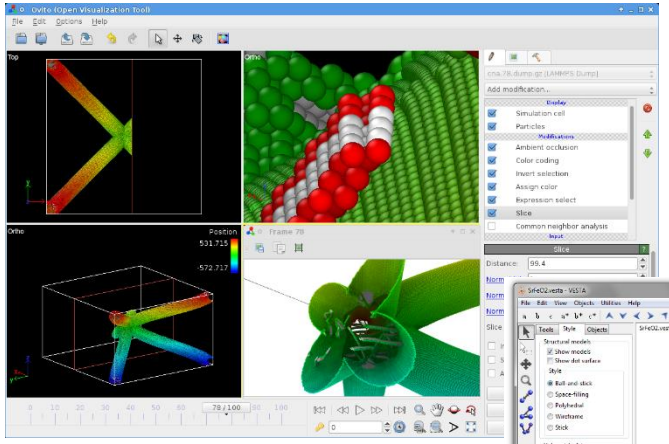
Optics

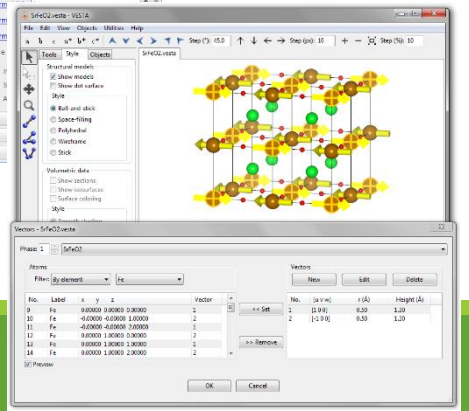
Birefringence
Dielectric constant
Abbe Number
Reflective Index



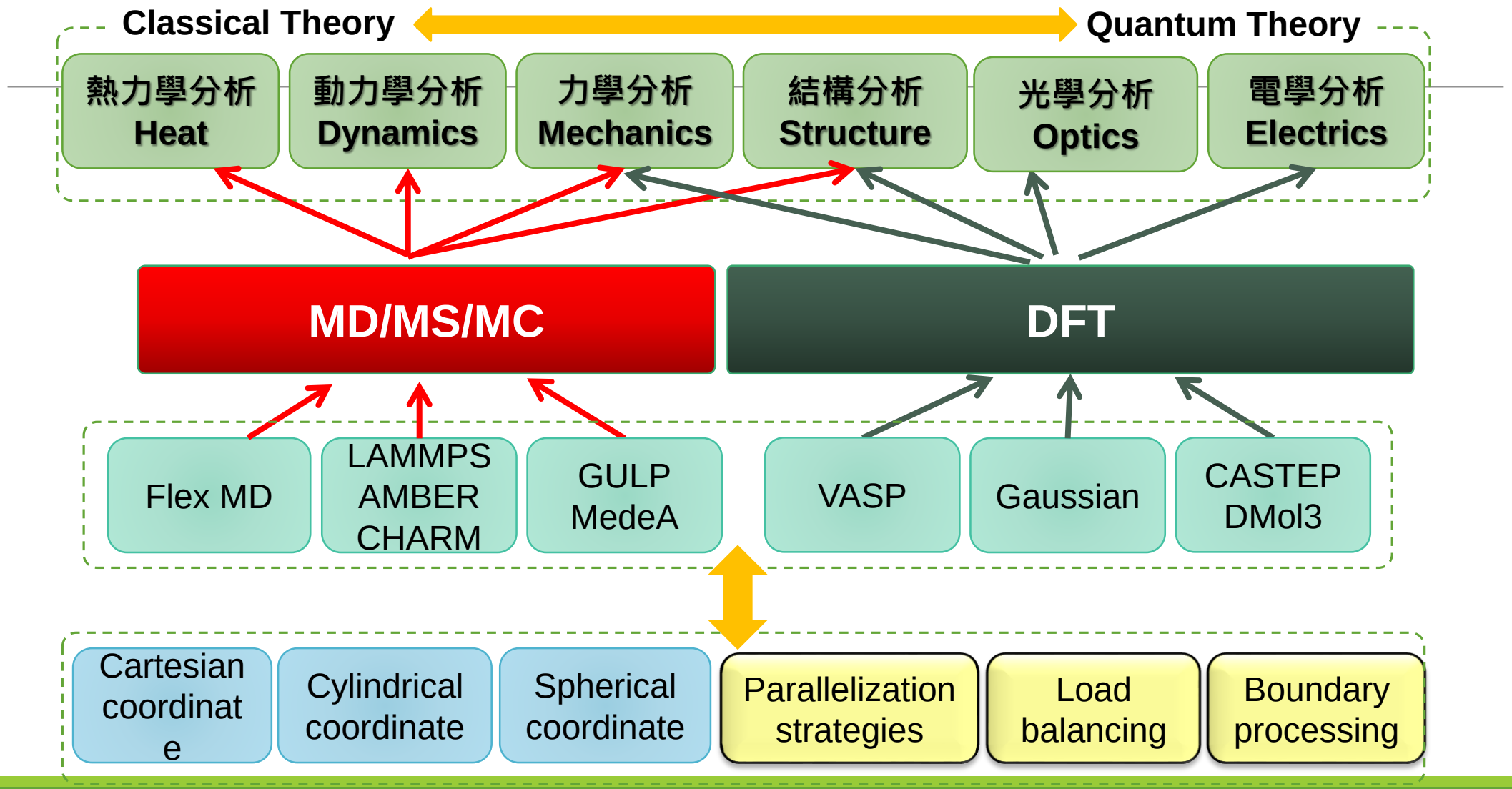
Electrics

Band structure/gap
Density of state

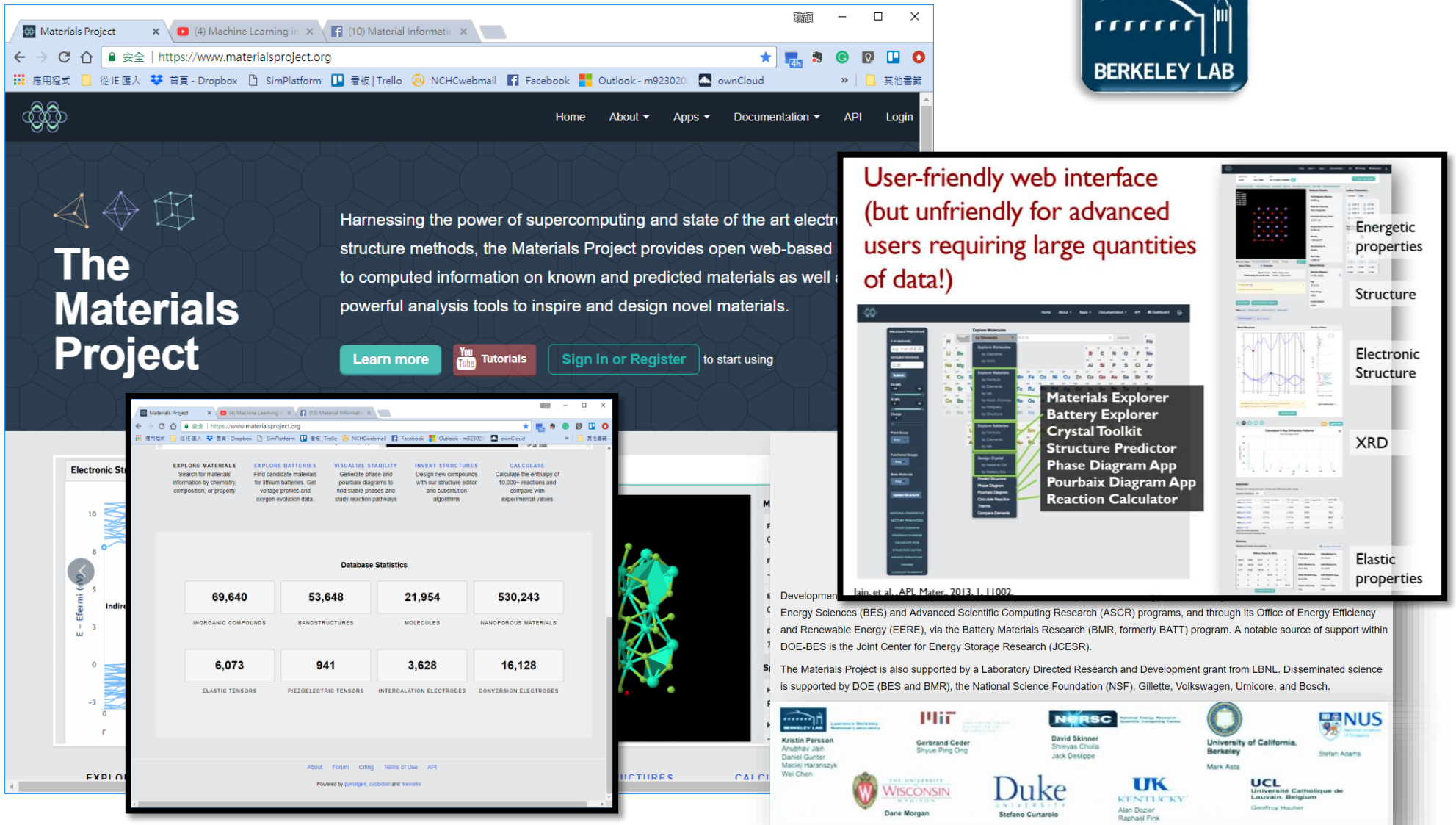




Molecular Modeling



Database
Material Project



About Material Project

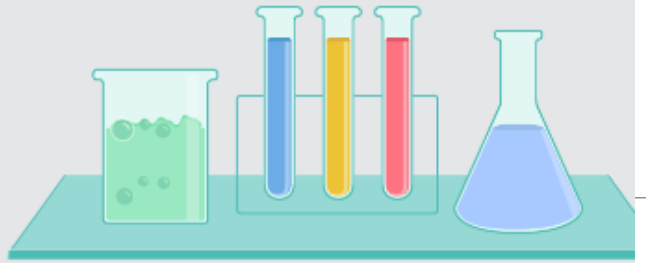
software

By computing properties of all known materials, the Materials Project aims to remove guesswork from materials design in a variety of applications. Experimental research can be targeted to the most promising compounds from computational data sets. Researchers will be able to data-mine scientific trends in materials properties. By providing materials researchers with the information they need to design better, the Materials Project aims to accelerate innovation in materials research.

supercomputers

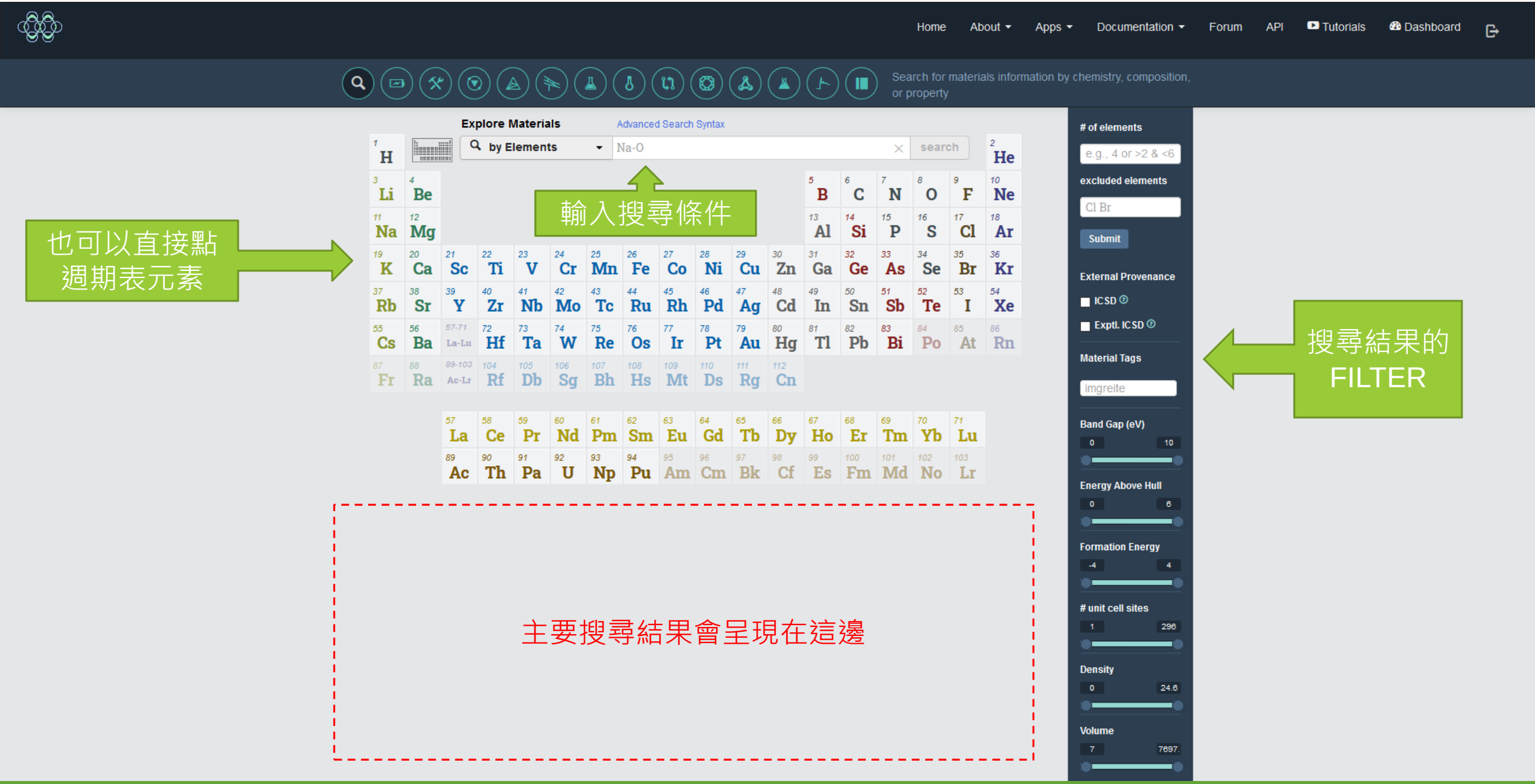
Supercomputing clusters at national laboratories provide the infrastructure that enables our computations, data, and algorithms to run at unparalleled speed. We principally use the Lawrence Berkeley National Laboratory's NERSC Scientific Computing Center and Computational Research Division, but we are also active with Oak Ridge's OLCF, Argonne's ALCF, and San Diego's SDSC.

rank1 rank28 rank13



screening

Computational materials science is now powerful enough that it can predict many properties of materials before those materials are ever synthesized in the lab. By scaling materials computations over supercomputing clusters, we have predicted several new battery materials which were made and tested in the lab. Recently, we have also identified new transparent conducting oxides and thermoelectric materials using this approach.


Home
About
Apps
Documentation
Forum
API
Tutorials
Dashboard
Search for materials information by chemistry, composition, or property
Explore Materials
Advanced Search Syntax
by Elements
Na-O
search
輸入搜尋條件
也可以直接點週期表元素
# of elements
e.g., 4 or >2 & <6
excluded elements
Cl Br
Submit
External Provenance
ICSD
Exptl. ICSD
Material Tags
imgreite
Band Gap (eV)
Energy Above Hull
Formation Energy
unit cell sites
Density
Volume
搜尋結果的FILTER
主要搜尋結果會呈現在這邊

Nelements | Elements

100 records per page | Batch Structures | Edit Structures | Show / hide columns | Print | Export

Materials Id	Formula	Spacegroup	Formation Energy (eV)	E Above Hull (eV)	Band Gap (eV)	Volume	Nsites	Density (gm/cc)
mp-1027525	MoS_2	$P\bar{3}m1$	-1.306	0	1.746	350.447	12	3.034
mp-1025874	MoS_2	$P\bar{6}m2$	-1.306	0	1.783	284.872	9	2.799
mp-1023939	MoS_2	$P\bar{3}m1$	-1.306	0	1.551	219.296	6	2.424
mp-1018809	MoS_2	$P6_3/mmc$	-1.305	0.001	1.338	123.449	6	4.306
mp-1023924	MoS_2	$P\bar{6}m2$	-1.305	0.001	1.658	153.721	3	1.729
mp-2815	MoS_2	$P6_3/mmc$	-1.304	0.002	1.229	118.065	6	4.503
mp-1434	MoS_2	R3m	-1.303	0.003	1.228	116.637	6	4.558
mp-2164	Mo_3S_4	$R\bar{3}$	-1.052	0.068	0.000	272.96	14	5.062
mp-31257	$Mo_{15}S_{19}$	$P6_3/m$	-1.011	0.084	0.000	1436.887	68	4.734
mp-1627	Mo_2S_3	$P2_1/m$	-1.08	0.096	0.000	167.797	10	5.702
mp-1104577	Mo_3S_4	$P\bar{1}$	-1.006	0.113	0.000	229.105	14	6.031
mp-673645	Mo_7S_8	P1	-0.919	0.126	0.000	543.214	30	5.674
mp-990083	MoS_2	Pmmn	-1.142	0.164	0.000	1790.583	54	2.672
mp-1210708	$Mo_{21}S_8$	I4/m	-0.289	0.251	0.000	431.006	29	8.75
mp-1239169	MoS_3	$P2_1/m$	-0.719	0.26	0.791	209.34	8	3.048
mvc-11780	MoS_2	$F\bar{4}3m$	-1.041	0.265	0.000	221.127	12	4.808
mp-558544	MoS_2	$R\bar{3}m$	-1.029	0.277	0.000	111.639	6	4.762

Formation Energy
-4 4

unit cell sites
1 296

Density
0 24.6

Volume
7 7697.

Crystal Systems
Any

Spacegroup Number
Any

Spacegroup Symbol
Any

Has properties →

Elasticity →

Piezoelectricity →

Dielectricity →

可以直接點進去看細節

也可以批次操作

將結構視覺化的區域

下載資料的地方

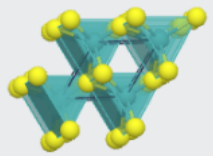
MATERIAL
MoS_2

ID:
mp-1434

DOI:
10.17188/1190621

Show Help Guides

Electronic Structure | X-Ray Diffraction | X-Ray Absorption | Substrates | Elasticity | Piezoelectricity | Dielectric Properties | Similar Structures | Synthesis Descriptions | Calculation Summary | User Contributions | Provenance/Citation



Mo S Atoms Unit Cell Bonds Polyhedra CIF

Zoom in/out Shift + Drag cursor
Rotate along the center axis Option + Drag cursor

Edit Crystal | Generate Phase Diagram

Tags: Molybdenite 3R | Molybdenum(IV) disulfide | Molybdenum(IV) sulfide

File Formats

Material Details

Final Magnetic Moment
0.000 μ_B

Magnetic Ordering
NM

Formation Energy / Atom
-1.303 eV

Energy Above Hull / Atom
0.003 eV

Density
4.56 g/cm^3

Decomposes To
MoS_2

Band Gap
1.228 eV

Space Group

Hermann Mauguin
R3m [160]

Hall
R 3 2"

Point Group
m

Crystal System
trigonal

Lattice Parameters

a	3.194 Å	α	97.687°
b	5.530 Å	β	76.529°
c	6.855 Å	γ	90.000°

Volume 116.637 A^3

Final Structure
Fractional Coordinates
CIF

Mo

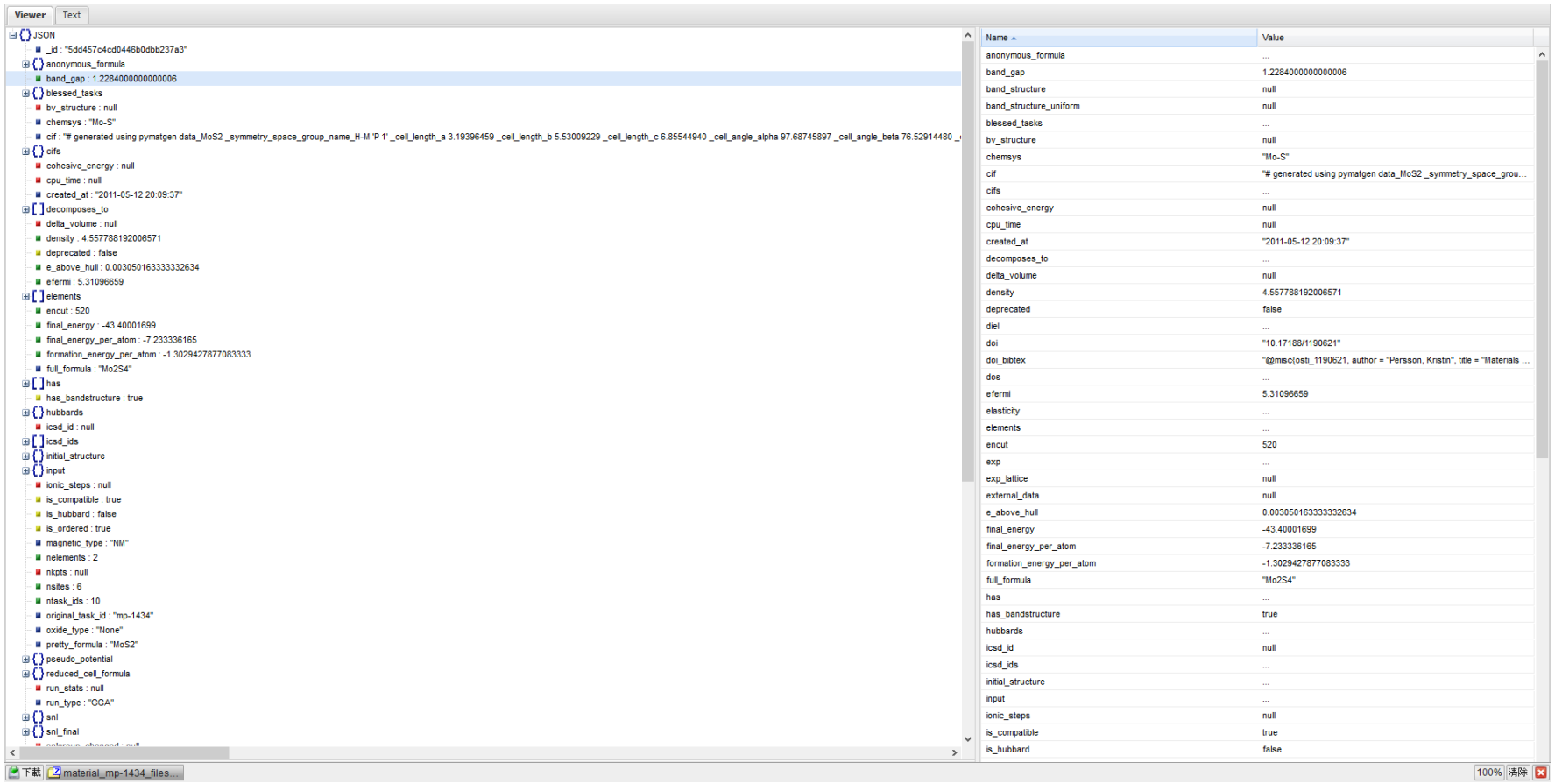
a	b	c
0	0	0.0001
0.5	0.5	0.0001

S

a	b	c
0.1187	0.6272	0.7626
0.3813	0.2061	0.2374
0.6187	0.1272	0.7626
0.8813	0.7061	0.2374

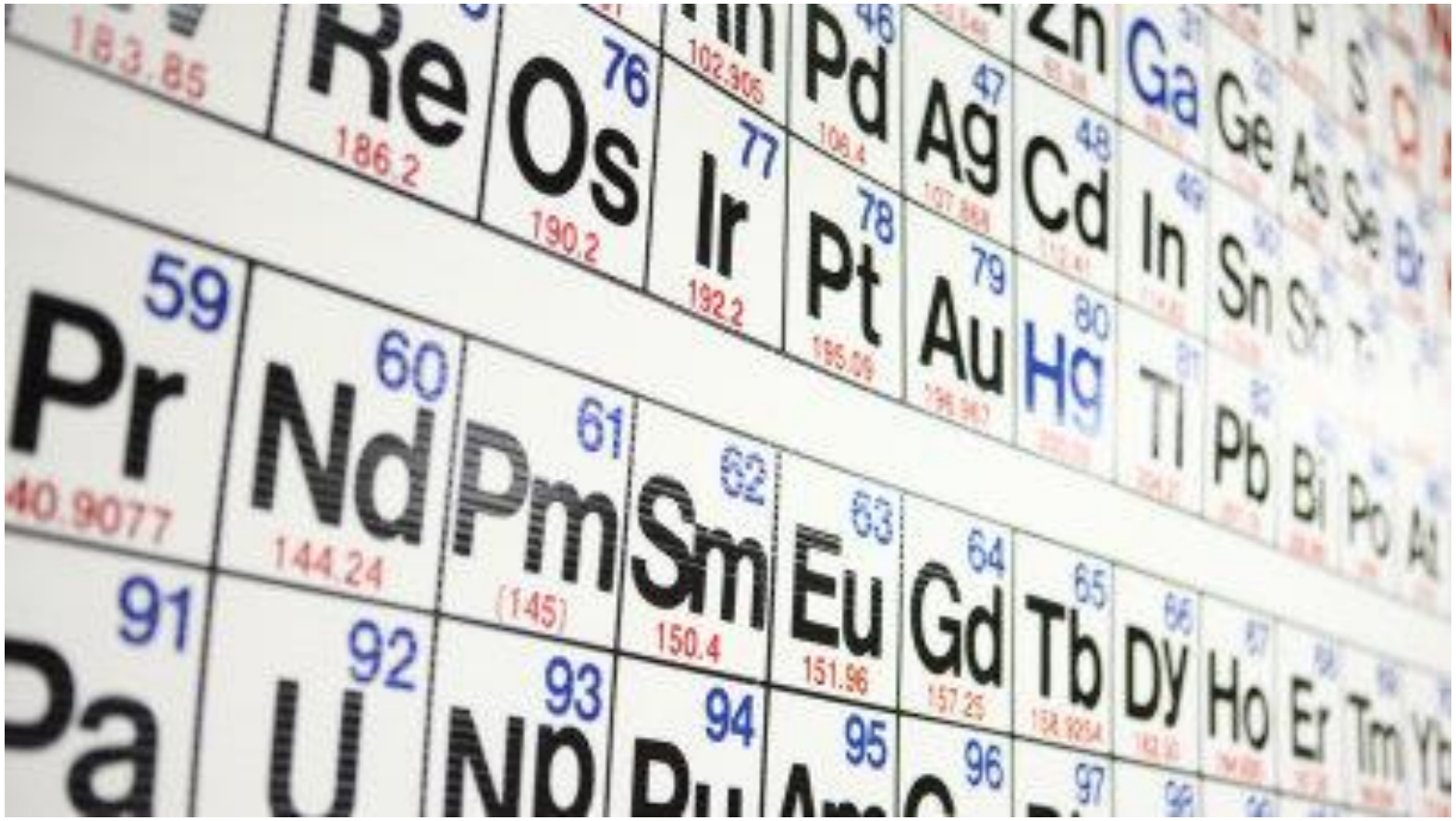
CIF:晶體結構的標準檔案交換格式
VASP:執行VASP計算的設定檔案(不包含最重要的POTCAR)
POSCAR:原子的座標輸入
CSSR: unit cell 定義和原子的fractional coordinates.
JSON:JSON格式的所有資訊

Online jsonviewer
http://jsonviewer.stack.hu/



Material Properties Prediction

digging into the periodic table



Problem Define

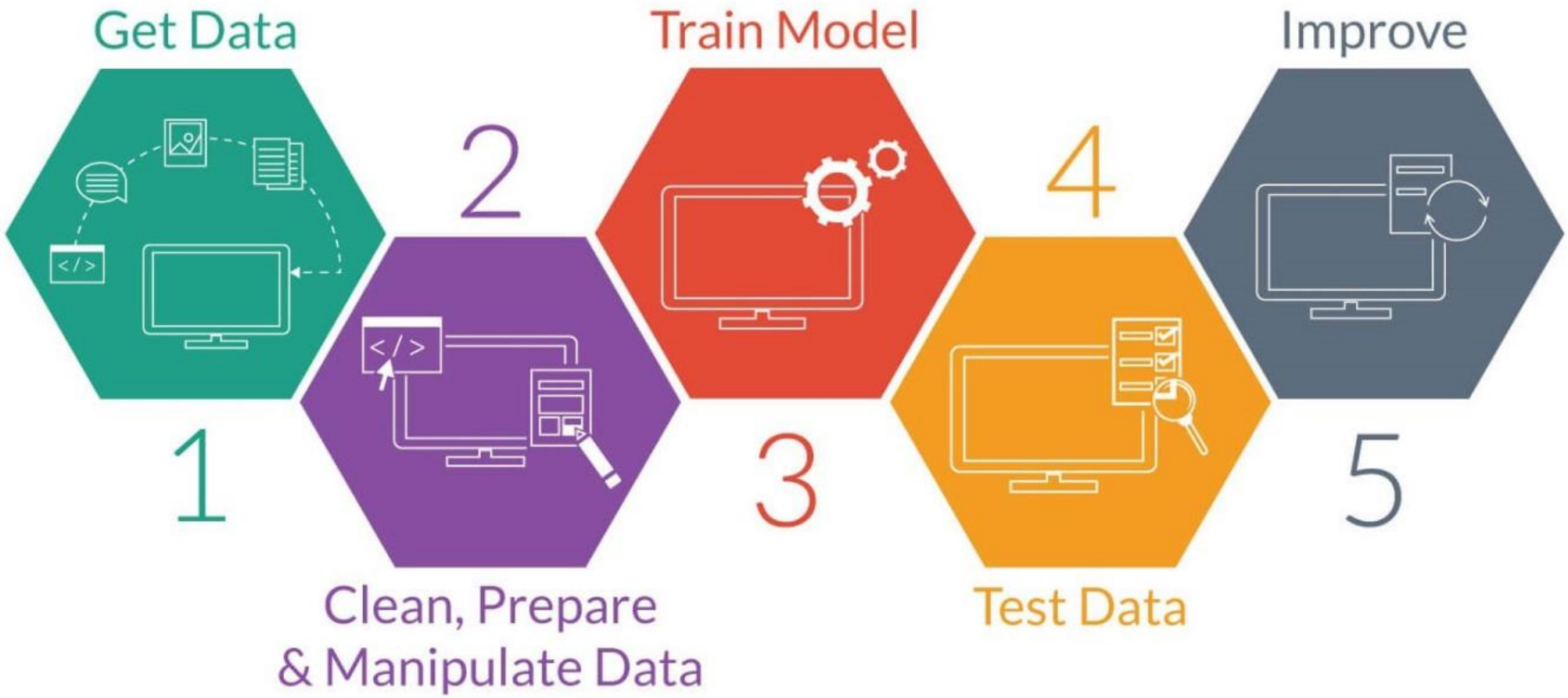
- 在早期，元素被發現的種類不多，化學家只能局部的對某些性質相近的元素進行歸類整理，例如1865年英國化學家伍德林（W. Woodling）按原子量排列元素順序，初步排出今日元素週期表中的鹵族、氮族、氧族。
- 俄國化學家門得列夫（Dmitri Ivanovich Mendeleev, 1834 – 1907）全面考慮了元素的各種性質，不僅根據元素的原子量，而且很重視元素的性質及其與其他元素的關係，他依原子量遞增的順序把元素排列成幾行，同時把各行中性質相似的元素左右對齊，這樣使得每一橫排化學元素的性質相近，每一縱列化學元素性質的變化也呈現着規律性，整個元素系列呈現出周期性變化。

資料的整理歸納

- 1869年2月，門得列夫發表了《元素性質和原子量的關係》論文，同時公布了他的第一張化學元素週期表，周期表中留下了四個空位，空位上沒有元素名稱，只有預計的原子量，表示尚待發現的元素。

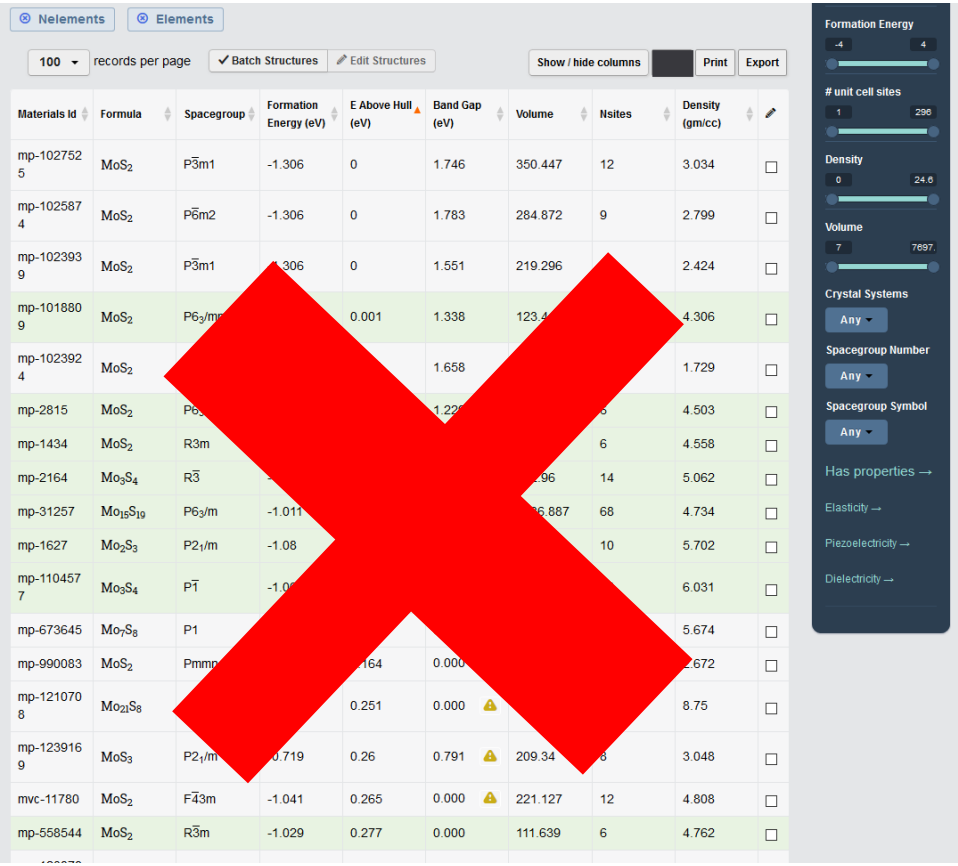
新元素的預測

- 那化合物的特性能不能找出週期性?
- 可不可以從既有的材料資料庫預測出新材料的性質?


Get Data
1
2
Clean, Prepare
& Manipulate Data
Train Model
3
4
Test Data
Improve
5

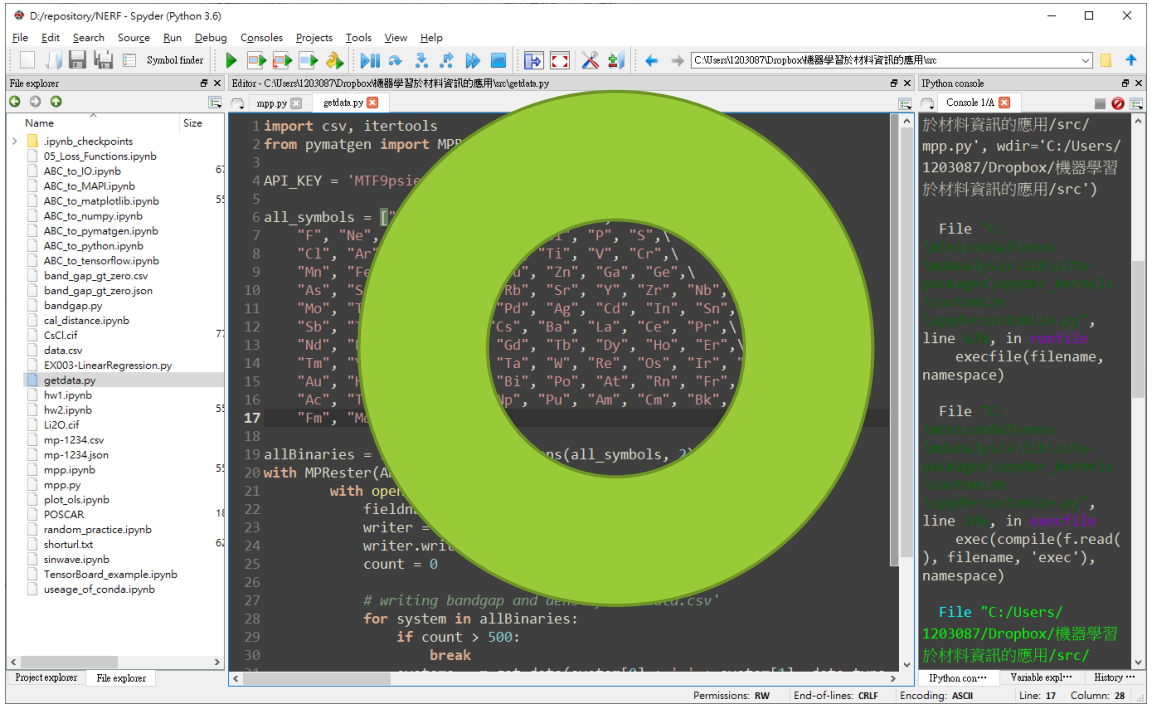
Get Data

手動下載



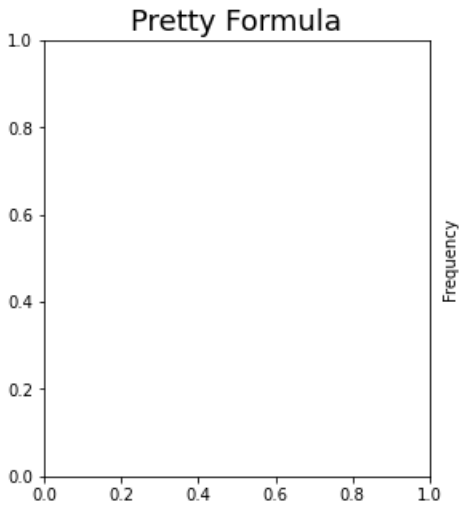
自動化下載

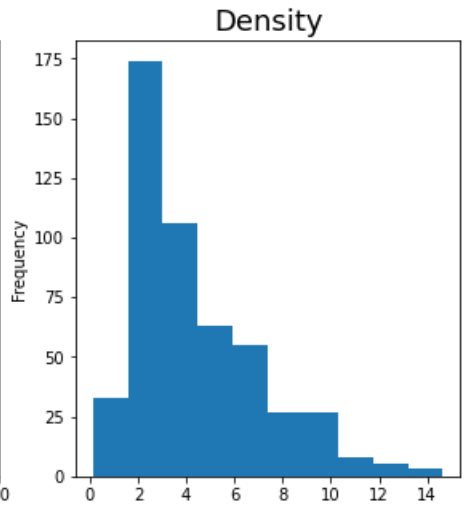
open Materials Application Programming Interface (API)

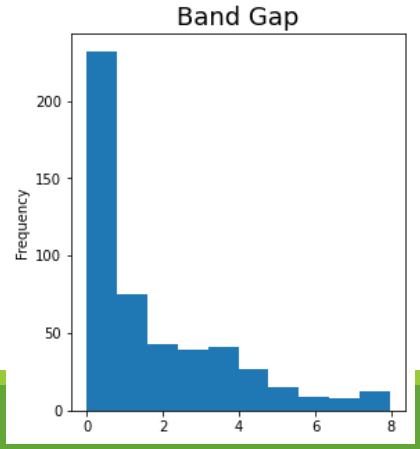


Clean, Prepare, Manipulate Data

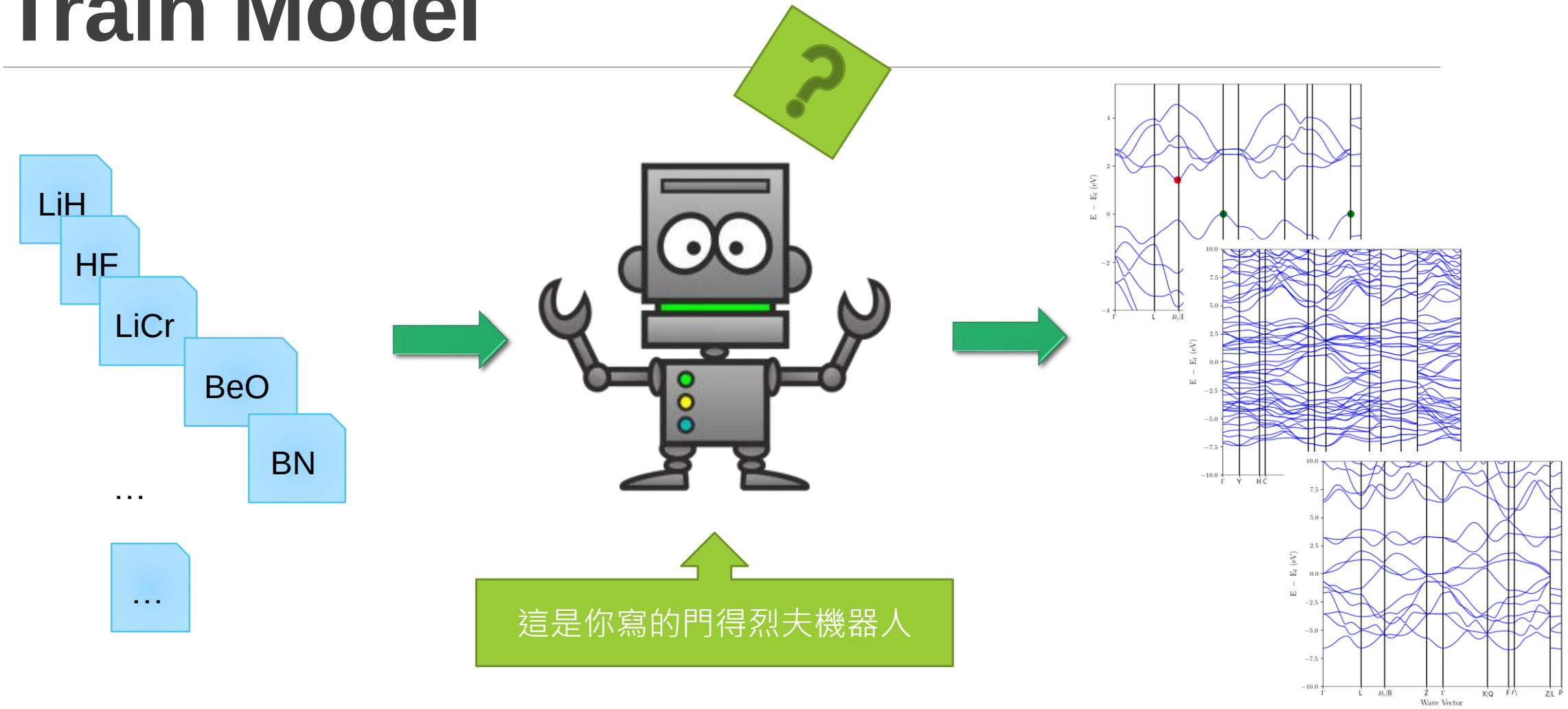
Pretty Formula	Density	Band Gap
LiH	0.814395727	3.018
BeH2	0.807534506	5.3418
B3H5	0.811598971	3.6983
HC	1.409269759	3.091
HN	1.336811255	4.1598
H2O2	1.834767366	4.1525
HF	1.721540912	6.7187
NaH	1.394380186	3.7974
MgH2	1.450102639	3.6284
AlH3	1.459511166	2.1855
…	…	…







Train Model



Model
Finding a function from data

$f_1($ LiH $)=$aaaaa $f_2($ LiH $)=$eeeee

$f_1($ HF $)=$bbbbb $f_2($ HF $)=$fffff

$f_1($ LiCr $)=$ccccc $f_2($ LiCr $)=$ggggg ……

$f_1($ BeO $)=$ddddd $f_2($ BeO $)=$hhhhh

訓練的過程說穿了就是找出一個合適的function來描述輸入和輸出的關係

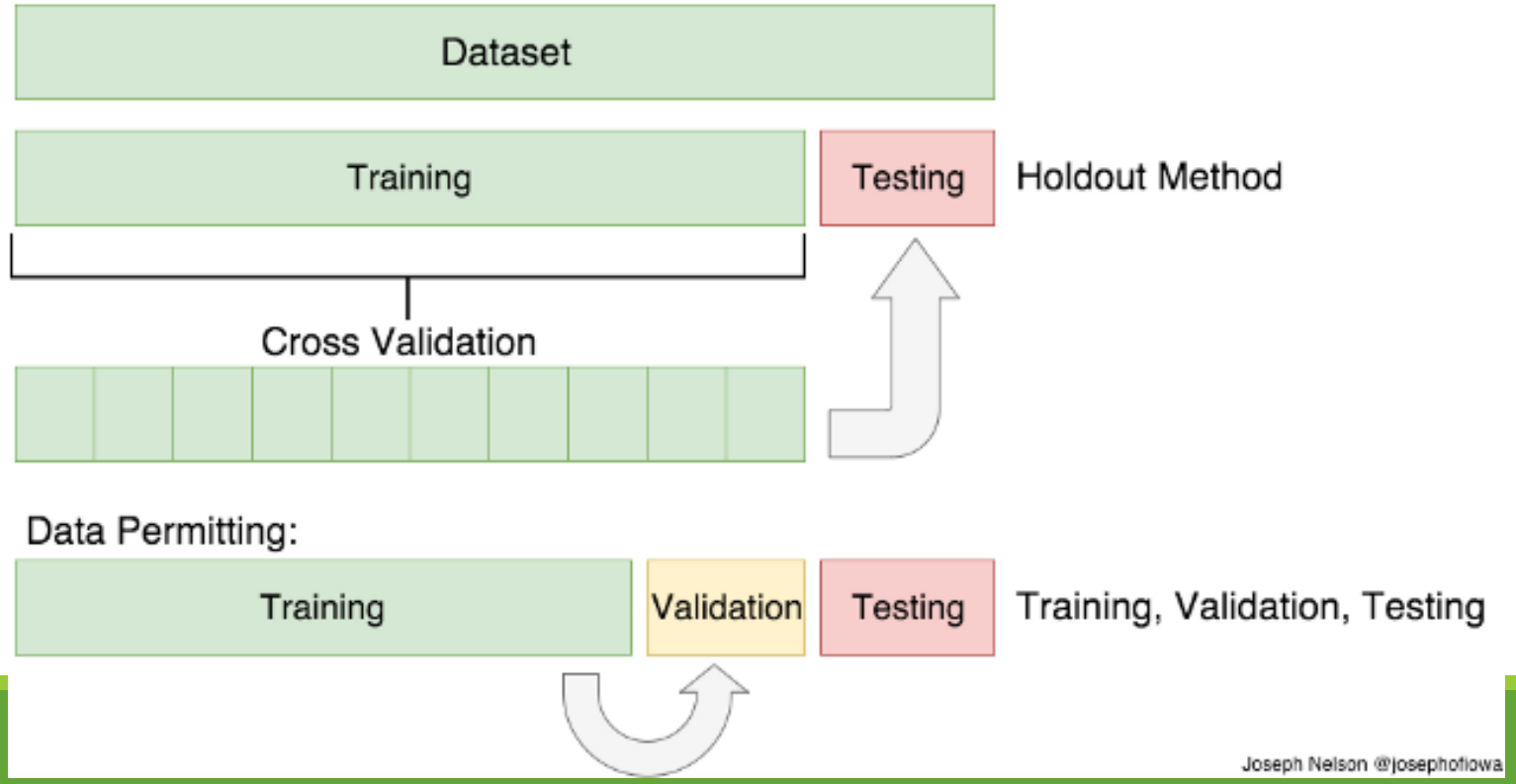
Scikit-Learn Regression algorithm

- Linear Models
- Kernel ridge regression
- Support Vector Machines
- Gaussian Processes
- Decision Trees
- Ensemble methods
- …

https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

Test Data (Test Model)

- 不要把所有所有的資料都餵進去給model，只要把一部分的資料餵進去(Training Dataset)訓練模型，需要保留一些資料拿來檢驗模型(Testing Dataset)。
- Cross Validation(交叉驗證)的部份之後會再講。



Train/Test Split

- sklearn.model_selection.train_test_split

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
```

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html