2019.09.06

Applications of Materials Informatics For High-productive Research:

Machine-learning potential for amorphous and spectrum-data analysis

Research Center for Computational Design of Advanced Functional Materials (CD-FMat), AIST

Senior Researcher

Yasunobu Ando

Self-Introduction

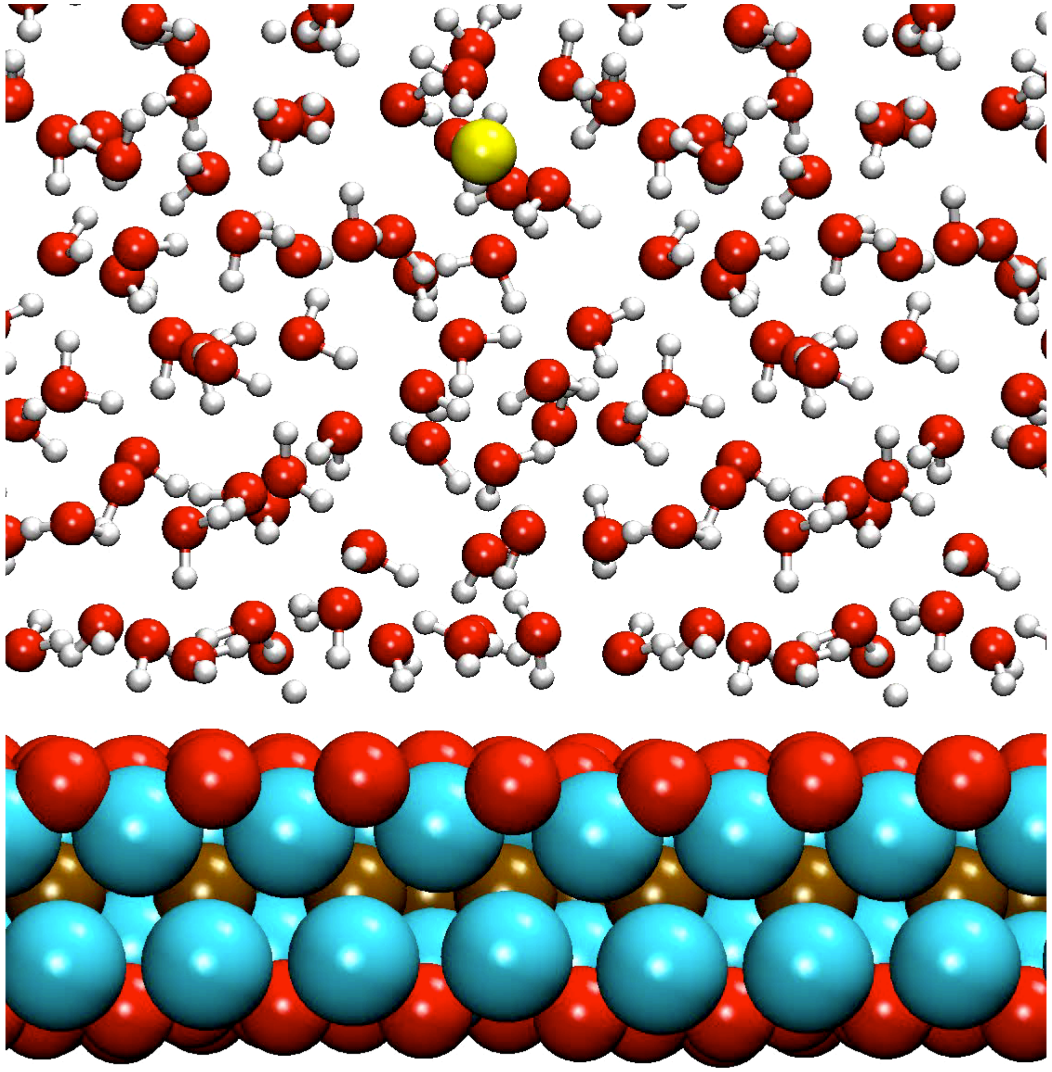
Name：Yasunobu ANDO（36）

CV

✓ 2012. 3.
Ph.D in physics @ Dept. Physics, UTokyo.
✓ 2012. 4 ~ 2013. 4
Posdoc @ AIST
✓ 2013. 5 ~ 2016. 3
Assistant professor @ Dept. Materials Eng., UTokyo
✓ **2016.4 ~**
Researcher @ CD-FMat,　AIST

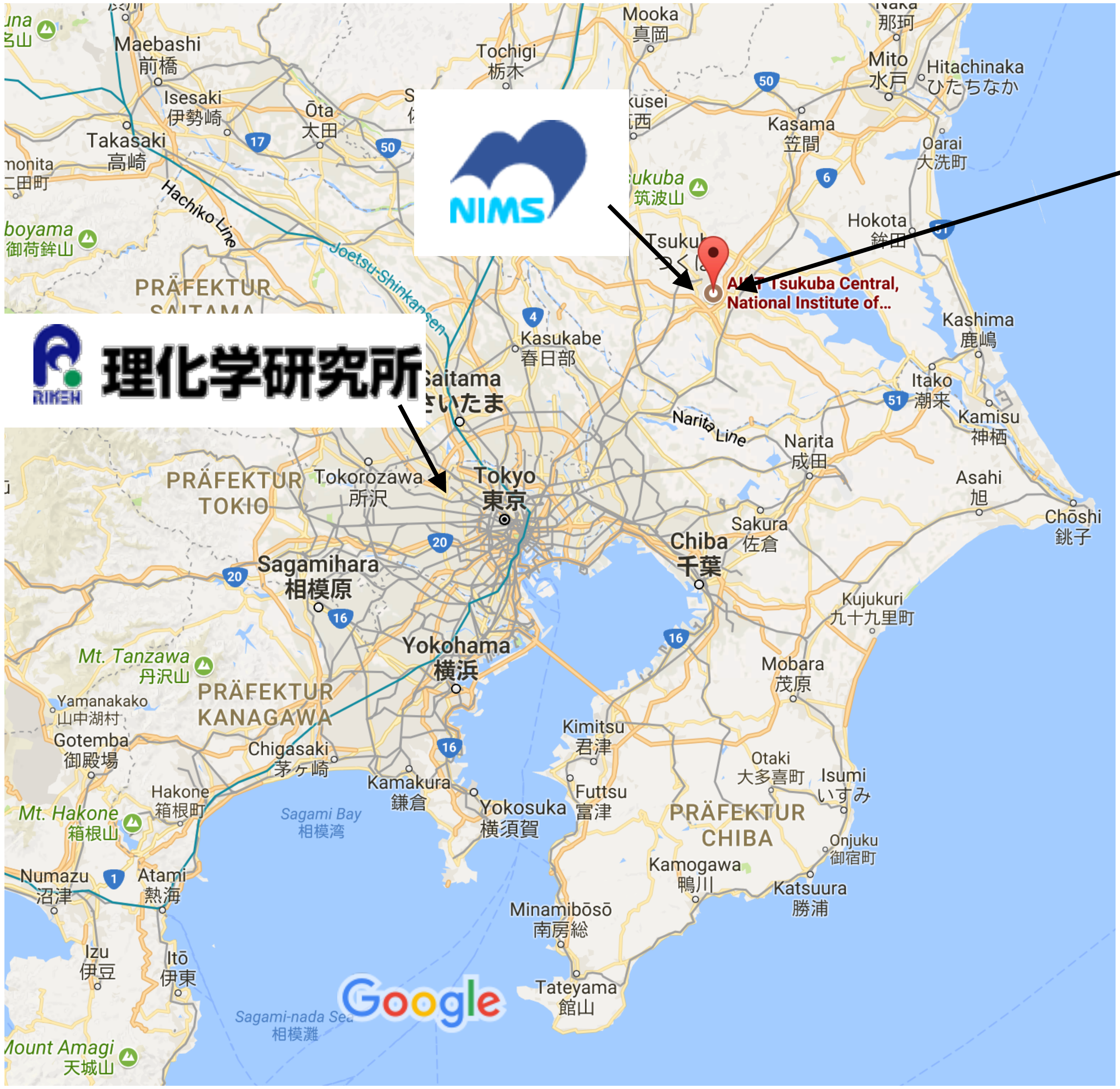
Major topics

Materials x Informatics
(Physics・Chemistry・Informatics)



***Ab initio* MD**
in Solid-Liquid Interfaces

Place where I come from



AIST

NATIONAL INSTITUTE OF
ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY (AIST)

● Researchers (foreign nationals)······	2,284(116)
[Permanent]	[1,925]
[Fixed term]	[359]
● Administrative employees (foreign nationals) ·········	686(1)
Total number of employees :	2,970(117)
● Executives (full time)···························	13
● Visiting researchers ·························	185
● Postdoctoral researchers ··················	190
● Technical staff ·······························	1,487

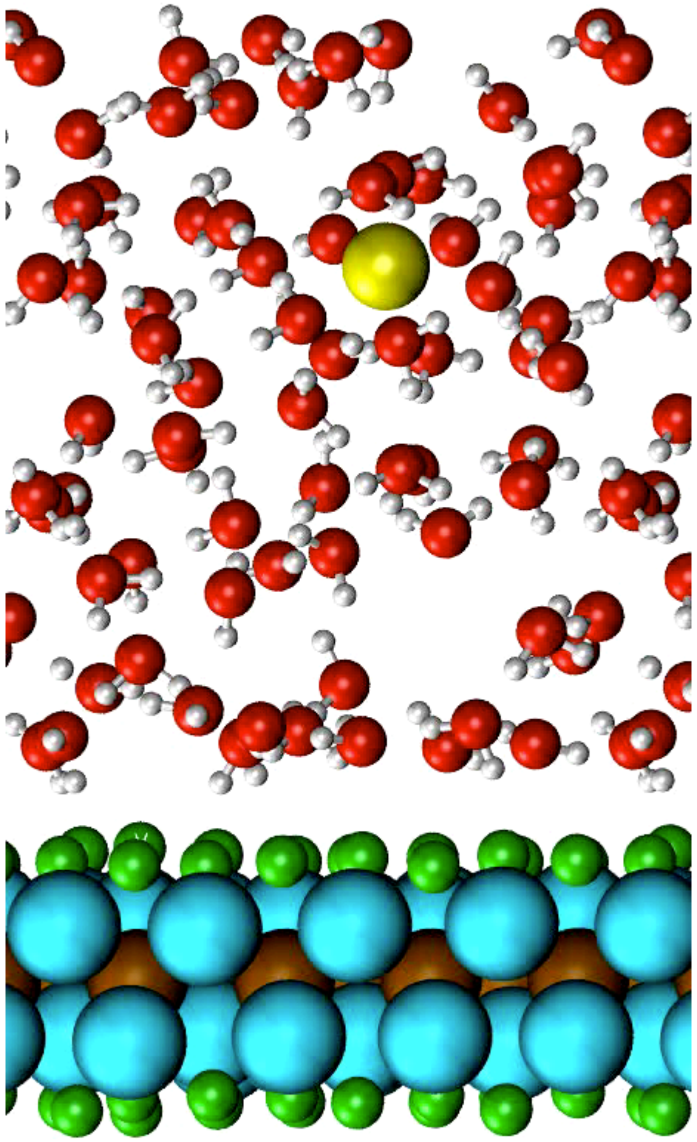
(As of July 1, 2016)

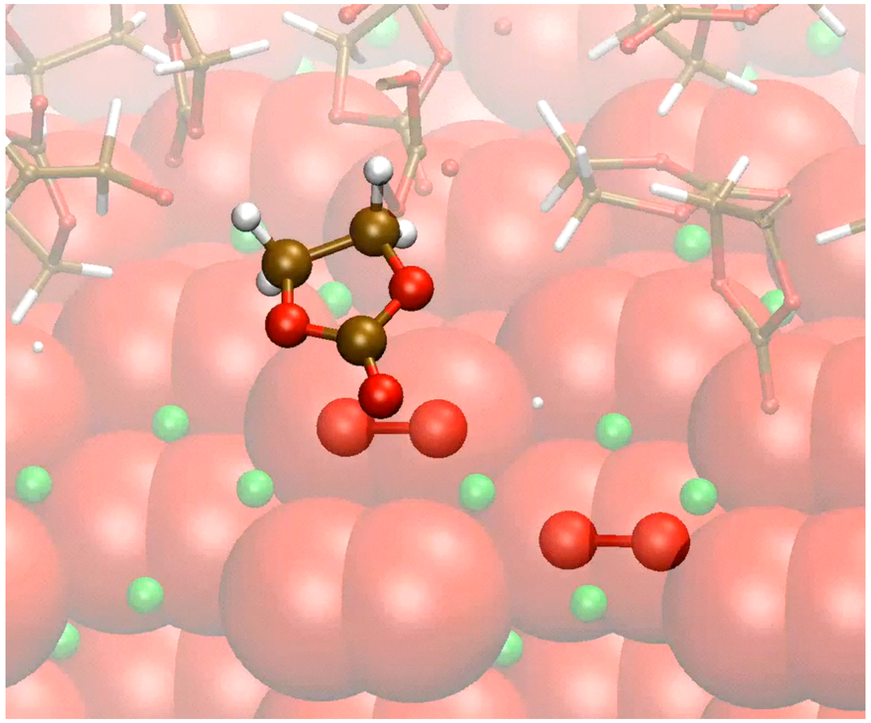
Number of researchers accepted through industry/academia/government partnerships

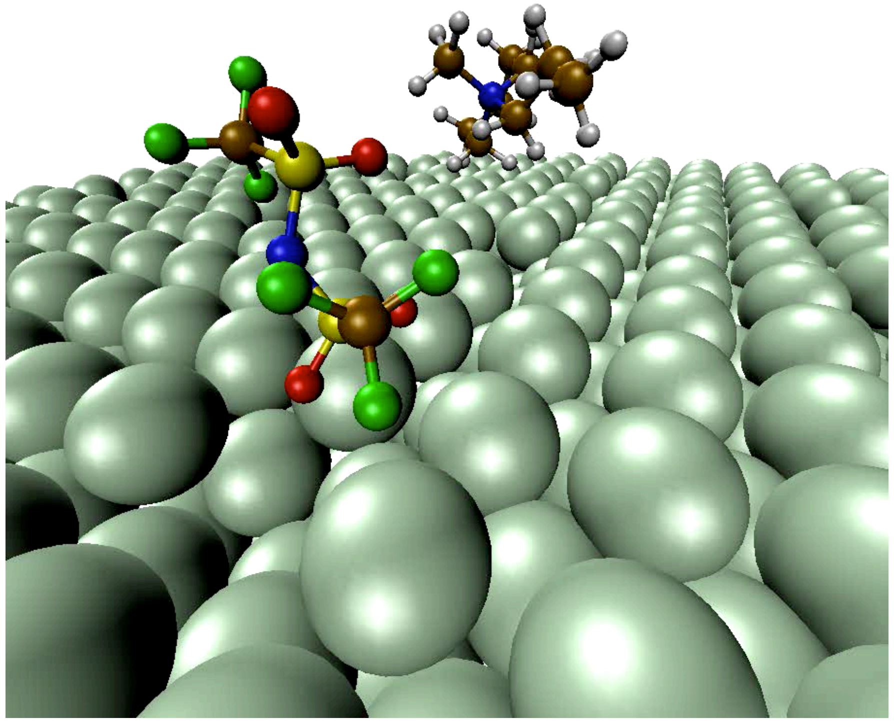
● Companies ··································	1,856
● Universities ·······························	1,924
● Other organizations ·······················	936
(foreign nationals :	456)

(Total number of researchers accepted in FY 2015)

Major research topics





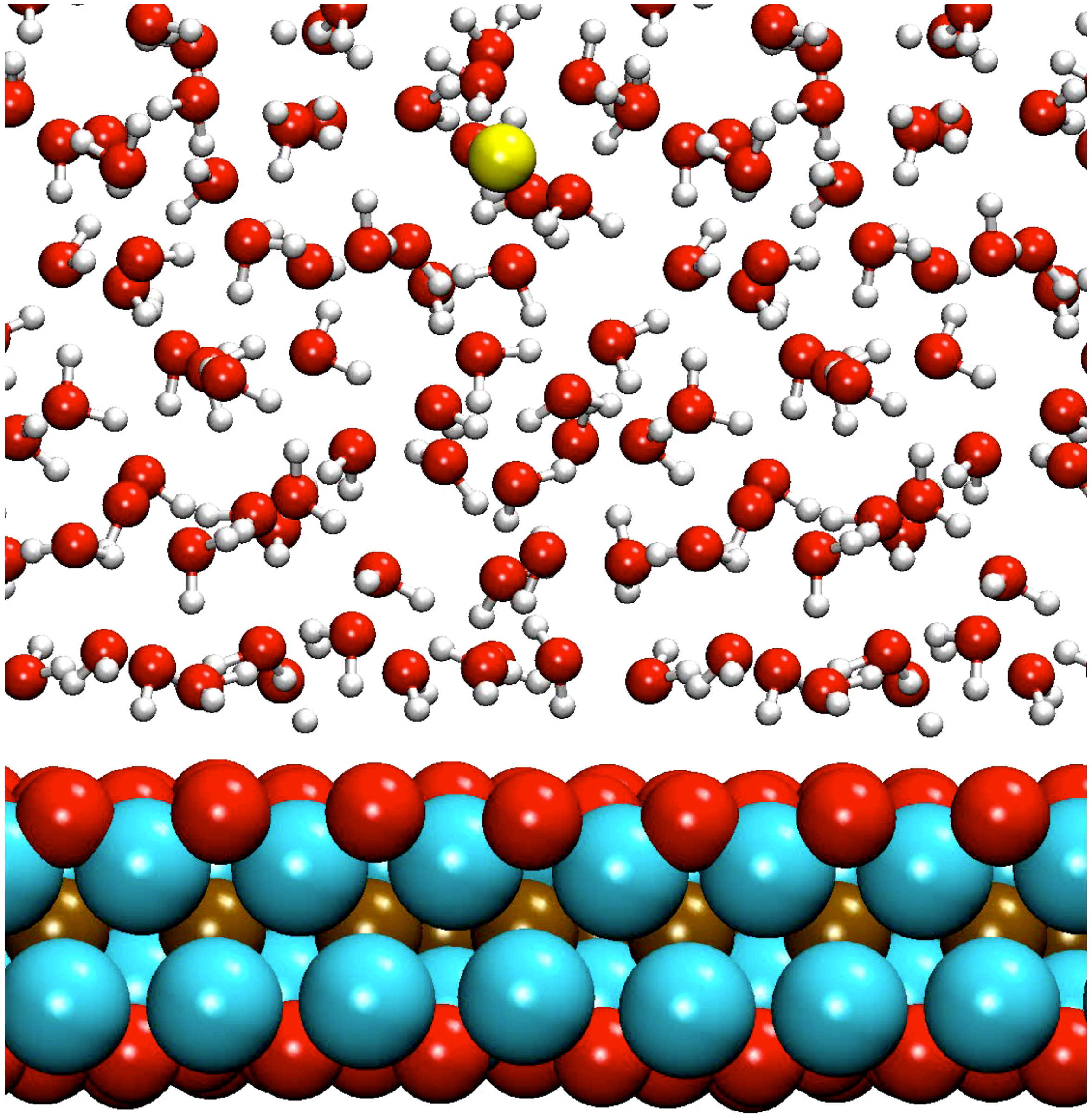


<u>My interests:</u>

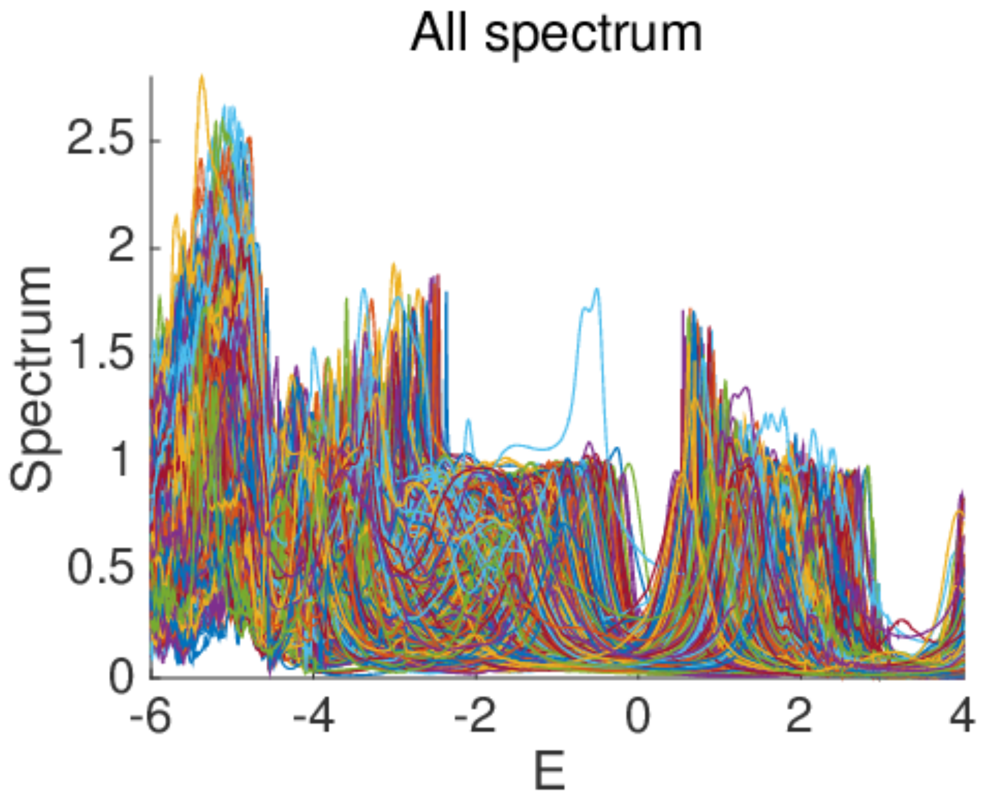
- Electric-double layer
- Reaction on surface and interfaces.

Tool: Density-functional-theory based (MD) simulation.

We need "NEW" tools



FPMD of electrochemical Interfaces



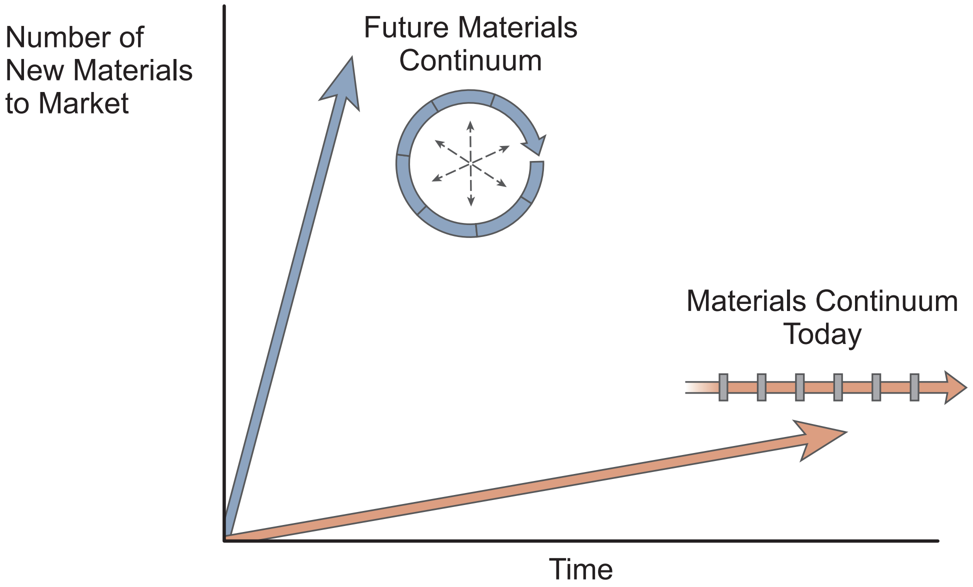
Spectra-data sets including 600 samples

Now we "can" obtain hundreds thousands spectra from simulation

Efficient Analysis is necessary.

Milestone project with Informatics

Materials Genome Initiative（MGI from 2011）



Main purpose

1. **Developing a Materials Innovation Infrastructure**
2. **Acheiving National Goals With Advanced Materials**
3. **Equipping the Next-Generation Materials Workforce**

→ **Using Database and machine-learning**

Our research in MI

1. Clustering and correlation analysis of transmission spectrum of molecular junction system

2. A new descriptor of perovskites for performance estimation of fuel cells

3. Materials Search for a well-worked 2D substrate for Germanene and Stanene

4. Unsupervised clustering of PDOS in Surface system

5. Development of machine-learning potentials for Amorphous research

6. Yield prediction in experiments from Simulation (Catalyst informatics)

7. High-throughput peak fitting on many XPS spectra

8. Model selection of equivalent circuits on impedance spectroscopy

9. Model selection of preferred orientation distribution of tourmaline-grains

10.Parameter optimization of equation of states for an inner earth environment

So, what is informatics?
What is machine-learning?

Short introduction

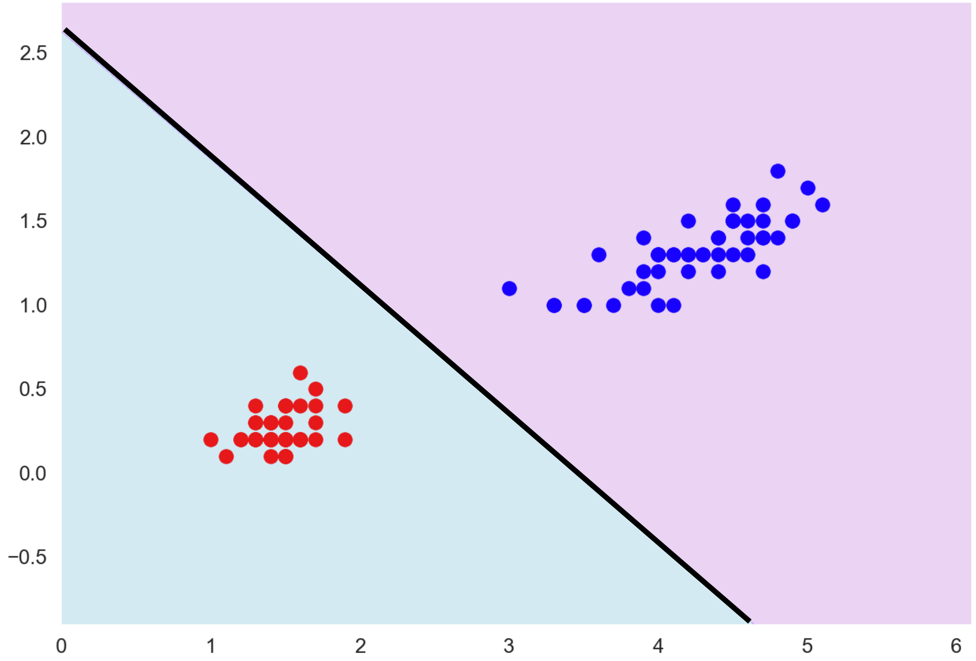
Functions of ML

- ✓ Prediction
- ✓ Characterization
- ✓ Classification
- ✓ Pattern Recognition

Linear Regression (PCA)

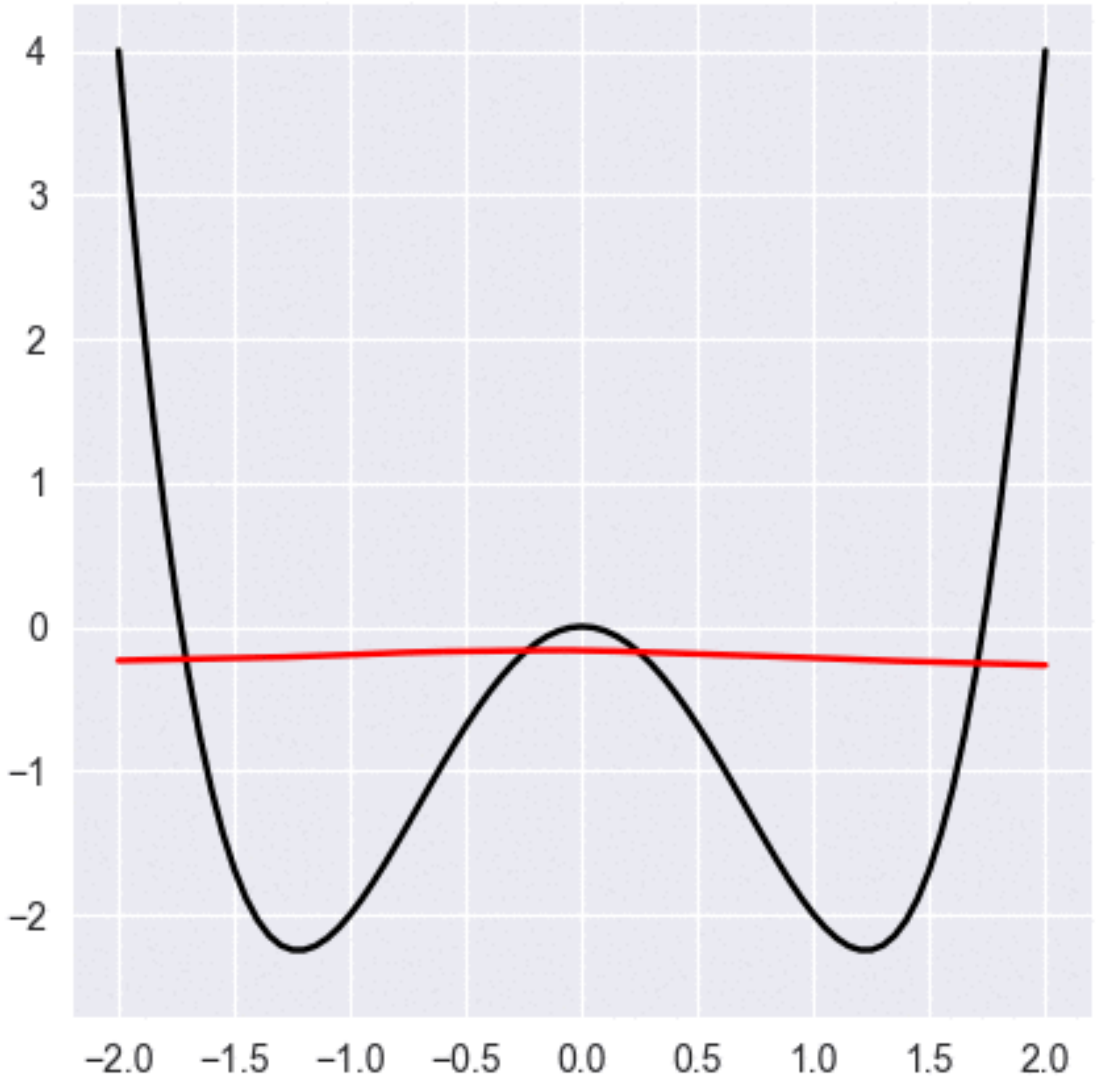
$$Y = \sum_i a_i X_i + b$$

Support Vector Machine (SVM)



What they do is just "Putting points and Drawing lines"

Representing Complex Situation



Fitting process by Neural network

NOT work lines ?
JUST draw the "curve"

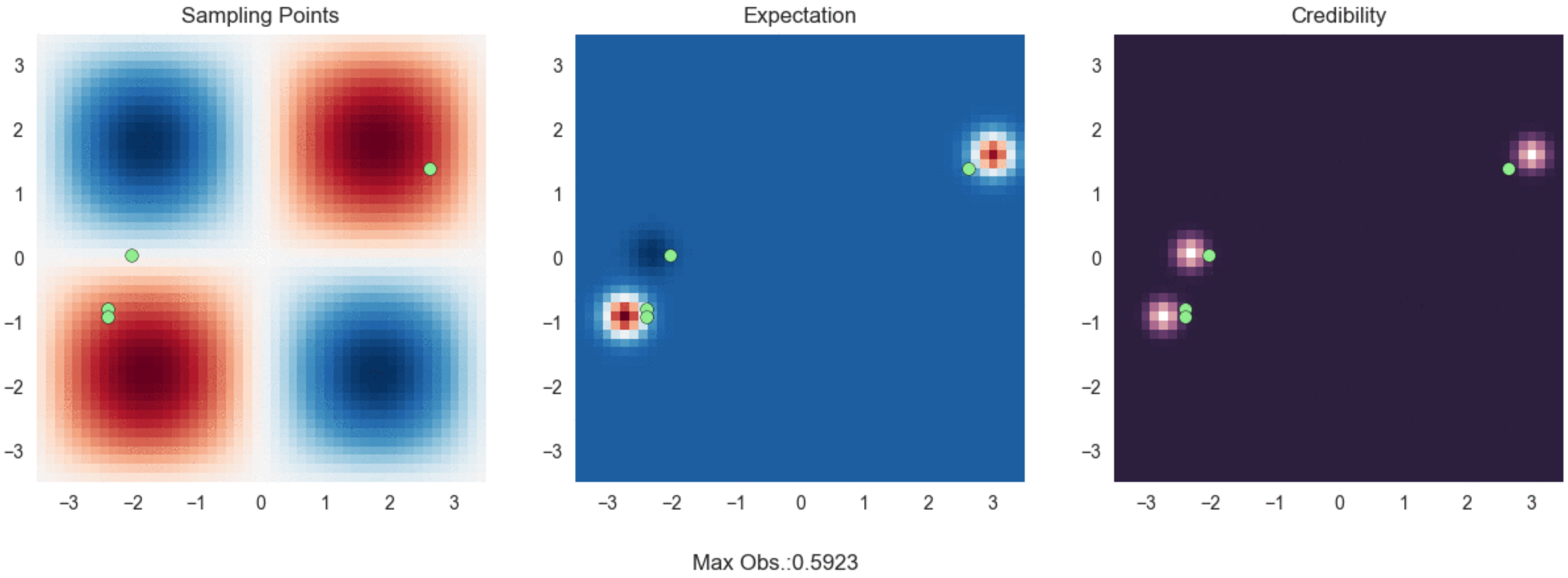
⬇

- ✓ Basis expansion (polynomial, spline etc.)
- ✓ Kernel Regression, Gaussian Process
- ✓ Neural Network

Applicable to high-dimensional data

Exploration and Exploitation

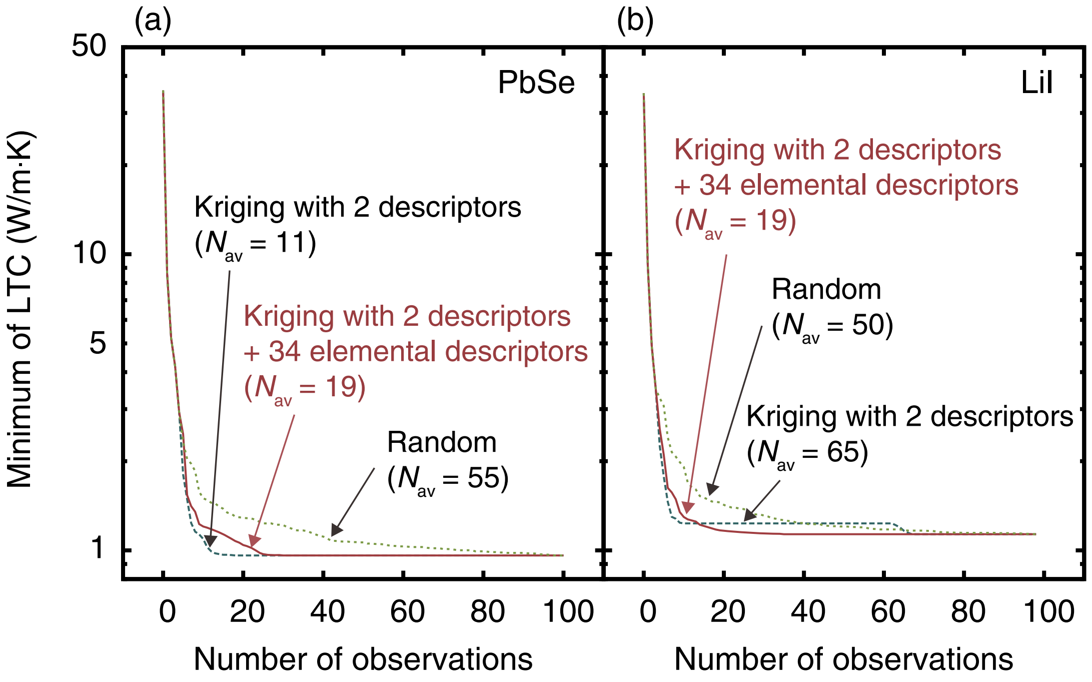
Bayesian Optimization



✓ *Based on the observation, predicting values and its "credibility" at unobserved points*

✓ *Automatically searching observable space considering predicted value and credibility.*

Material Search with Bayesian Opt.



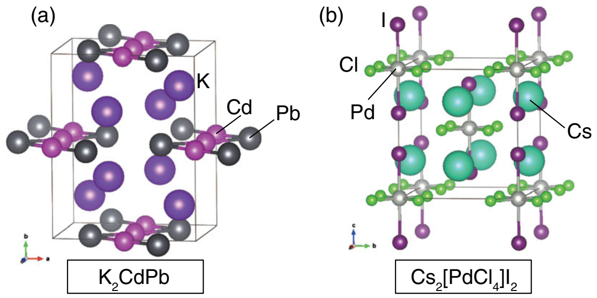


FIG. 4 (color online). Crystal structures of K_2CdPb and $Cs_2[PdCl_4]I_2$ predicted to show the low LTC of < 0.5 W/mK (at 300 K) and narrow band gap of < 1 eV.

Thermoelectric material

PRL **115**, 205901 (2015).

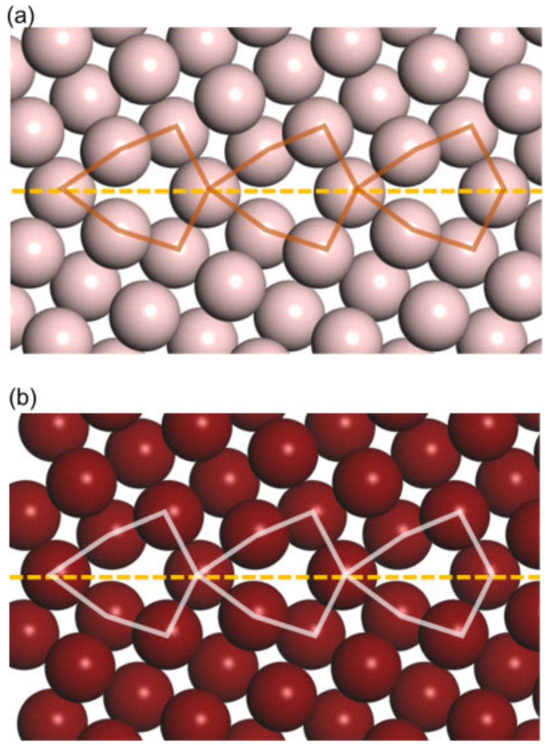


Fig. 2. (Color online) Stable atomistic arrangements of Σ5[001]/(210) with (a) the conventional method and (b) the kriging method. Dashed line represents the position of the GB. Polyhedron shows the six-membered structure unit.

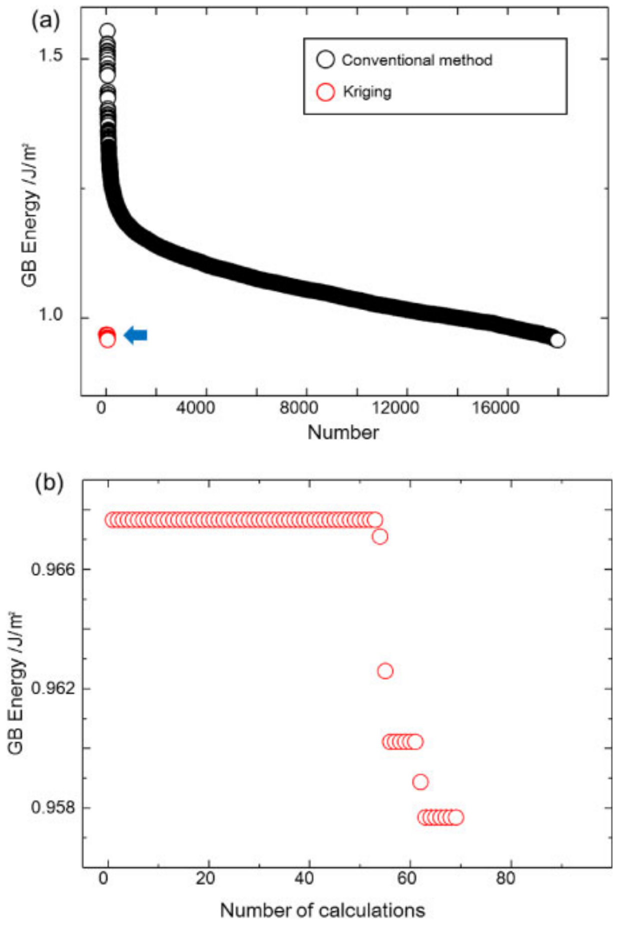


Fig. 4. (Color online) (a) Trajectory of the calculated GB energy to the convergence. The "conventional method" means the computation of all configurations. To understand the calculation efficiency easily, the calculated result by the conventional method (black) was plotted by order of the GB energy, whereas it was plotted by order of the trial number in the kriging method (red). (b) Magnified image around the area pointed by blue arrow in the (a).

Interface matching

Jpn. J. Appl. Phys. **55** (2016) 2–6.

Science Adv., **2**, e1600746-1-7.

Appl. Phys. Lett. **91**, 132102 (2007). Appl. Phys. Rev. 4, 011105 (2017).

High-Throughput Observation

High-Throughput Experimental (HTE) methodologies

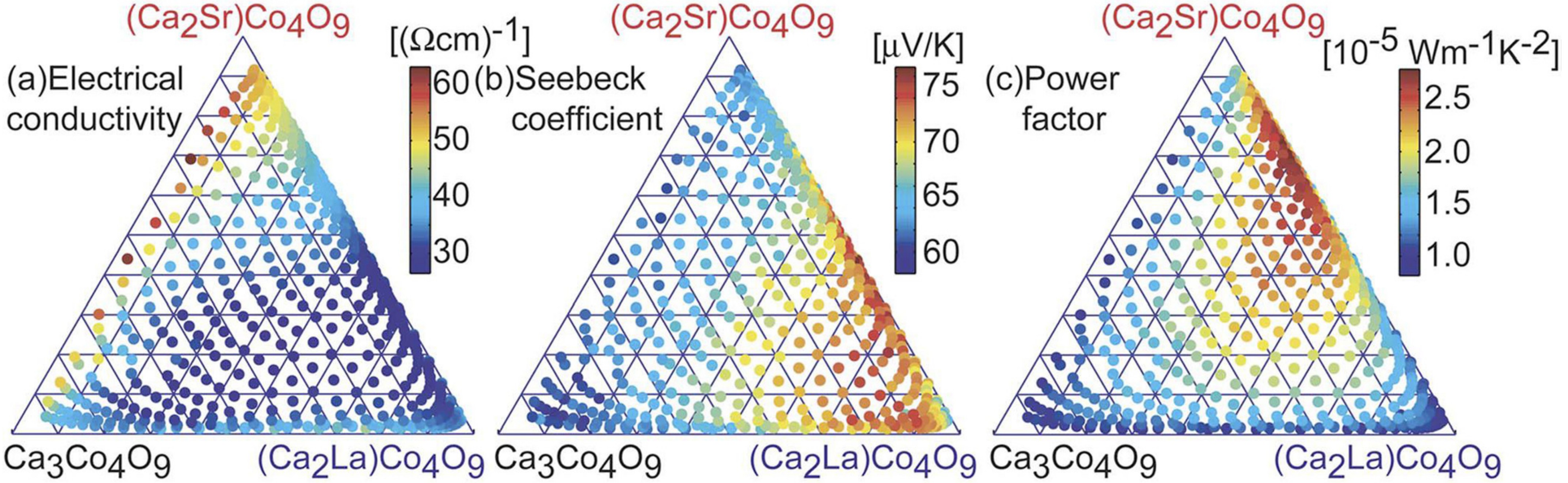


FIG. 2. (a) Electrical conductivity, (b) Seebeck coefficient, and (c) power factor of the composition-spread $(Ca_{1-x-y}Sr_xLa_y)_3Co_4O_9$ film ($0 < x < 1/3$ and $0 < y < 1/3$). Reproduced with permission from Appl. Phys. Lett. **91**, 3 (2007). Copyright 2007 AIP Publishing LLC.[56]

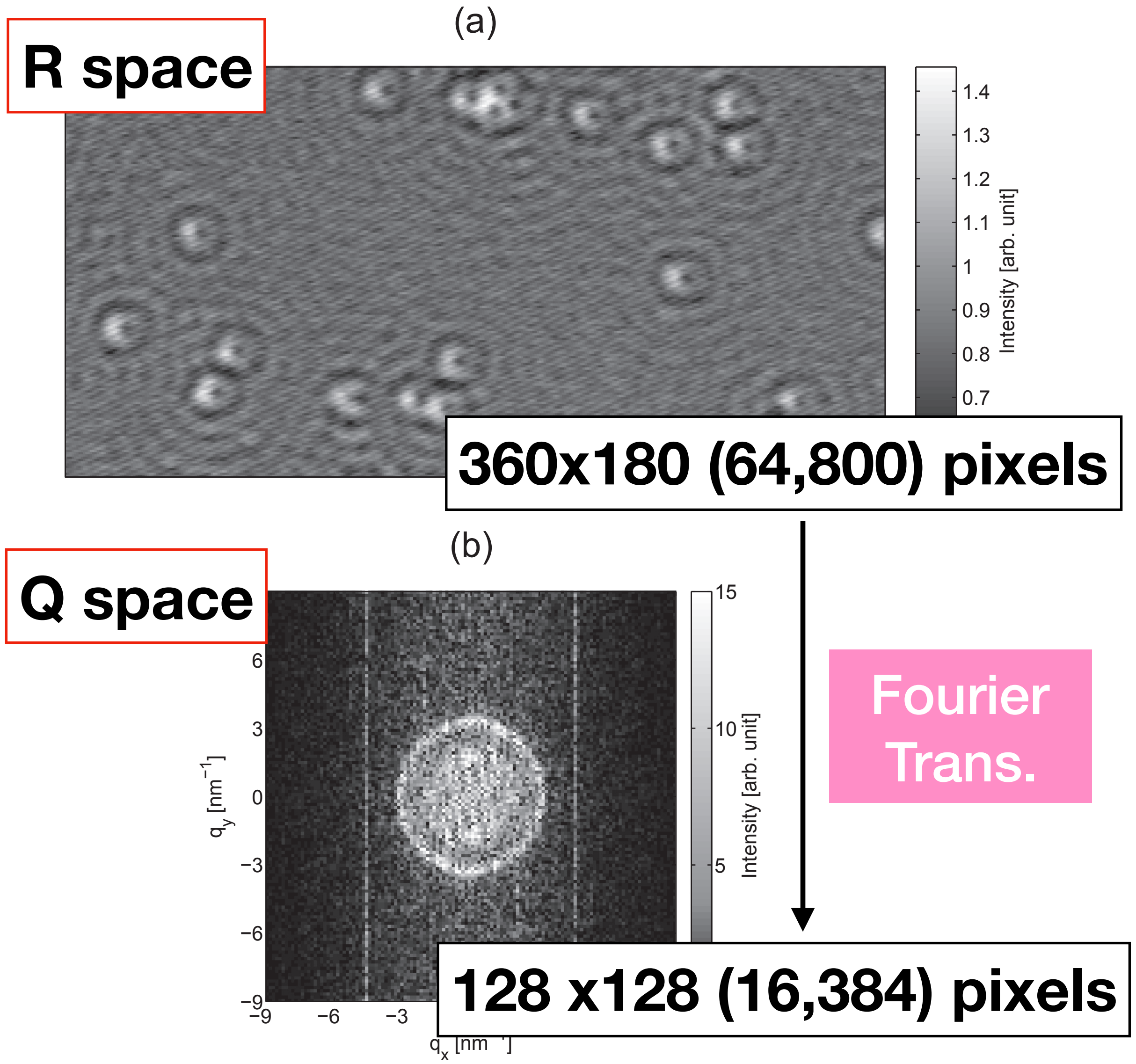
Materials "Library" : Dispersing the several compositions on a single sheet

Observing Big Materials Space at once, Finding optimum one

Y. N.-Ohno, M. Haze, Y. Yoshida, K. Fukushuma, Y. Hasegawa, and M. Okada, J. Phys. Soc. Jpn. 85, 093702 (2016).

Downsampling method for quasi-particle Ohs.

Quasi-particle interference (QPI) on Ag(111) surface



R space

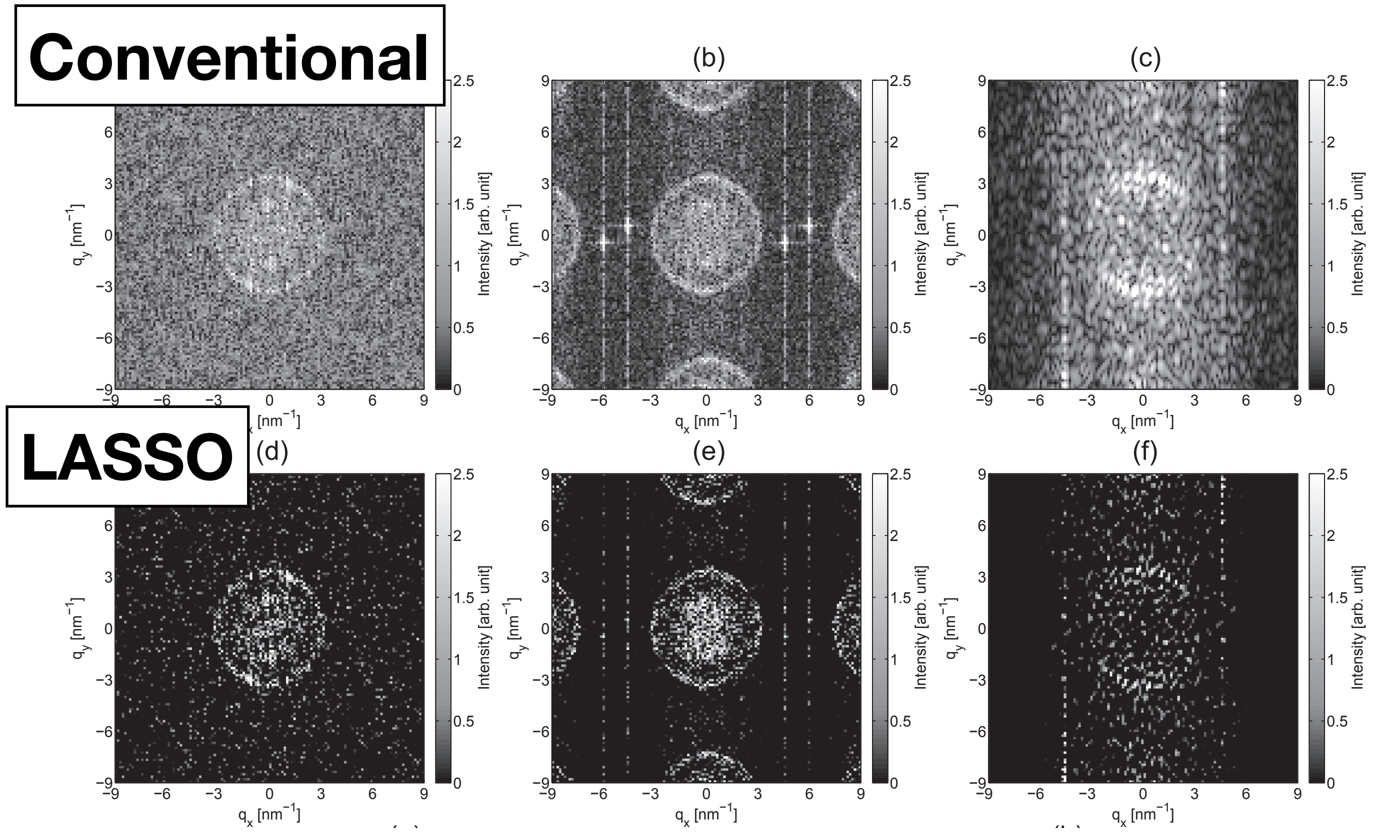
360x180 (64,800) pixels

Fourier Trans.

Q space

128 x128 (16,384) pixels

Fig. 1. (a) dI/dV map of Ag(111) surface. (b) FT of (a) obtained by conventional method.



Conventional

LASSO

Random sampling

Every third pixels

Using central region

Q space is really sparse

Sampling in R space can be reduced.

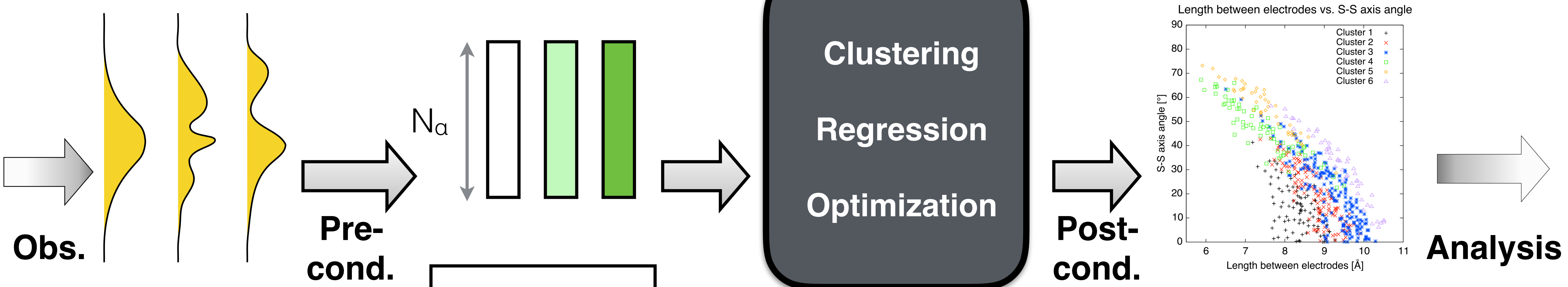
Basic flow chart of Application of MI

Always Start with an Issue

The most important

- ✓ What do you want to know from data?
- ✓ What benefits are obtained by applying machine-learning?
- ✓ Can you breakdown the issue enough to solve?
- ✓ **NEVER** just USING the machine-learning.

Basic Flow



Raw Data

Low-dim. Descriptors

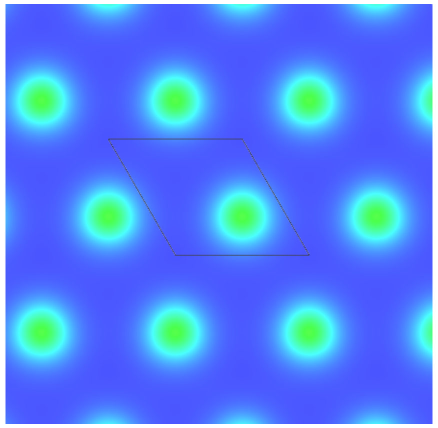
ML

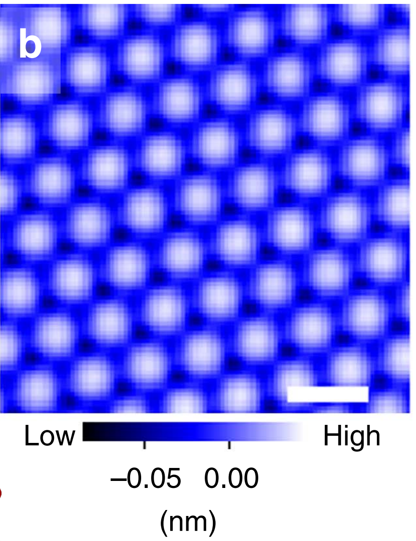
Visualization

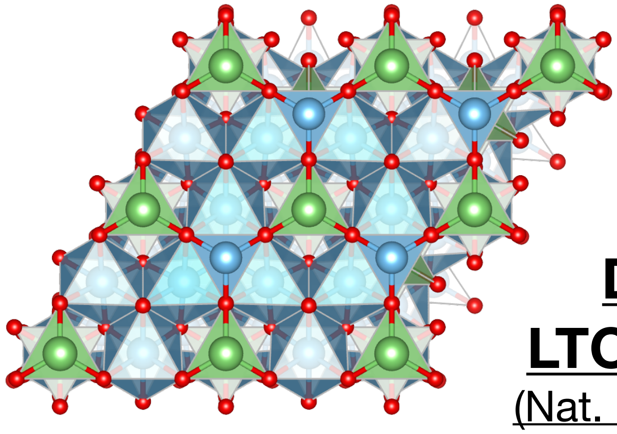
Machine-learning potentials

Molecular dynamics simulations

Structure Optimization







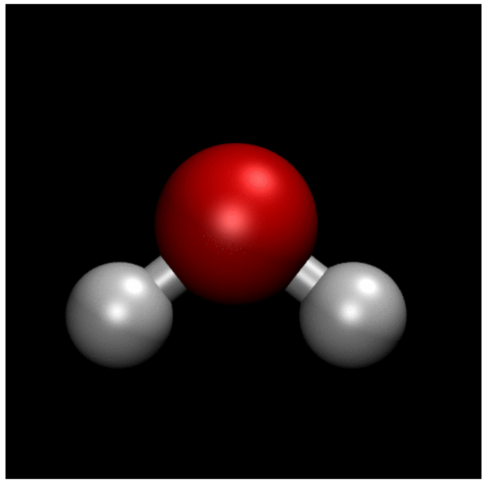
Determining the LTO surface structure

(Nat. Commun. 8, 15975 (2017))

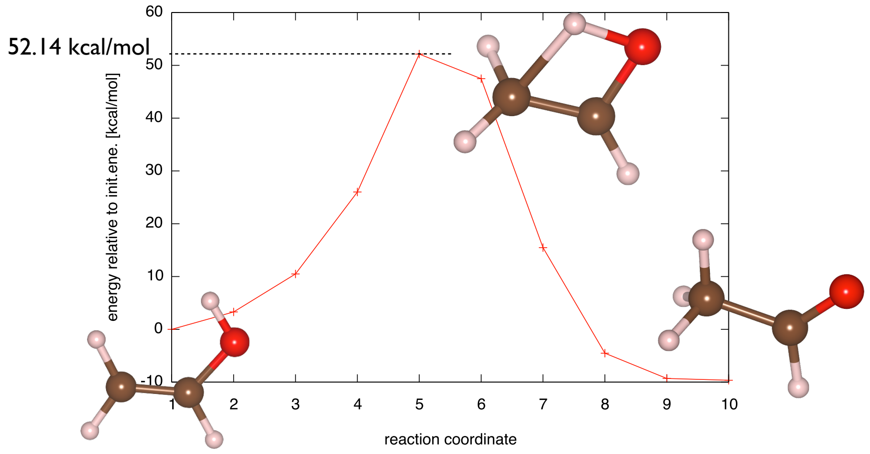
- From stable structure to dynamical behaviors
- Must tools for research in nano sciences

We need to model interaction between atoms

Normal modes and transition states

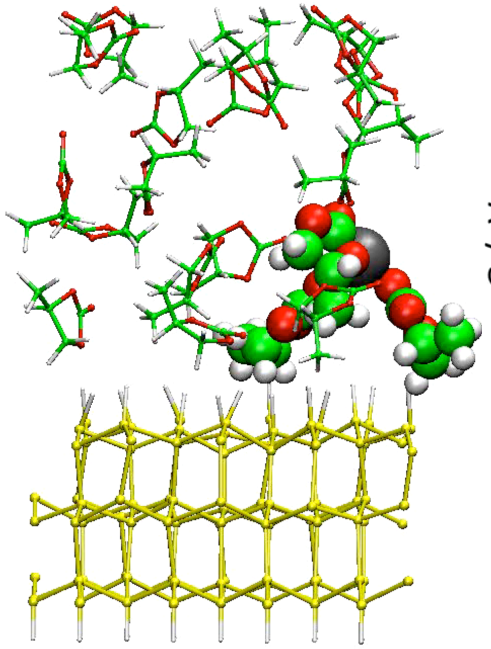


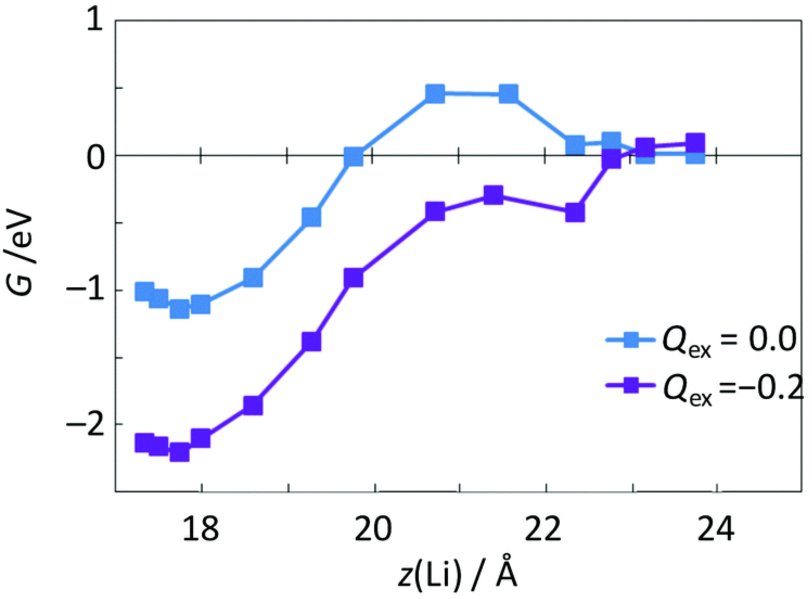
vinyl alcohol to acetaldehyde (NEB method)



DMol3 TS search (LST/QST) 51.473 cal/mol, -10.851 cal/mol (PBE)

Sampling for Free energy





Phys. Chem. Chem. Phys., 2018, 20, 11586-11591

Potential modeling by fitting

TTAM (Tsuneyuki-Tsukada-Aoki-Matsuda) Potential

Parametrizing potentials by fitting on Hartree-Fock calculations of SiO_4^{4-} + 4 point charge model

Phys. Rev. Lett. **61**, 86 (1988).
Nature 339, 209 (1989).

- **Dividing the electrostatic interaction to long-range, short-range parts.**
- **Fitting on cluster model, applying bulk.**
- **Reproducing SiO_2 polymorphs**

$$U_{ij}(r) = U_{ij}^{\text{Coulomb}}(r) + f_0(b_i+b_j)\exp[(a_i+a_j-r)/(b_i+b_j)] - c_ic_j/r^6,$$

$$U_{ij}^{\text{Coulomb}} = \tilde{Q}_i\tilde{Q}_j[1-g_{ij}(r)]/r + Q_iQ_jg_{ij}(r)/r,$$

$$g_{\text{SiO}}(r) = (1+\zeta r)\exp(-2\zeta r), \quad g_{\text{OO}}(r) = [1+11(\zeta r)/8+3(\zeta r)^2/4+(\zeta r)^3/6]\exp(-2\zeta r).$$

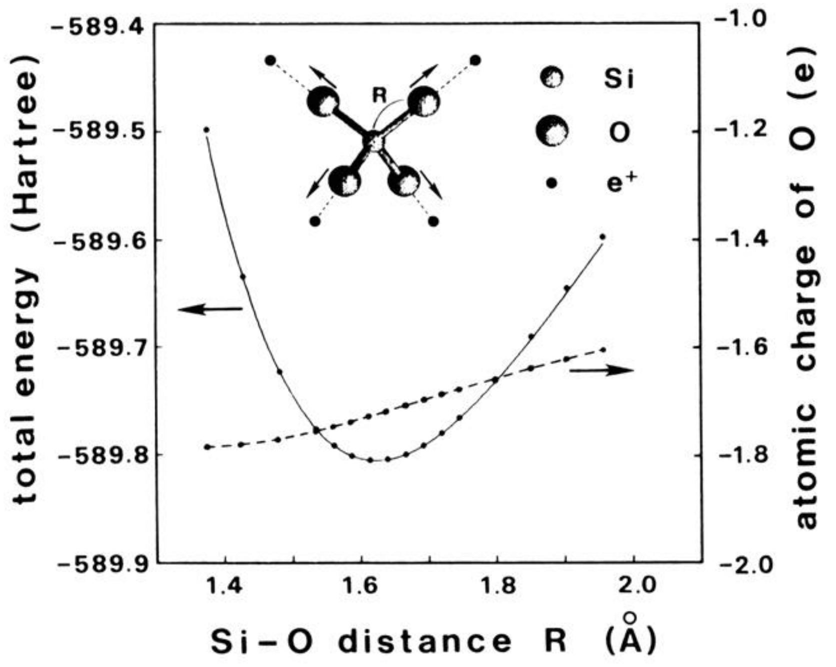


FIG. 1. Total energy and the Mulliken charge on an oxygen atom for T_d deformation of a SiO_4^{4-}-$4e^+$ cluster shown in the inset. The solid circles are the cluster calculation, full curve is the fitted potential, and the broken curve is a guide to the eye.

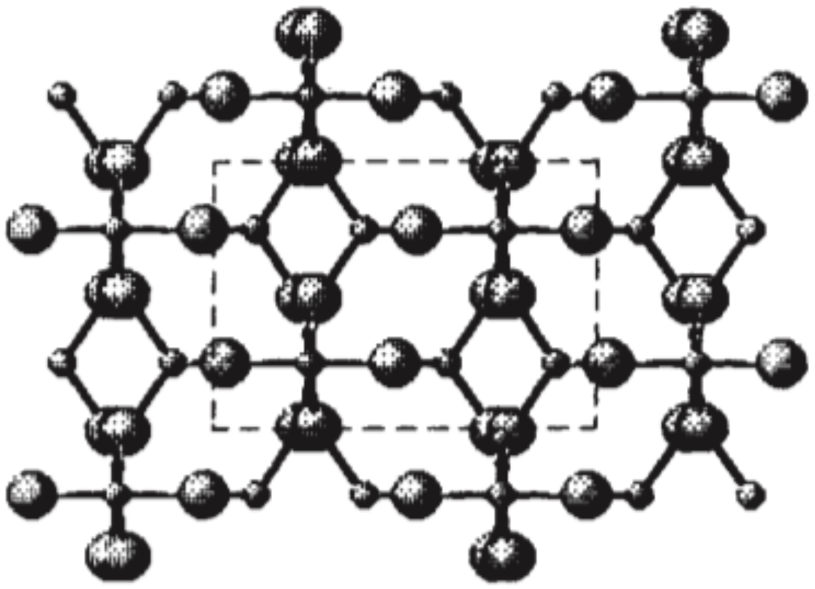


FIG. 2 The new *Cmcm* phase predicted by this study, involving both four- and sixfold coordination, as seen from the *a* axis, which corresponds to the *c* axis of low-cristobalite and stishovite. Small spheres represent silicon, and large spheres represent oxygen. The unit cell is shown by dashed lines.

Discovery of new SiO_2 high-pressure phase

[1] J. Chem. Phys. **103** (10), 8 (1995) .

Problem of NNP modeling

In 1995, NN fitting for interatomic potential has been reported [1]. But…

Construction of proper and simple descriptors

e.g.) Descriptor based on Fourier expansion on fcc(111) surface structure[2]

[2] J. Behler, S. Lorenz, and K. Reuter, J. Chem. Phys. **127**, 014705(2007)

HOWEVER… Such an expansion is needed to construct one by one and is not is not applicable to a complex surfaces and.

3-dimensional coordinates are not appropriate.

- Input layer depends on the total amount of atoms
- Impossible to apply for extended systems
- Permutation symmetry for the same particles are broken
- The other symmetries are not considered

Descriptor, symmetry are the keywords.

Behler-Parrinello Ansatz

1. Introduction of Symmetry Functions describing local environments

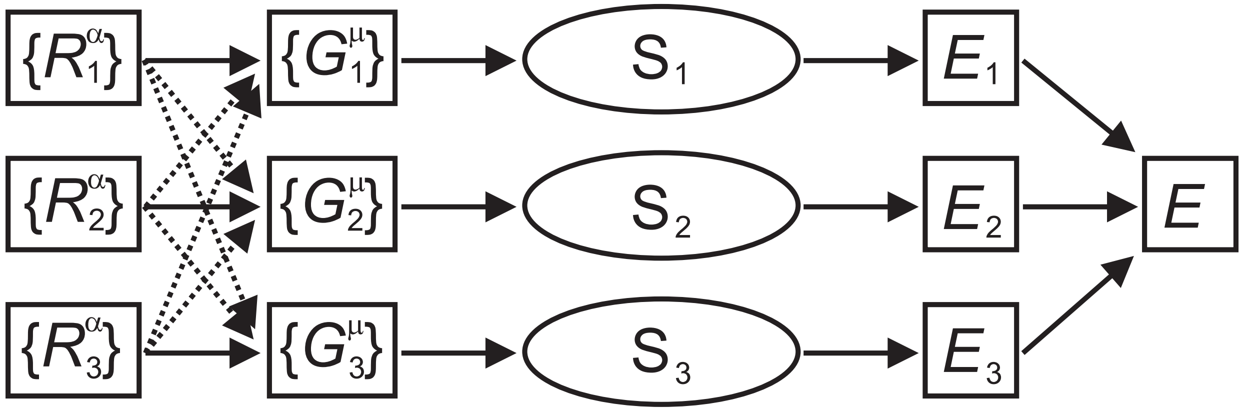
$$G_i^1 = \sum_{j\neq i}^{\text{all}} e^{-\eta(R_{ij}-R_s)^2} f_c(R_{ij}), \qquad (2)$$

$$G_i^2 = 2^{1-\zeta} \sum_{j,k\neq i}^{\text{all}} (1+\lambda\cos\theta_{ijk})^\zeta \, e^{-\eta(R_{ij}^2+R_{ik}^2+R_{jk}^2)} f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}), \qquad (3)$$

- Indexing the local environments based on lengths and angles
- They have a invariance for translational and rotational operations

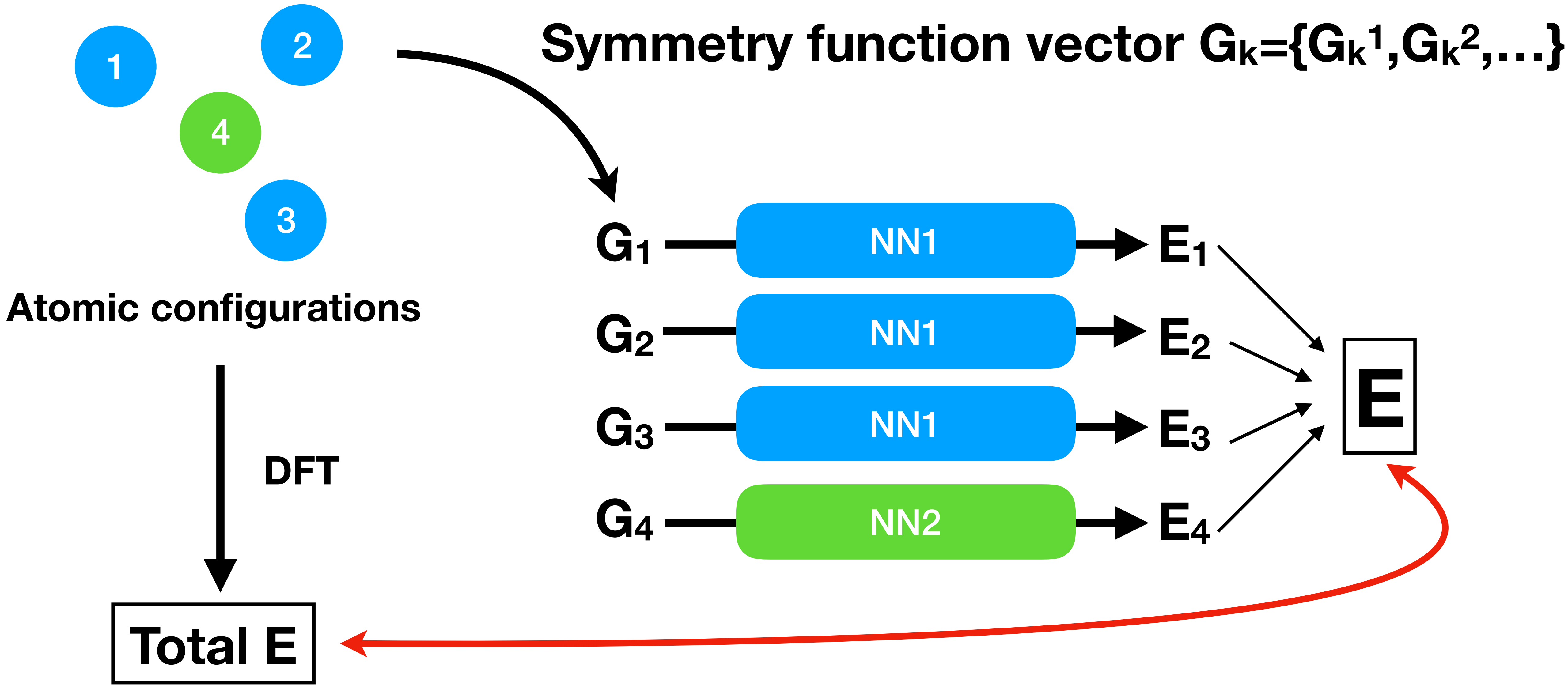
2. Dividing the total energy into atomic energies

- Keeping permutation symmetry for the same particle
- Easy to extend a system size (or number of particles) by adding subnets



Flow of constructing NNP

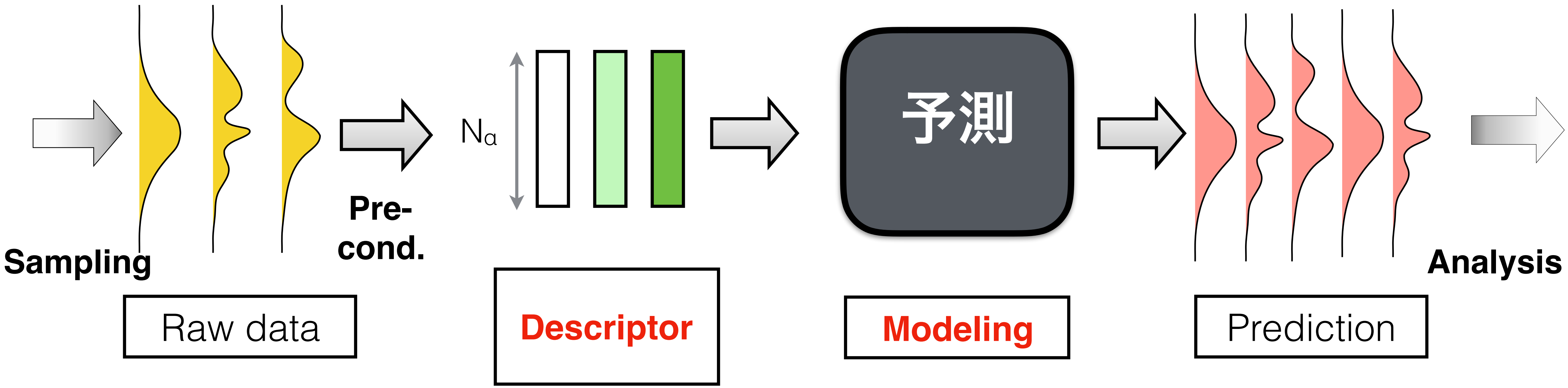
Describing the local environments



Symmetry function vector $G_k = \{G_k^1, G_k^2, \ldots\}$

- Total INPUT dimension: input of subnetworks x number of atoms
- Same parameters on the subnetworks are used for the same atomic species
- It is easy to expand the number of atoms by just adding the subnetwork.

Total flow chart of NNP



Sampling

Pre-cond.

N_α

予測

Analysis

Raw data

Descriptor

Modeling

Prediction

DFT results

（X） atomics configurations

（Y） total energy (or forces)

- Symmetry Function
- Structural fingerprint
- SOAP
- Coulomb Matrix
- MBTR

Neural network
Polynomial exp.
Gaussian process
Kernel ridge
Mixture models

High-Throuput MD

- Structure optimization
- Materials screening
- Large-scale simulation

Type of ML potentials: **（1） Descriptors （2） modeling**

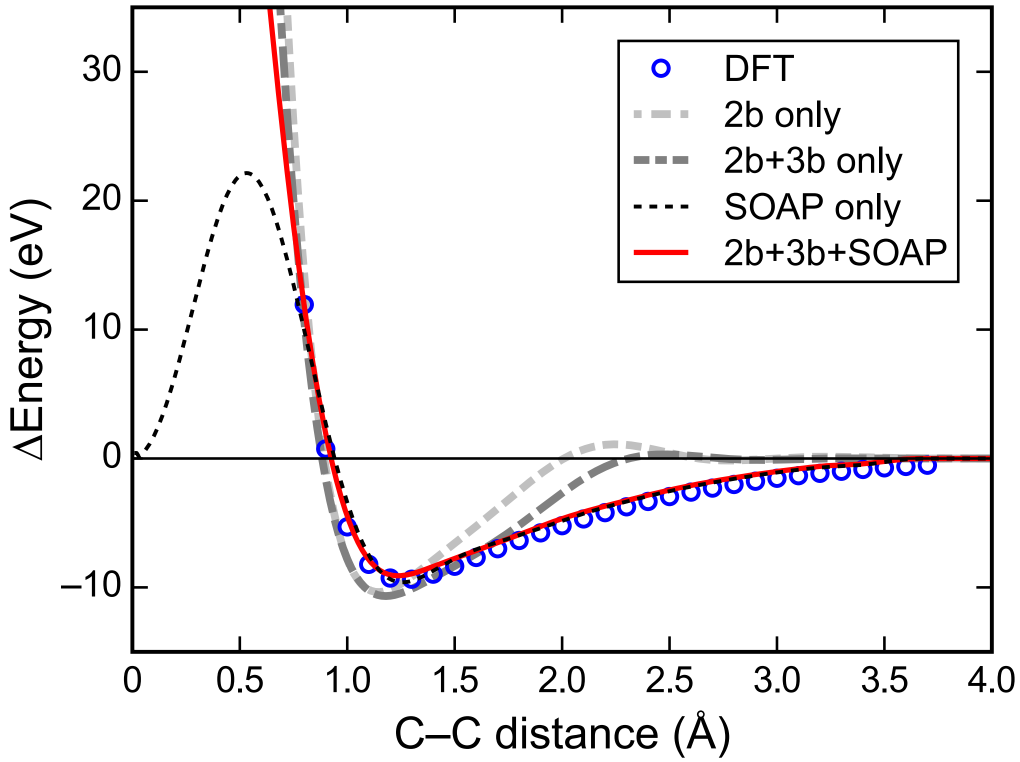
Keeping in mind their a merit, a purpose, and a issue

PHYS. REV. B **95**, 094203 (2017) .

GAP and SOAP

GAP: Gaussian Approximation Potential

✓ Fitting by Gaussian Process Regression
✓ Decomposing 2-, 3-, many-body terms
✓ 2- and 3-body: Gaussian kernel
✓ Many-body: Simple dot product kernel



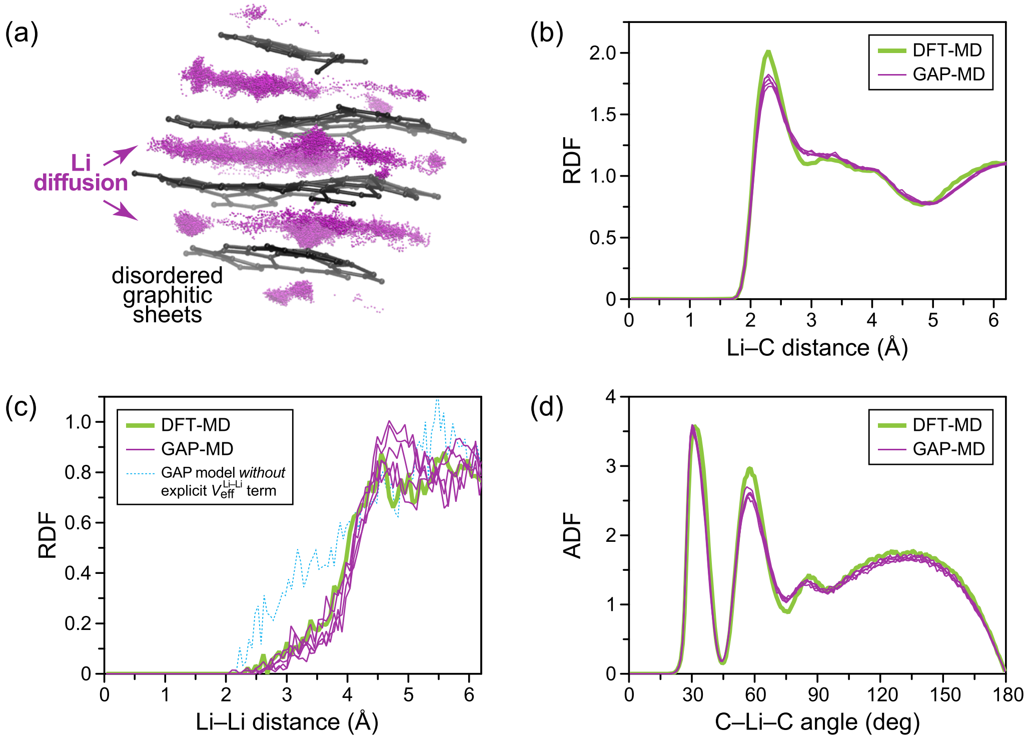
Descriptor

2-body: $$q^{(2b)} = |\mathbf{r}_2 - \mathbf{r}_1| \equiv r_{12}$$

3-body: $$\mathbf{q}^{(3b)} = \begin{pmatrix} r_{12} + r_{13} \\ (r_{12} - r_{13})^2 \\ r_{23} \end{pmatrix}.$$

Many-body: **SOAP**

(Smooth overlap of atomic position)



c.f. : Li Diffusion in graphites

Fujikake et al., J. Chem. Phys. 148, 241714 (2018).

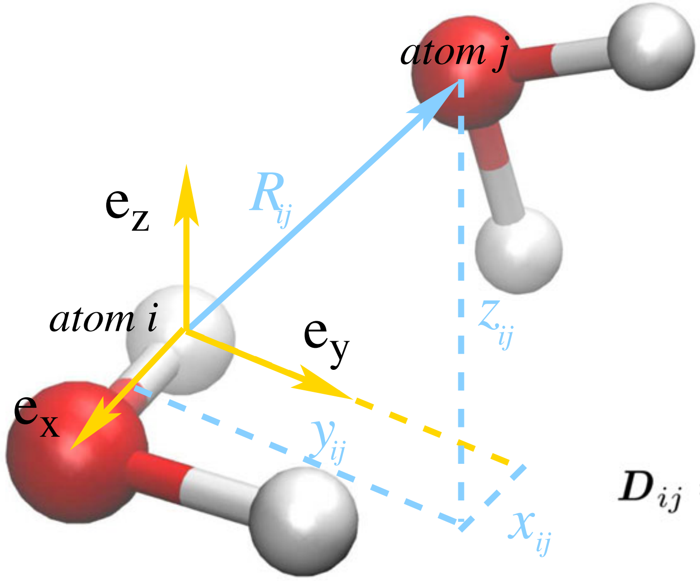
Phys. Rev. Lett. **120**, 143001 (2018).

DeePMD

[modeling]
Framework of Behler-Parrinello NNP

[Descriptors]

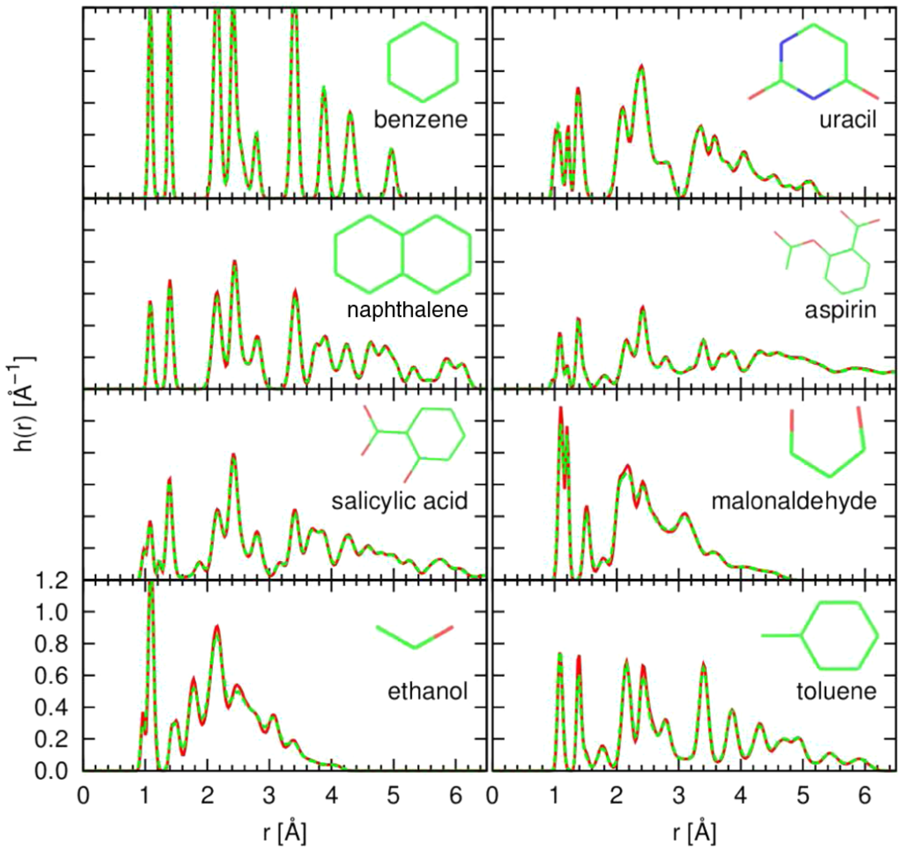
- Inner coordinates (invariance for rotation)
- Based on Inverse distance

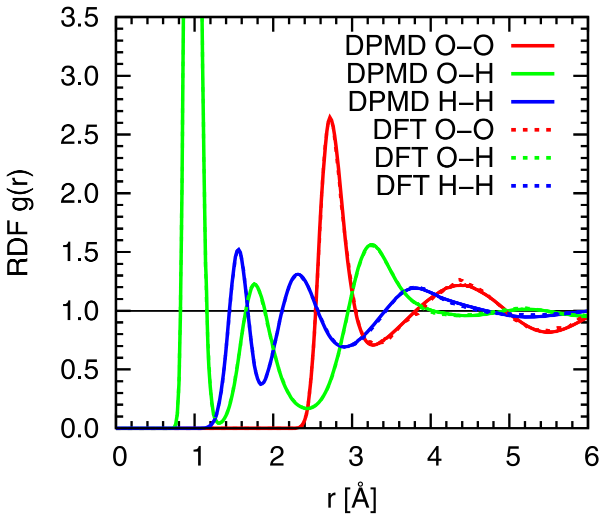


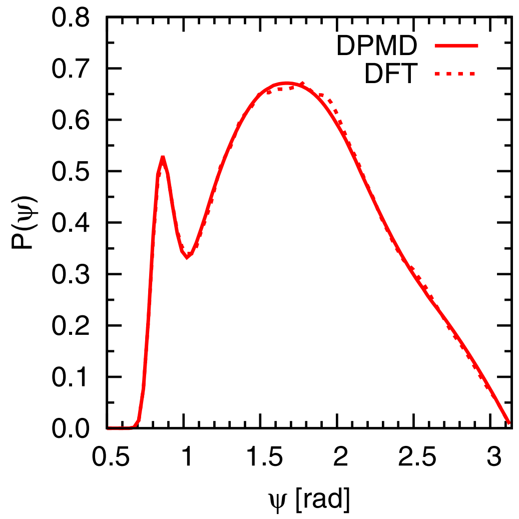
$$\boldsymbol{D}_{ij} = \{1/R_{ij}\}$$

or

$$\boldsymbol{D}_{ij} = \{1/R_{ij}, x_{ij}/R_{ij}^2, y_{ij}/R_{ij}^2, z_{ij}/R_{ij}^2\}$$







It works well in molecular systems.

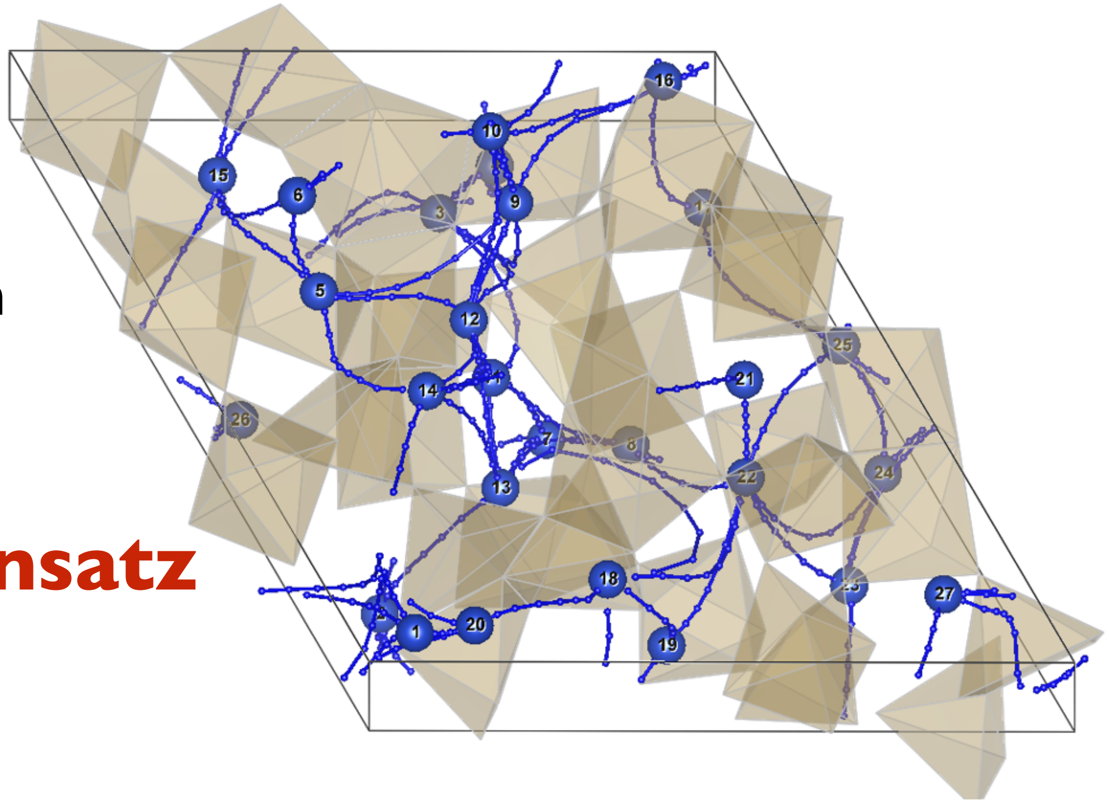
Diffusion pathways in amorphous

Problem：Too expensive to calculate with DFT

(N^3 times optimization x simulation time T)   N~50, T > 1 h => 13 year

NEB calculation is also necessary to estimate the activation energy…

$$E_{\mathrm{tot}} = E_{\mathrm{amorph}} + \underset{\substack{\downarrow \\ \text{Cu insertion term}}}{\Delta E_{\mathrm{Cu}}} + \underset{\substack{\downarrow \\ \text{Relaxation term}}}{\Delta E_{\mathrm{opt}}}$$

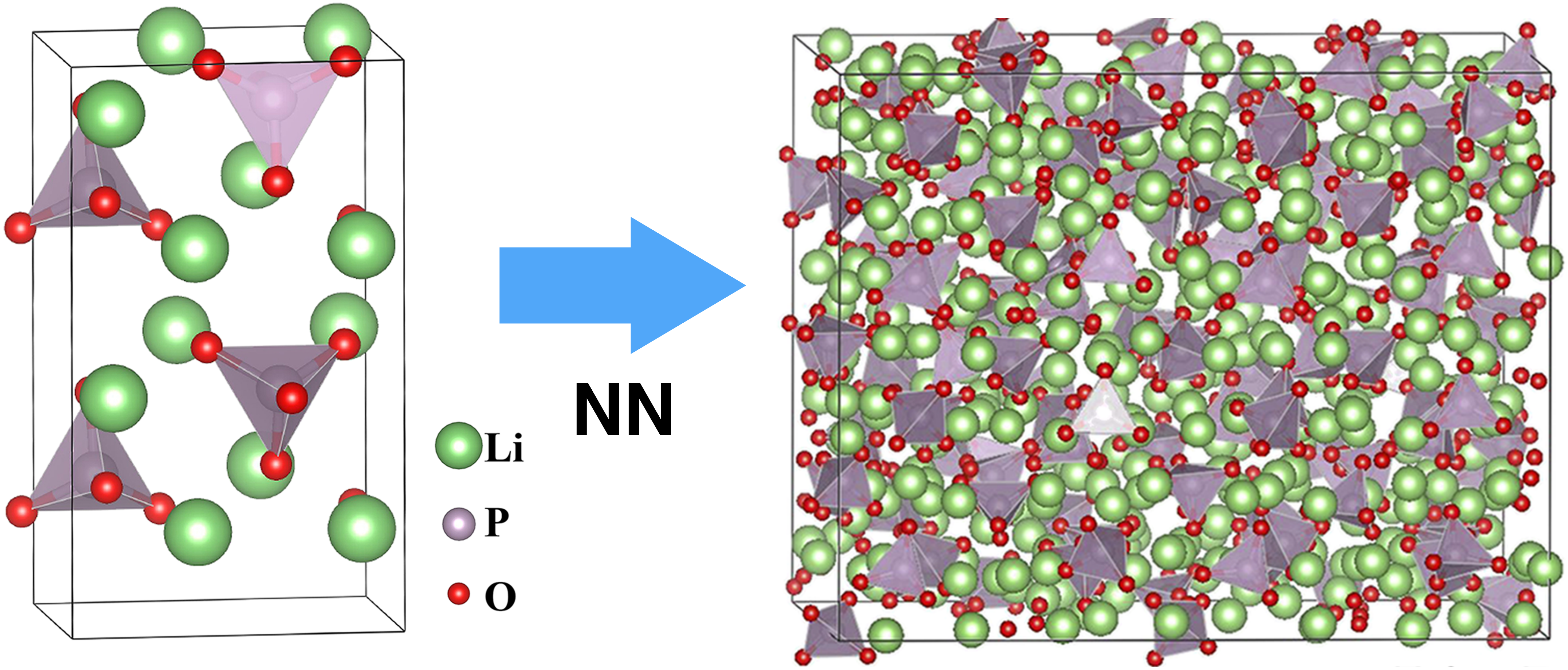


Found diffusion pathways

Construction with Behler-Parrinello ansatz

- ✓ Considering amorphous matrix as "a field"
- ✓ Learning only energy of Cu atom with its surroundings
- ✓ Able to simplify the NN even for ternary systems
- ✓ The calculation cost is cheaper than that of full NN
- ✓ Impossible to execute the molecular dynamics

Small system to Large system



DFT for small cell

Simulation for large cell

Small, accurate training data-sets

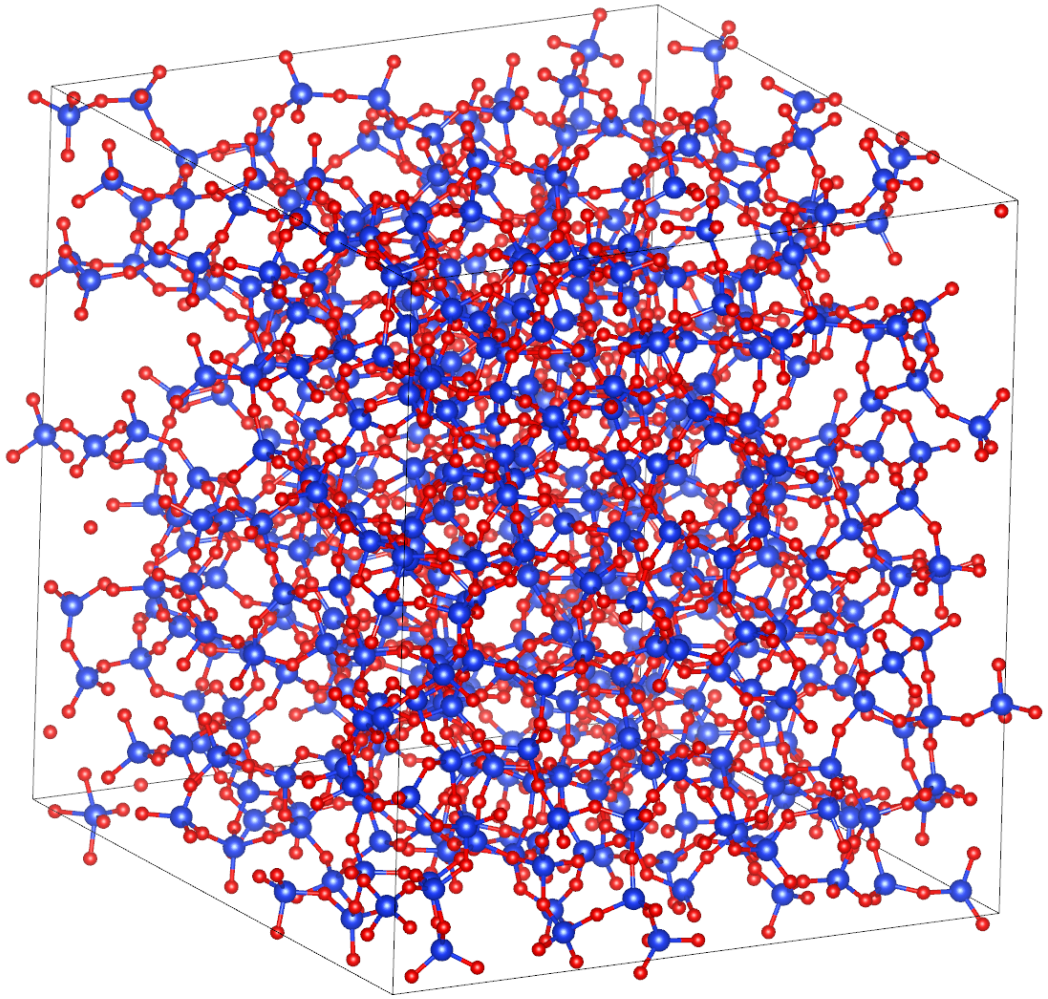
Scale up with ML potential

Reproduce the RDF and Diffusion constant estimated by DFT

W. Li, and Y. Ando, Phys, Chem. Chem. Phys. accepted.

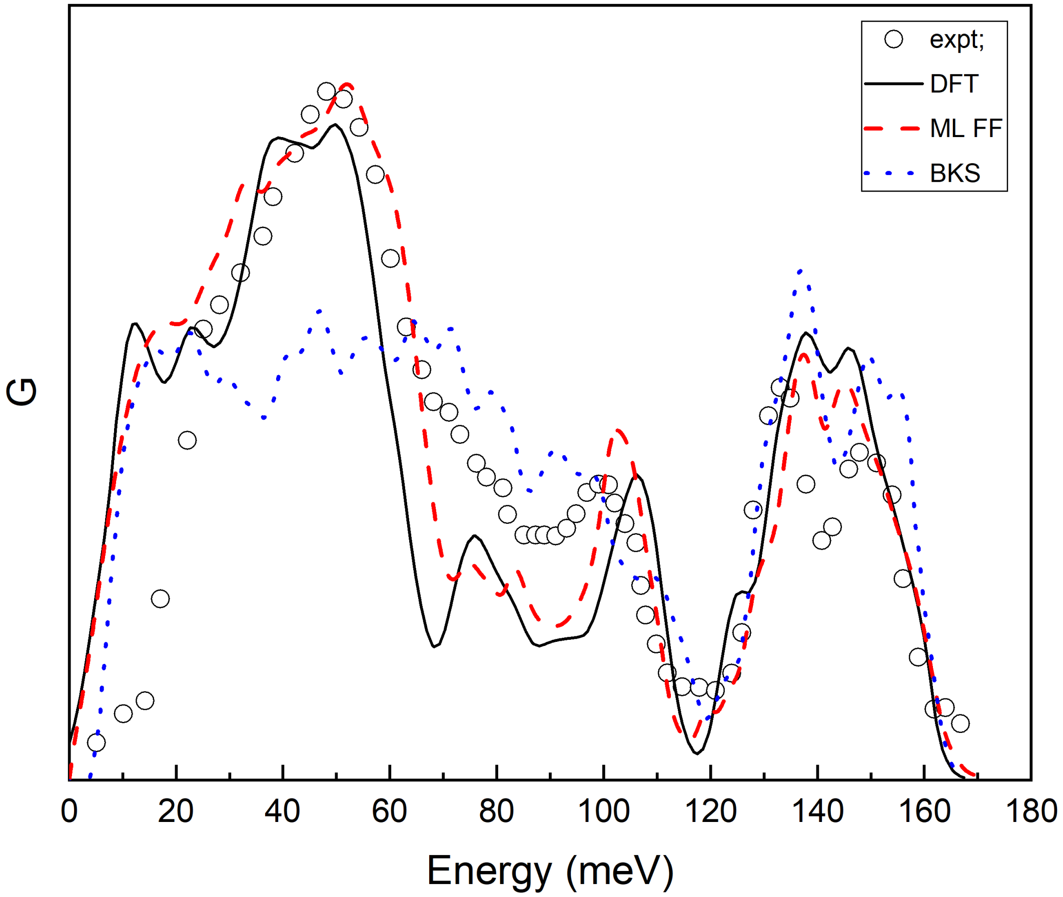
Vibration property depending on structure

Amorphous model



1728 atoms

Phonon density of state



The phonon density of state (DOS) calculated from ML force fields. [Science, 1997, 275, 1925–1927; Europhys. Lett., 2002, 60, 269–275; Phys. Rev. Lett., 1985, 54, 441–443]

It is possible to anneal slowly for large-scale system by ML potential（0.01K/fs）

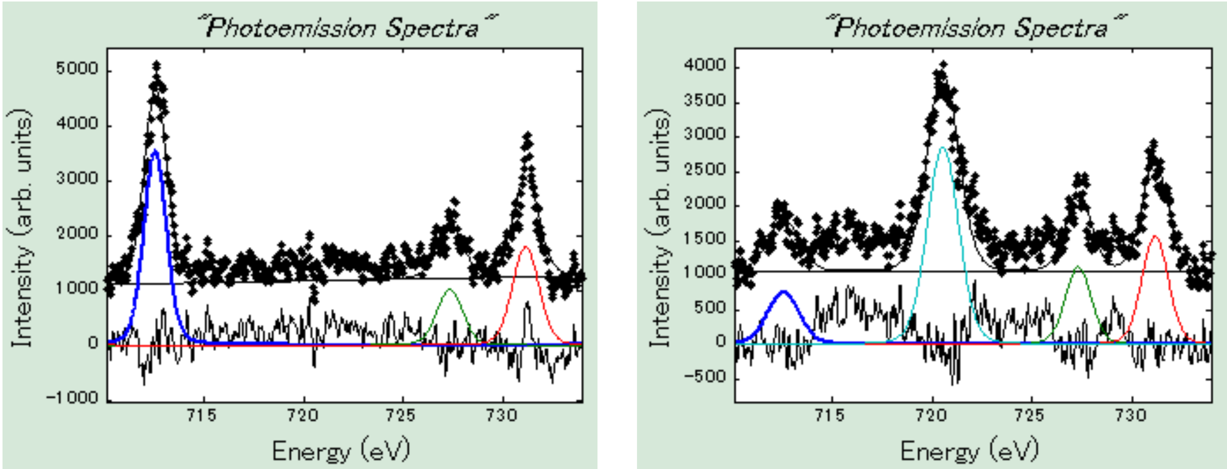
Improving ring statistics of Si-O networks/Phonon DOS also agrees well.

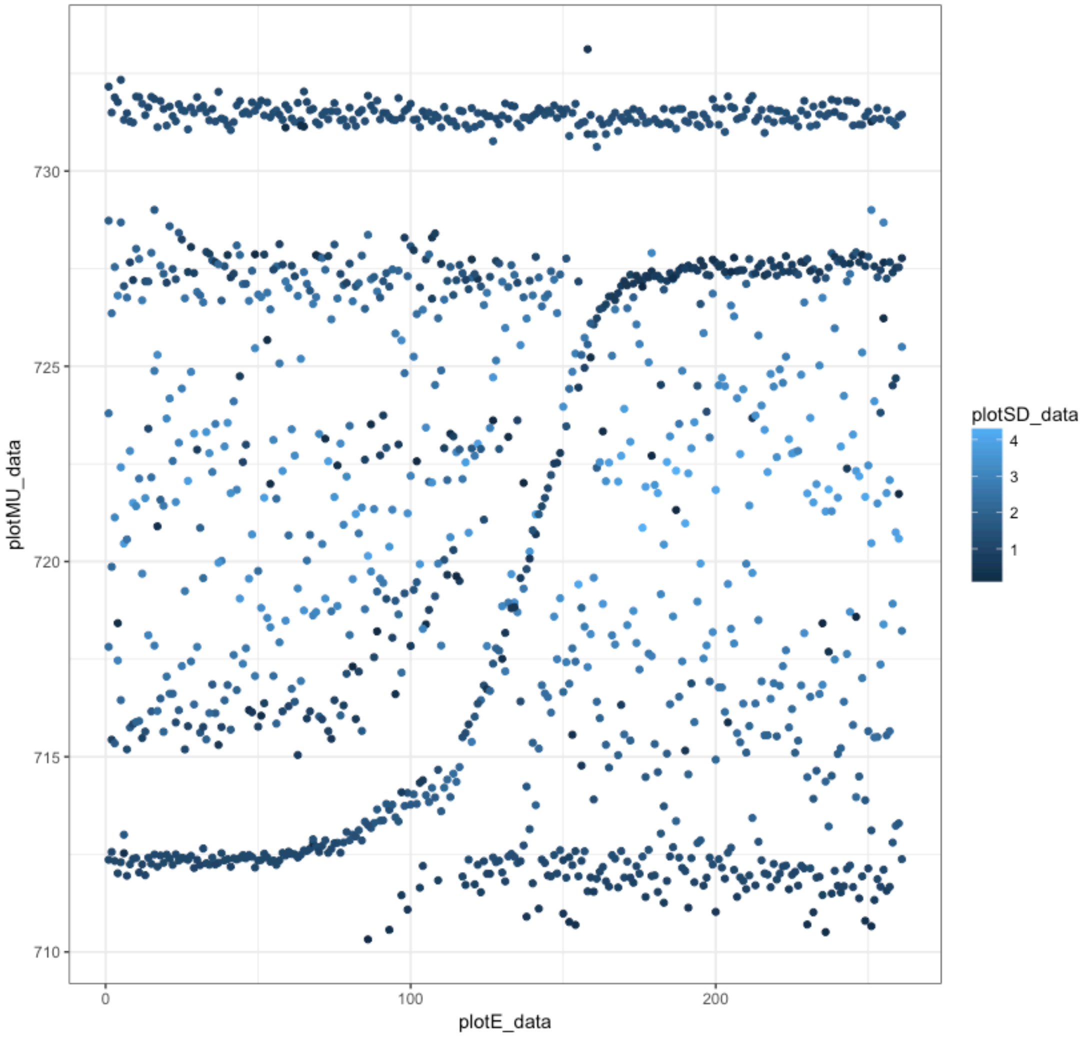
High-throughput Spectrum analysis

Collaborated with 永村直佳（NIMS）

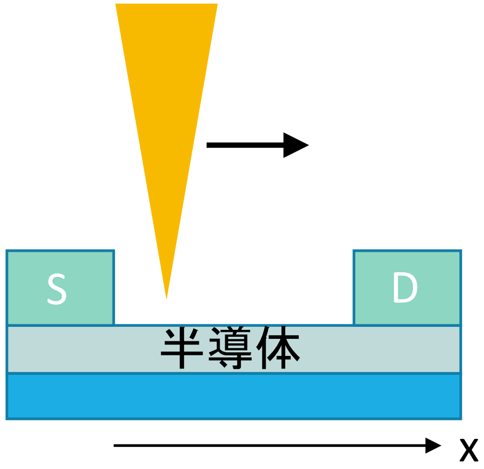
Automatic estimation of peak position

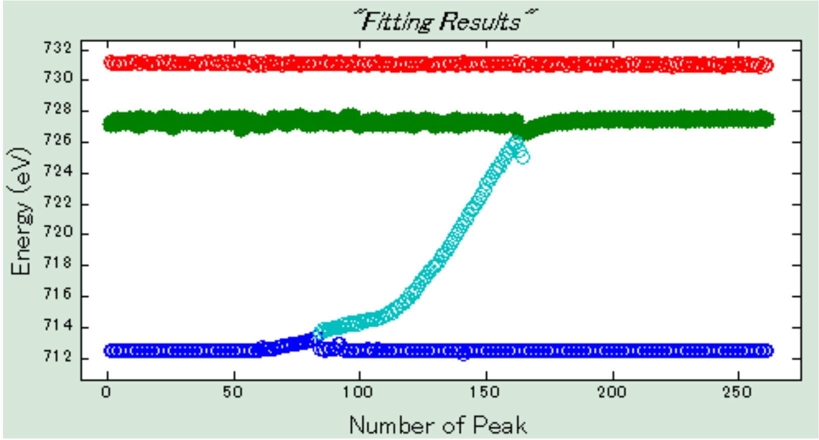
Experimental data-sets of spectrum





By hand





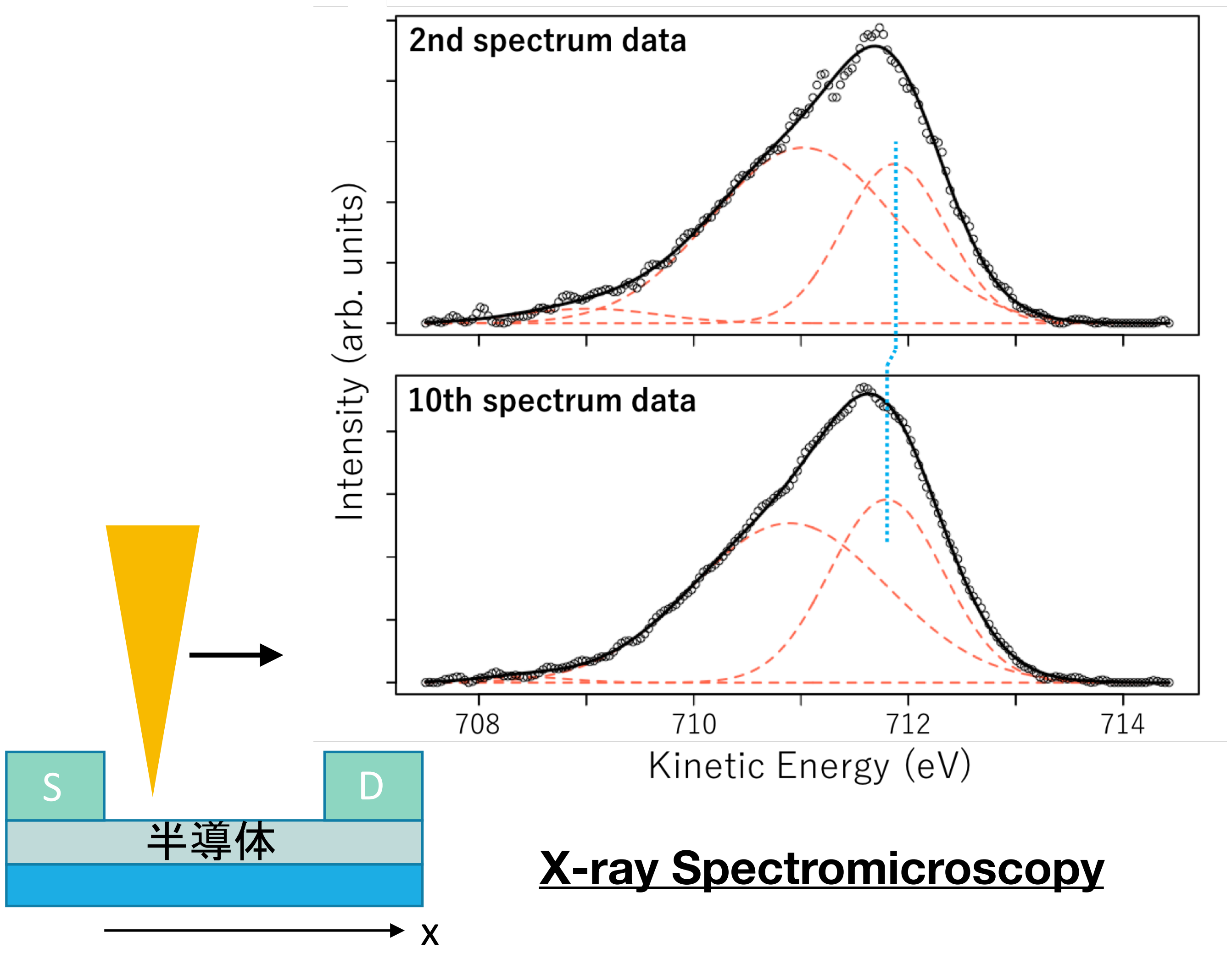
Peak estimation by EM algorithm

X-ray spectrum imaging

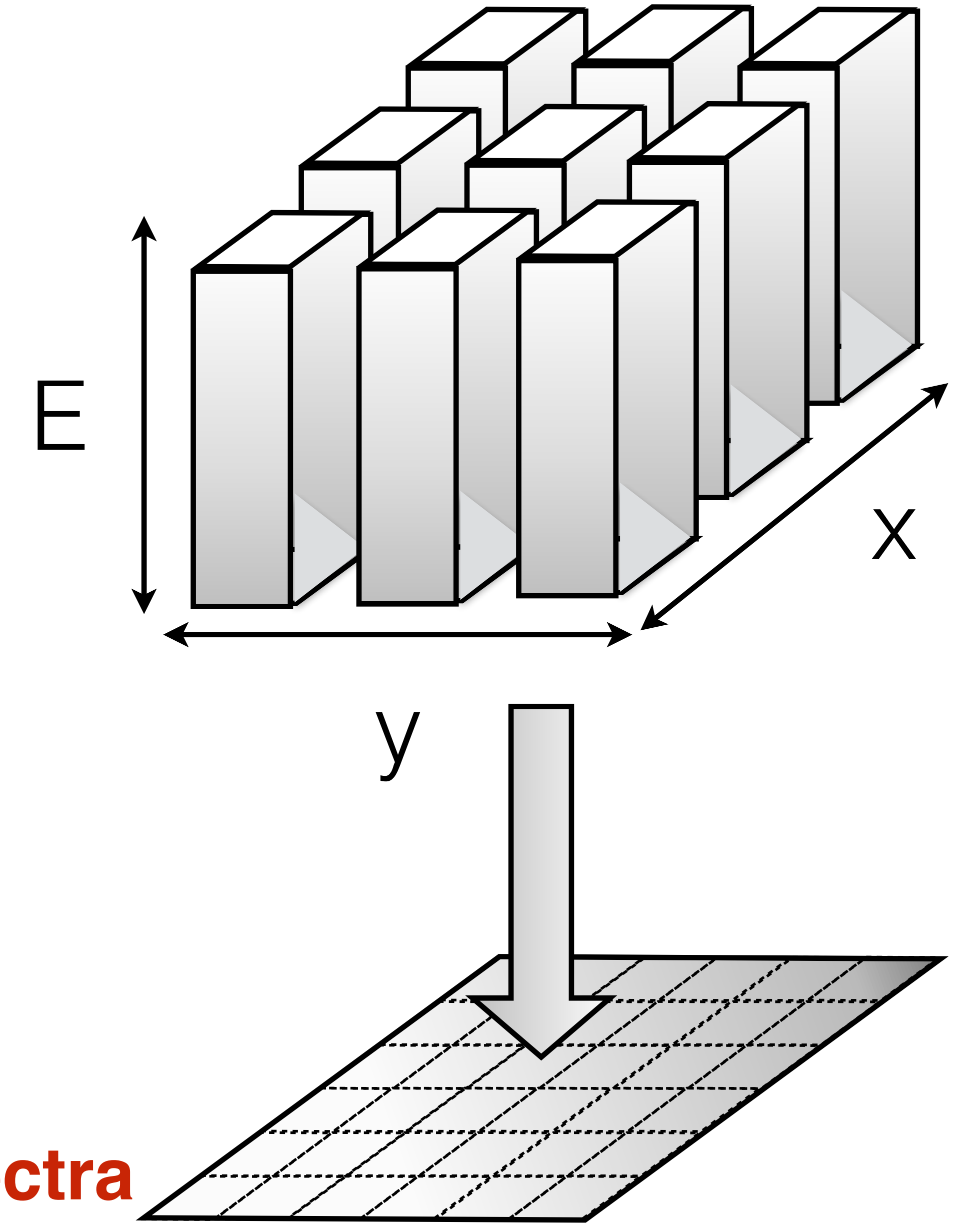
Collaborated with 永村直佳（NIMS）, 松村太郎次郎（PD）, 永田賢二(AIST), 赤穂昭太郎(AIST)

Auto-estimation of peak position

Issue：”**Automatic fitting to finding peaks in many spectra**”



X-ray Spectromicroscopy

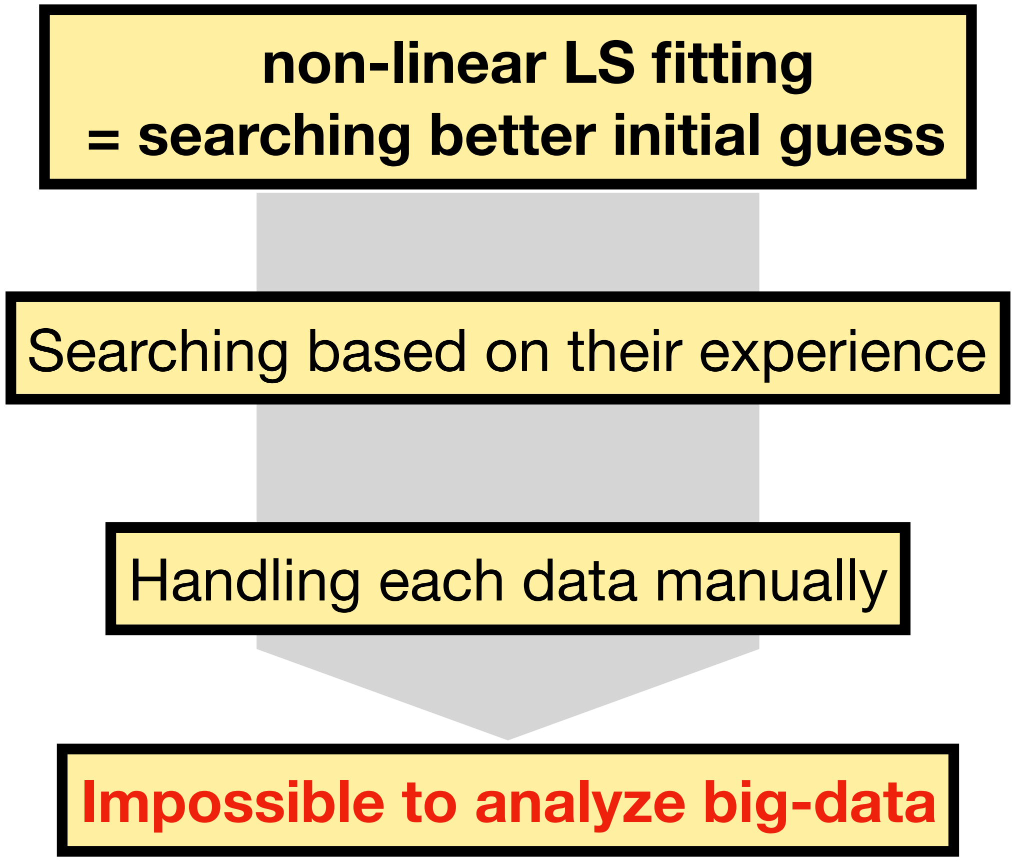


Character mapping

Need to extract peak positions of whole spectra
Fitting by gaussian mixture model?

Difficulty of parameter fitting

Hard to use non-linear fitting scheme.



non-linear LS fitting
= searching better initial guess

Searching based on their experience

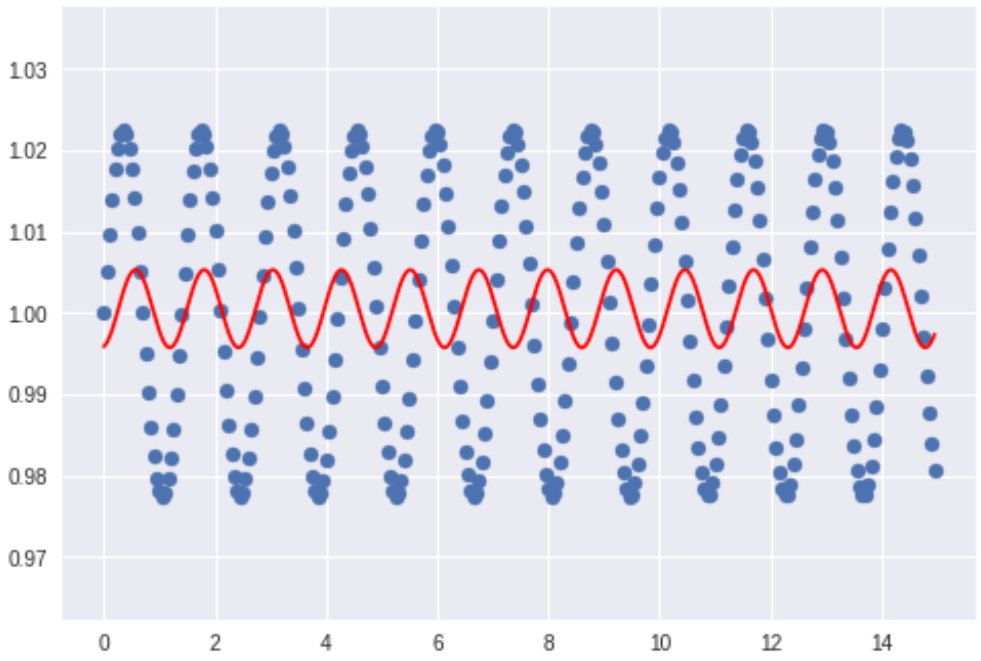
Handling each data manually

Impossible to analyze big-data

解決策

Effective way to find（EM algorithm）

Stochastic sampling（monte-carlo method）



$$y = A\sin(Bx + C) + D$$

Even though there are no noise,
It is not work with bad initial guess.

Maximum likelihood approach

ML estimation For Gaussian distribution

$$p(\{x_1,\ \cdots,\ x_N\}) = \prod_{n=1}^{N} N(x_n|\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right)$$

Likelihood: probability of obtaining observed data

Maximizing it

ML estimation for Gaussian mixture model

- Mixing ration(sum. 1)

$$p(x) = \sum_{k=1}^{K} \pi_k N(x\,|\,\mu_k, \sigma_k)$$

- **Summation are in log.**
- **Difficult to solve analytically**

$$\ln p(\{x_1,\ \cdots,\ x_N\}) = \sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k N(x_n\,|\,\mu_k, \sigma_k)\right)$$

Expectation-Minimization algorithm

Gaussian mixture model

Estimation
1. mean μ$_k$, variance σ^2_k, mixing ration π$_k$
2. Latent variable r$_{nk}$ for data n

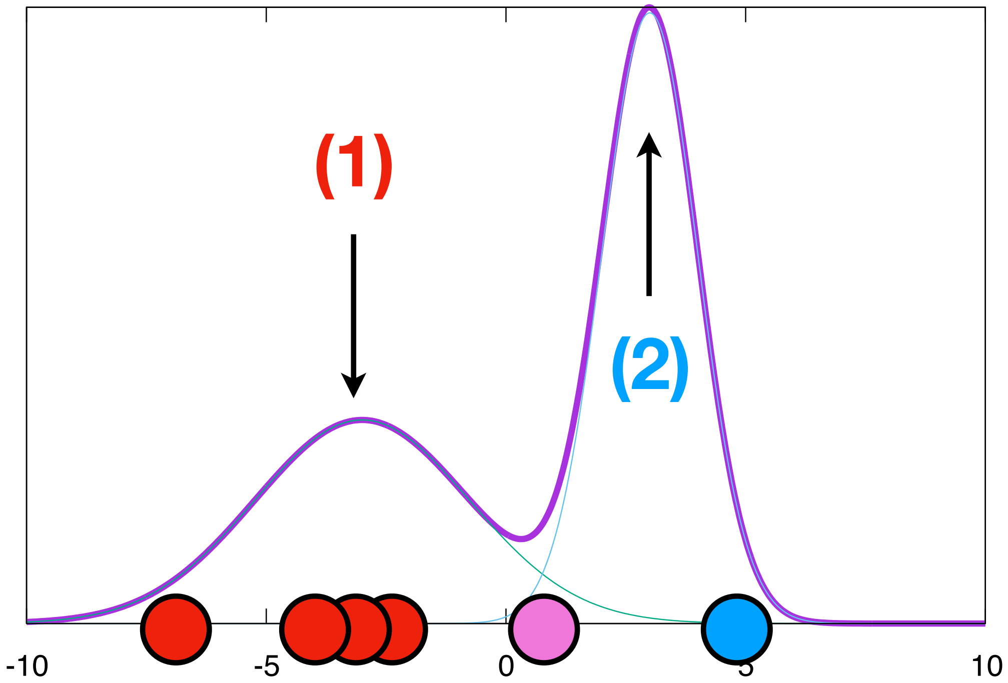


図３－１ 混合ガウス分布の例

E (Expectation)**-step**：Estimate (2) from (1).

expectation of r$_{nk}$（: responsibility γ(z$_{nk}$)）

$$E[z_{nk}] = \frac{\pi_k N(x_n, \theta_k)}{\sum_j \pi_k N(x_n \mid \theta_k)} = \gamma(z_{nk})$$

Depend on (1) and x$_n$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

M (Maximization)**-step**：Estimate (1) from (2).

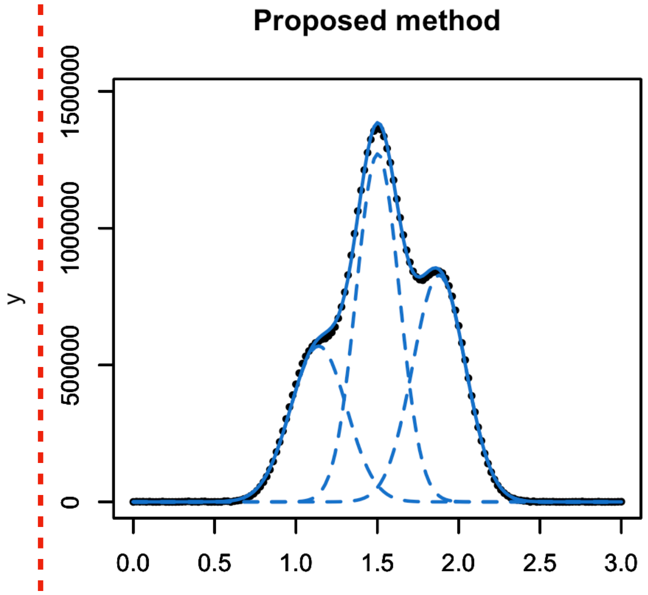
Maximizing expectation of log. Likelihood from complete data（pairs of x$_n$ and r$_{nk}$）

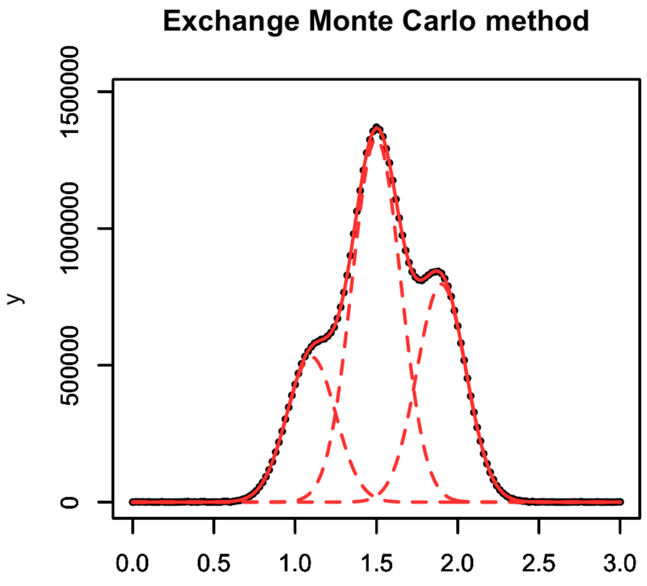
$$\mu_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n, \quad \sigma_k^{2\ new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n - \mu_k^{\text{new}})^2, \quad \pi_k^{\text{new}} = \frac{N_k}{N}$$

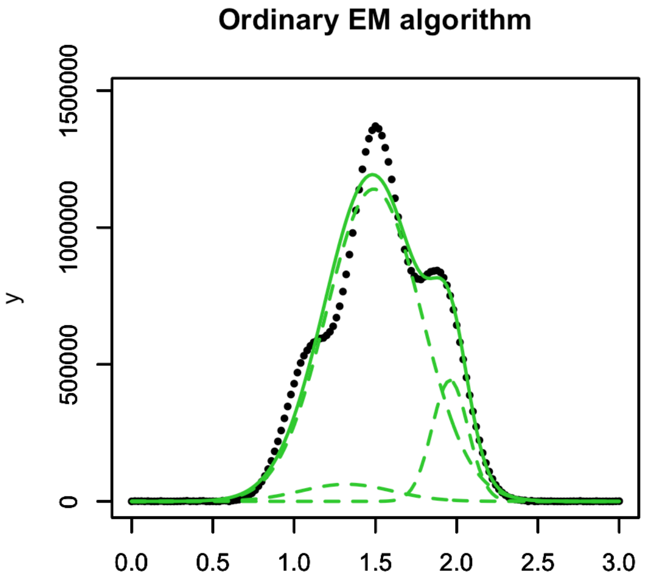
Likelihood monotonically increase!

Spectrum modeling of EM algorithm

Conventional EM algorithm: the cost depends total data number ⇐ **improved！**







Time [s]	**0.8**	481.0	19748.5
RMSE	8707.4	933.8	66086.0

- Robust for selection of initial guess, cheap computational cost.
- Multi-trial of initialization makes improve the accuracy.
- Noisy and peak overlapping case is difficult.

現在はこの手法を拡張して「ピーク本数」「pseudo Voigt関数による自動フィット」まで可能

Key consideration for using Informatics

“Finding” does not mean “Understanding”

- ✓ Many researchers have Internet of “Mechanism”.
- ✓ Machine-learning does not take into account the “physical law” generally

Analysis and Interpretation is must to do.

Trade of “Accuracy” and “White box”

Non-linear formalism

$$E[f(x)|\mathrm{Y}] = \mathrm{k}^T(K + \sigma^2 I_n)^{-1}\mathrm{Y}$$

High-Accurate one
BUT BLACK BOX

↔

Linear formalism

$$Y = \sum_i a_i X_i + b$$

NOT so high-accurate
BUT WHITE BOX

Several problem on Materials Informatics

“Small data” rather than “Big data”

- ✓ Conventional Experiment: less than 100, maximally 1,000.
- ✓ Systemic error depending on labs. (human, equipment, *etc.*)
- ✓ Spectra and Simulation data is close to “Big data”.

↔ Sequencer

NO universal representation

- ✓ object: polymer・semiconductor・metal etc.
- ✓ Scale: 1 nm ～ 1 m （9 order!!!)
- ✓ Diversity of representation for systems and scales

↔ Genomics

Key aspects of Materials sciences → **“controllability (reproducibility)” and ”theory (prior knowledge)”**

Prior cases are important

Always starts with an issue

Not necessary to be an expert of ML.

- ✓ Algorithm development is too difficult for materials researchers.
- ✓ Basic knowledge helps us to follow the cutting-edge algorithms.
- ✓ Important things is communication with experts.

Making a issue in materials science IS YOUR WORK

- ✓ ML experts cannot make an issue from the aspect of Materials science.
- ✓ Our duty is making an issue by ourselves.
- ✓ Applying basic algorithm by ourselves initially if possible.

Understanding basics of ML makes us to communicate with experts.
Communication makes us to solve our issue!

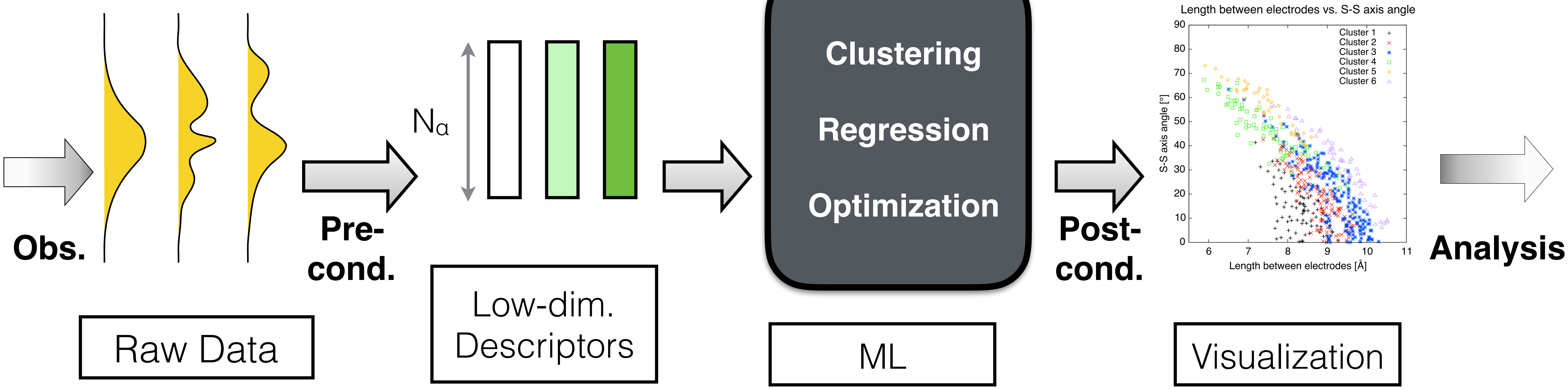
Take-home messages

Always Start with an Issue

The most important

- ✓ What do you want to know from data?
- ✓ What benefits are obtained by applying machine-learning?
- ✓ Can you breakdown the issue enough to solve?
- ✓ **NEVER** just USING the machine-learning.

Basic Flow



Obs.

Pre-cond.

N_α

Clustering

Regression

Optimization

Post-cond.

Analysis

Raw Data

Low-dim. Descriptors

ML

Visualization