Osaka University/NANOPROGRAM
June 24, 2024

Introduction to Materials Informatics

Tamio Oguchi
Center for Spintronics Research Network
Osaka University

アウトライン

- 科学のパラダイム
- 計算科学 → データ科学
- 物質・材料の研究開発

背景・経緯

- 物質・材料分野におけるデータ
- 仮想スクリーニング
- 機械学習
- 記述子

基礎事項

- 物質探索
- 構造探索
- 構造のマップ

研究事例

- まとめ

Paradigm of Science

共有される思考法

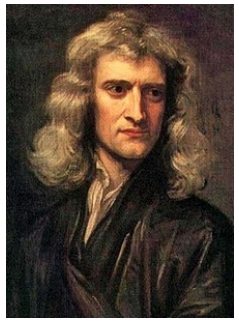
Inductive 帰納的 ***Deductive*** 演繹的

Experiment **Theory**

Johannes Kepler
1571-1630

Tycho Brahe
1546-1601

Kepler's laws of planetary motion



Isaac Newton
1642-1727

Galileo Galilei
1564-1642
Telescope [1609]

Newton's law of universal gravitation and law of motion

Observations → ***Laws***
→ ***Principles***
→ ***Applications***

https://en.wikipedia.org/wiki/

The 5th Solvay Conference 1927.10
“Electrons and photons”

https://en.wikipedia.org/wiki/Solvay_Conference

"Most of Physics and

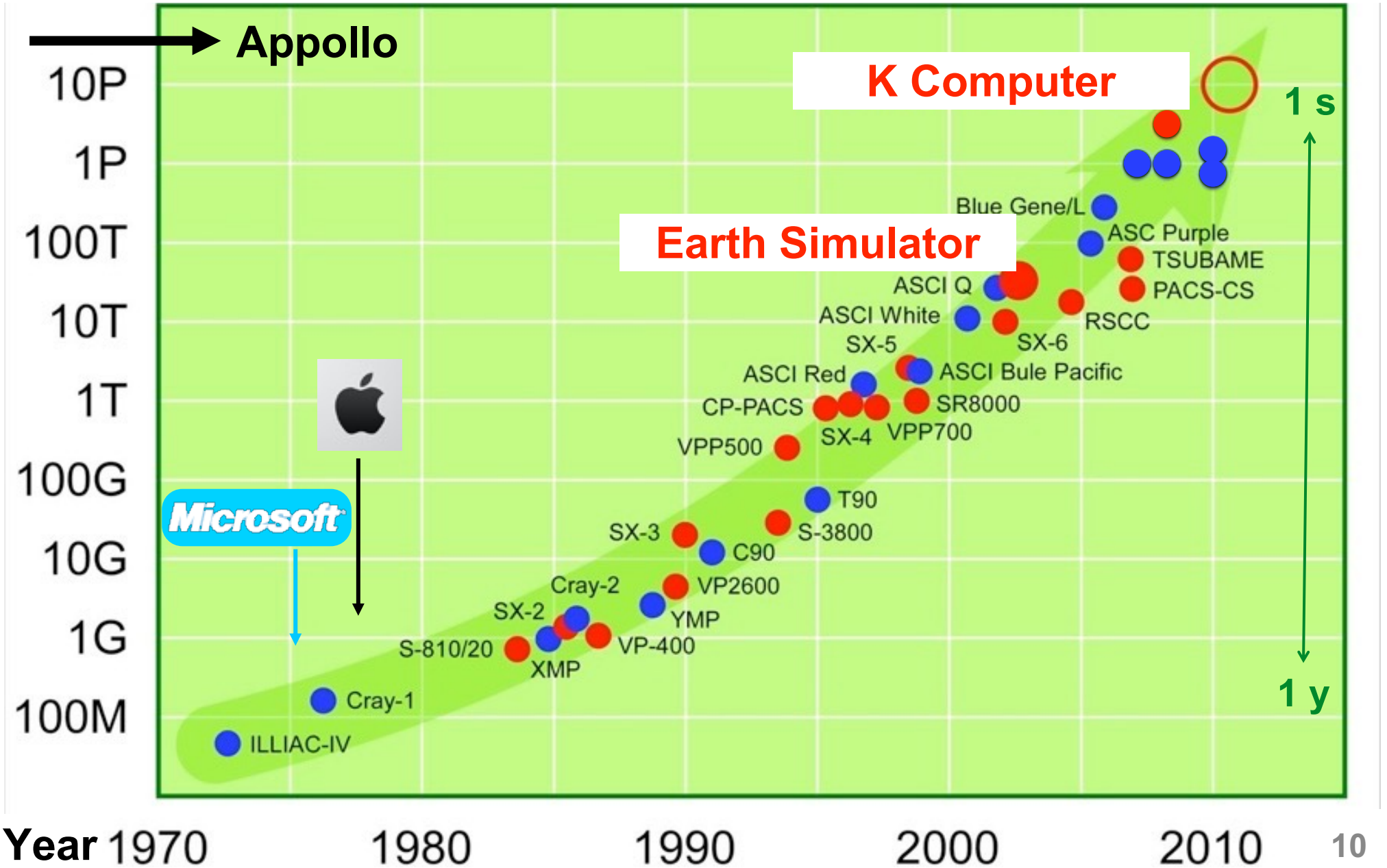
all of Chemistry are solved."

Paul Dirac 1902-1984

"You can not understand it,
until you know how to calculate it."
J. C. Slater 1900-1976

Growing Power of Supercomputer

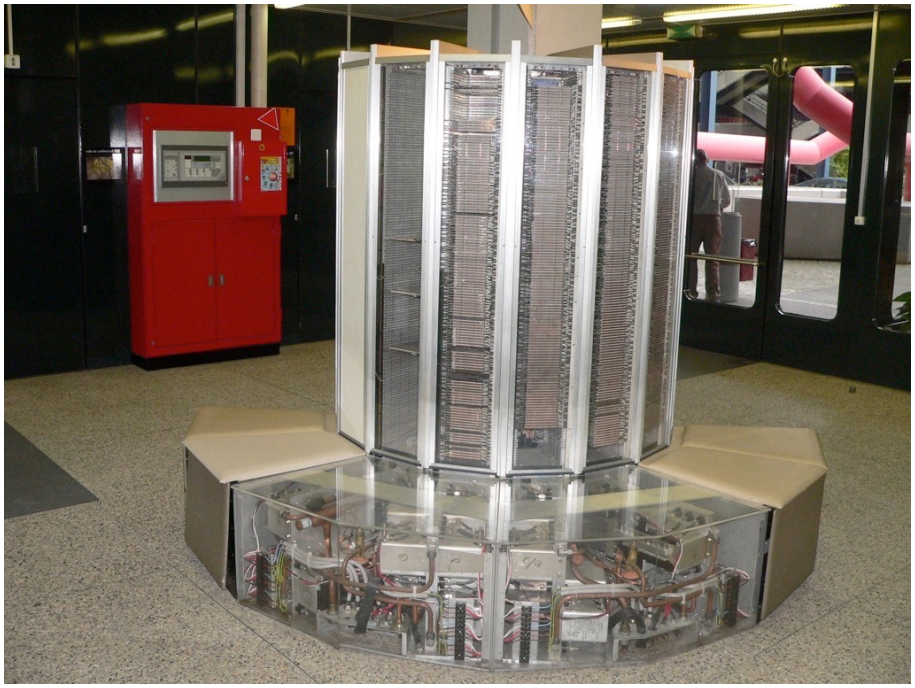
2021
Fugaku



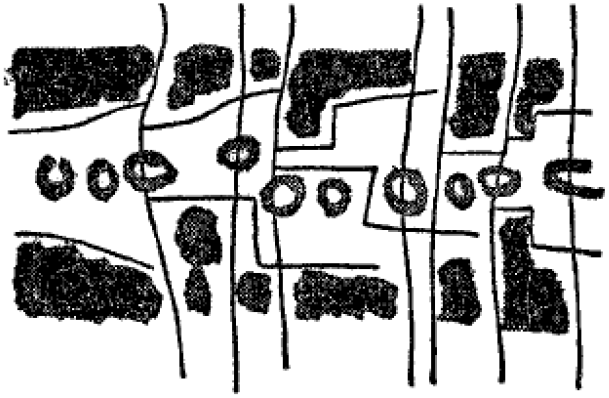
Cray-1

- First Supercomputer in the World
  - Installed in Los Alamos National Lab in 1976
  - Peak Performance:100MFLOPS
  - Main Memory: 8MB

Seymour Cray



Los Alamos National Laboratory



解　　説

第三の物理学としての計算物理学†

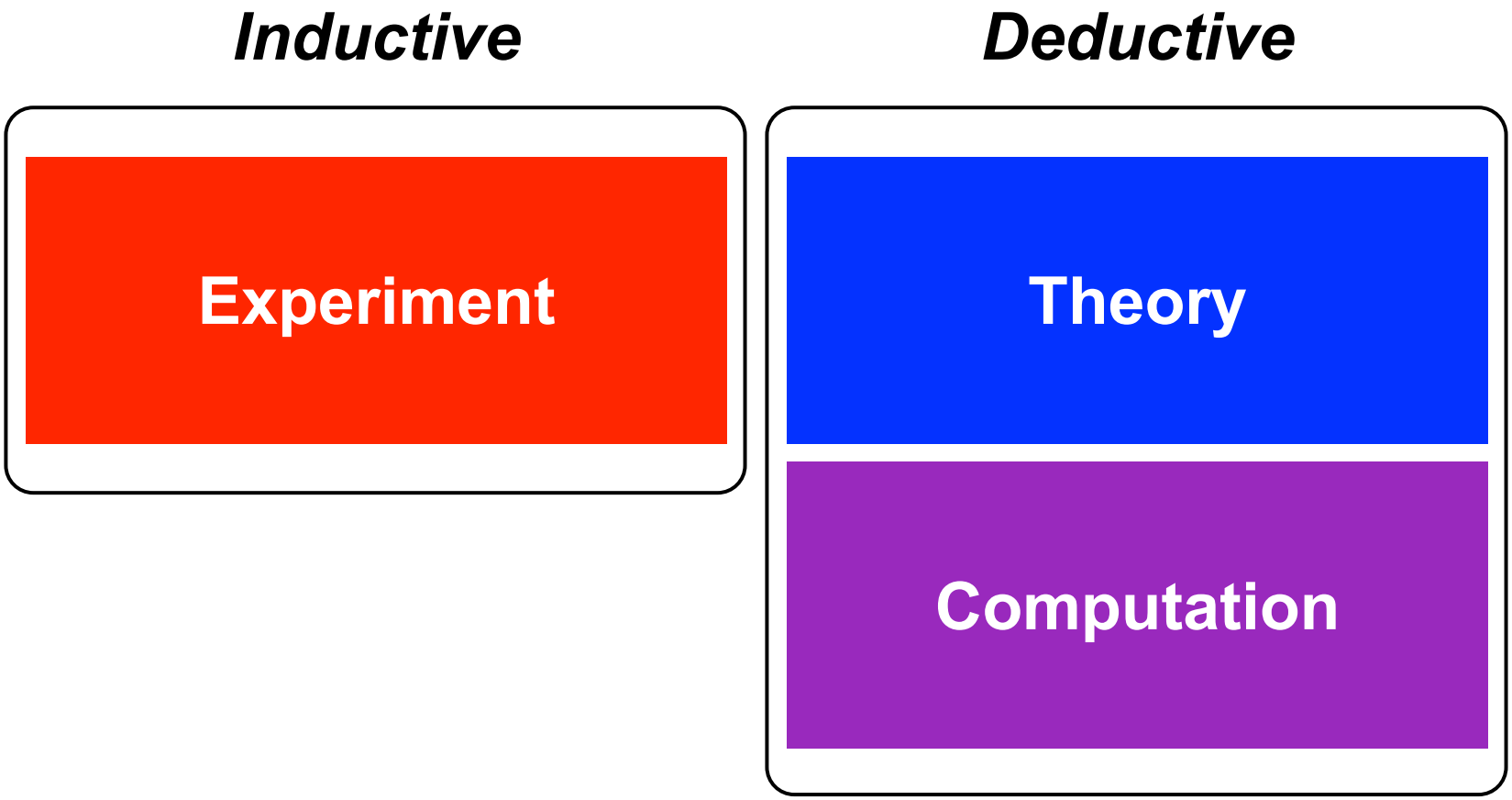
A. J. Freeman

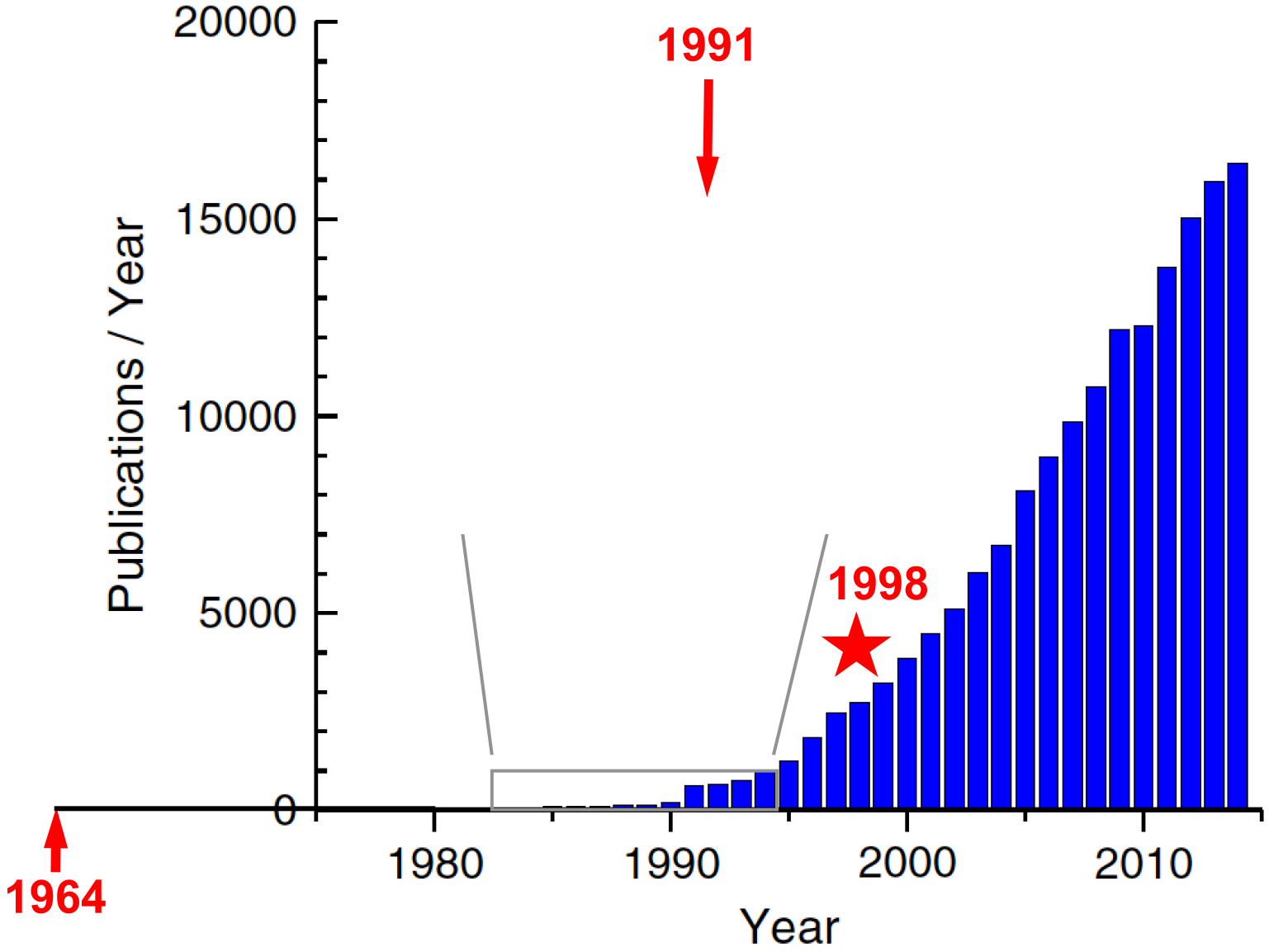
訳　大　西　楢　平*

本稿は，著者が山田財団の援助により2カ月間日本に滞在した機会に行った第37回日本物理学会年会(1982年10月)の特別講演から，計算物理学に関する一般的考察の部分を新たに展開，加筆したものである．米国における計算物理学の進展の現状と物理学の諸分野とのかかわりあいを多くの例をあげて紹介している．複雑な系をとり扱う物理学としての計算物理学，その理論科学的，実験科学的，及び応用科学的側面，また計算物理学者の特性などが議論されている．

日本物理学会誌 1983

Paradigm of Science



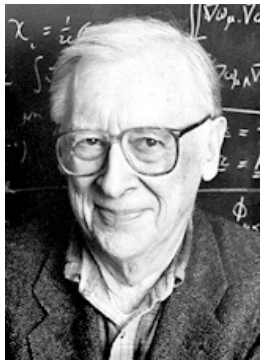


Number of publications per year (1975–2014) on topics ("density functional" or "DFT"), according to the Web of Science Core Collection (February 2015)

Jones (2015)

The Nobel Prize in Chemistry 1998

Walter Kohn



John A. Pople

The Nobel Prize in Chemistry 1998 was divided equally between Walter Kohn "for his development of the density-functional theory" and John A. Pople "for his development of computational methods in quantum chemistry".

Current Issues in Materials Science and Engineering

- **Complexity in Materials**
  - **Multi-scale, multi-physics, hierarchal structures in time and space, leading to the importance of top-down and bottom-up connections in many contexts**
  - **Strong requirements not only to get high performance for cost and efficiency with sustainability for environment and energy problems, and but also to accelerate R&D**

✓ **We are confronted with limitations in the accuracy and efficiency of traditional research strategies such as experimental, theoretical, and computational approaches.**

✓ **Importance of "Data-Intensive Scientific Discovery" as the fourth paradigm of science has been rapidly rising quite recently in materials science & engineering, as called**

★ ***MATERIALS INFORMATICS***

SCIENCE

The New York Times

A Deluge of Data Shapes a New Era in Computing

DEC. 14, 2009

In a speech given just a few weeks before he was lost at sea off the California coast in January 2007, Jim Gray, a database software pioneer and a Microsoft researcher, sketched out an argument that computing was fundamentally transforming the practice of science.

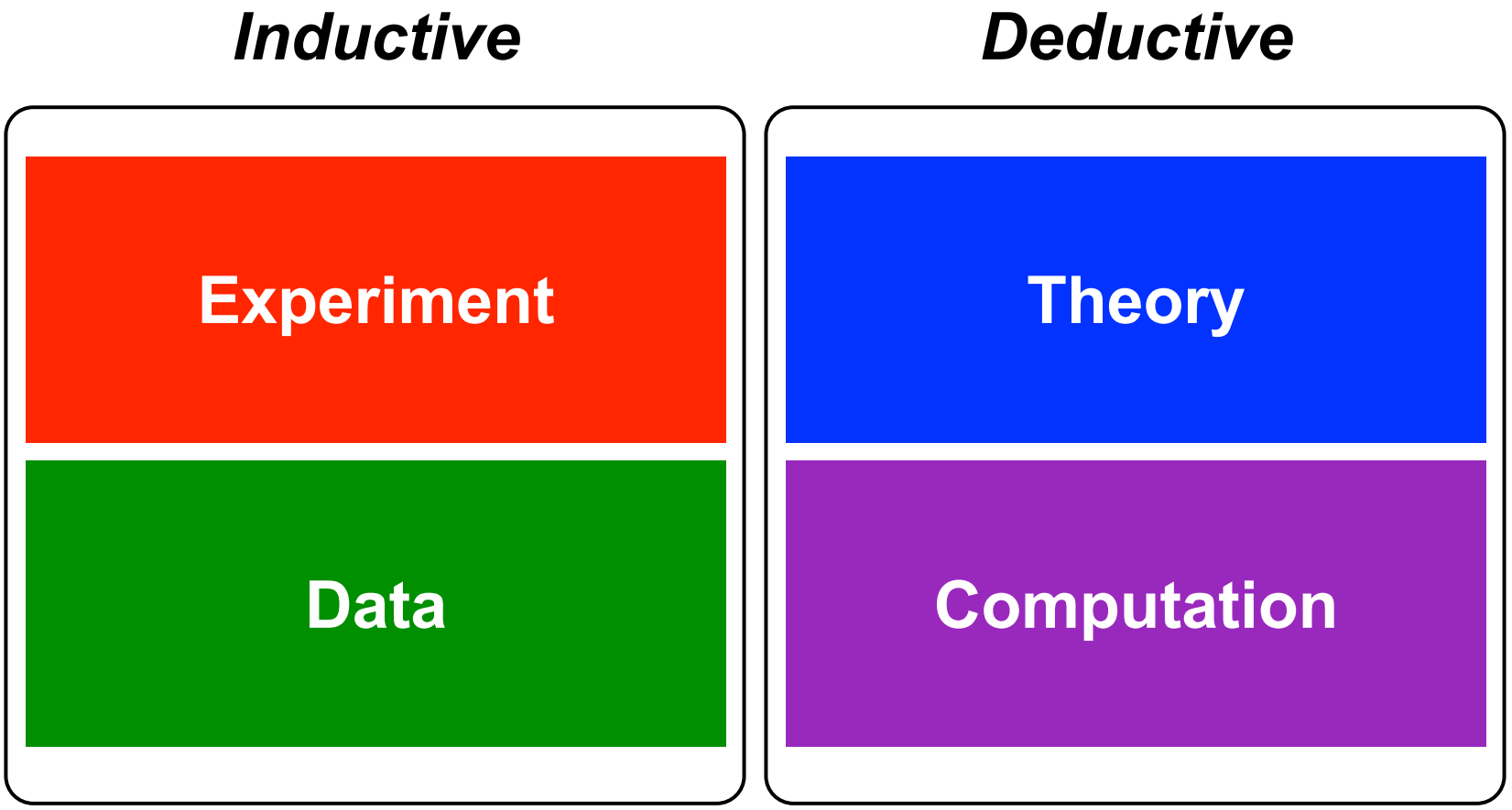
Dr. Gray called the shift a "fourth paradigm." The first three paradigms were experimental, theoretical and, more recently, computational science. He explained this paradigm as an evolving era in which an "exaflood" of observational data was threatening to overwhelm scientists. The only way to cope with it, he argued, was a new generation of scientific computing tools to manage, visualize and analyze the data flood.

.....

Now, as a testimony to his passion and vision, colleagues at Microsoft Research, the company's laboratory that is focused on science and computer science, have published a tribute to Dr. Gray's perspective in "The Fourth Paradigm: Data-Intensive Scientific Discovery." It is a collection of essays written by Microsoft's scientists and outside scientists, some of whose research is being financed by the software publisher.

.....

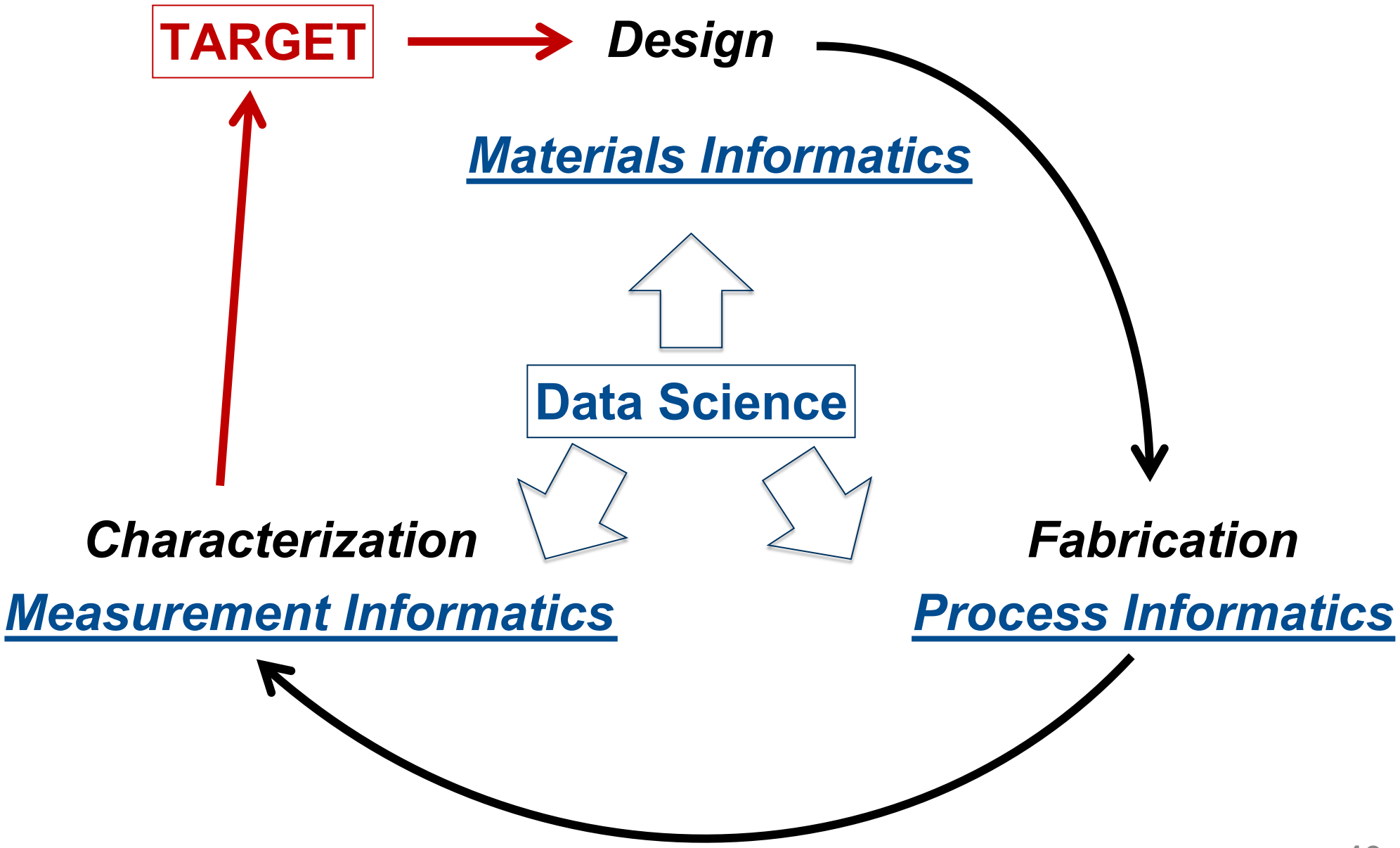
Paradigm of Science



"The Fourth Paradigm: Data-Intensive Scientific Discovery" in Materials Science

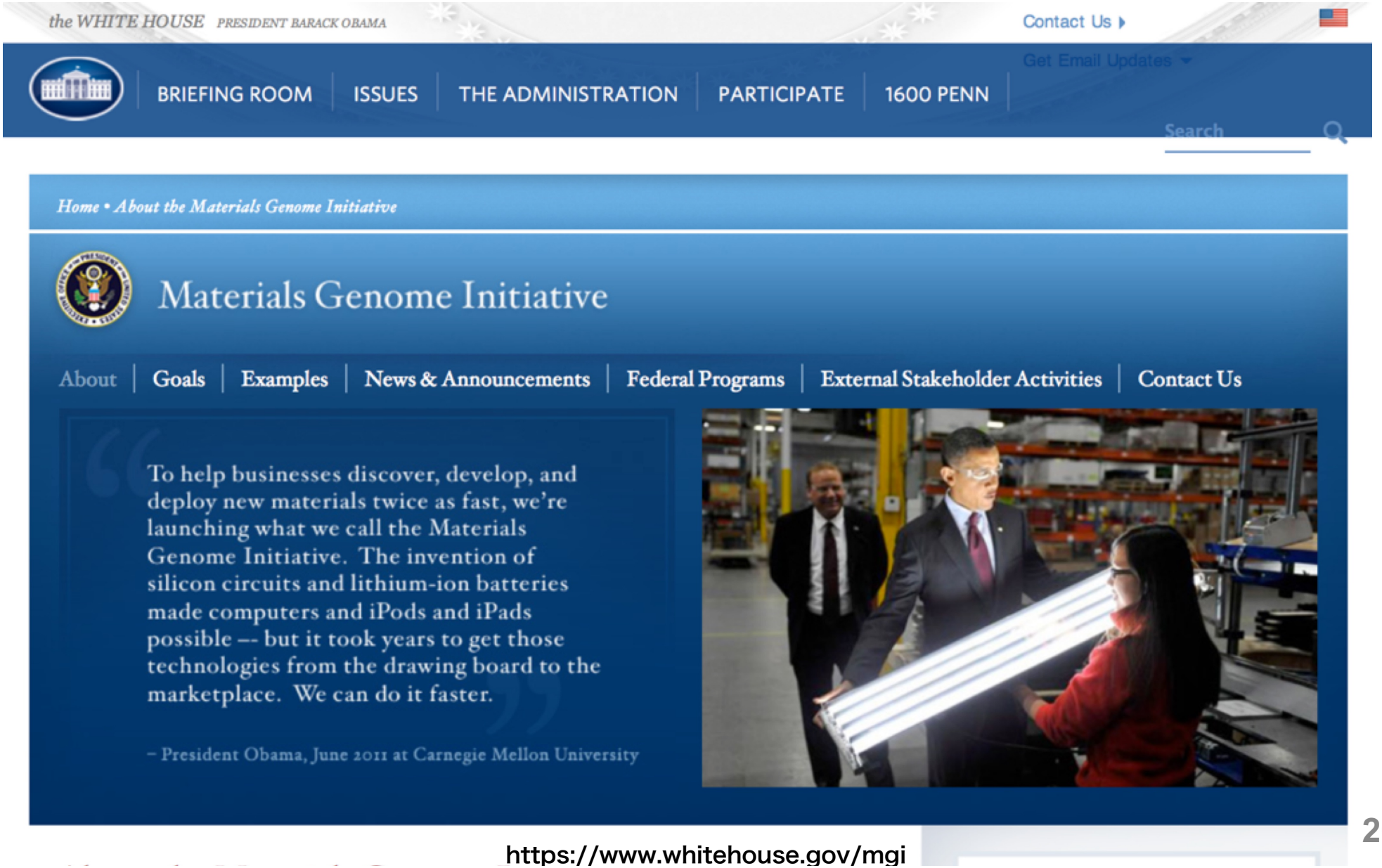
MATERIALS INFORMATICS

Materials Science & Engineering



Materials Genome Initiative

https://www.mgi.gov



https://www.whitehouse.gov/mgi

Materials Informatics

What can we provide ?

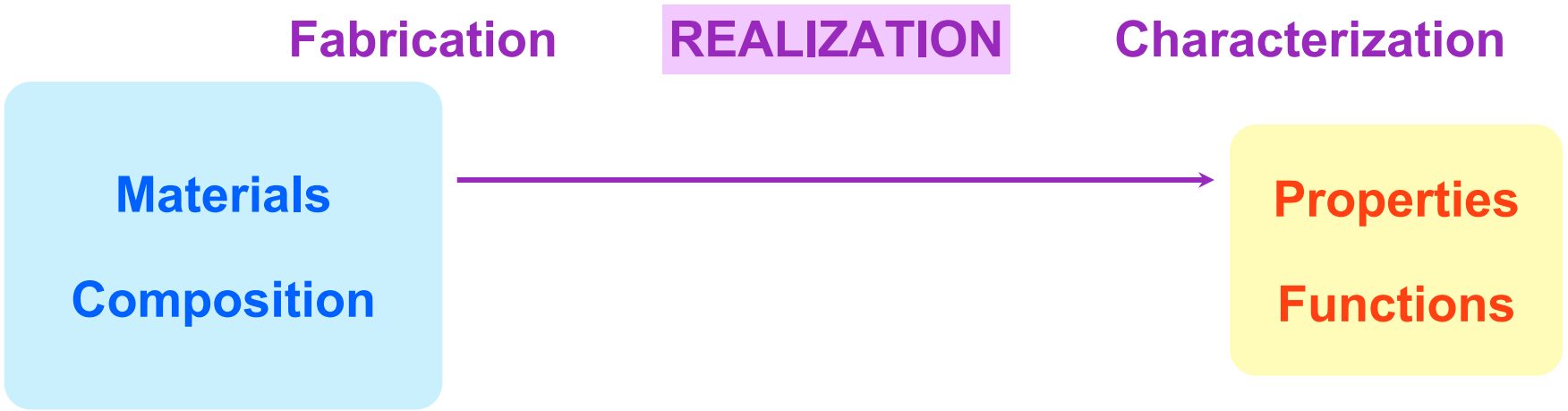
Materials Science

Data Science

Guiding Principles

What can we get ?

Materials Research



x_i P_i

(x_6,P_6) (x_1,P_1) (x_2,P_2) (x_3,P_3)

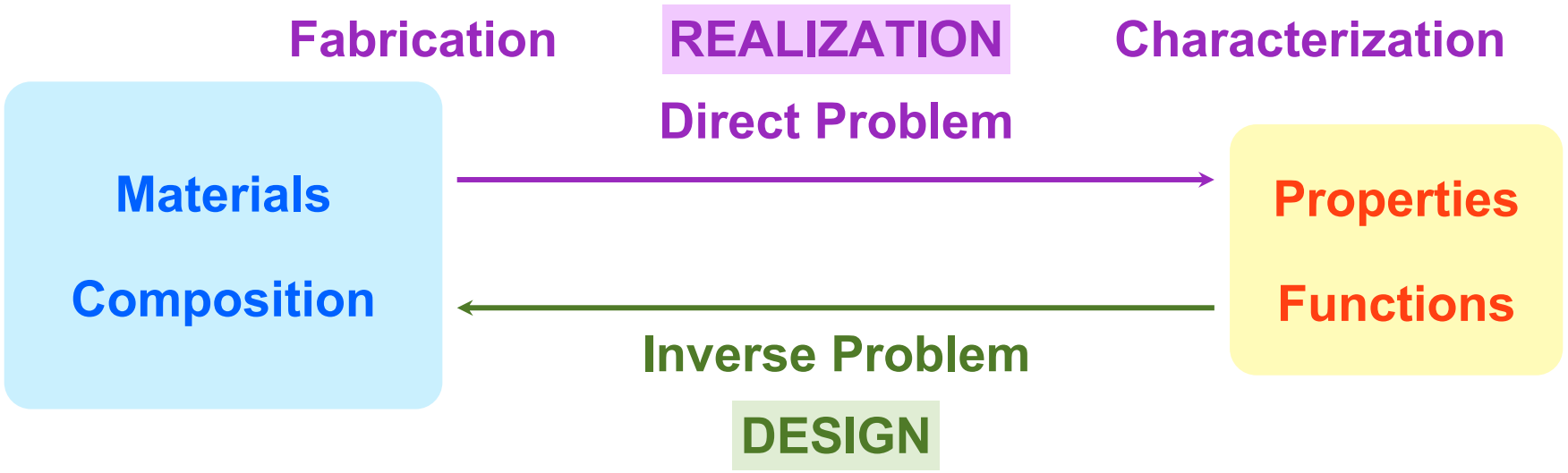
(x_{10},P_{10}) (x_5,P_5)

(x_7,P_7) (x_4,P_4)

(x_9,P_9) (x_{11},P_{11}) (x_{12},P_{12})

(x_8,P_8)

Materials Research

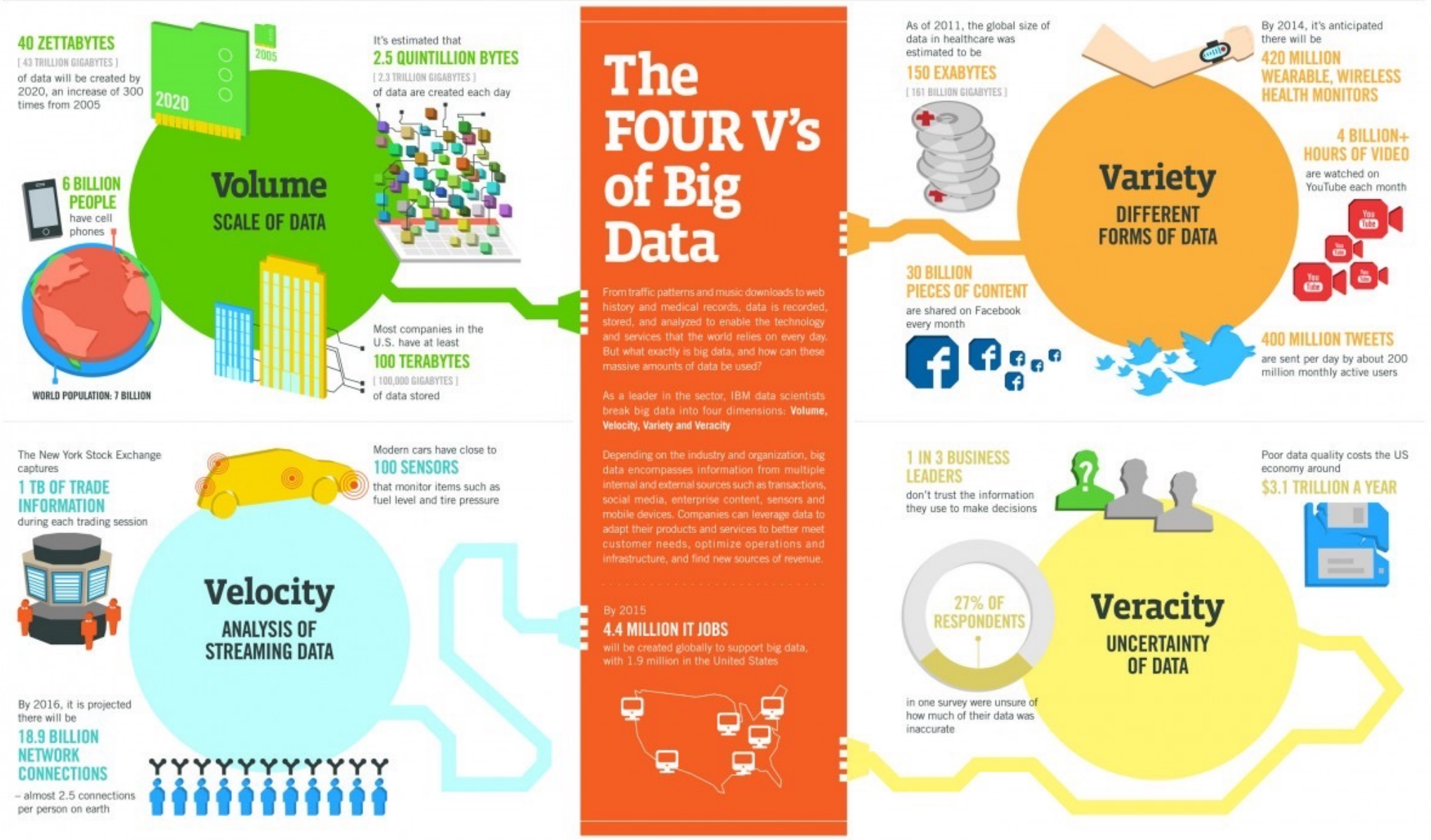


Data Science

$\{x_i, P_i\} \longrightarrow P(x) \longrightarrow x = P^{-1}(P)$

Discovery of Correlation

Big Data



http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Materials Database

- **MatNavi** http://mits.nims.go.jp
  - **AtomWork: inorganic material DB** ***Free***
    - **Structure: 82,000, Property: 55,000, Phase diagram: 15,000**
  - **…**

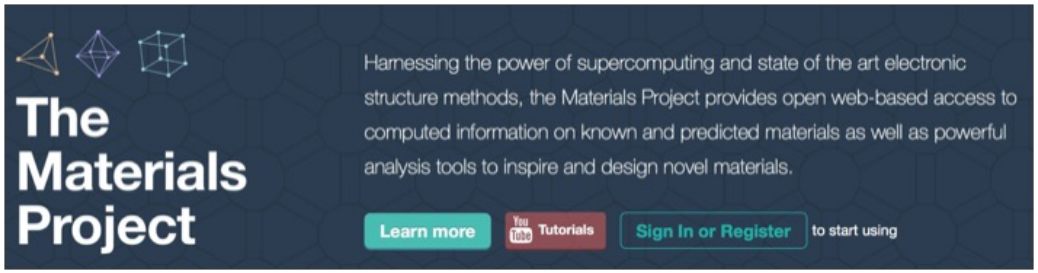
★ **AtomWork-Adv** ***Fee-Based***

- **Structure: 273,830, Property: 298,021, Phase diagram: 40,301**

Limited number of data for particular purpose !

Computational Materials Design with Machine Learning

Computational Materials Database



https://materialsproject.org

OQMD:

The Open Quantum Materials Database

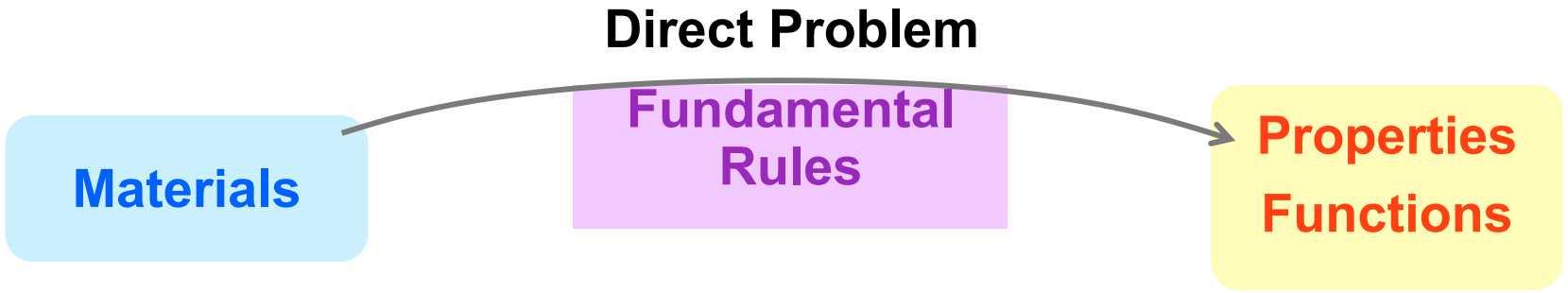
http://oqmd.org

https://nomad-coe.eu

http://aflowlib.org

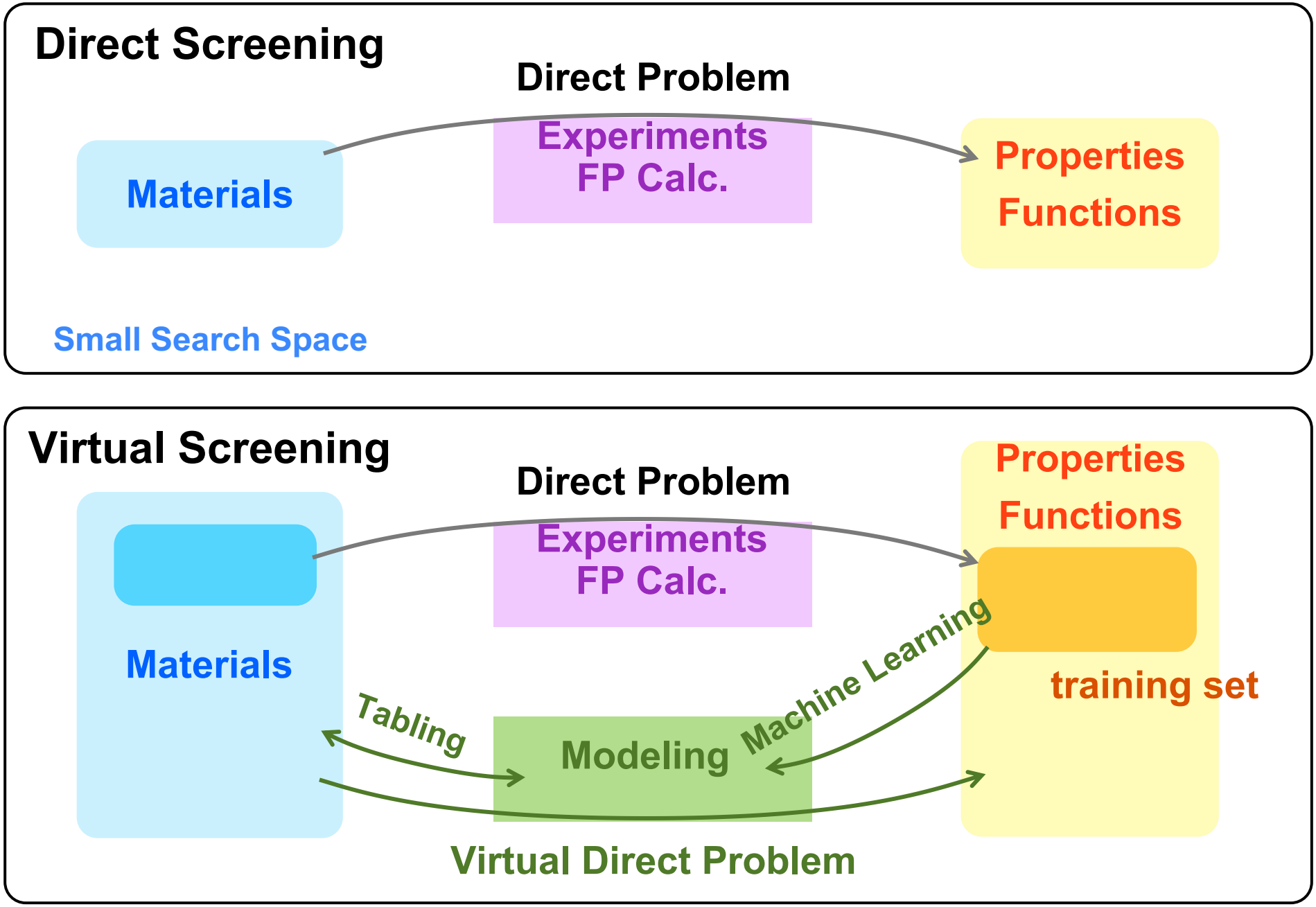
Role of First-Principles Calculation

- Most of material properties are governed by the electronic states described by the quantum theory.
- For concrete materials, many of properties can be predicted from first principles, offering another route to solve the direct problem.



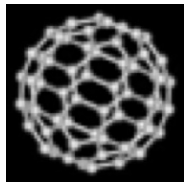
- Electron theory may provide new axes x (descriptors) that P spans.
- First-principles calculations can give some valuable pieces of information at less cost, even for non-existing and unstable materials.

Direct and Virtual Screening



Machine Learning

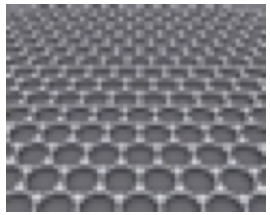
- **Modeling by Statistical Procedure and Evaluation**



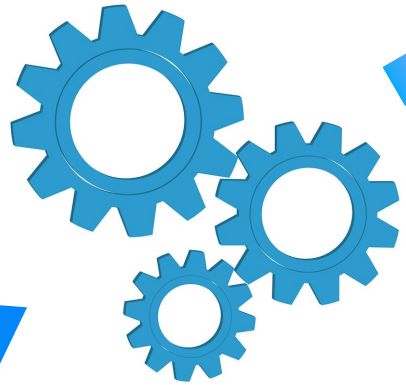
Prediction

Exploration

Extraction of Law



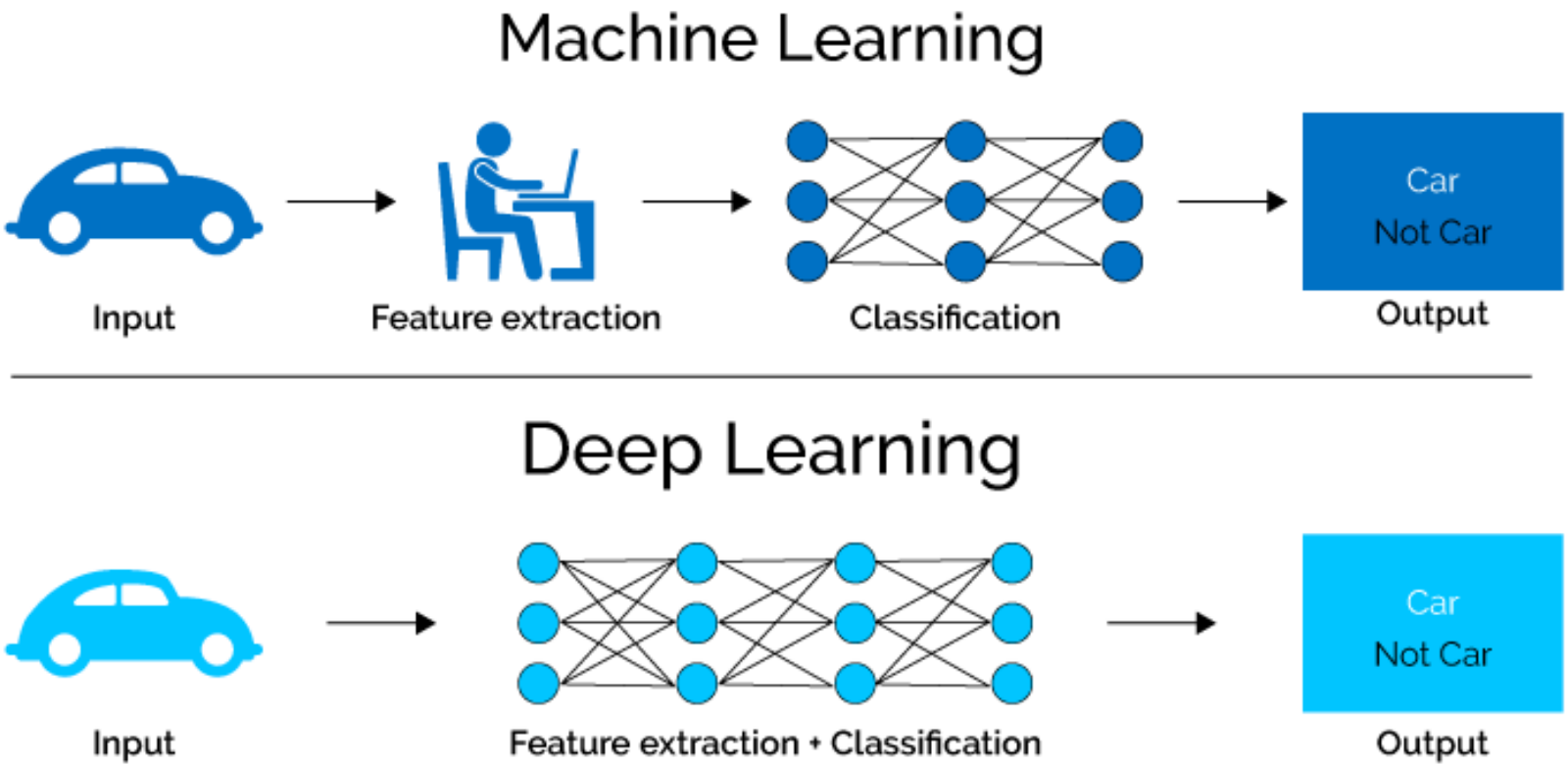
Modeling



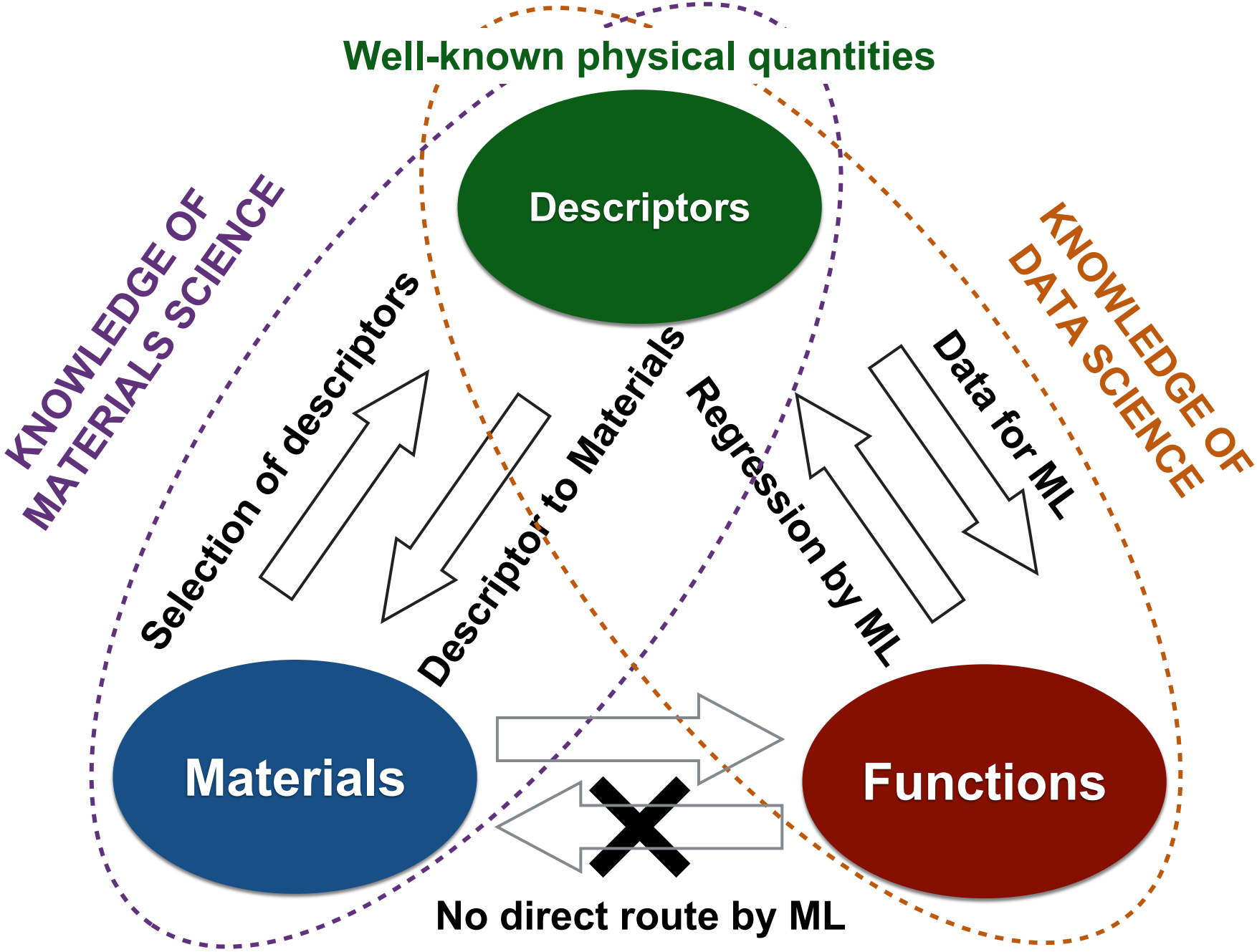
Database

Descriptors/Features

Machine Learning and Deep Learning

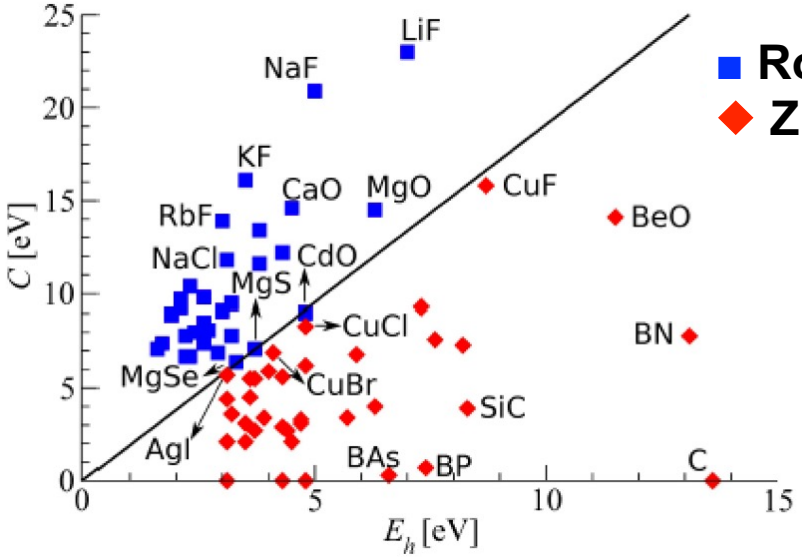


https://www.xenonstack.com/blog/log-analytics-with-deep-learning-and-machine-learning


Well-known physical quantities
Descriptors
KNOWLEDGE OF MATERIALS SCIENCE
KNOWLEDGE OF DATA SCIENCE
Selection of descriptors
Descriptor to Materials
Regression by ML
Data for ML
Materials
Functions
No direct route by ML

Descriptor for Structural Stability in Binary Compounds

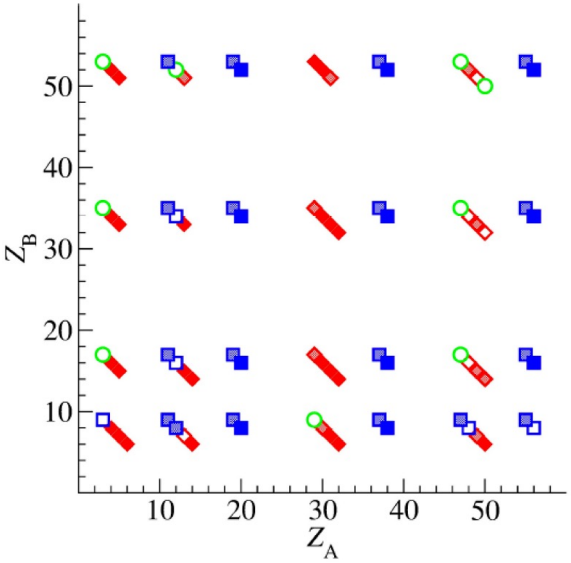
- **Empirical 2D descriptors by van Vechten and Phillips**



■ **Rock Salt**
◆ **Zinc Blende/Wurtzite**

J.A. van Vechten, PR (1969).
J.C. Phillips, RMP (1970).

- **Atomic number: not a good descriptor**



L.M. Ghiringhelli et al. PRL (2015).

Conditions of Good Descriptor

- **A descriptor d_i uniquely characterizes the material *i* as well as property-relevant elementary processes.**
- **Materials that are very different (similar) should be characterized by very different (similar) descriptor values.**
- **The determination of the descriptor must not involve calculations as intensive as those needed for the evaluation of the property to be predicted.**
- **The dimension of the descriptor should be as low as possible (for a certain accuracy request).**

L.M. Ghiringhelli et al. PRL (2015).

- **説明変数（explanatory variable）、特徴量（feature）**

Sparse Modeling by LASSO and Exhaustive Search

- **Best Descriptors**

highest occupied KS level of B

lowest unoccupied KS level of B

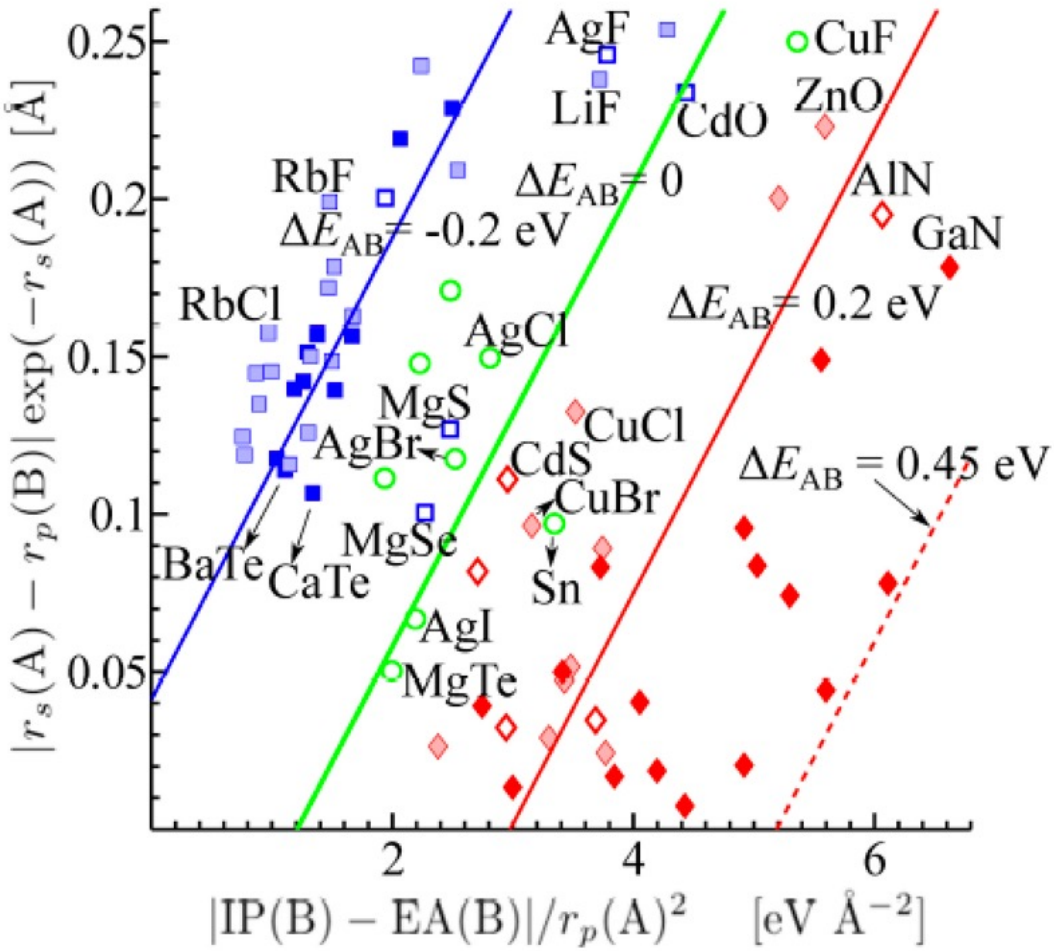
$$\frac{\mathrm{IP(B)} - \mathrm{EA(B)}}{r_p(\mathrm{A})^2}$$

$$\frac{|r_s(\mathrm{A}) - r_p(\mathrm{B})|}{\exp(r_s(\mathrm{A}))}$$

$$\frac{|r_p(\mathrm{B}) - r_s(\mathrm{B})|}{\exp(r_d(\mathrm{A}))}$$

density maximum radius of A-d orbital

$$\Delta E_{\mathrm{AB}} = E(\mathrm{AB_{RS}}) - E(\mathrm{AB_{ZB}})$$

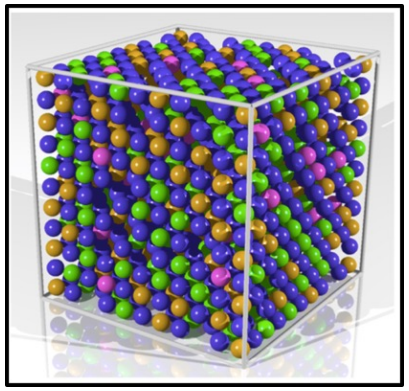


L.M. Ghiringhelli et al. PRL (2015).

First-Principles Calculations

Elements & Composition

$A_mB_n\ldots X$

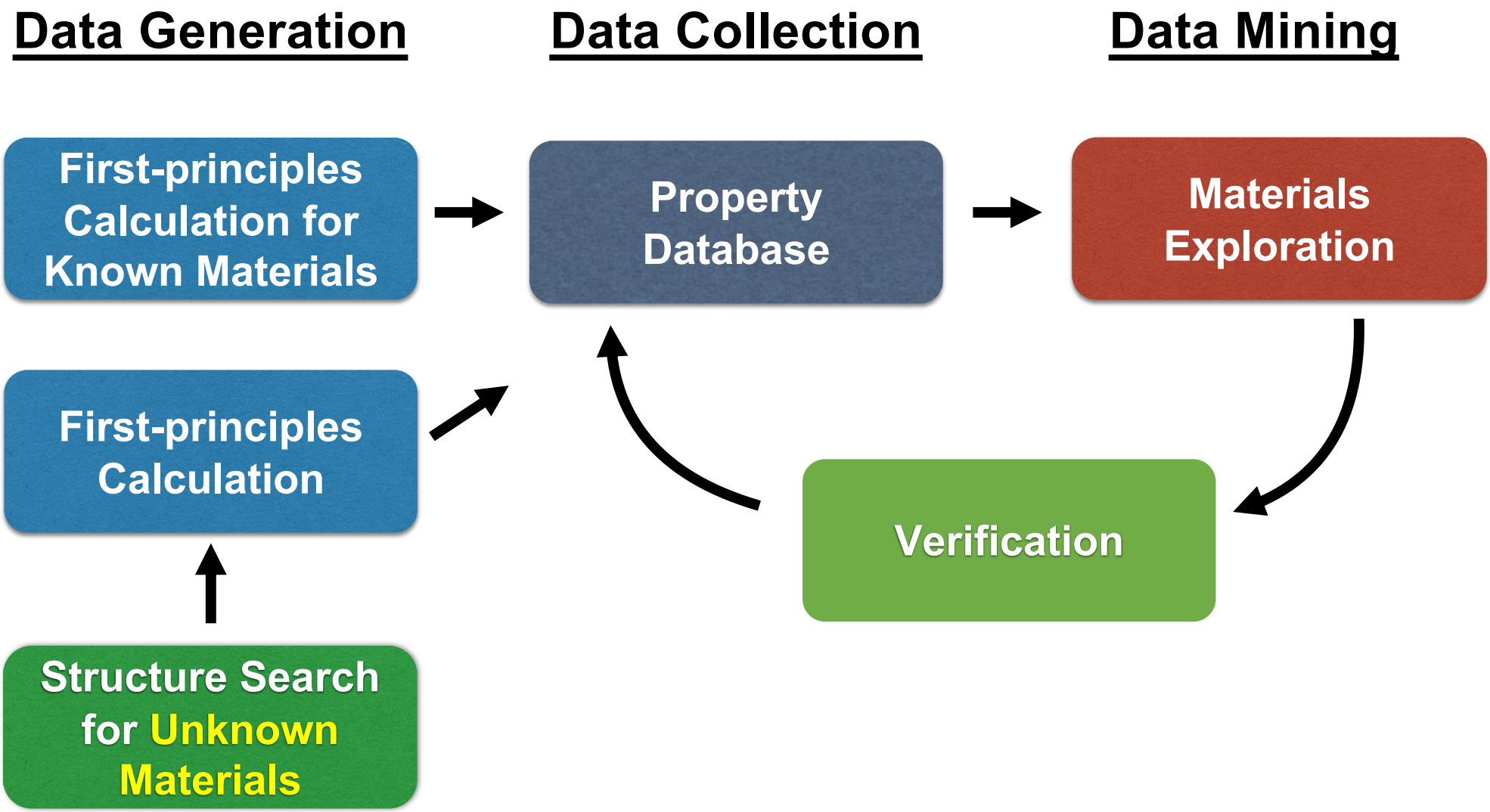


Structure

First-Principles Calculations

Property

Computational Materials Design with Machine Learning



Crystal Structure Search for Materials Discovery and Design

Acc. Chem. Res. **1994,** *27,* 309−314

Are Crystal Structures Predictable?

ANGELO GAVEZZOTTI*

Dipartimento di Chimica Fisica ed Elettrochimica, Università di Milano, Milano, Italy

Received May 16, 1994

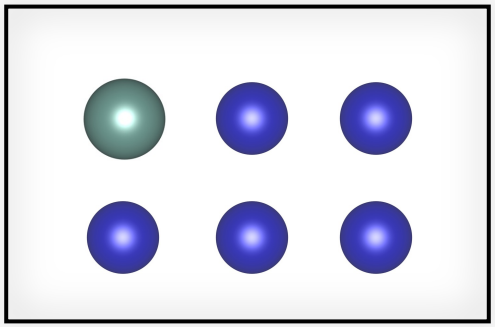
"No": by just writing down this concise statement, in what would be the first one-word paper in the chemical literature, one could safely summarize the present state of affairs, earn an honorarium from the American Chemical Society, and do a reasonably good service to his or her own reputation. In the mainstream of academic tradition, one could then concede a "maybe", or even a conditional "yes", thus making a good point for discussion; and then, in the mainstream of publication policy tradition, proceed eventually to have his or her papers rejected by referees taking the opposite stand.

Crystal structure prediction

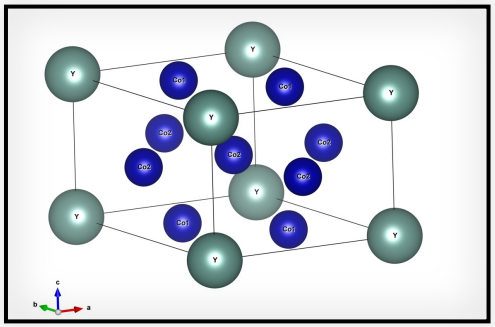
Crystal Structure ⇨ **First-Principles Calculations** ⇨ **Properties**

- First-principles calculations provide a powerful tool to predict properties for realistic materials with the information of crystal structure.
- Therefore, crystal structure prediction is indispensable for materials discovery without prior knowledge on structure.

Elements
Composition



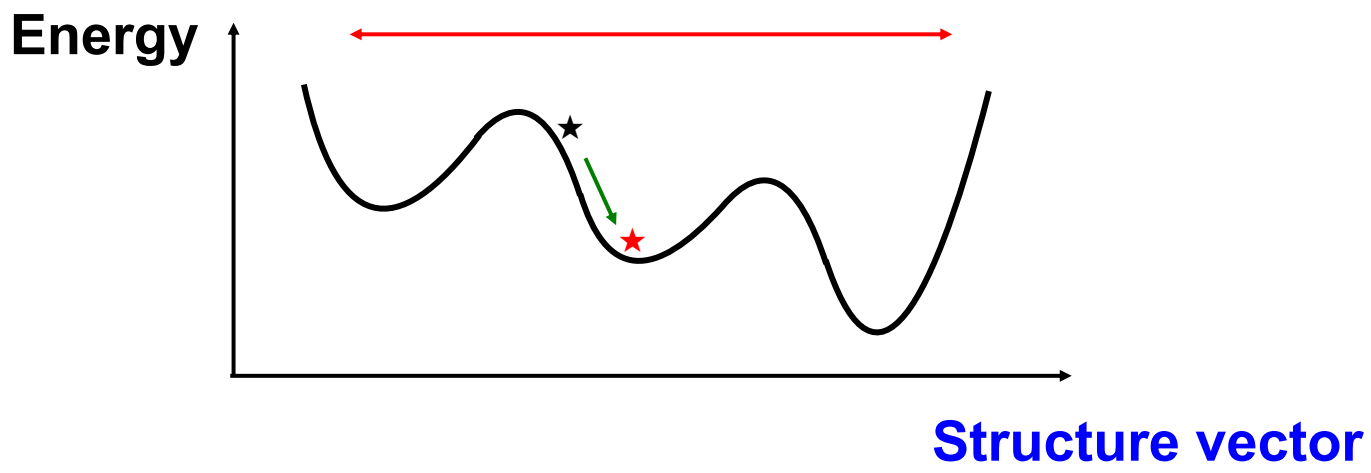
Size and shape of unit-cell
Atomic positions



Issues in crystal structure prediction

★ Huge Structure Space

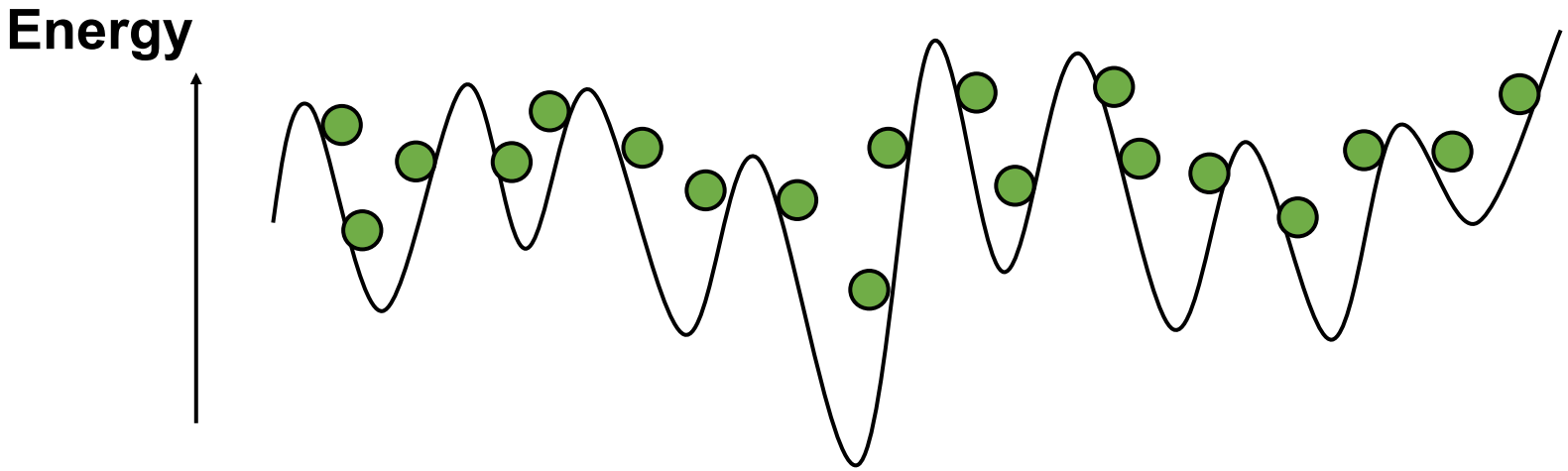
1. How to search the structure space globally
   - Local optimization by first-principles calculations or classical MD simulations
2. How to represent structure space (descriptor)
   - Easy and efficient representation of structure vector



1. Global structure search

Difficult due to many local minima

- **Many computationally demanding structure optimization processes are required.**



Development of an efficient global search algorithm is highly desired.

Structure search algorithm

- **Random search algorithm**
  - ✓ Random selection of lattice parameters and atomic positions
- **Evolution-type algorithm**
  - ✓ Evolutionary algorithm (USPEX)
  - ✓ Particle swarm optimization (CALYPSO)
- **Learning-type algorithm**
  - ✓ Bayesian optimization

→ A sequential design strategy for global optimization of black-box functions that doesn't require derivatives. [Wikipedia]

★ The target functions may not be analytic nor continuous.

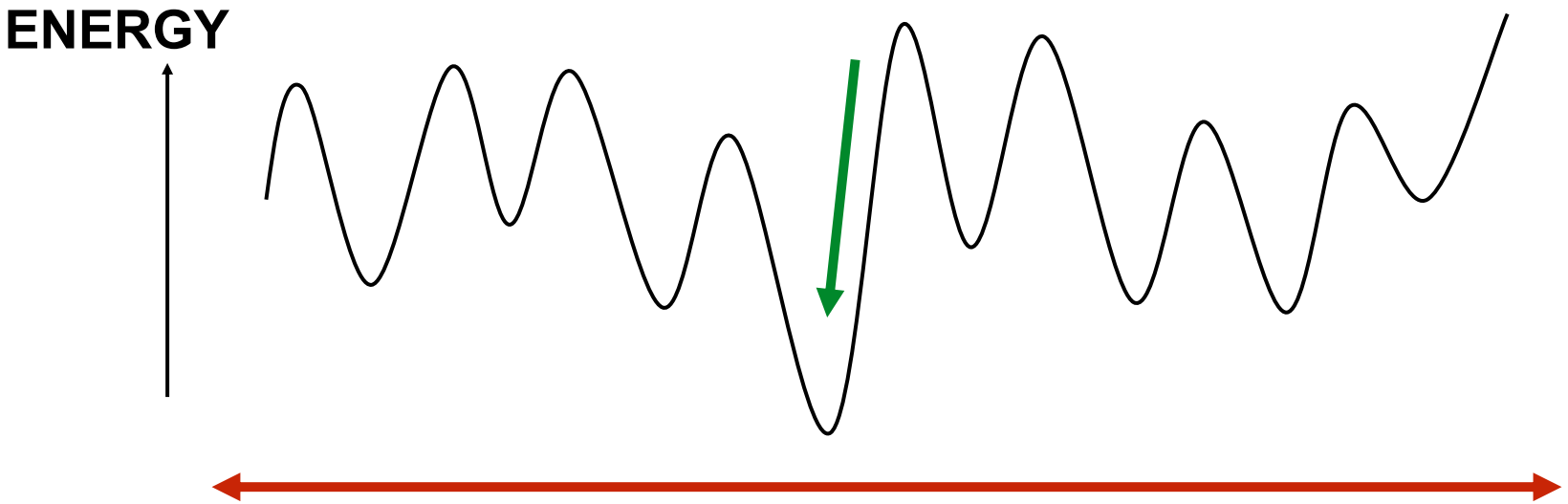
Key features for structure prediction

Exploration

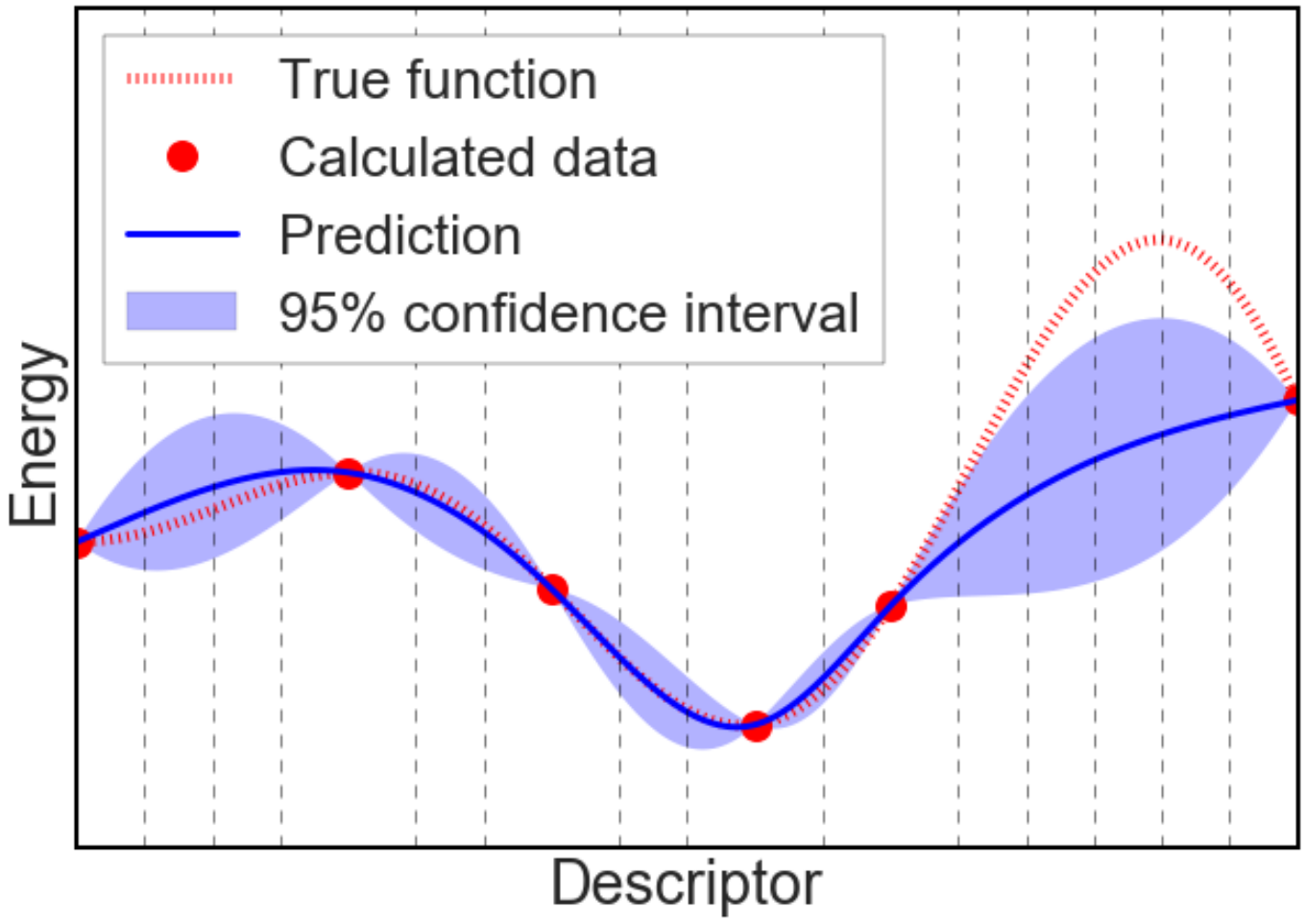
- to search the structure space as globally as possible

Exploitation

- to search the structure space without missing important minima



Bayesian optimization (BO)



BO enables to balance trade-off between exploration and exploitation of the search algorithm.

2. Descriptor for crystal structure

Fingerprint: based on radial distribution function

A. R. Oganov and M. Valle, J. Chem. Phys. 130, 104504 (2009).

$$F_{AB}(R) = \sum_{A_i,\text{cell}} \sum_{B_j} \frac{\delta(R - R_{ij})}{4\pi R_{ij}^2 \frac{N_A N_B}{V} \Delta} - 1$$

$$F_{AB}(0) = -1$$

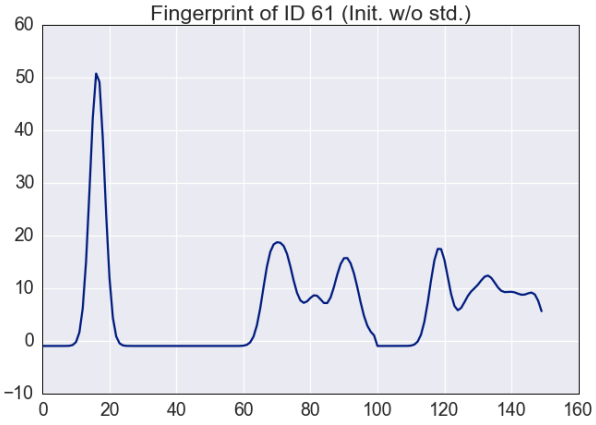
$$F_{AB}(\infty) = 0$$

- ➢ *i*-th atom of type A within the unit cell
- ➢ *j*-th atom of type B within the distance R_{max} (~5 Å)
- ➢ discretized over bins of width Δ (50 discretized R points)

Example of Y_2Co_{17}

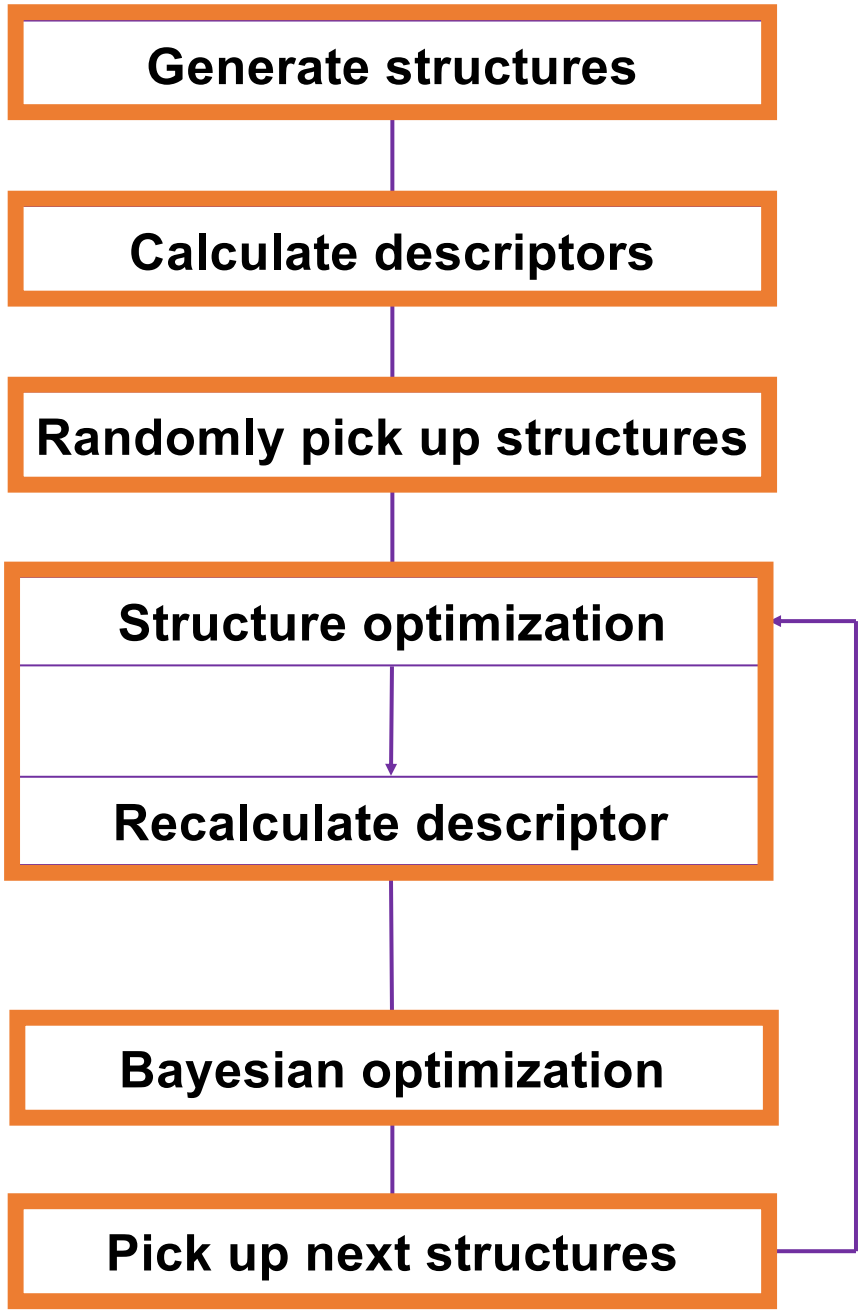
Y-Y, Y-Co, Co-Co

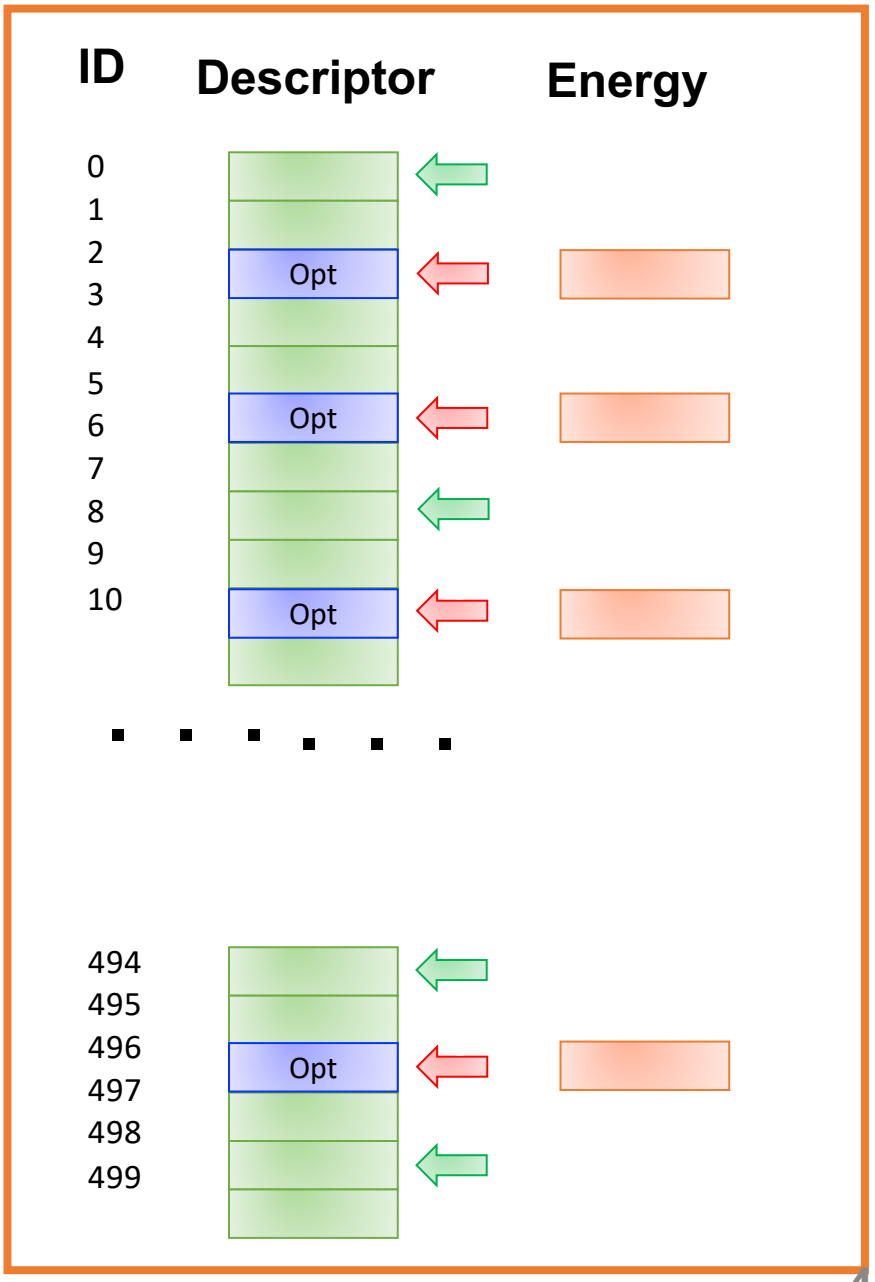
50, 50, 50



150 dimensional vector

Flowchart of Bayesian optimization





Lattice constants

Triclinic

$a < b < c$

$90° - \theta \leq \alpha, \beta, \gamma < 90°$ (Type 1)

$90° \leq \alpha, \beta, \gamma \leq 90° + \theta$ (Type 2)

Monoclinic

$a < c$

$\alpha = \gamma = 90°$

$90° \leq \beta \leq 90° + \theta$

Orthorhombic

$a < b < c$

$\alpha = \beta = \gamma = 90°$

Tetragonal

$a = b \neq c$

$\alpha = \beta = \gamma = 90°$

Rhombohedral

$a = b = c$

$\alpha = \beta = \gamma$

$90° - \theta \leq \alpha \leq 90° + \theta$

Hexagonal

$a = b \neq c$

$\alpha = \beta = 90°$

$\gamma = 120°$

Cubic

$a = b = c$

$\alpha = \beta = \gamma = 90°$

Constraints

Min. length

Max. length

$\theta = 20°$

Code development

CrySPY

https://github.com/Tomoki-YAMASHITA/CrySPY
distributed under the MIT License

Space group by find_wy
H. Kino, https://github.com/nims-hrkn/find_wy

Bayesian optimization library, COMBO
T. Ueno *et al.*, Materials Discovery 4, 18 (2016).
https://github.com/tsudalab/combo

- **Algorithm**
  - ✓ Random search
  - ✓ Bayesian optimization
  - ✓ Evolutionary algorithm
  - ✓ LAQA: Look Ahead based on Quadratic Approximation
    A fine-grained method to reduce local optimization steps
- **Local optimization**

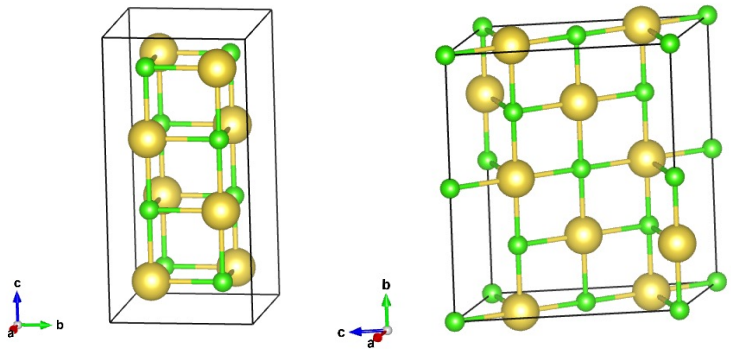
  Interfaced with
  - ✓ VASP
  - ✓ Quantum Espresso
  - ✓ soiap
  - ✓ LAMMPS

Test simulations

- Both random search and Bayesian optimization are applied to two known systems

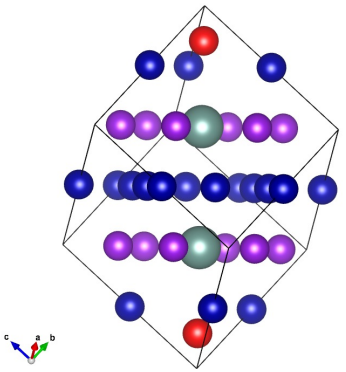
  - Na_8Cl_8 (16 atoms/cell)

    Rocksalt structure



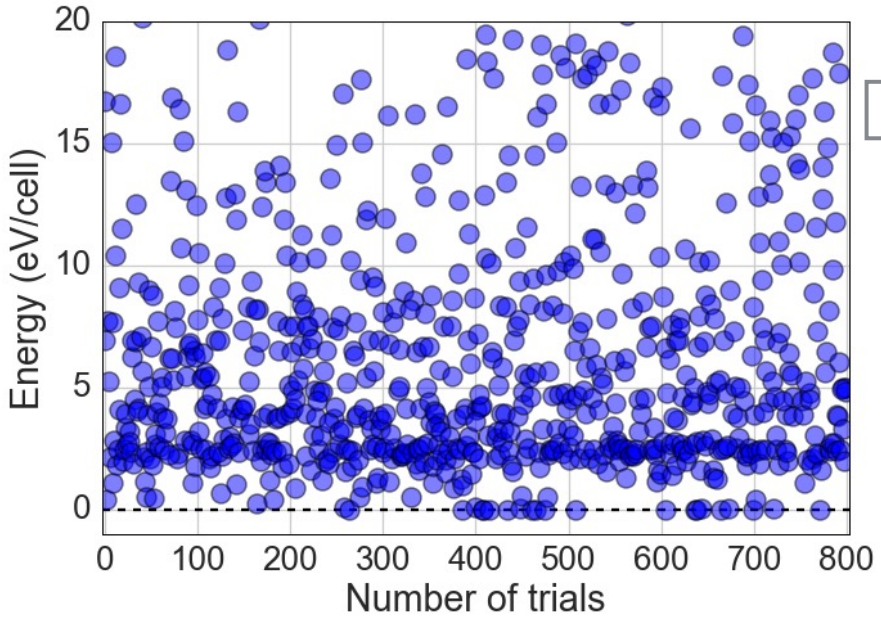
  - Y_2Co_{17} (19 atoms/cell)

    Th_2Zn_{17}-type structure (R-3m)
    Ferromagnetic



Test simulation: Na_8Cl_8

Random search

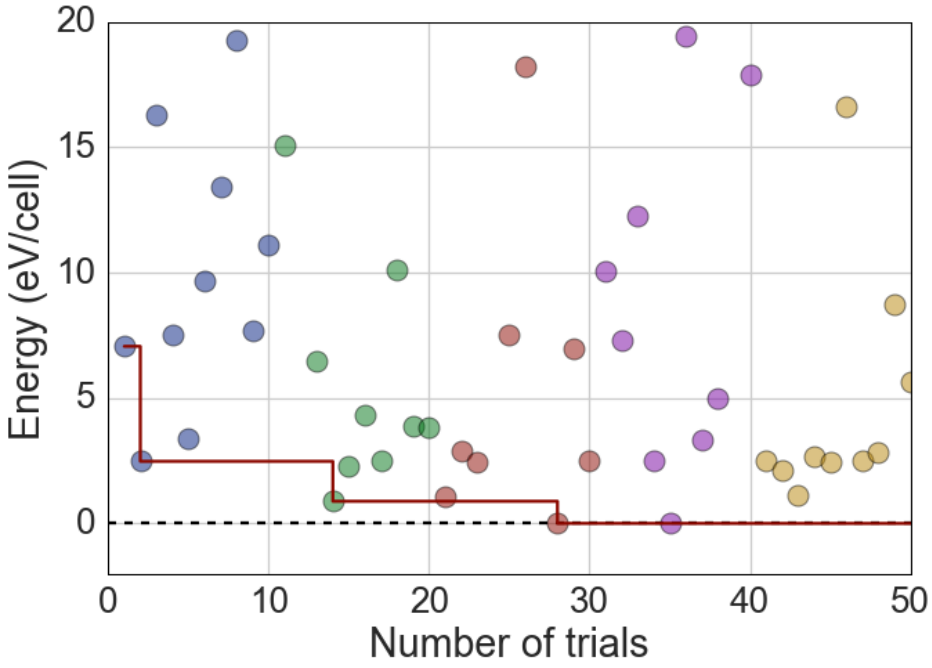


Success rate: 23/800

Bayesian optimization

Same data

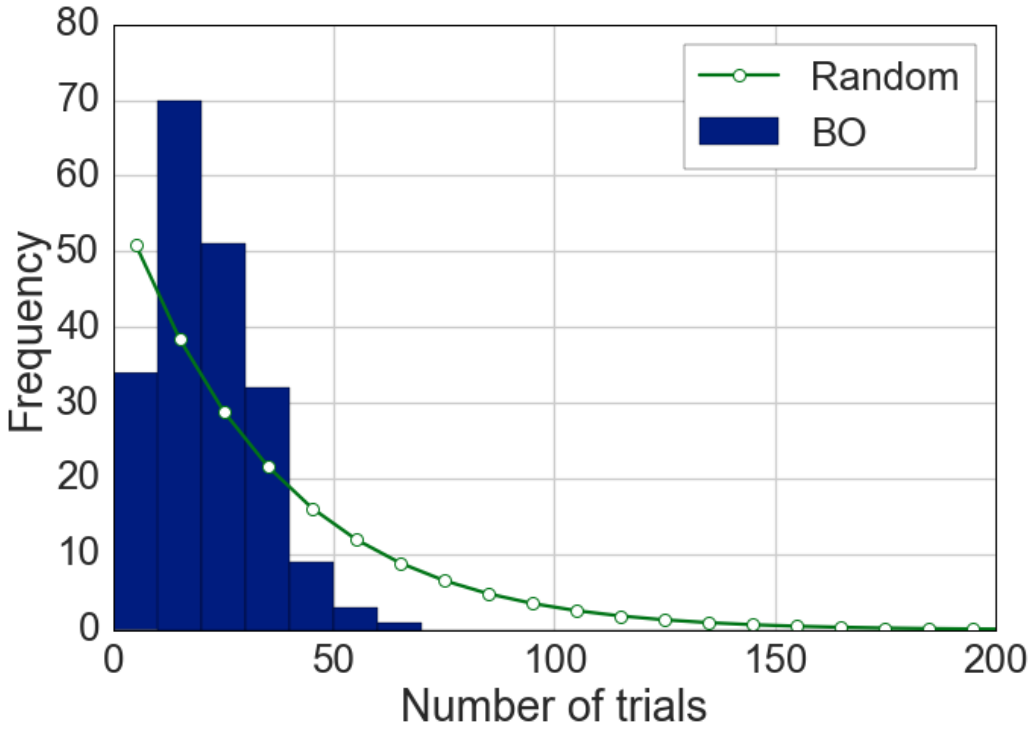
- **800 structures**
- **23 good candidates**



Selection: every 10 trials

Test simulation: Na_8Cl_8

To investigate the statistical efficiency, 200 BO simulations were carried out.



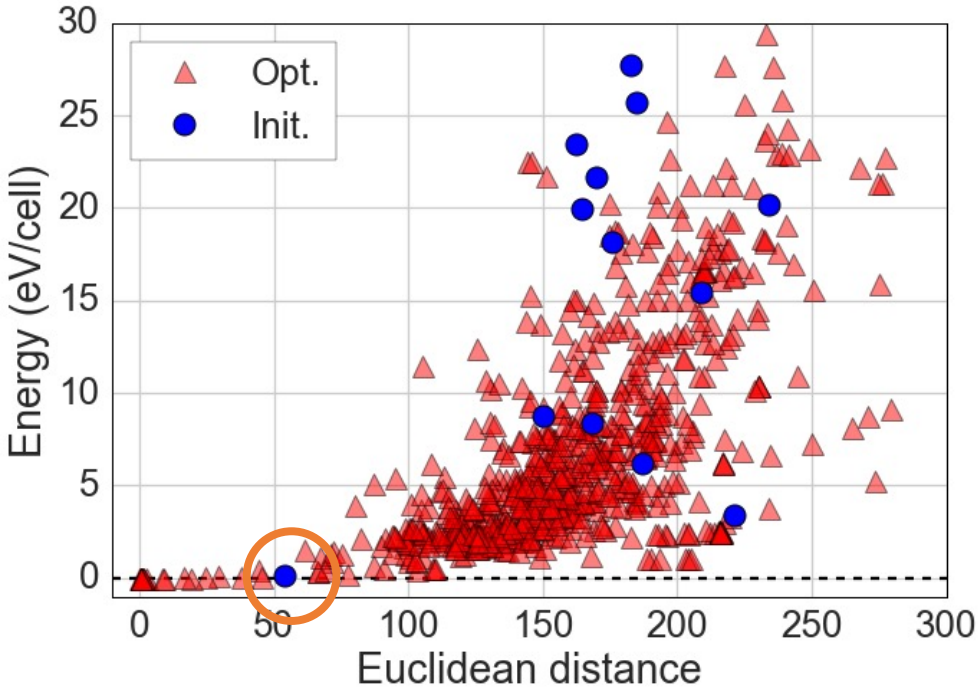
Frequency distribution of number of trials required to find the rocksalt structure

★ BO reduces trials by 31% compared with RS.

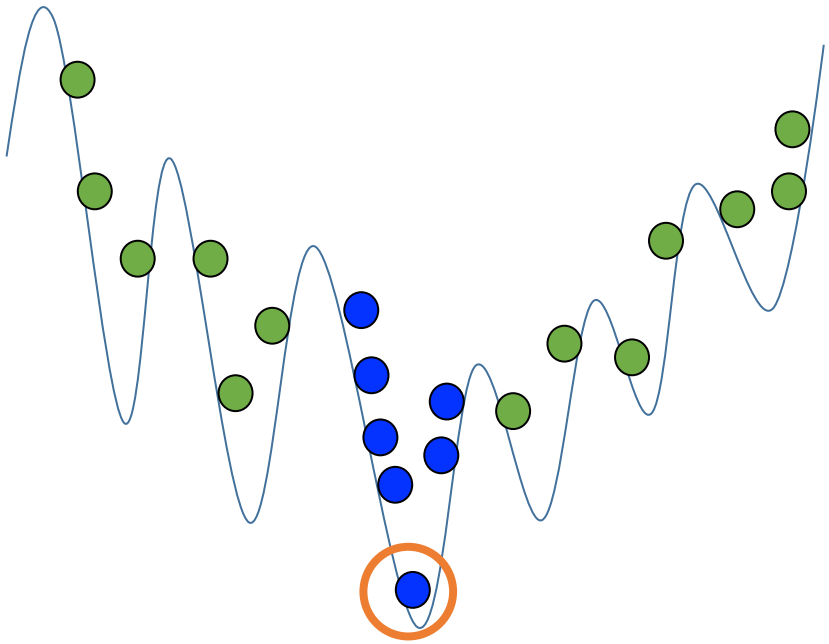
Test simulation: Na_8Cl_8

800 structure data

● : 23 good candidates



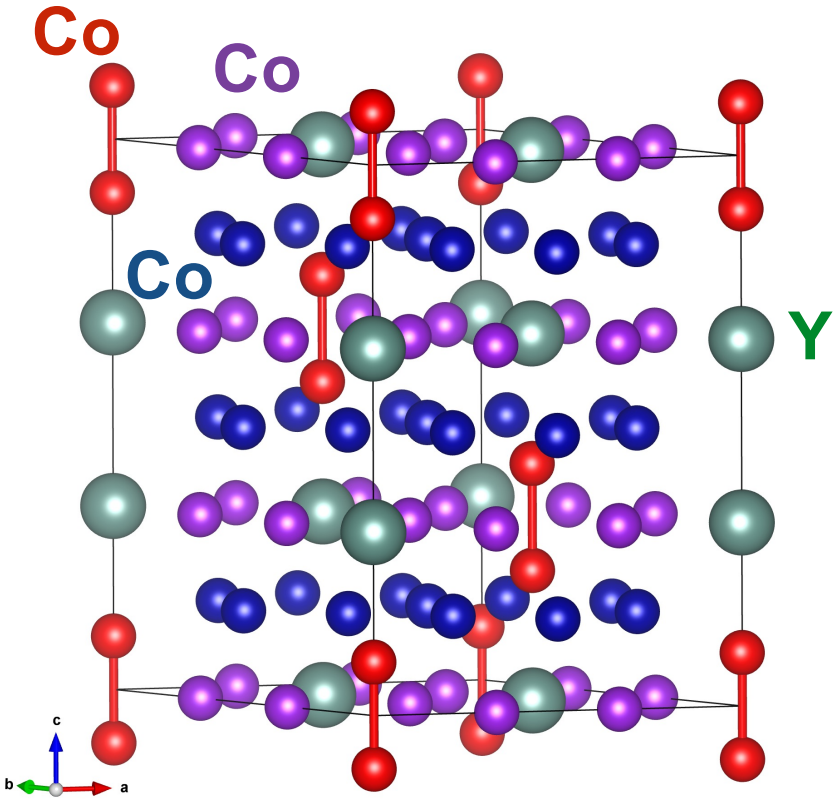
Learned correlation between energy and structure distance

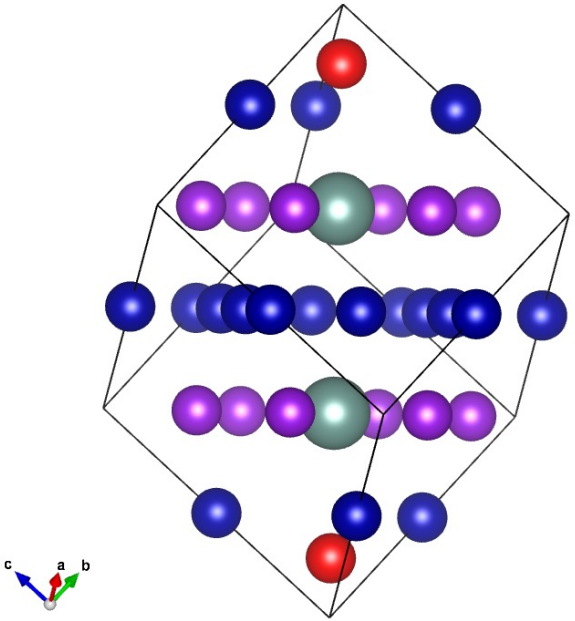


One of the good candidates in the global-minimum valley is selected within 70 trials, so BO works well in the exploitation phase.

Test simulation: Y_2Co_{17}

Th_2Zn_{17}-type structure (R-3m)





[Conventional] Hexagonal

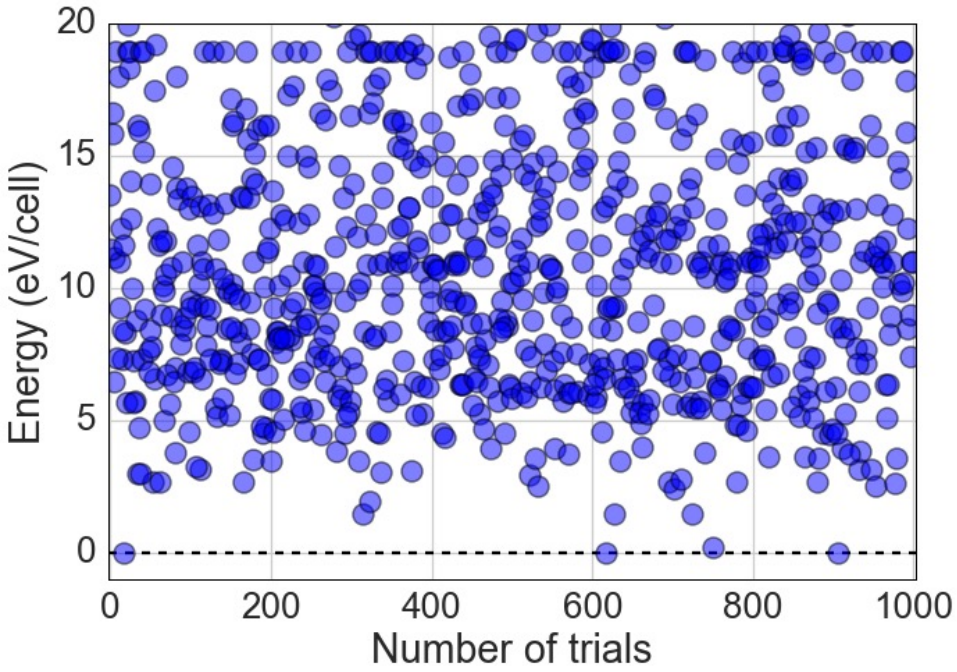
[Primitive] Rhombohedral

19 atoms/cell

Test simulation: Y_2Co_{17}

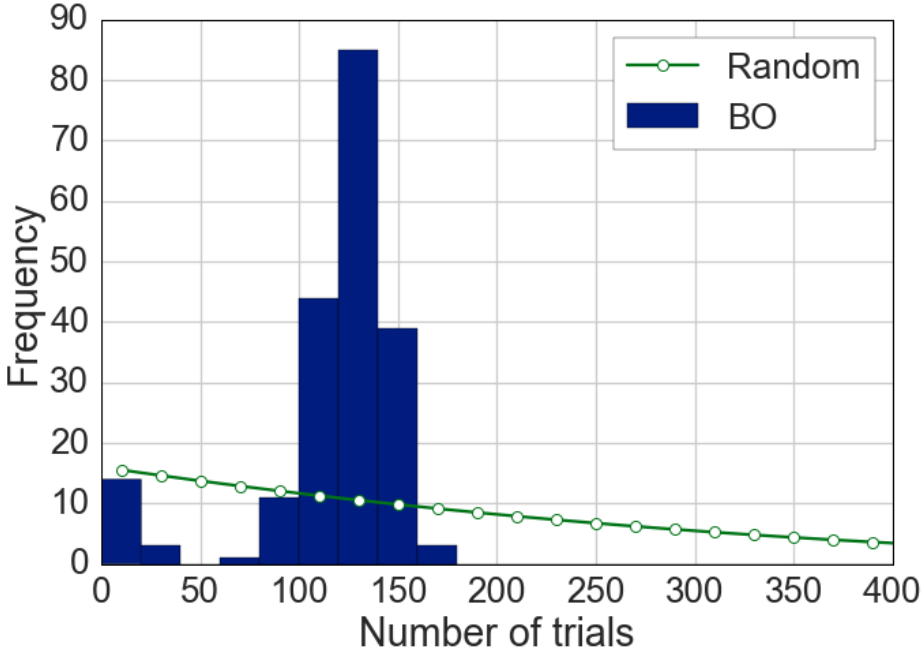
Random search

Collinear magnetic order



Success rate: 4/1000

200 BO simulations

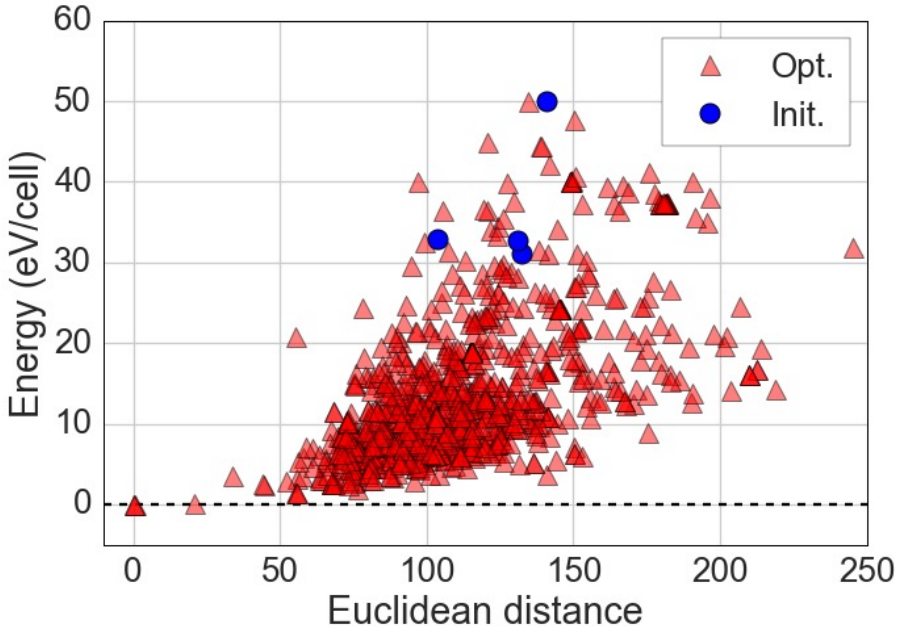


Selection: every 20 trials

★ BO reduces trials by 39% compared with RS.

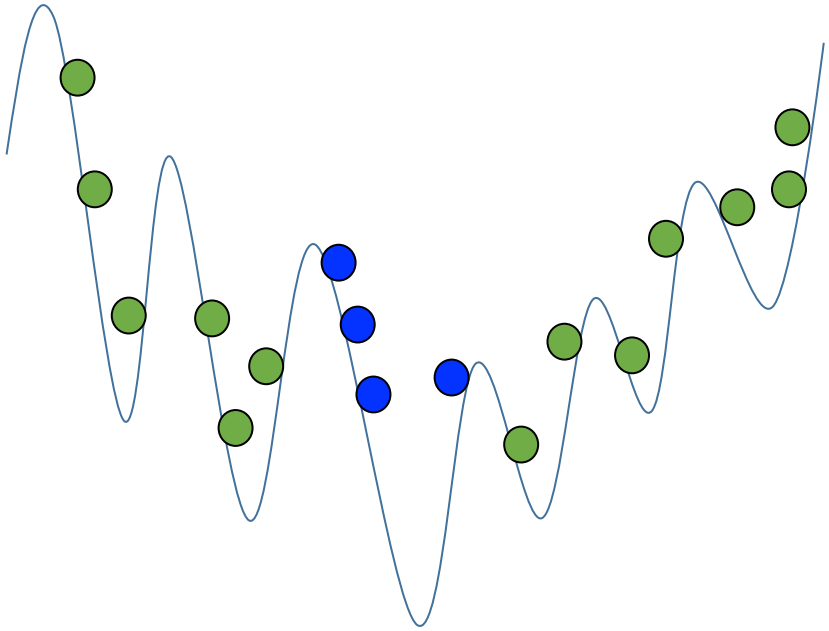
Test simulation: Y_2Co_{17}

1000 structure data



Learned correlation between energy and structure distance

● **: 4 good candidates**



It's quite difficult to search for the global minimum, but BO works well in the exploration phase.

Summary

- A tool for crystal structure search
  - ➢ Random search
  - ➢ Bayesian optimization (BO)
- Test simulations for Na_8Cl_8 and Y_2Co_{17}
  - ➢ Bayesian optimization is highly efficient and significantly reduce the number of searching trials required to find the global minimum structure.

Phys. Rev. Materials 2, 013803 (2018).
Phys. Rev. Materials 4, 033801 (2020).
Sci. Technol. Adv. Mater.: Methods 1, 87 (2021).
Sci. Technol. Adv. Mater.: Methods 2, 67 (2022).

Crystal Structure Map
for Materials Classification and Modeling

Current Hot Issue in Materials Research

- **Materials Discovery Assisted by Data Science Approach**

  ✓ **Today, a wealth of materials data are being accumulated, experimentally and computationally.**

  ✓ **Where and how is each existing materials system placed for design and discovery ?**

"Materials Map"

Map

Q: How to define such a low-dimensional axis system and the coordinate of materials targets ?

Structure as Axes of the Map

- **Most fundamental attribute in condensed matter systems, especially crystalline materials**
- **Governing the electronic states as under the Born-Oppenheimer approximation.**

$$\mathcal{H}_{\text{electron}} = \mathcal{H}_{\text{electron}}\left[\{\boldsymbol{R}_n\}\right]$$

★ **Our understanding of materials properties is often based on the correlation with structure.**

Crystal Structure Map

✓ **Coordinates classify huge related types of structures.**

✓ **Axes provide structure features for modeling properties.**

Crystal Structure DB

Crystallography Open Database

COD Home
- Home
- What's new?

Accessing COD Data
- Browse
- Search
- Search by structural formula

Add Your Data
- Deposit your data
- Manage depositions
- Manage/release prepublications

Documentation
- COD Wiki
- Obtaining COD
- Querying COD
- Citing COD
- COD Mirrors
- Advice to donators
- Useful links

Search results

Result: there are 60 entries in the selection

Switch to the old layout of the page

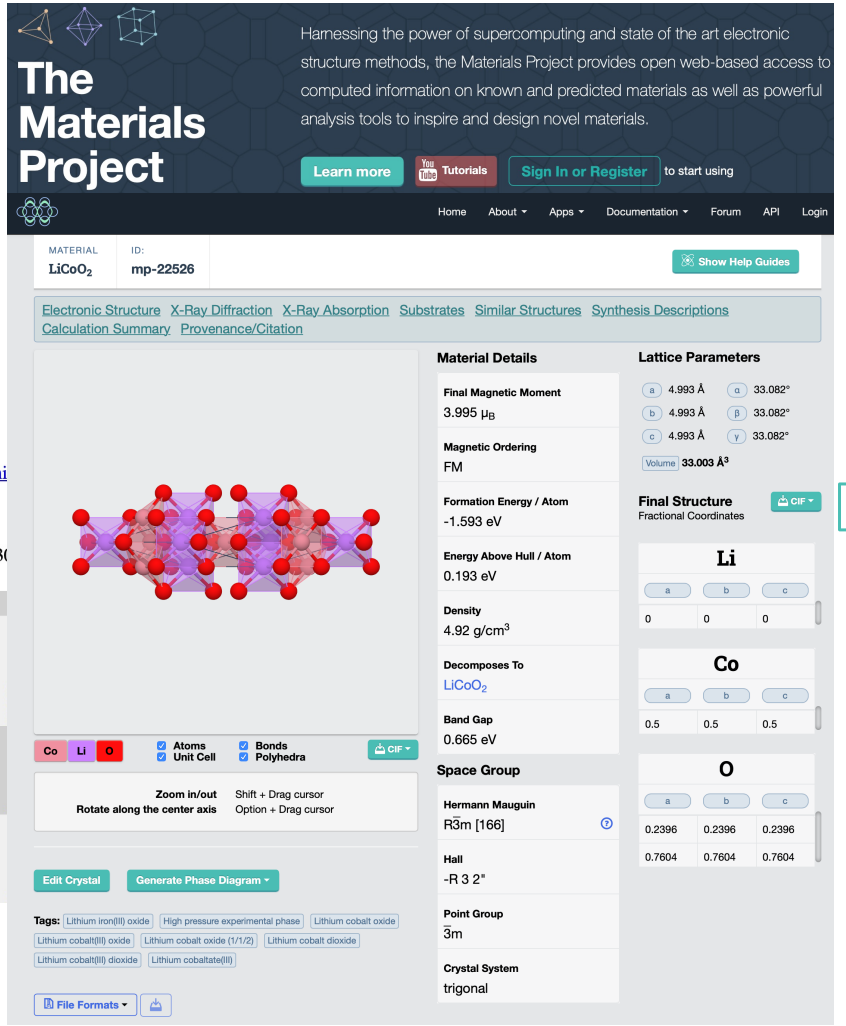
Download all results as: list of COD numbers | list of CIF URLs | data in CSV format | archi

Searching elements including Ba, Ti, O number of elements between 1 and 3

◄◄ First | ◄ Previous 20 | Page 1 of 3 | Next 20 ► | Last ►► | Display 5 **20** 50 100 200 3(

COD ID ▲	Links	Formula ▲	Space group ▲	Cell parameters	Cell volume ▲
1507755	CIF	Ba2 O13 Ti6	C 1 2/m 1	15.004; 3.953; 9.085 90; 98.01; 90	533.582
1507756	CIF	Ba O3 Ti	P 4 m m	3.9999; 3.9999; 4.017 90; 90; 90	64.269
1507757	CIF	Ba O3 Ti	P m -3 m	4.0073; 4.0073; 4.0073 90; 90; 90	64.351

CIF



http://www.crystallography.net/cod/

https://materialsproject.org

What's CIF ?

- Crystallographic Information Framework
- Established by International Union of Crystallography (IUCr) – Dictionary: CoreCIF
- Widely used in structure database such as COD, ICSD, Materials Project, MatNavi, etc.
- Containing all information about crystal structure necessary for electronic structure calculations
- Conversion apps from CIF to VASP, QE, etc.:

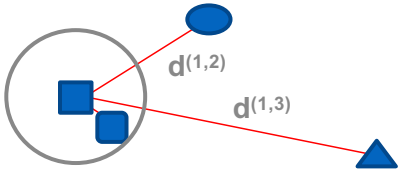
  cif2cell*, C-Tools$, cifconv#

*: T. Björkman, Comput. Phys. Commun., 182, 1183 (2011).
$: https://sourceforge.net/projects/c-tools/
#: F. Izumi and K. Miyazaki, Ceramics 54, 473 (2019).

***de facto* standard for crystal structure format in experiment and theory/computation**

Procedure of Structure Classification for a Set of CIFs

- **Step 1:** To **represent a (often highly dimensional) vector** as structure feature
  - ✓ Simple, tractable, objective, and translationally and rotationally invariant enough to apply to general structures
- **Step 2:** To **define distance** between feature vectors for measuring similarity
  - ✓ Degree of similarity
    - $d^{(i,j)} < d_{th}$: equivalent within a given *threshold*
    - $d^{(i,j)} \sim d_{th}$: close to equivalent but minor dissimilarity
    - $d^{(1,2)} < d^{(1,3)}$ means that $d^{(1,2)}$ is closer in similarity than $d^{(1,3)}$.

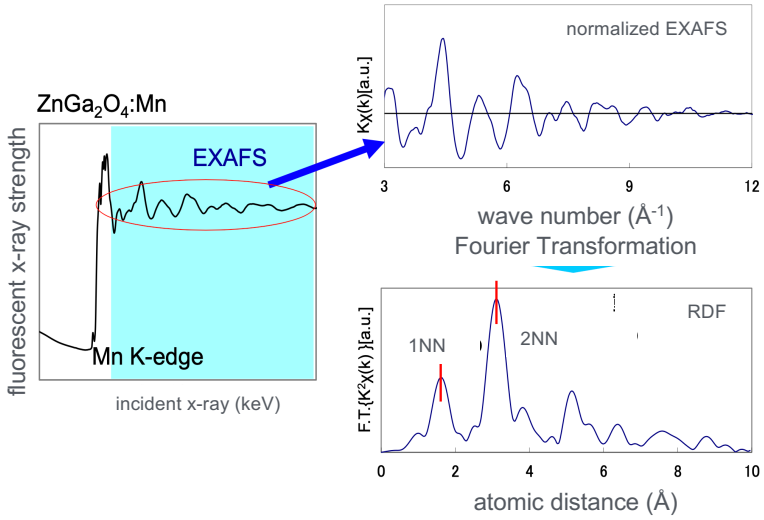


- **Step 3:** To **map** the position of each structure **in a low-dimensional space** so that the distances calculated from it well approximate the original ones

➢ Classification (clustering) of crystal structures for further analysis

Step 1: Feature Vector for Crystal Structure

- ✓ **Coordination Number (CN)**
  - simple feature of local structure around each atomic site
  - no unique definition of neighbors in general structures
- ✓ **Cluster Expansion**
  - useful for considering different configurations like alloy ordering
- ✓ **Radial Distribution Function (RDF)**
  - complete pair-wise information
  - very simple
  - measurable in EXAFS



http://titan.nusr.nagoya-u.ac.jp/Tabuchi/BL5S1/lib/exe/fetch.php?media=tabuchi:gairon-20181210-v2.pdf

F-Fingerprint

Oganov and Valle, J. Chem. Phys. 130, 104504 (2009)

- **For *AB* element pair in a compound**

$$F_{AB}(R) = \frac{V}{4\pi R^2 N_A N_B} \sum_{A_i,\text{cell}} \sum_{B_j} \delta(R - R_{ij}) - 1 = g_{AB}(R) - 1$$

RDF (pair correlation)

$$F_{AB}(0) = -1, \quad F_{AB}(\infty) = 0$$

This "-1" makes fingerprint short-ranged and components distributed around zero.

A_i : i-th site of element A
R_{ij} : distance between A_i and B_j
N_A : number of sites of A in cell
V : volume of cell
d : smoothened with Gaussian

FP parameters
RMAX : R_m
Δ: $R_i - R_{i-1}$
σ²: Gaussian variance

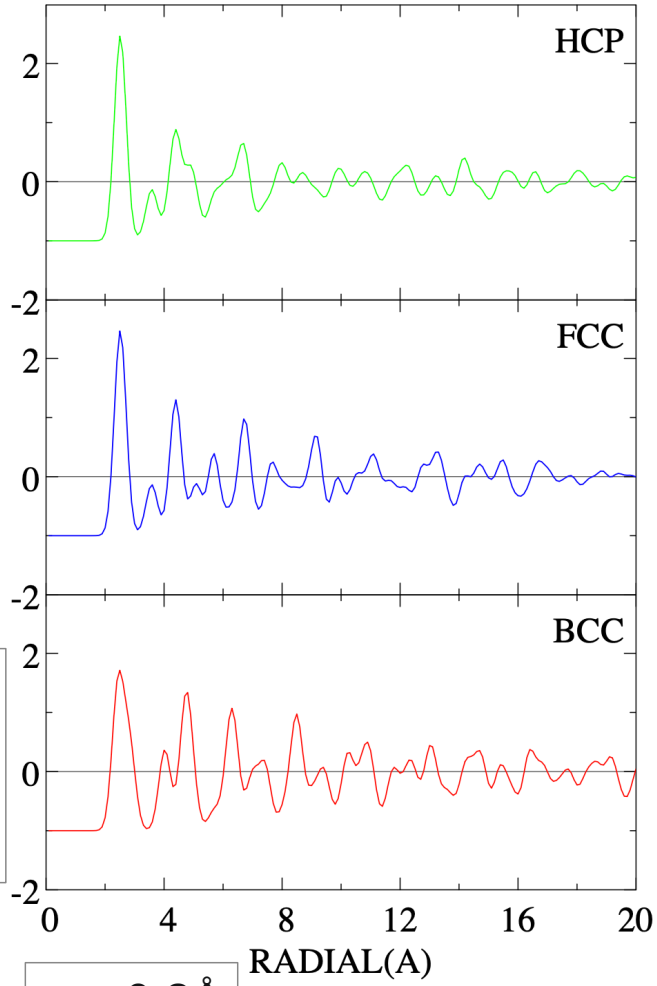
Discretizing $F_{AB}(R_l)$ l=1, ..., m

⇩

m-D vector for element-pair AB

⇩

n(=pm)-D vector for p pairs



σ ~ 0.2Å
Δ ~ 0.1Å

Step 2: Distance as a Measure of Similarity

- **Distance between two *n*-D vectors *X* and *Y***

1. Euclidean distance (ED)

$$d(\boldsymbol{X}, \boldsymbol{Y}) = \left[\sum_{i=1}^{n}(X_i - Y_i)^2\right]^{1/2}$$

Minkowski distance

$$d(\boldsymbol{X}, \boldsymbol{Y}) = \left[\sum_{i=1}^{n}|X_i - Y_i|^p\right]^{1/p}$$

2. Pearson correlation coefficient (PCC)

$$r(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right)^{1/2}}$$

$$-1 \le r \le 1$$

→ Cosine distance (CD)

$$d_{\text{cosine}} = \frac{1}{2}(1 - r) \qquad 0 \le d_{\text{cosine}} \le 1$$

When *X* and *Y* are normalized

$$d = 2(1 - r) = 4d_{\text{cosine}}$$

https://en.wikipedia.org/wiki/Cosine_similarity

Step 3: Mapping Structure

W. S. Torgerson, Psychometrika 17, 401 (1952)

- **Multidimensional Scaling**
  - ✓ **For a given pair-wise (Euclidean/non-Euclidean) distance matrix D of *N* points, we seek *N*-point coordinates X in a low ($k \ll n$) dimensional space so that pair-wise Euclidean distance matrix D^X calculated using X is the closest approximation to D.**
  - ✓ **Often called "Dimension Reduction" for data mapping**
  - ✓ **Equivalent to "Principal Component Analysis (PCA)" for Euclidean distance**
  - ✓ **Implemented in MATLAB, R, Scikit-learn, ...**

Multidimensional Scaling (MDS)

W. S. Torgerson, Psychometrika 17, 401 (1952)

- **Algorithm**

squared distance

1. $\mathbf{B} = \mathbf{X}\mathbf{X}^T = -\frac{1}{2}\mathbf{J}\mathbf{D}^{(2)}\mathbf{J}, \quad \mathbf{J} = \mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^T$ ← ***double centering***

2. $\mathbf{B} = \mathbf{Q}\Lambda\mathbf{Q}^T$ **eigenvalue decomposition**

3. $\mathbf{X}_+ = \mathbf{Q}_+\Lambda_+^{1/2}$ **first *k* positive-eigenvalue part only**

($N \times k$)

$\mathbf{I}$: identity matrix
$\mathbf{1}$: identity vector
$\mathbf{X}$: ($N \times d$)
d-D coordinates centered

k ≪ ***d*** : ***dimension reduction***

What dimensionality *k* one should choose?

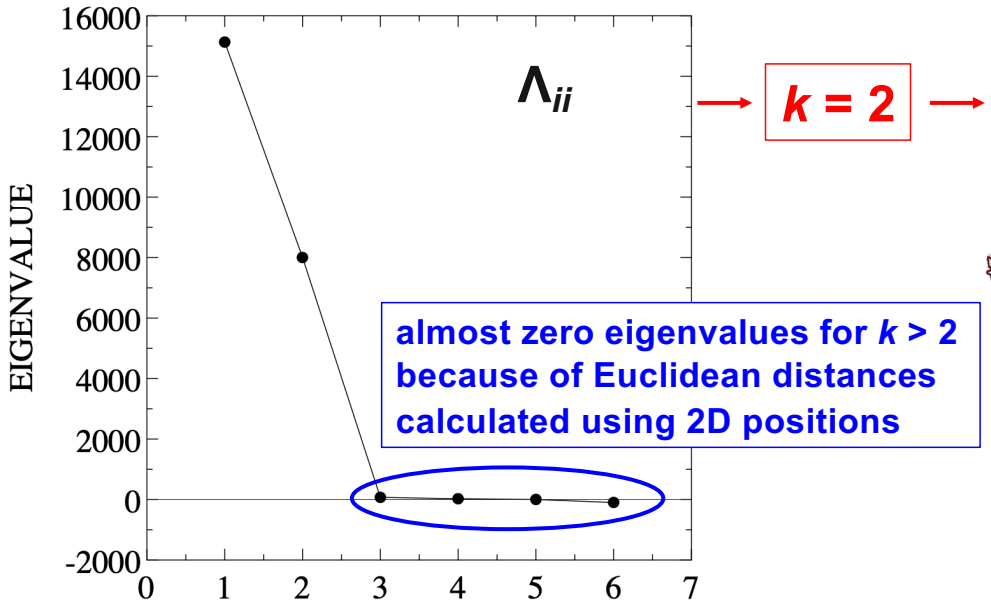
The larger eigenvalues, the more important coordinates, because the sum of the eigenvalues in Λ_+ should approximate the sum of all eigenvalues in Λ, so that small negative eigenvalues cancel out small positive eigenvalues.

R. Sibson, J. R. Statist. Soc. B 41, 217 (1979)

$$\mathbf{B} \approx \mathbf{B}_+ = \mathbf{X}_+\mathbf{X}_+^T = (\mathbf{Q}_+\Lambda_+^{1/2})(\mathbf{Q}_+\Lambda_+^{1/2})^T$$

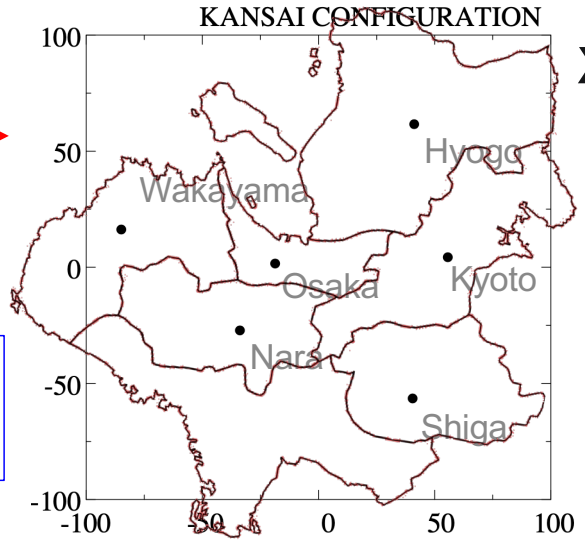
MDS Example: Kansai-Area Prefecture Center Distance

	Hyogo	Wakayama	Osaka	Nara	Shiga	Kyoto
Hyogo	0	134	85	116	118	60
Wakayama	134	0	68	66	145	141
Osaka	85	68	0	32	83	75
Nara	116	66	32	0	79	95
Shiga	118	145	83	79	0	63
Kyoto	60	141	75	95	63	0



$k = 2$

almost zero eigenvalues for $k > 2$ because of Euclidean distances calculated using 2D positions



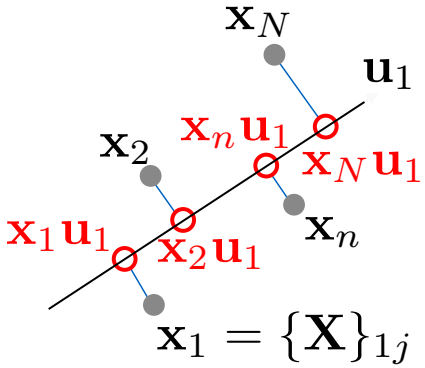
$X_+ = Q_+ \Lambda_+^{1/2}$

MDS and Principal Component Analysis (PCA)

- **New coordinates are span in the axes given as eigenvectors of principal component on original *n*-D space**

$$\underset{(N\times N)}{\mathbf{B}} = \underset{(N\times d)(d\times N)}{\hat{\mathbf{X}}\hat{\mathbf{X}}^T} = \underset{(N\times N)}{\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T} \approx \underset{(N\times N)}{\mathbf{Q}_+\boldsymbol{\Lambda}_+\mathbf{Q}_+^T} \equiv \underset{(N\times k)\ (k\times N)}{\mathbf{X}_+\mathbf{X}_+^T}$$

$$\hat{\mathbf{X}} = \mathbf{X} - \mathbf{1}\bar{\mathbf{X}} \quad \text{centered}$$



$$\underset{(N\times N)\ (N\times k)}{\hat{\mathbf{X}}\hat{\mathbf{X}}^T\mathbf{Q}_+} \approx \mathbf{Q}_+\boldsymbol{\Lambda}_+\mathbf{Q}_+^T\mathbf{Q}_+ = \underset{(N\times k)}{\mathbf{Q}_+\boldsymbol{\Lambda}_+}$$

$$\underset{(d\times d)\ (d\times k)}{\hat{\mathbf{X}}^T\hat{\mathbf{X}}(\hat{\mathbf{X}}^T\mathbf{Q}_+)} \approx \underset{(d\times k)}{(\hat{\mathbf{X}}^T\mathbf{Q}_+)}\boldsymbol{\Lambda}_+$$

PCA: $\mathbf{S} = \mathbf{X}^T\mathbf{X}$ **: covariance**

$$\frac{\delta}{\delta\mathbf{u}^T}\left[\underline{\mathbf{u}^T\mathbf{S}\mathbf{u}} + \lambda(1-\mathbf{u}^T\mathbf{u})\right] = 0$$

variance

$$\mathbf{S}\mathbf{u} = \mathbf{u}\boldsymbol{\Lambda}$$

$$\underset{(d\times k)}{\mathbf{u}_+} = \hat{\mathbf{X}}^T\mathbf{Q}_+\boldsymbol{\Lambda}_+^{-\frac{1}{2}} = \hat{\mathbf{X}}^T\mathbf{X}_+\boldsymbol{\Lambda}_+^{-1} = \mathbf{X}^T\mathbf{X}_+\boldsymbol{\Lambda}_+^{-1}$$ **: *d*-D eigenvectors of PC**

$$\mathbf{Y}_+ = \hat{\mathbf{Y}}\mathbf{u}_+$$ **: projection on the PC space**

$$\hat{\mathbf{Z}} = \mathbf{Z}_+\mathbf{u}_+^T$$ **: inverse projection**

C. M. Bishop, Pattern Recognition and Machine Learning (Springer-Verlag, New York, 2006)

Generation of Structure with CIF

*findsym**

```
!useKeyWords
!title
Fe4Co4 alloy
!atomicPositionTolerance
0.00001
!latticeParameters
5.72  5.72  5.72  90.0  90.0
90.0
!unitCellCentering
I
!atomCount
8
!atomType
4*Fe  4*Co
!atomPosition
     0.00     0.00     0.00
     0.25     0.25    -0.25
     0.25    -0.25     0.25
    -0.25     0.25     0.25
     0.25     0.25     0.25
     0.00     0.50     0.00
     0.00     0.00     0.50
     0.50     0.00     0.00
```

Fe_4Co_4 Alloy

$_8C_4$=70

atls

```
2 Fe 4 Co 4
```

```
4*Fe 4*Co
3*Fe Co Fe 3*Co
3*Fe 2*Co Fe 2*Co
3*Fe 3*Co Fe Co
3*Fe 4*Co Fe
2*Fe Co 2*Fe 3*Co
2*Fe Co Fe Co Fe 2*Co
2*Fe Co Fe 2*Co Fe Co
2*Fe Co Fe 3*Co Fe
2*Fe 2*Co 2*Fe 2*Co
.......
```

CIF

cif2esc → distance → diagnosis

```
4*Fe 4*Co
3*Fe 2*Co Fe 2*Co
Fe 4*Co 3*Fe
```

Only three inequivalent structures

Equiatomic quaternary cases

4 Mn 2 Fe 2 Co 2 Ni 2

$_8C_2 \cdot {}_6C_2 \cdot {}_4C_2$=2,520

4 Mn 3 Fe 3 Co 3 Ni 3

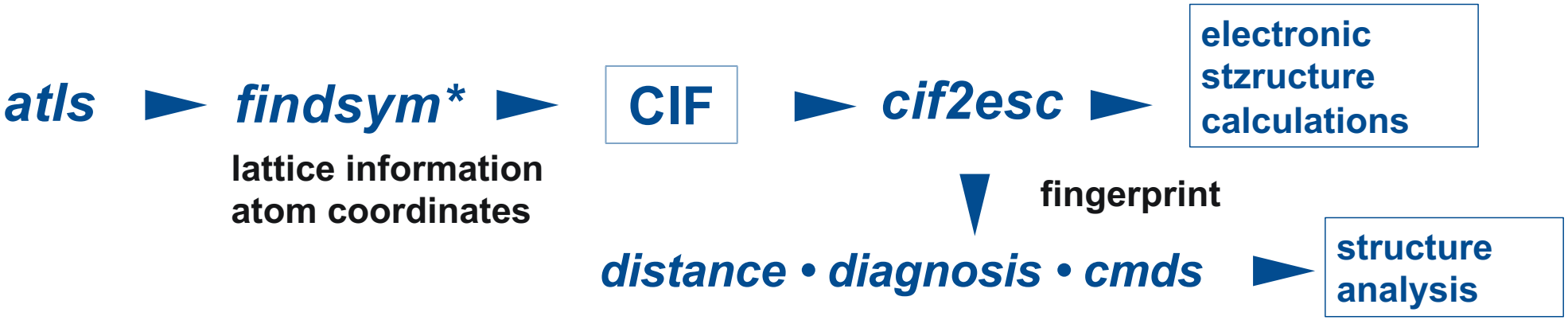
$_{12}C_3 \cdot {}_9C_3 \cdot {}_6C_3$=369,600

4 Mn 4 Fe 4 Co 4 Ni 4

$_{16}C_4 \cdot {}_{12}C_4 \cdot {}_8C_4$=63,063,000

*: H. T. Stokes, D. M. Hatch, and B. J. Campbell, ISOTROPY Software Suite, iso.byu.edu.
H. T. Stokes and D. M. Hatch, J. Appl. Cryst. 38, 237 (2005).

cif2esc Project



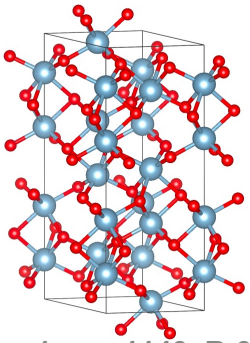
- "*cif2esc*" converts crystal structure information from CIF files to input data of several DFT codes for electronic structure calculations and to fingerprint data for structure analysis.
- "*diagnosis*" classifies crystal structure according to fingerprint distance as its similarity computed by "*distance*" and "*cmds*" extracts principal features by using dimension reduction.
- "*atls*" generates list of possible atom-type combinations from given atom compositions, passing to input for the "*findsym**" application with lattice information and atom coordinates to get CIF files.

*: H. T. Stokes, D. M. Hatch, and B. J. Campbell, ISOTROPY Software Suite, iso.byu.edu.
H. T. Stokes and D. M. Hatch, J. Appl. Cryst. 38, 237 (2005).

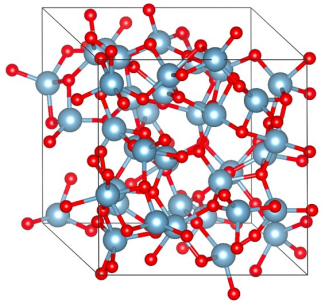
Demonstration: Al_2O_3 polymorph

Crystal Structure of Al_2O_3

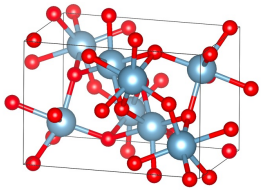
https://materialsproject.org



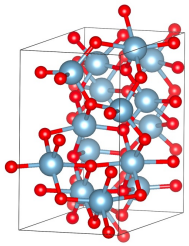
1 mp-1143_R-3c



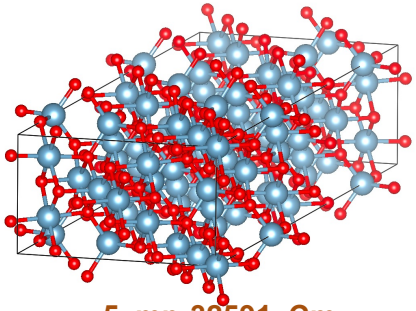
2 mp-1245063_P1



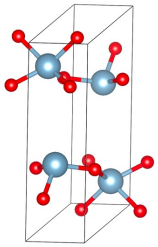
3 mp-1938_Pbcn



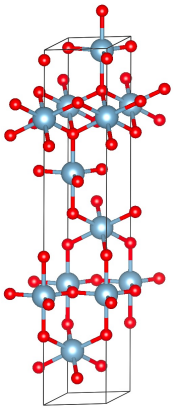
4 mp-2254_Pna2



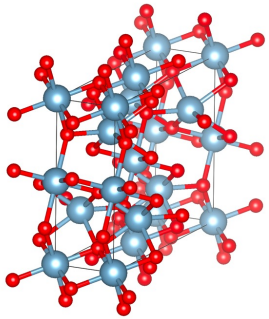
5 mp-32591_Cm



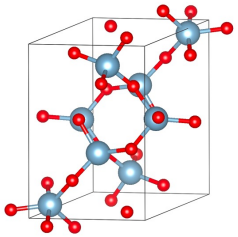
6 mp-638765_P-1



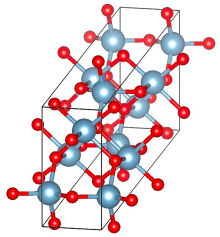
9 mp-755483_R3m



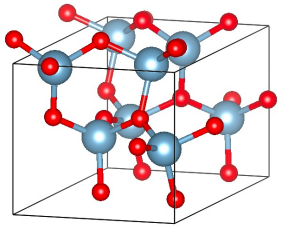
7 mp-642363_Cmcm



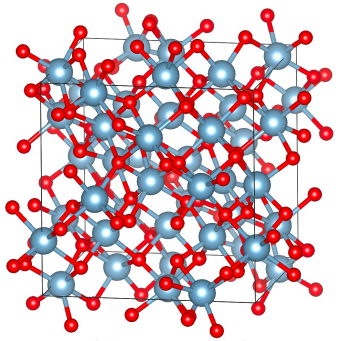
8 mp-754531_P21/c



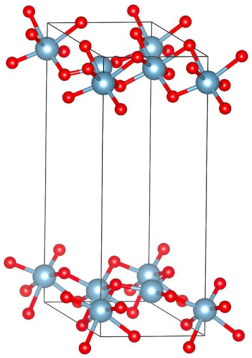
12 mp-7048_C2/m



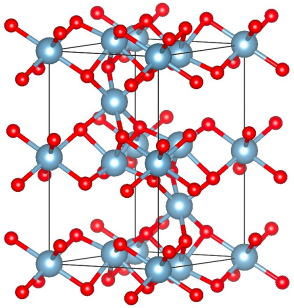
13 mp-754401_Cmc21



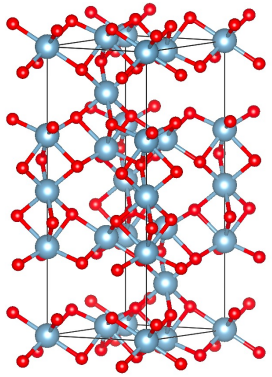
10 mp-776475_Ia-3



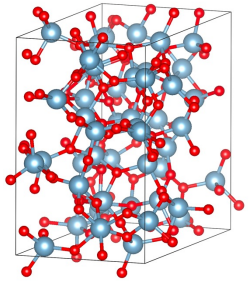
11 mp-985587_P321
artificial bilayer film



15 mp-755175_P-31c



14 mp-754624_R-3



a-pristine by Momida*

*: H. Momida *et al.*, The 81st JSAP Autumn Meeting 2020, 11p-Z07-13, 11 Sept. 2020.

Crystal Structure of Al_2O_3

	Materials Project	SG	HOF [meV/atom]	Gap [eV]	Comments
1	mp-1143	R-3c	0 (0)	5.599 (6.044)	most stable
2	mp-1245063	P1	+218 (+204)	3.137 (3.326)	amorphous model
3	mp-1938	Pbcn	+92 (+94)	5.047 (5.313)	
4	mp-2254	Pna21	+18 (+17)	4.592 (4.830)	
5	mp-32591	Cm	+63 (+57)	2.854 (3.075)	
6	mp-638765	P-1	+683 (+809)	1.119 (1.136)	
7	mp-642363	Cmcm	+266 (+281)	4.009 (4.249)	$C_{44}<0$, $C_{55}<0$
8	mp-754531	P21/c	+98 (+77)	4.141 (4.374)	
9	mp-755483	R3m	+93 (+91)	3.863 (4.098)	
10	mp-776475	Ia-3	+44 (+30)	4.982 (5.216)	In_2O_3 stable phase
11	mp-985587	P321	+208 (+210)	4.686 (4.807)	artificial film [excluded]
12	mp-7048	C2/m	+17 (+10)	4.239 (4.605)	Ga_2O_3 stable phase
13	mp-754401	Cmc21	+74 (+46)	4.021 (4.235)	
14	mp-754624	R-3	+72 (+66)	5.386 (5.797)	
15	mp-755175	P-31c	+79 (+72)	5.227 (5.494)	
	a-pristine	P1	+179	3.725	amorphous model*

data with parentheses:
https://materialsproject.org

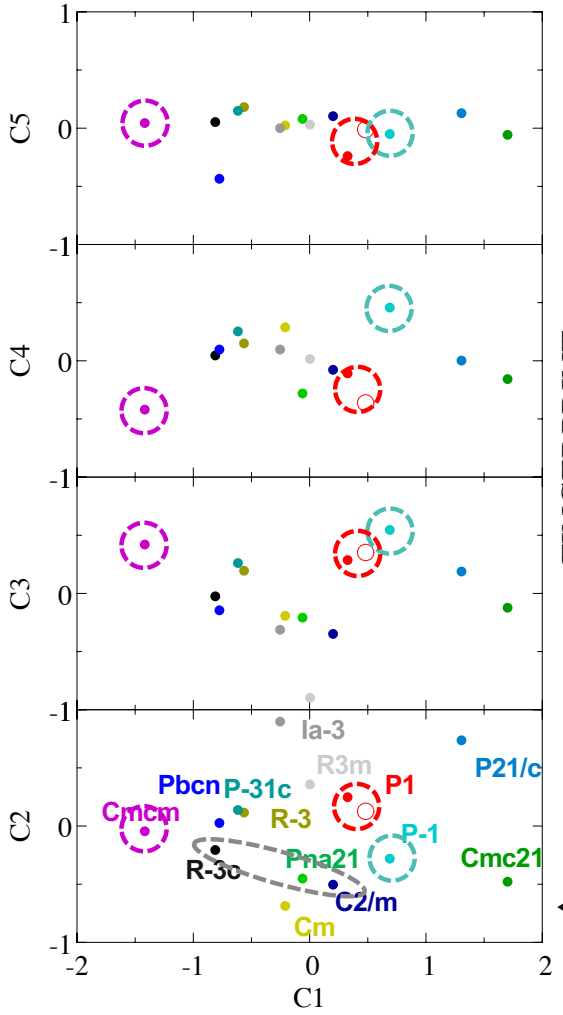
*: H. Momida *et al.*, The 81st JSAP Autumn Meeting 2020, 11p-Z07-13, 11 Sept. 2020.

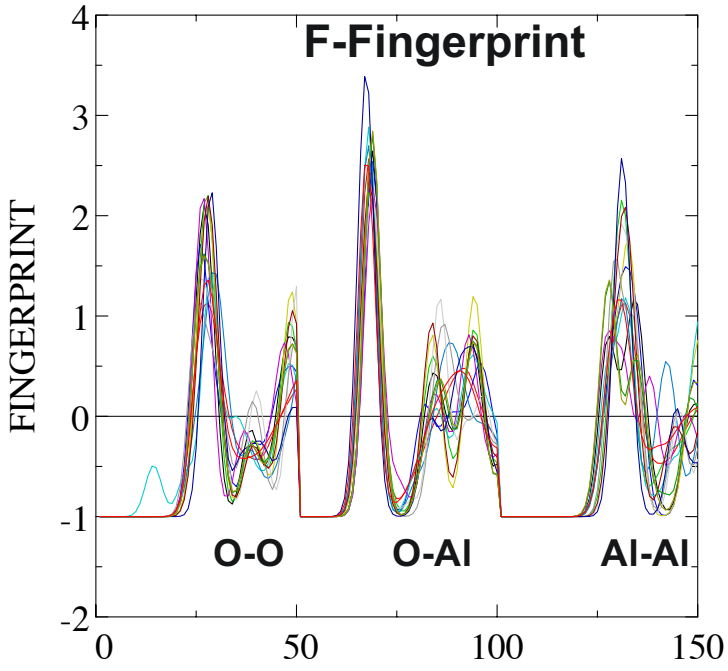
Where is the position of amorphous structure?

Structure Map of Al_2O_3 w/o P321

R_{max} = 5Å
Δ = 0.1Å
σ = 0.2Å

$$F_{AB}(R) = \underbrace{g_{AB}(R)}_{\text{RDF}} - 1$$





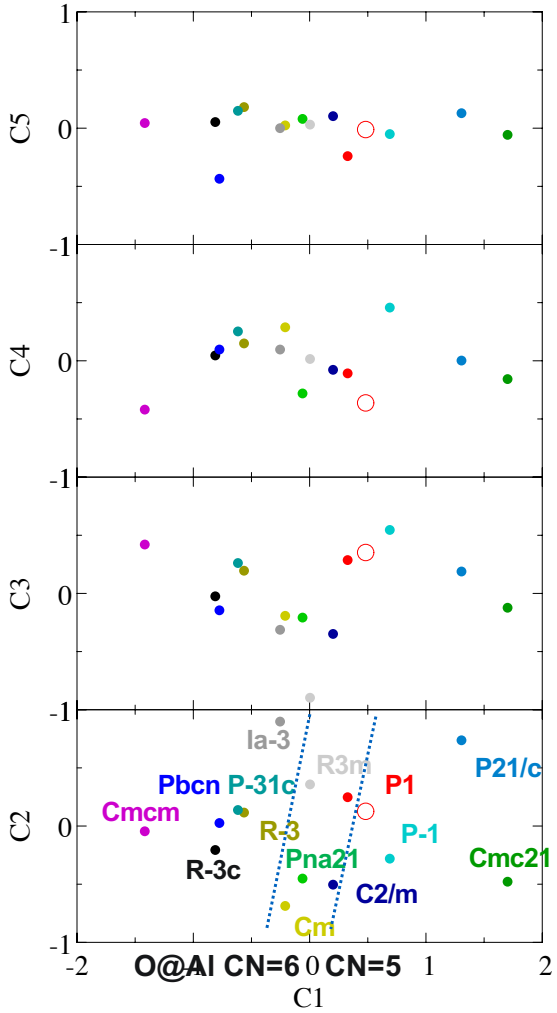
Distance as Similarity

Dimension Reduction

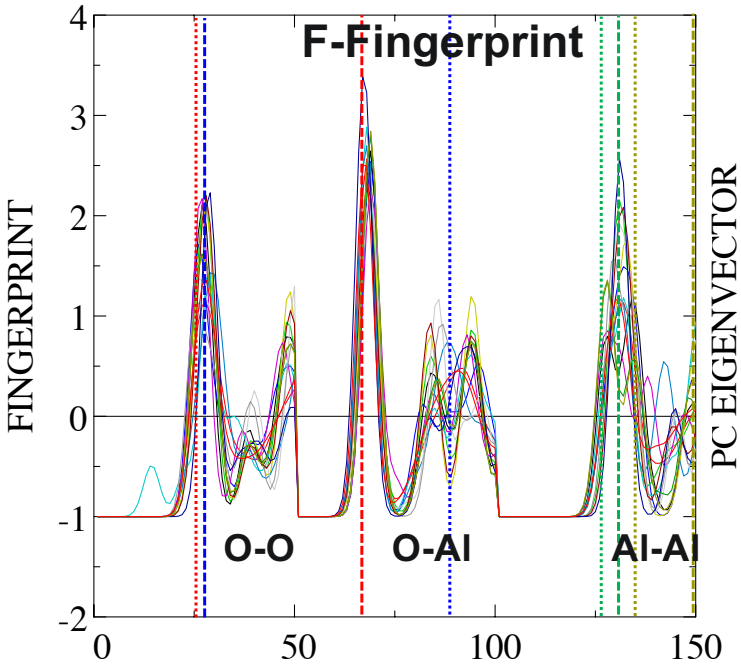
	System	SG	HOF (meV/atom)
1	mp-1143	R-3c	0
12	mp-7048	C2/m	+17
4	mp-2254	Pna21	+18
10	mp-776475	Ia-3	+44
5	mp-32591	Cm	+63
14	mp-754624	R-3	+72
13	mp-754401	Cmc21	+74
15	mp-755175	P-31c	+79
3	mp-1938	Pbcn	+92
9	mp-755483	R3m	+93
8	mp-754531	P21/c	+98
	a-pristine	P1	+179
2	mp-1245063	P1	+218
7	mp-642363	Cmcm	+266
6	mp-638765	P-1	+683

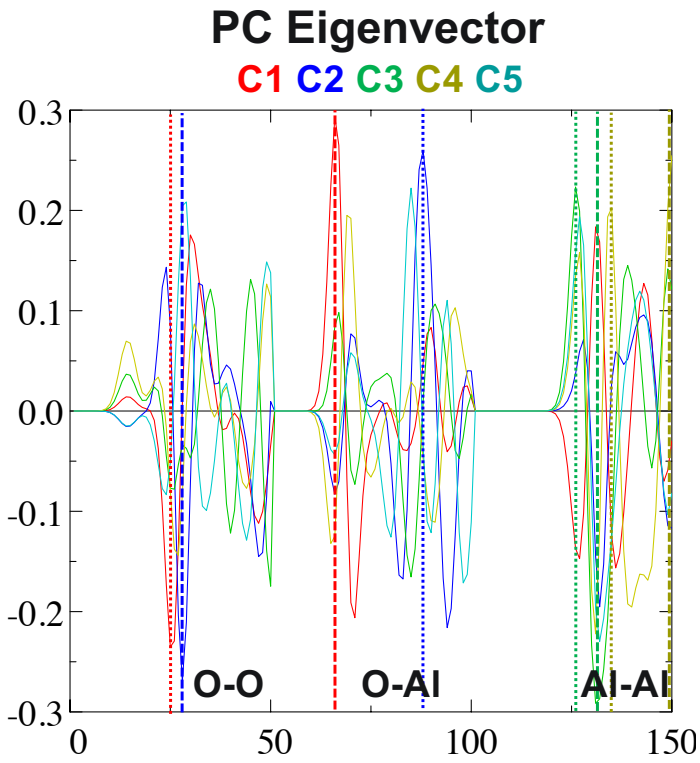
- Amorphous model structues (P1) can be distinguished from P-1 by the PC C4.

Structure Map of Al_2O_3 w/o P321



R_{max} = 5Å
Δ = 0.1Å
σ = 0.2Å

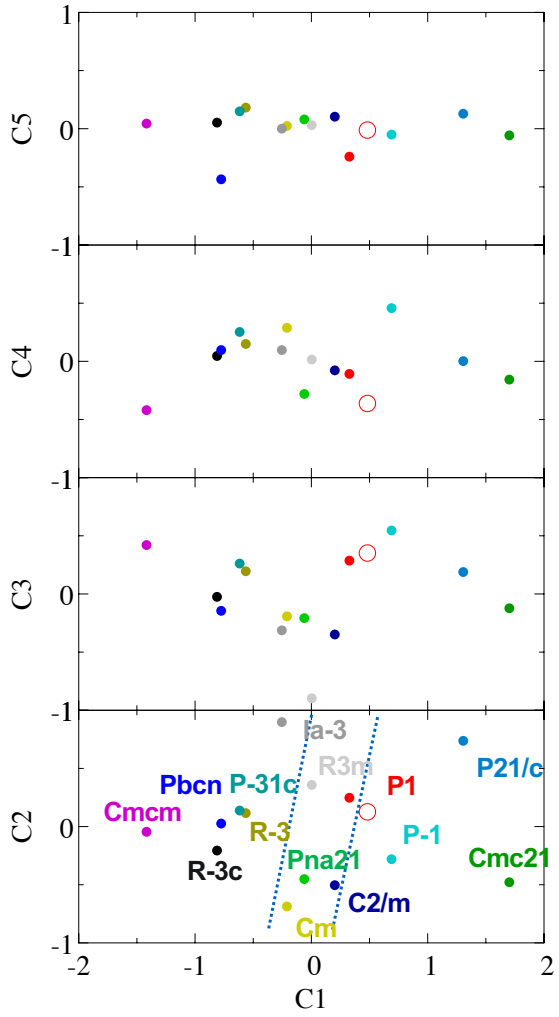




- The PCs C1 & C2 contain radial distribution information between neighboring O-Al and O-O while C3 & C4 do that between Al-Al.

Summary

- **Features of crystal structure are extracted by dimension reduction from the fingerprint distances.**
- **The principal components C1 & C2 may be related to the coordination numbers of O@Al and O@O, while C3 & C4 might be so to Al@Al.**
- **The features can be used as descriptors to model the total energies and energy gaps by regression analysis.**



TO: Sci. Technol. Adv. Mater.: Methods 4, 2355860 (2024).

計算機によって物性科学はどう変わるか

...インプットからアウトプットまでの膨大な計算の労力から解放されることによって、我々は何をインプットすべきかという創造的な部分と、何をアウトプットから読み取るかという判断力・想像力を要する部分とに専念することができる。...

Materials Informatics

豊沢 豊　日本物理学会誌1985