Machine learning on molecular simulation

Hongbin Ren （任宏斌） | 2016.11.30 | IOP. CAS

Historic way
v.s.
Modern way

ALPHA

how can machine learning help physicist, based on my understanding, I think it makes material simulation based on DFT possible to be implemented in large scale calculation

现阶段工业和制药中采用的材料计算方法：

经典分子动力学，力常数方法，经验参数

机器学习使得**更精确的**大规模

材料计算成为现实……

Benefits ……

Medicine

Industry

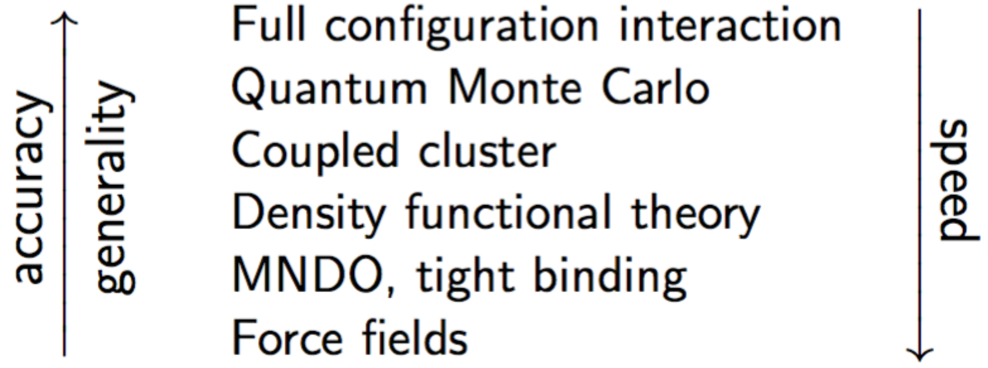
Outline

- DFT in nutshell
- Why machine learning (**Supervised learning**) ?
- Physics+Technology=?
- Summary

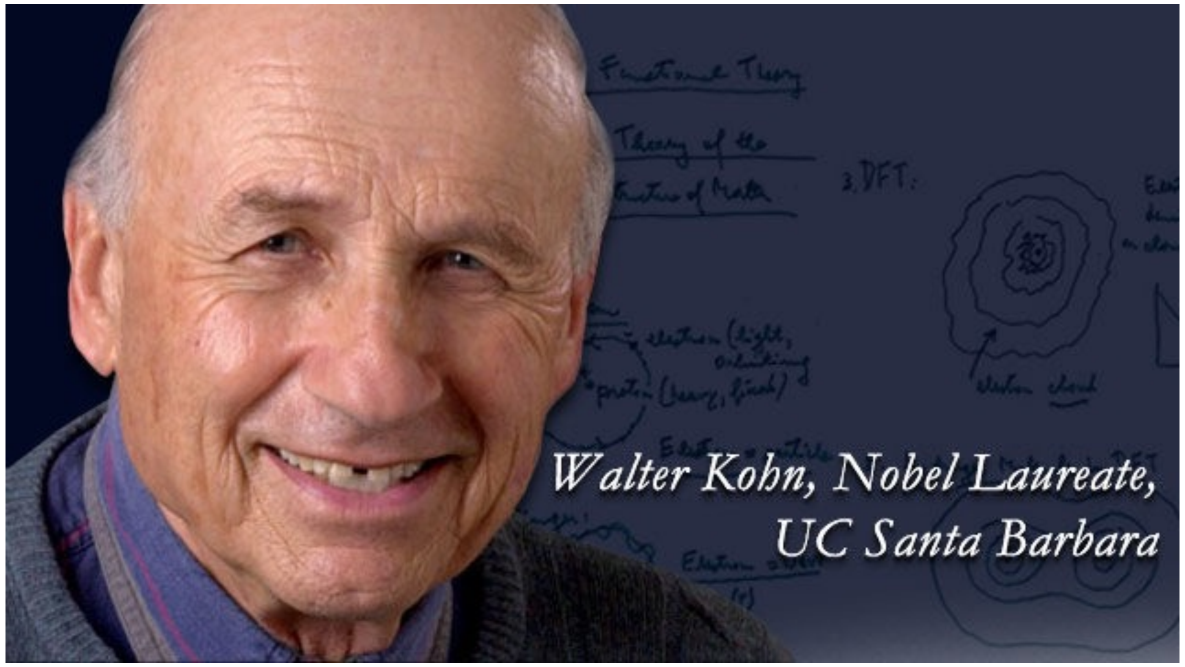
DFT in nutshell

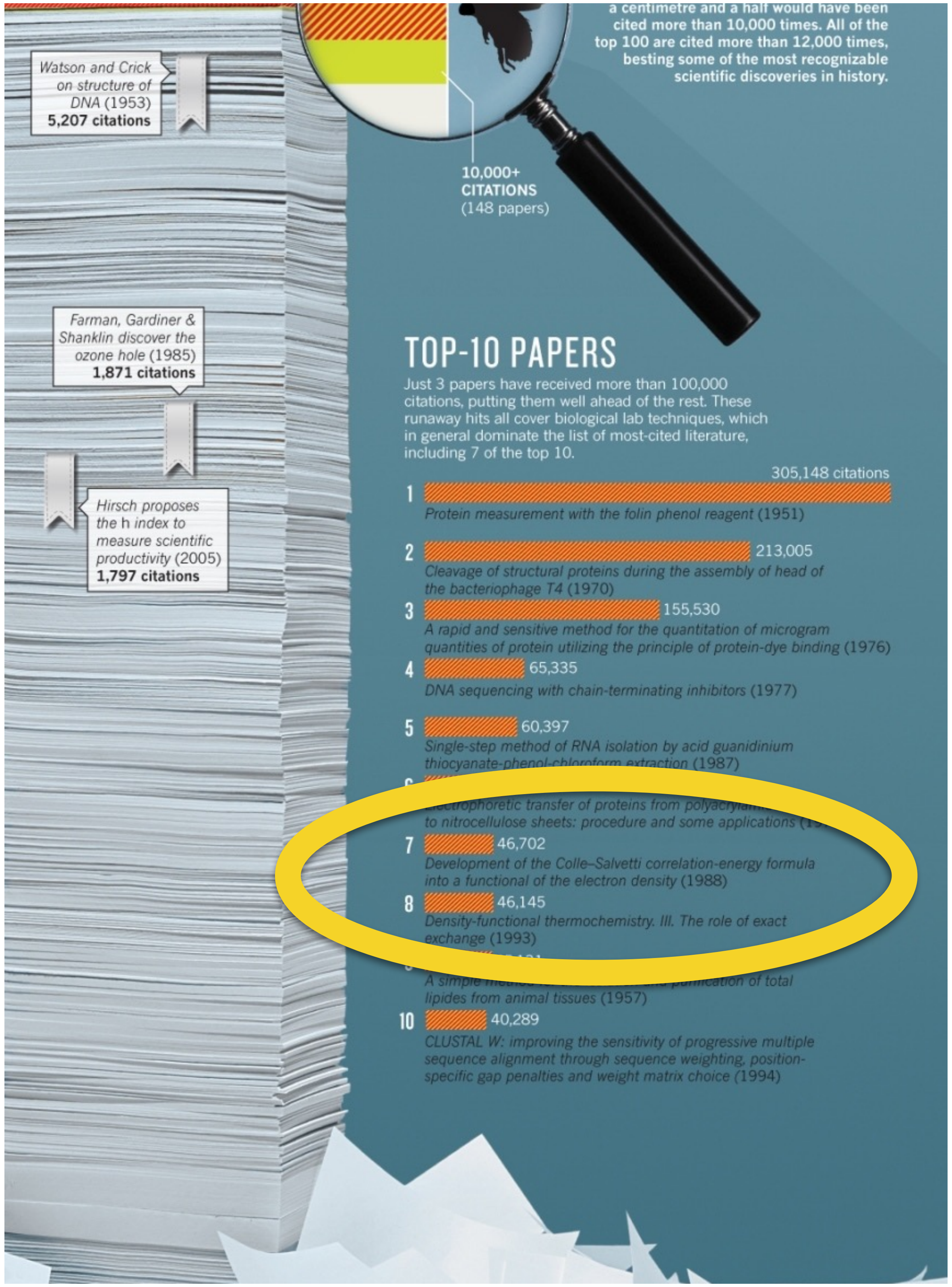
DFT:

(Density functional theory)



Full configuration interaction
Quantum Monte Carlo
Coupled cluster
Density functional theory
MNDO, tight binding
Force fields


Walter Kohn, Nobel Laureate,
UC Santa Barbara


a centimetre and a half would have been cited more than 10,000 times. All of the top 100 are cited more than 12,000 times, besting some of the most recognizable scientific discoveries in history.

TOP-10 PAPERS

Just 3 papers have received more than 100,000 citations, putting them well ahead of the rest. These runaway hits all cover biological lab techniques, which in general dominate the list of most-cited literature, including 7 of the top 10.

305,148 citations
1
Protein measurement with the folin phenol reagent (1951)
2 213,005
Cleavage of structural proteins during the assembly of head of the bacteriophage T4 (1970)
3 155,530
A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding (1976)
4 65,335
DNA sequencing with chain-terminating inhibitors (1977)
5 60,397
Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction (1987)
ectrophoretic transfer of proteins from polyacryla
to nitrocellulose sheets: procedure and some applications
7 46,702
Development of the Colle–Salvetti correlation-energy formula into a functional of the electron density (1988)
8 46,145
Density-functional thermochemistry. III. The role of exact exchange (1993)
A simple
purification of total lipides from animal tissues (1957)
10 40,289
CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice (1994)

DFT overview

self-consistent loop (on the board)
approximation

$$\begin{aligned}\hat{H} \;=\; & -\frac{\hbar^2}{2}\sum_i \frac{\nabla^2_{\vec{R}_i}}{M_i} - \frac{\hbar^2}{2}\sum_i \frac{\nabla^2_{\vec{r}_i}}{m_e} \\ & - \frac{1}{4\pi\epsilon_0}\sum_{i,j}\frac{e^2\, Z_i}{|\vec{R}_i - \vec{r}_j|} + \frac{1}{8\pi\epsilon_0}\sum_{i\neq j}\frac{e^2}{|\vec{r}_i - \vec{r}_j|} + \frac{1}{8\pi\epsilon_0}\sum_{i\neq j}\frac{e^2\, Z_i Z_j}{|\vec{R}_i - \vec{R}_j|}\end{aligned}$$

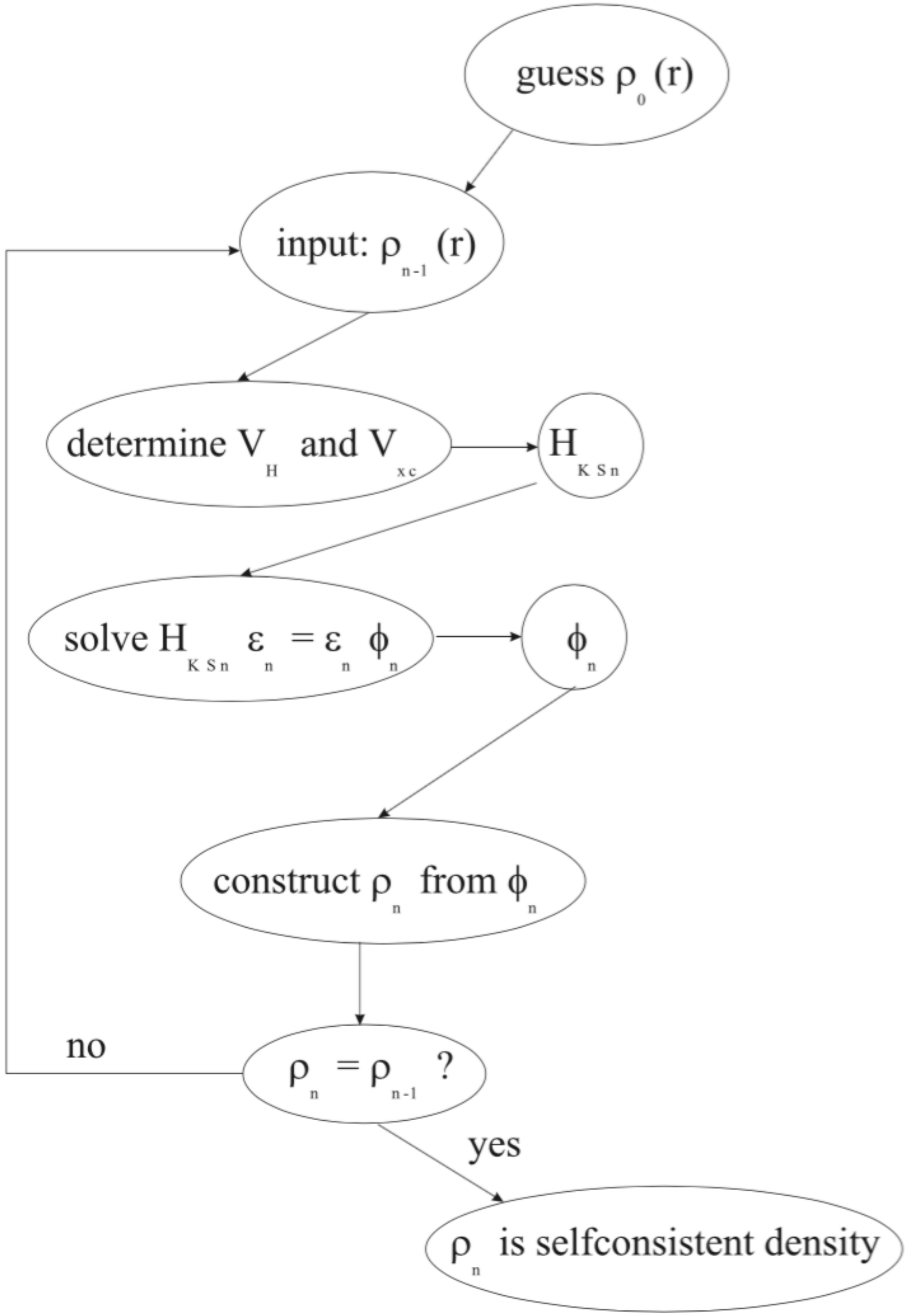
Many-body Hamiltonian

Interacting ground state

Non-interacting ground state

Single particle Hamiltonian

$$\begin{aligned}\hat{H}_{KS} \;&=\; \hat{T}_0 + \hat{V}_H + \hat{V}_{xc} + \hat{V}_{ext} \\ &=\; -\frac{\hbar^2}{2m_e}\vec{\nabla}^2_i + \frac{e^2}{4\pi\epsilon_0}\int \frac{\rho(\vec{r}\,')}{|\vec{r} - \vec{r}\,'|}\, d\vec{r}\,' + V_{xc} + V_{ext}\end{aligned}$$


guess $\rho_0(r)$
input: $\rho_{n-1}(r)$
determine V_H and V_{xc}
H_{KSn}
solve $H_{KSn}\,\varepsilon_n = \varepsilon_n\,\phi_n$
ϕ_n
construct ρ_n from ϕ_n
no
$\rho_n = \rho_{n-1}$?
yes
ρ_n is selfconsistent density

For any new and complex system, it need vast amount of time to reformulate the information needed.

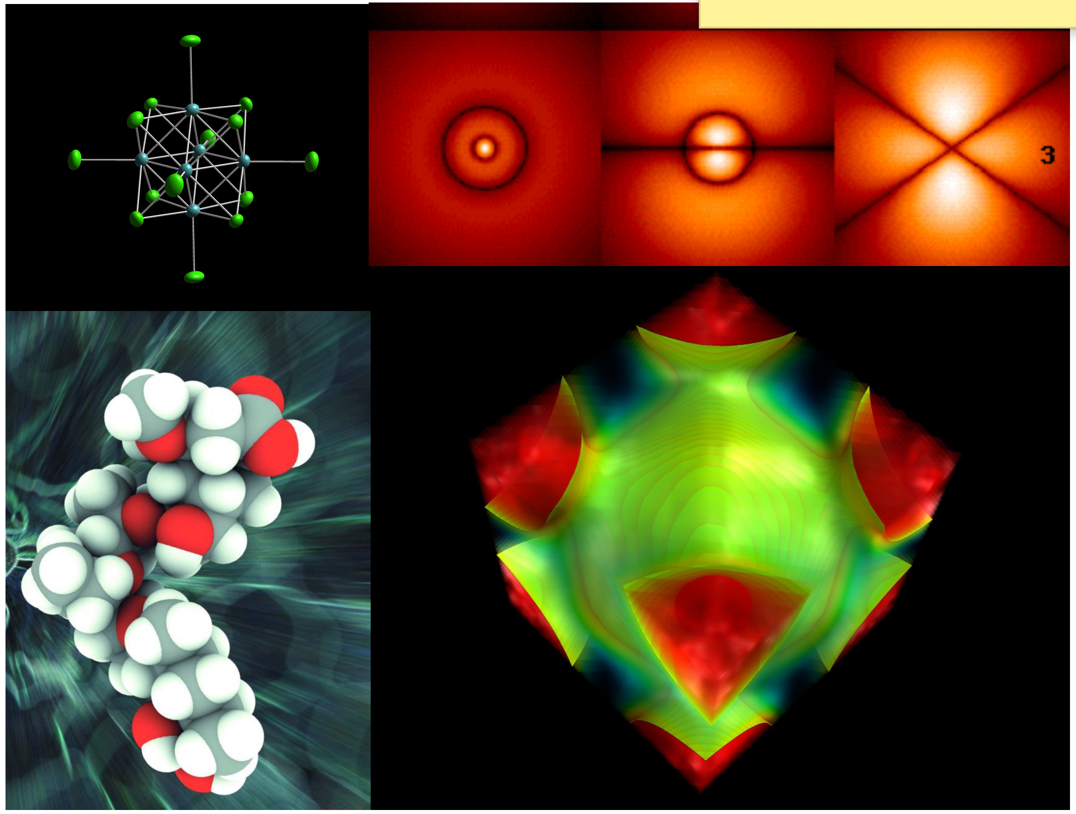
It is called “Ab init”, because every time when you start a new calculation, you need to do all the initial settings again.

Setting parameter is an art more than science, relies heavily on experience. Therefore, beginner may be “too young, too simple, sometime naive”

acters of traditiona

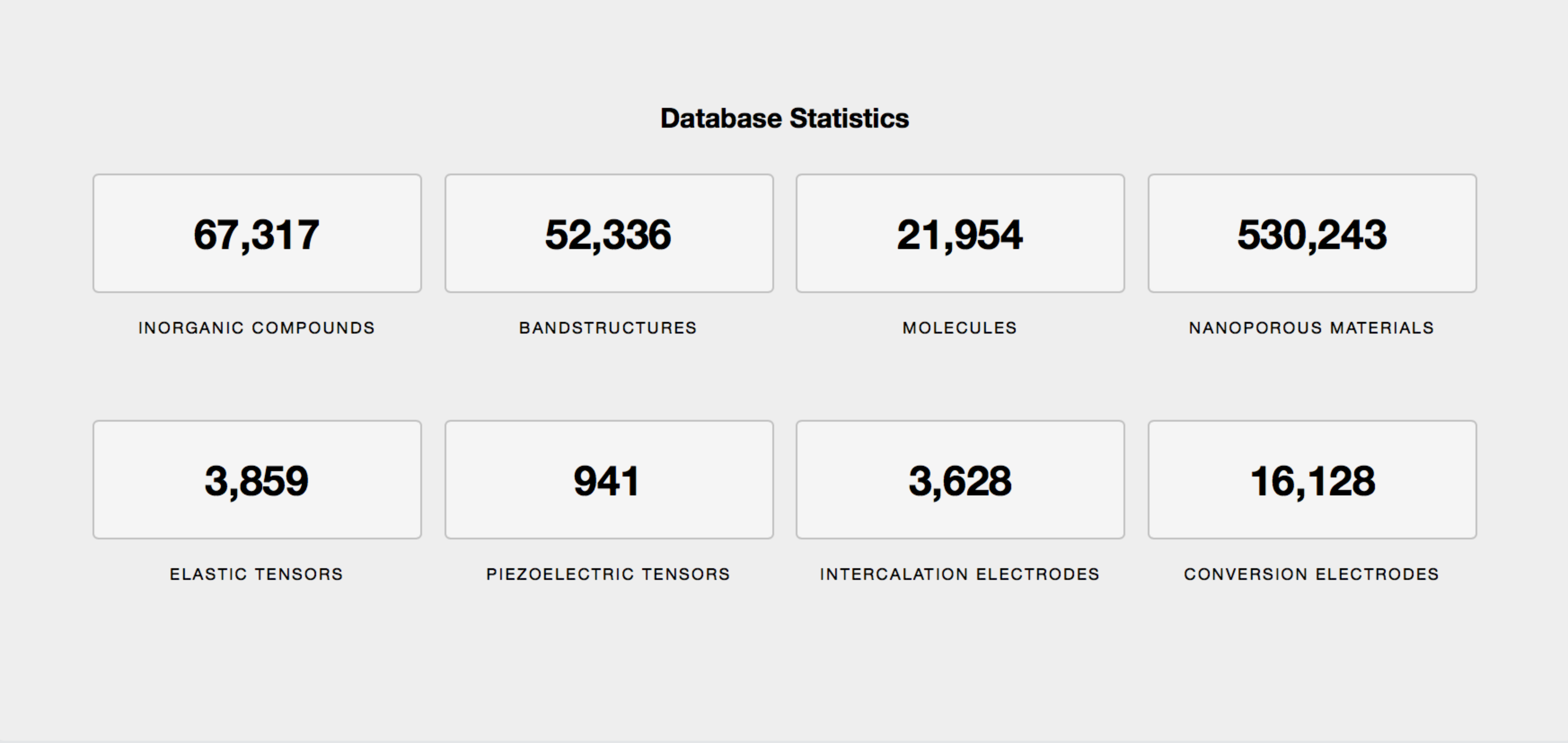
1. DFT calculations cover from single atom all way up to solid
2. It is called first principle because it is based on the quantum mechanic principle with nearly no or fewer empirical factors
3. However, due to the limited computing power, DFT can’t work efficiently when system contains more than ? of atoms
4. Doing DFT calculation is really time consuming and therefore is not applicable to fields that need instant results.

- Generally applicable
- No or few parameters
- Limited system size
- Time consuming



Why machine learning ?

Because ……



Database Statistics

67,317	**52,336**	**21,954**	**530,243**
INORGANIC COMPOUNDS	BANDSTRUCTURES	MOLECULES	NANOPOROUS MATERIALS
3,859	**941**	**3,628**	**16,128**
ELASTIC TENSORS	PIEZOELECTRIC TENSORS	INTERCALATION ELECTRODES	CONVERSION ELECTRODES

https://materialsproject.org

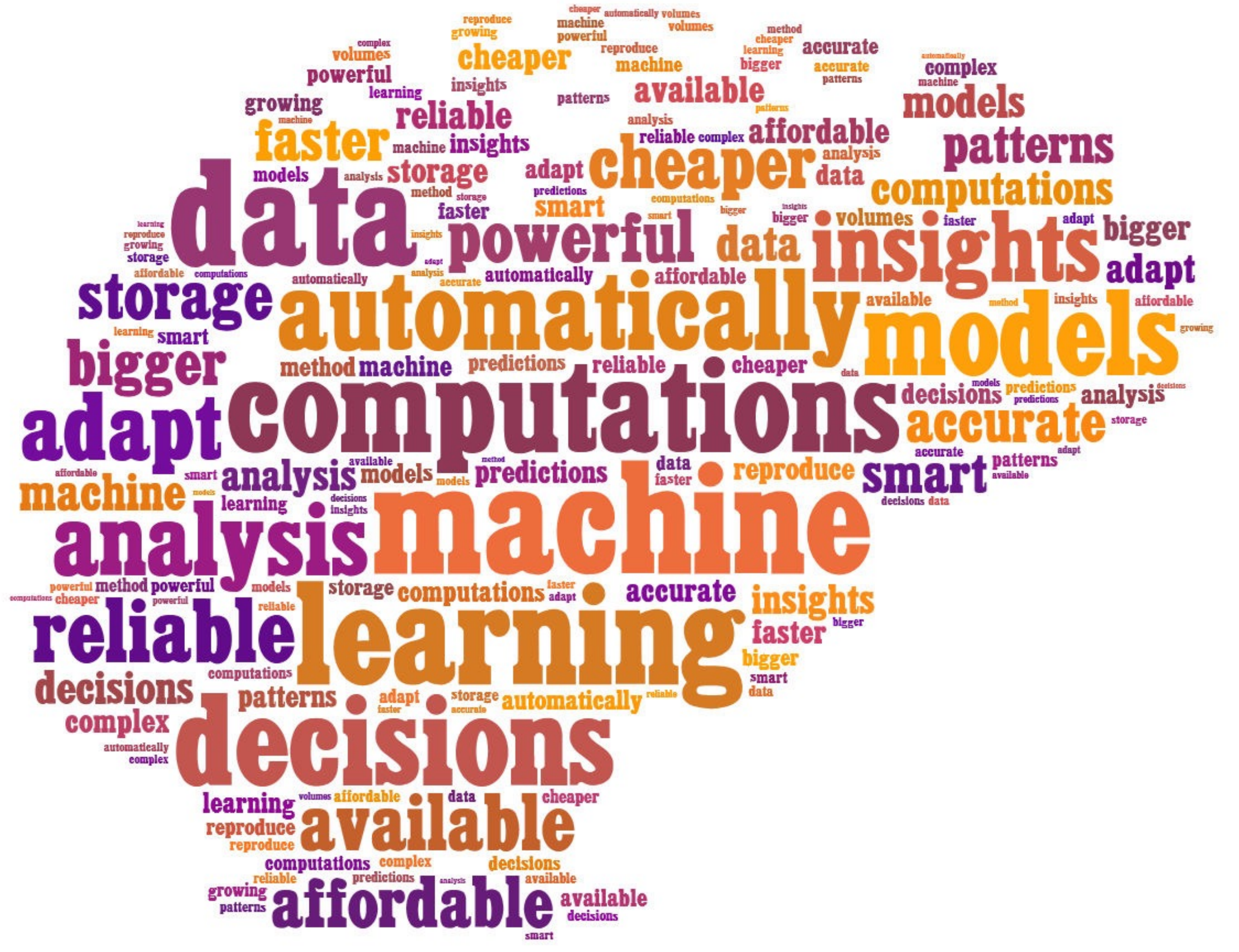
Power of machine lea

Since the first demonstration of the DFT method, it has be used in fields involve chemical compounds. Therefore, accumulate large amount of data, e.g. Material Project

We want our DFT calculation to be more smart, and more adaptive to various of conditions.

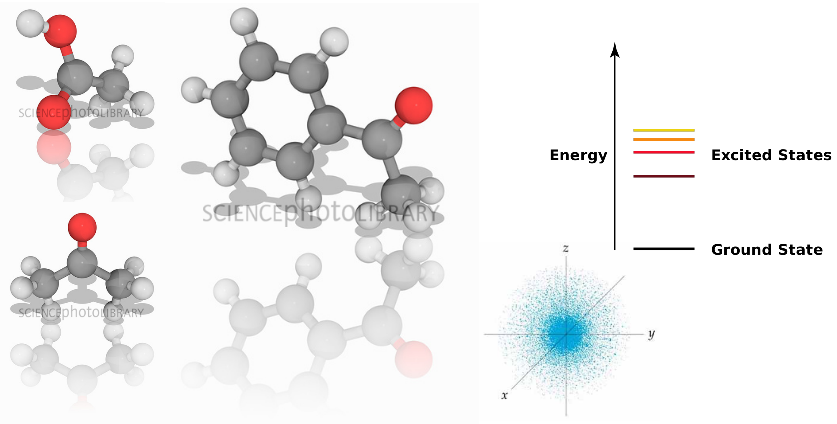
The prediction is able to be made within O(N) time.

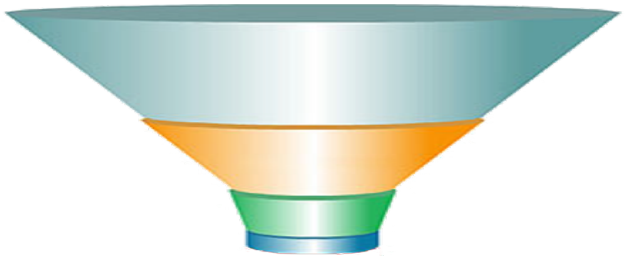
- **Experienced** learn from the existing knowledge
- **Accurate** training and testing scheme
- **Efficient** predict from the model learned



In the following content, machine learning is regarded as an efficient tool, physics ends in the DFT part.

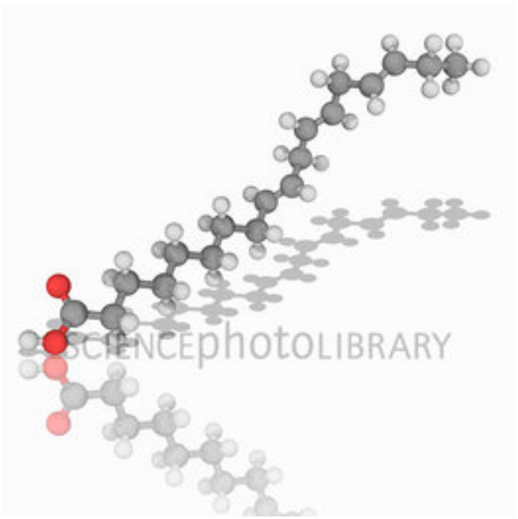
Physics+Technology=?



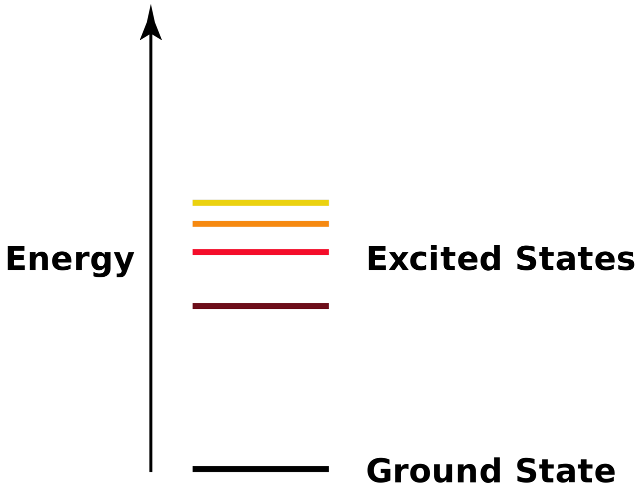


database construction:
(maybe update)
1. wave function orientation
2.

New molecule

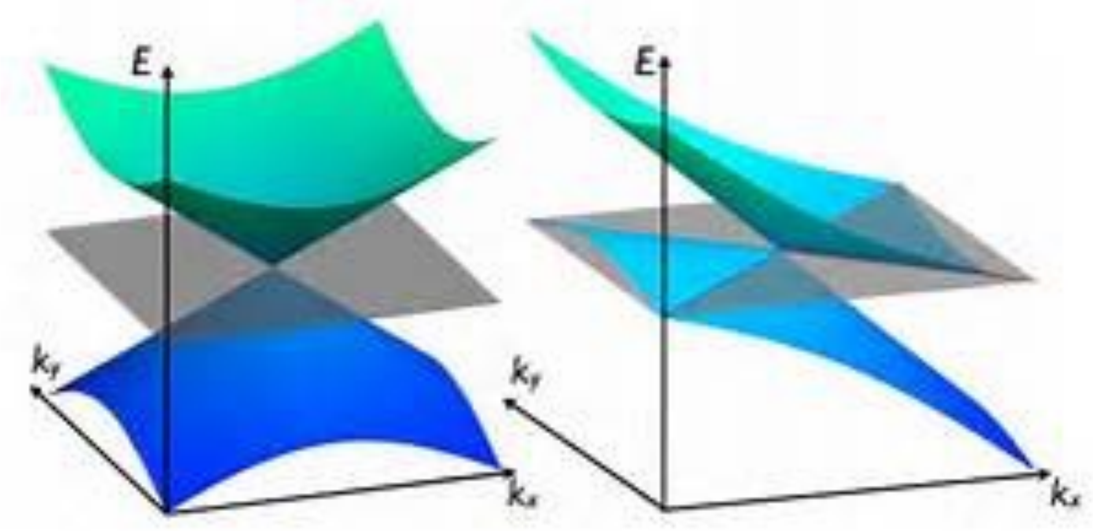


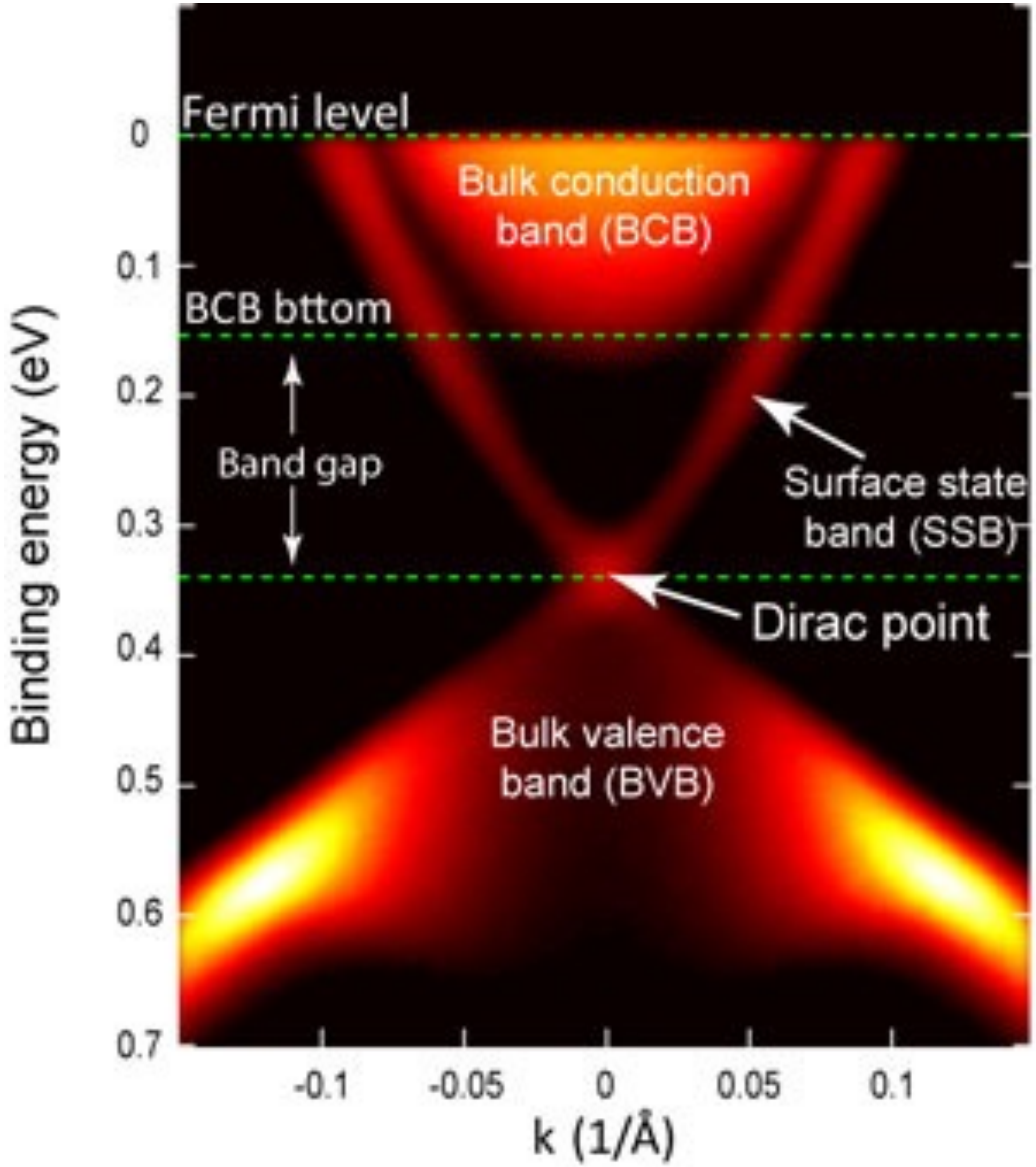
Properties

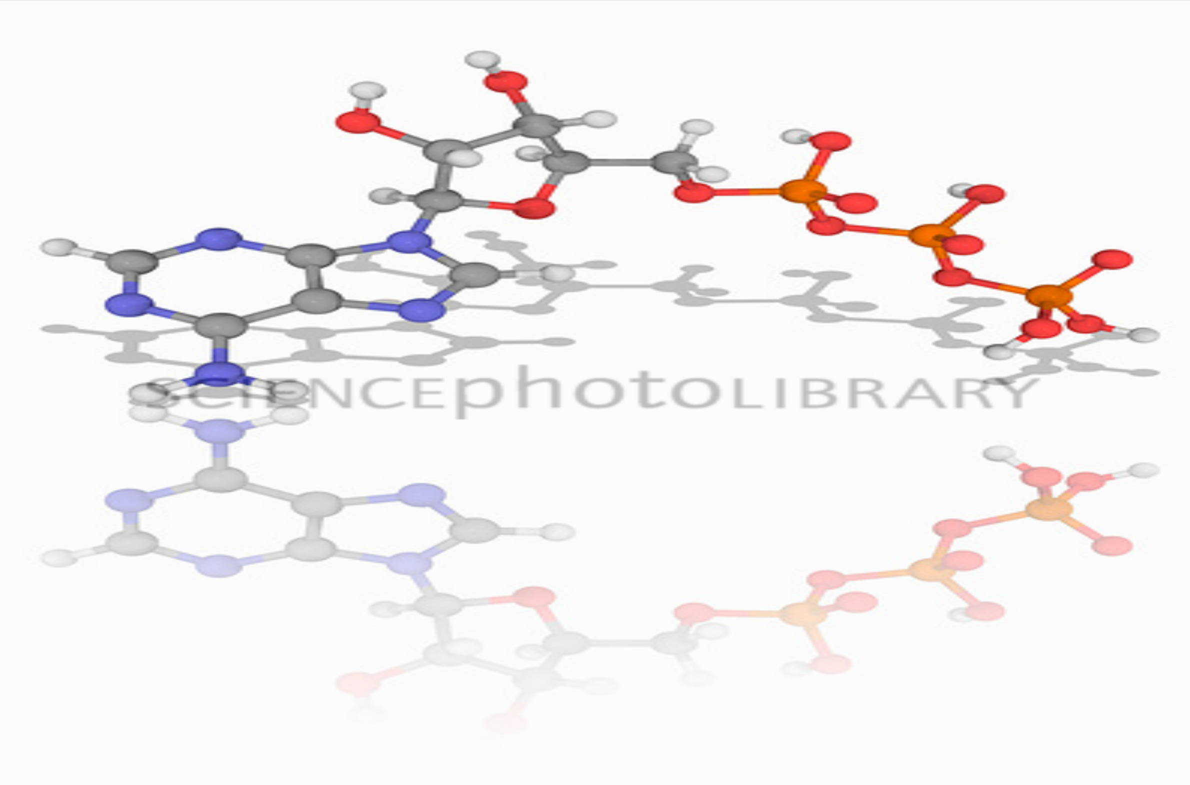


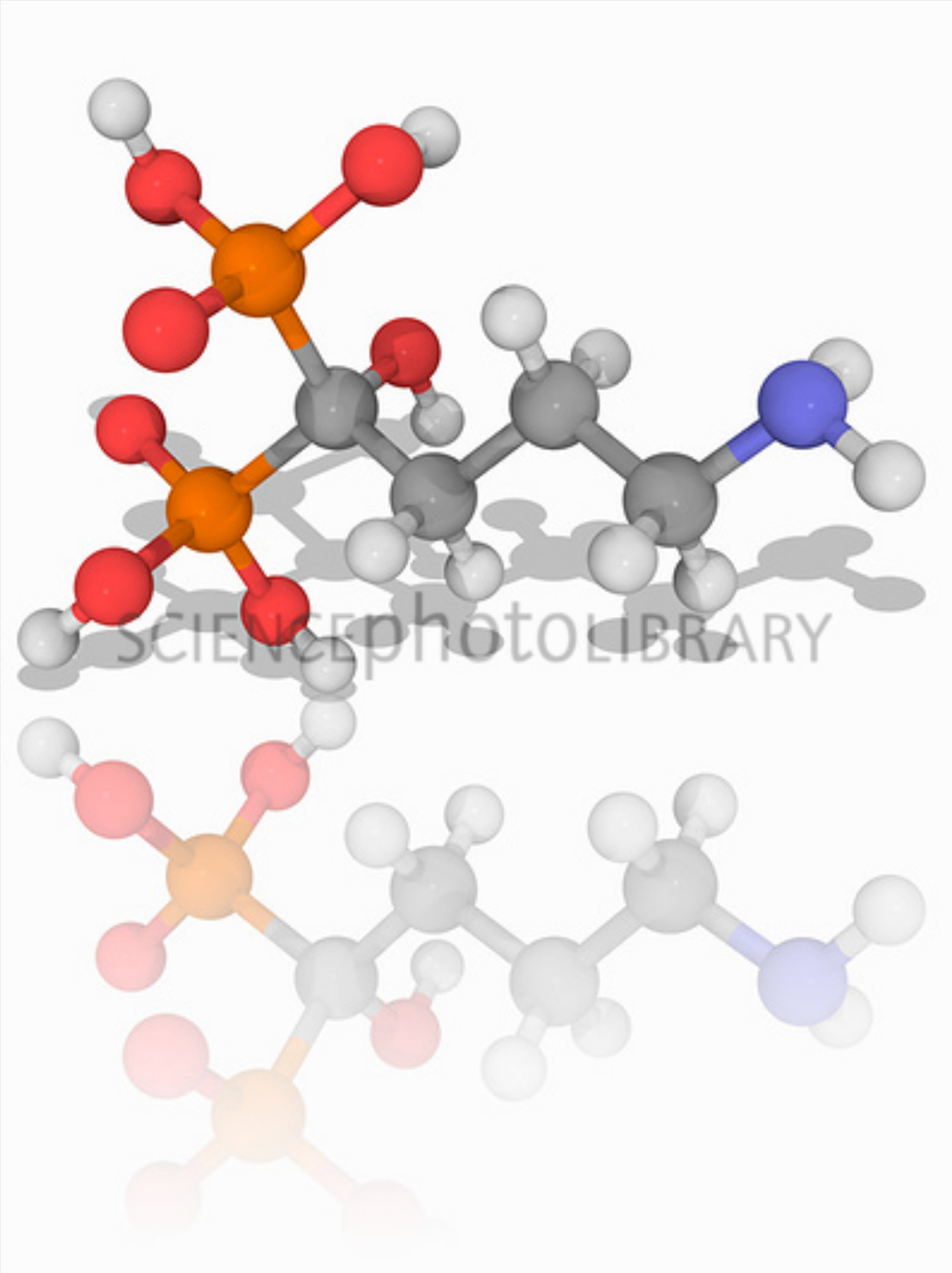
physicsworld

TOP 10 BREAKTHROUGH 2015









Simple ORGANIC COMPOUNDS:

There structure and ground state energy

Database

Regression

Prediction

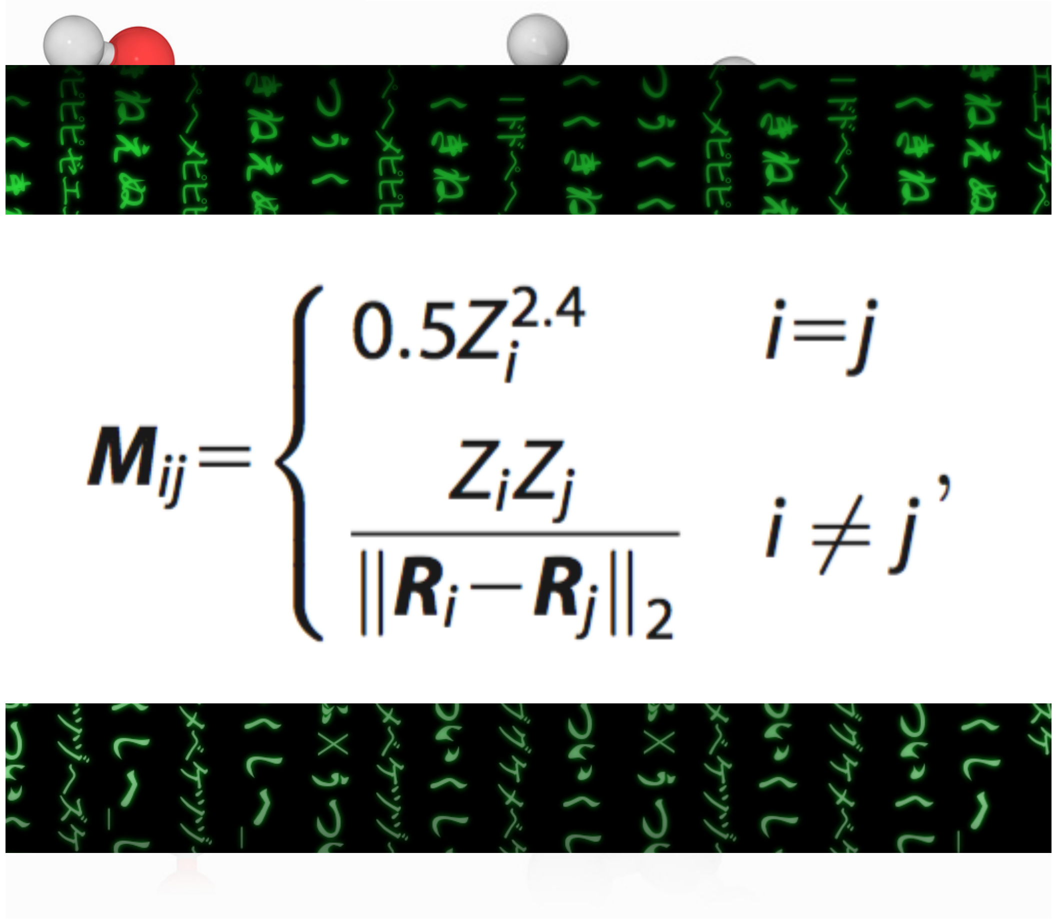
Database

extract **GEOMETRIC information**

动画效果
tell machine learning it’s a 3D system
add on lattice

+features: Coulomb matrix
+targets: ground state energy

- **Basic Feature** atoms’ position & charge
- **Interaction** Coulomb interaction
- **Other** symmetry



$$\mathbf{M}_{ij} = \begin{cases} 0.5 Z_i^{2.4} & i = j \\ \dfrac{Z_i Z_j}{\|\mathbf{R}_i - \mathbf{R}_j\|_2} & i \neq j \end{cases},$$

M. Rupp. Int. J. Quantum Chem. 2015, 115, 1058–1073. DOI: 10.1002/qua.24954

Kernel: ?
1. kernel theory
2. functional form
3. linear

Regression

build **GEOMETRY-ENERGY map**

how to validate it's a right kernel

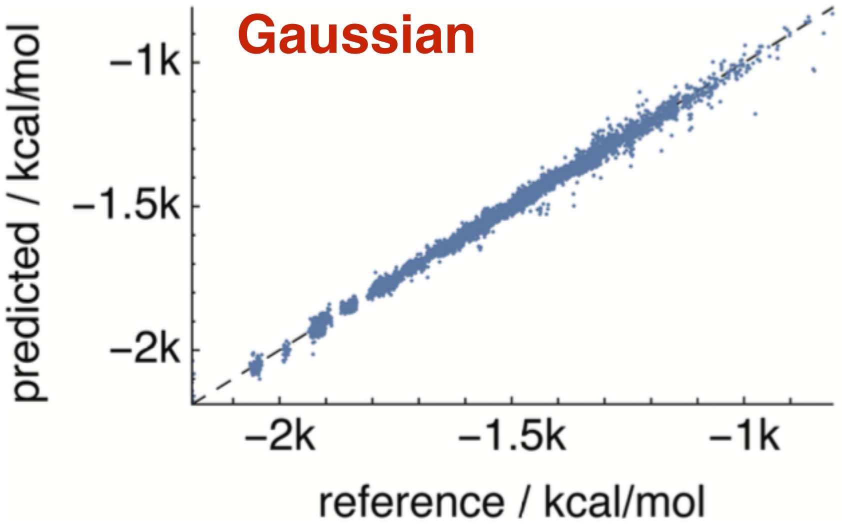
the relationship between geometry and energy is complex, maybe non-linear, therefore, we use KRR as a tool to build our map

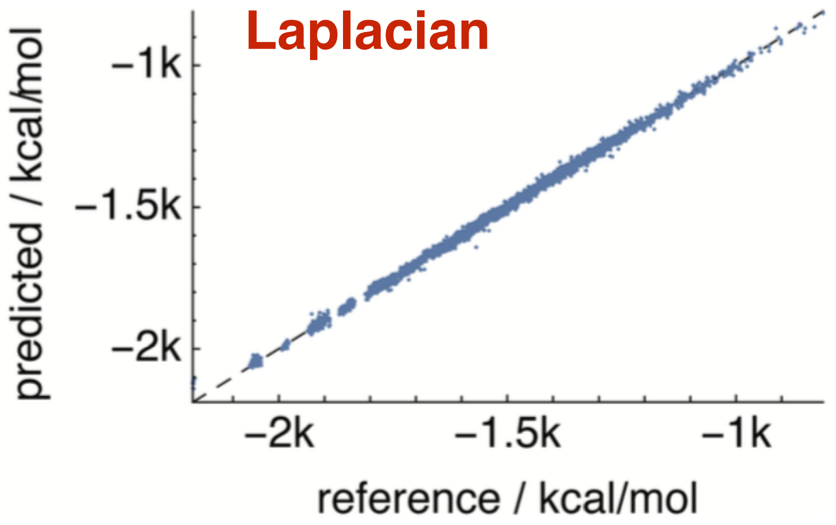
acceptance ratio ?

time complexity ?

- **Non-linear** kernel trick (Gaussian kernel, Laplacian kernel ……)
- **K**ernel **R**idge **R**egression

M. Rupp. Int. J. Quantum Chem. 2015, 115, 1058–1073. DOI: 10.1002/qua.24954





Limitation & Outlook

- Limits:
    - the training data (DFT generated)
    - interpolation
- Future work:
    - universal functional (electron only term)
    - connections among different materials (kernel)

instead of considering a single microscopic system, we should pay more attention to multiple such system and find their connections

rethink the red part

Summary

- DFT: generally implementable, time consuming
- ML: experienced, accurate, efficient
- Able to do large scale computation more precisely

Reference

- S. Cottenier, Density Functional Theory and the family of (L)APW-methods: a step-by-step introduction, 2002-2013 (2nd edition), ISBN 978-90-807215-1-7 (freely available at http://www.wien2k.at/reg user/textbooks).
- M. Rupp. Int. J. Quantum Chem. 2015, 115, 1058–1073. DOI: 10.1002/qua.24954
- Rasmussen, Williams: Gaussian Processes for Machine Learning, MIT Press, 2006.
- Justin Domic: Kernel method and SVM, Lecture notes.
- The Top 100 papers,Nature vol.514,Oct.30,2014

Thank you for your attention