**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**

> > 1

## Self-Introduction

#### **Name**:**Yasunobu ANDO**(**36**)

### **CV**

 $\sqrt{2012.3}$ . Ph.D in physics @ Dept. Physics, UTokyo.  $\sqrt{2012.4}$  ~ 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



 $A$  and  $A$  is unceded industrial Science and Technology -  $G$  and  $T$ 







## Major research topics



• Reaction on surface and interfaces.

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

#### We need "NEW" tools TABLE I. Explanatory variables: The first 55 variables are structural variables and the ones for *>* 55 are derived by quantum  $\frac{m}{\sqrt{2}}$  computations from the structural variables.



**FPMD of electrochemical Interfaces**



#### Milestone project with Informatics compare data must be implemented to allow greater This new integrated design continuum  $\bullet$  integrated design continuum  $\bullet$  in  $\bullet$  in  $\bullet$ tone project with i coupled with a

#### Materials Genome Initiative (MGI from 2011)  $M$ otoriale  $\epsilon$  $\frac{1}{2}$  all  $\frac{1}{2}$  $\mathcal{L}$ anomo Initiativo (MCI fro <u>and in chative the indice</u>

experiment — will significantly accelerate the time and



(see Figure 2).



#### **Figure 2: Initiative acceleration of the materials continuum of the materials continuu**

materials behavior must be developed to supplement

attention of industry and consumers is the recyclability

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

*analytical tools already in use in engineering*  **fig. 5 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**

### **Linear Regression (PCA)**

- ✓ Prediction
- ✓ Characterization
- ✓ Classification
- 

## *i*

 $Y = \sum a_i X_i + b$ 

## ✓ Pattern Recognition **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. PRL 115, 205901 (2015) PHYSICAL REVIEW LETTERS week ending coefficient, should also be optimized. Although the should also be optimized:  $\mathcal{S}$ be the band gap, the band gap, the band gap, the band gap  $\mathcal{L}$



discriminate in a simple way between materials that can be

144 0.29 Cs2½PdCl4"I<sup>2</sup> I4=mmm 0.31 0.88

Fig. 2.  $(Colo)$  (210) elemental descriptors is found to improve the robustness of  $\begin{array}{c} \n 1.5 \\
 \end{array}$  $\leq$  1.4,  $\frac{14}{415}$  [33],  $\geq$   $\geq$  $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$  can be found for parameters that  $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$  $\frac{2}{3}$  and  $\frac{1}{2}$  a  $\frac{1}{2}$  do not use such phonon parameters as predictors in the such phonon parameters as predictors in the such parameters  $\frac{1}{2}$  $p = \frac{1}{2}$  $\begin{array}{|c|c|c|c|c|}\n\hline\n\text{8.0} & & & & \\
\hline\n\text{7.} & 6.0 & & & \\
\hline\n\end{array}$  $\frac{4.9}{4.0}$   $\frac{4.9}{4.0}$   $\frac{2.0}{4.0}$  0.0 0.0 0.0 1.0  $\frac{2.0}{4.0}$  and  $\frac{2.0}{4.0}$  Number of prediction  $\frac{1}{2}$ composed of 34 and 24 elemental descriptors on top of V and *R* (b)  $\frac{367}{260}$  compounds over compounds over compounds over compounds over compounds over compounds in a compound over compounds large library is carried out using a GPR prediction model.  $\sum_{n=1}^{\infty}$ "virtual screening" in biometers  $\frac{1}{2}$ . For  $\frac{1}{2}$  $\sum_{n} w_n = w_n^{\text{max}}$  and  $w_n^{\text{max}}$  $\mathcal{O}(358)$ structure data available in ICSD  $\frac{1}{8}$ . This means that most  $\frac{1}{20}$  available in ICSD  $\frac{1}{20}$  $\frac{1}{2}$  them have been synthesized experimentally at least once. The  $\overline{A}$  (Coloradina) (a) The intervals of  $\overline{B}$ 





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 respectively the conventional method (black) was plotted by the conventional method (black) was plotted by order of the GB<br>34 elemental description of the GB and the 101 LTC data of the 101 LTC data of the 101 ltc data is energy, whereas it was plotted by order of the trial number in the kriging energy, whereas a wave precise by start of the article and include the area pointed by blue arrow in the  $(c_0)$ . (b) Magnified image around the area pointed by blue arrow in the (a).

#### and the structure was investigated. Using the convention  $\mathbf{f}_{\text{max}}$  $t = 1$ ntertac most stable point was determined from the search space  $\mathbf{r}$ calculations. Since its efficiency depends on the first random  $\mathsf{S}$  data points, we repeated the kriging operation by  $\mathsf{S}$  $\blacksquare$  times for the same GB. The results of all trials are shown as  $\blacksquare$ been reported previously.26) The kriging method was applied  $\mathbf{m}$  $\blacksquare$ was that we succeeded in determining the structure afternoon  $\omega$  $\mathbf{F}$  shows the distribution of  $\mathbf{F}$  shows the  $\mathbf{F}$ compounds along with  $\Gamma$  and  $\Gamma$  and  $\Gamma$  and  $\Gamma$ score is plotted in panels corresponding to constituent

Jpn. J. Appl. Phys. 55, 045502 (2016) S. Kiyohara et al.

13 NOVEMBER 2015

Jpn. J. Appl. Phys. **55** (2016) 2–6. Jpn. J. Appl. Phys. **55** (2016) 2–6.  $\sum_{i=1}^n$ 

*Science Adv.*, **2**, e1600746-1-7.  $\sum$  and  $\sum$  are lowest  $\sum$  orders  $\sum$  and  $\sum$ 

PRL **115**, 205901 (2015). bers of compounds required for the optimization using the

Appl. Phys. Rev. 4, 011105 (2017). with United States is the United States in the States in the States is the States in the States in the States I  $R_{\rm P}$ Energy storage materials such as in Li ion batteries rep-Appl. Phys. Lett. **91**, 132102 (2007). Appl. Phys. Rev. 4, 011105 (2011)

nential growth of computer processing power, combined

tional methods and software infrastructure was pioneered

#### High-Throughput Observation heat could be in the multi-terawatt range.<sup>48</sup> Although thermoelectric devices have significant potential to recover the significant potential to recover the significant waste heat from industrial processes in the set of the devices are only about 5% efficient. Therefore, discovery of  $m_{\rm s}$  mechanics, has made it possible to design  $m_{\rm s}$  $f(x)$  value physics value  $f(x)$ mid-2000s, the development of high-throughput computations of high-throughput computations of high-

#### $L_{\text{obs}}$  the problem to  $\Gamma_{\text{obs}}$  are covered to  $\Gamma$ <u>Material (i large Sepermini a large Sepermini a la</u>  $\Gamma$  and  $\Gamma$  applied to the discovery of  $\Gamma$ High-Throughput Experimental (HTE) methodologies

annual potential for electrical energy recovery from waste



FIG. 2. (a) Electrical conductivity, (b) Seebeck coefficient, and (c) power factor of the composition-spread  $(Ca_{1-x-y}Sr_xLa_y)_{3}Co_4O_9$  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.<sup>56</sup>

<u>Materials "Library"</u>: Dispersing the several compositions on a single sheet **Observing Big Materials Space at once, Finding optimum one**

## Downsampling method for quasi-particle Ohs.

Downloaded from journals.jps.jp by on 10/26/17

#### Quasi-particle interference (QPI) on Ag(111) surface  $J$ . Soc.  $J$



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

the phase noise that are no separate that are  $\mathsf R$ scattering in the pace can l are shown in the top and middle rows, respectively. (a) and (d) are obtained by using randomly chosen pixels. (b) and (e) are obtained by using every third pixel and the both the horizontal and vertical directions. (c) and (f  $\alpha$  in the radial region **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.



## Machine-learning potentials

#### $\blacksquare$   $\blacksquare$ temperature of 600 !C (a) and 400 !C (b). (c,d) STM images after post-deposition annealing (PDA) for films deposited at 400 !C (c) and 300 !C (d). Molecular dynamics simulations

8 12 16

#### **Structure Optimization**

Low Low

High

epita $\mathcal{L}_{\mathcal{A}}$  thin films claims  $\mathcal{L}_{\mathcal{A}}$ 

smaller than that of all the previous reports. Furthermore, the





roughness (RRMS) values: as-grown samples (blue symbols) and after PDA (red symbols). The value of RRMS is evaluated from topographic images

 $\overline{\phantom{a}}$ 

#### **Structure Optimization Exercise Structure Structure Optimization** observed at a sample-bias voltage of  $10$  pA (scan area of  $400$  nm). (f) Temperature dependence of the field-cooled-cooled-cooled-cooled-cooled-cooled-cooled-cooled-cooled-cooled-cooled-cooled-cooled-cooled-cooled-cooled



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

**Determining the** 

 $\overline{\phantom{a}}$  Prometable structure to dynam  $\frac{1}{2}$  from stable structure to ayrian • From stable structure to dynamical behaviors  $\begin{bmatrix} x_1 & x_2 & x_3 & a & -1 \end{bmatrix}$ 

#### Coherence length on the surface. To further study the super-**Conductivity on the value of the value of the value of x from the value of x from the value of x from the value of**  $\mathbf{w}$  $t_{\text{intra}}$  structures around a magnetic vortex core. **interaction between atoms**



[Phys. Chem. Chem. Phys.,](https://doi.org/10.1039/1463-9084/1999) 2018, 20, 11586-11591

#### VOLUME 61, NUMBER <sup>7</sup> PHYSICAL REVIEW LETTERS <sup>15</sup> AUGUST 1988 oxygen. We assume no electron orbitals around the The potential energy surface is then obtained by our changing the Si-0 distances or 0-Si-0 angles with three tained the results for the C3",mode, in which only one modeling by fitti 0-Si-0 angles are varied with constant Si-0 distance. It has turned out that these three potential energy sur-Potential modeling by fitting

#### TTAM (Tsuneyuki-Tsukada-Aoki-Matsuda) Potential <u>f TAM (Tsuneyuki-Tsukada-Aor</u> ed in Fig. 1, in the Indeed-<u>inatsuuajit oteituai</u> t  $\frac{1}{2}$  Transferred (b) and b) and the boundary reserved to  $\frac{1}{2}$ Mateuda) Potential <u>Ivialoudaj i Oldifiliai</u>

**Parametrizing potentials by fitting on Hartree-Fock calculations of SiO<sub>4</sub>4- + 4 point charge model** oxygen. We assume no electron orbitals around the tained the results for the C3",mode, in which only one Parametrizing potentials by fitting on Hartree-Fock calculations of SiO $_4^\prime$ 

faces can be fitted well by a sum of pairwise interatomic interatomic interatomic interatomic interatomic interatomic

Phys. Rev. Lett. **61**, 86 (1988). ditlet modes. One potential energy surface is depicted in the effective charge is depicted in the effective charge is depicted in the effective charge in the cluster, because the cluster, because the effective charge in th  $\mathbb{R}^n$  surface is the potential equation of  $\mathbb{R}^n$ Phys. Rev. Lett. **61**, 86 (1988).<br>Neture 330, 200-(1080). sive interaction. Here r is the distance between atoms and a; (b;) is the effective radius (softness parameter) of the ith

functions (with a  $\sim$  0.059) in for negative-ion states of  $\sim$ 

different modes. One potential energy surface is depicted in  $\mathcal{O}(n)$ 

$$
U_{ij}(r) = U_{ij}^{\text{Coulomb}}(r) + f_0(b_i + b_j) \exp[(a_i + a_j - r)/(b_i + b_j)] - c_i c_j/r^6,
$$
 range, short-range parts.

- **Example 12 Constraining the electrostatic interaction to long-**
- **• Fitting on cluster model, applying bulk.** <br>• **Fitting on cluster model, applying bulk.**
- $U_{\text{c}}^{\text{Coulomb}} = \tilde{O} \cdot \tilde{O} \cdot (1 \sigma_{\text{c}}(\mathbf{r})) / \mathbf{r} + O \cdot O \cdot \sigma_{\text{c}}(\mathbf{r}) / \mathbf{r}$  **• Reproducing SiO<sub>2</sub> polymorphs**  $U_{ij}^{\text{Coulomb}} = \tilde{Q}_i \tilde{Q}_j [1 - g_{ij}(r)]/r + Q_i Q_j g_{ij}(r)/r,$  Peproducing SiO<sub>2</sub> polymor

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



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



FIG. 2 The new Cmcm phase predicted by this and sixfold coordination, as seen from the a ax  $c$  axis of low-cristobalite and stishovite. Small-cluster results are insufficient the best crystal parameters. to determine the long-range coulomb interaction, and the fitting procedure for a since the fitting procedure f<br>Since the fitting procedure for a since the fitting procedure for a single procedure for an and concern to the

#### **Discovery of new SiO<sub>2</sub> high-pressure phase**  $\mathbf{E}$  choice of  $\mathbf{y}$  of how only in  $\mathbf{f}(\mathbf{c})$ more than  $\overline{\text{Disc}}$  of new SiO<sub>2</sub> chosen the one (Table II) which one (Table II) which one structure is the structure of  $\sim$

## 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**

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

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

- 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 layer are connected to the same particle to the adjacent layers by a set of the adjacent layer
- Easy to extend a system size (or  $r = \frac{1}{2}$  is directed a given set of  $\frac{1}{2}$  $N_{\rm H}$  is the expansion by the expression by the expression by the expression by the expression of  $P_{\rm H}$ number of particles) by adding subnets



## Flow of constructing NNP



- 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



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

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

for a sharper distinction between different environments. This is a sharper different environment  $\mathcal{L}_\text{max}$ 

#### **GAP and SOAP** Both for 20 and 3 and 3 and 30 an <sup>⎣</sup> <sup>−</sup><sup>1</sup>  $\overline{a}$  ,  $\overline{b}$   $\overline{c}$   $\overline{d}$   $\$

⎦ *,* (3)

representative points differs for each descriptor and must be

Both for 2b and 3b contributions, we use a squared

23

#### GAP: Gaussian Approximation Potential *K*(*<sup>d</sup>* ) " **q**(*<sup>d</sup>* ) # .<br>İ <u>ו חסו</u>:<br>∍∍∍∍  $\frac{1}{2}$ *k* Gaussian Approximation Potential discussion over the control of the

#

exponential kernel (35) and (35)

"

**q**(*<sup>d</sup>* )

*<sup>i</sup> ,***q**(*<sup>d</sup>* ) *t*

*K*(*<sup>d</sup>* )

✓ Fitting by Gaussian Process Regression ✓ Decomposing 2-, 3-, many-body terms  $\sqrt{2}$ - and 3-body: Gaussian kernel √ Many-body: Simple dot product kernel *<sup>i</sup> ,***q**(*<sup>d</sup>* ) *t* = exp  $\overline{R}$ ξ  $\mathcal{A}$  Fitting by Gaussian Process Regression  $\mathcal{A}$   $\mathbb{R}^{20}$  $\sqrt{2}$  Decomposing  $2 - 3$ - many-body terms  $\frac{12}{2}$  between  $\frac{12}{2}$ 

 $=$   $\frac{1}{2}$ 

### **Descriptor**

2-body:  $q^{(2b)} = |\mathbf{r}_2 - \mathbf{r}_1| \equiv r_{12}$  $r_{12} + r_{13}$  (a) **3-body:**  $\mathbf{q}^{(30)} = | (r_{12} - r_{13})^2 |$  $r_{23}$  ${\bf q}^{(3b)} =$  $\sqrt{2}$  $\mathbf \Gamma$  $r_{12} + r_{13}$  $(r_{12} - r_{13})^2$ *r*<sup>23</sup>  $\sum_{i=1}^{n}$ 2-body: 3-body:

Many-body: **SOAP SOAP SOAP SOAP SOAP SOB** 

th overlap **AP**  $\epsilon$  is the second to  $\epsilon$  is the second is and the second is and  $\epsilon$  is an induced in  $\left[\begin{array}{ccc} \text{O} & \$ (Smooth overlap of atomic position)

**SOAP** 

#### **c.f.**:**Li Diffusion in graphites**  denotes constraints on the specific function  $\frac{d}{d\theta}$  o.g.  $\frac{1}{\theta}$ **C.T.**  $\cdot$  LI DITUSION IN Graphites  $\begin{bmatrix} 0.4 \end{bmatrix}$   $\begin{bmatrix} 0.4 \end{bmatrix}$

Fujikake et al., J. Chem. Phys. 148, 241714 (2018). UJINANC CL AL, J. CHCHI. FIIYS. 140, Z417 14 (ZUTO).<br><sup>01</sup>12334 smooth overlap of atomic positions (SOAP) [53] descriptor,



 $\sum_{k=0}^{\infty} \frac{1}{k}$  and  $\sum_{k=0}^{\infty} \frac{1}{k}$  and  $\sum_{k=0}^{\infty} \frac{1}{k}$  and  $\sum_{k=0}^{\infty} \frac{1}{k}$ Phys. Rev. Lett. **120**, 143001 (2018).

and  $\Delta$  IMD trajectories. In the case of the molecules, we can expect the m

#### DeePMD  $p$  at the same temperature of the same temperature of the original  $\mathcal{D}$ data, using a Langevin thermostat with a damping time  $\sim$  1 keV  $\sim$  0.1 ps. The corresponding distributions of interactions of interatomic distributions of interactions of inter

brackets are the AIMD results. The numbers in parentheses are Framework of Behler-Parrinello NNP [modeling]

[Descriptors]

- tes water • Inner coordinates (invariance for rotation)
- **PHYSICAL REVIEW LETTERS**  $\mathcal{I}$   $\mathcal{$ · Based on Inverse distance



#### FIG. 1. Schematic plot of the neural network input for the **ENVIRANMENT III, THOIECUIDE SYSTER** in modenulou overtence It works well in molecular systems.



## Diffusion pathways in amorphous

### **Problem**:**Too expensive to calculate with DFT**

( $N<sup>3</sup>$  times optimization x simulation time T)  $N~50$ , T  $> 1 h = > 13$  year NEB calculation is also necessary to estimate the activation energy…



#### Small system to Large system we can can obtain 12 different Library formation of the contract of the contra  $\mathcal{F}$  shows the comparison of the vacancy  $\mathcal{F}$ potential in various calculations, we do not discuss the cases y Lary<del>e</del> by



**Small, accurate training data-sets Scale up with ML potential**

a Li atom and then relaxed the structure. Since the supercell



be formed rather than an isolated vacancy. Since that  $\theta$  is obtained vacancy. Since the purpose the purpose  $\theta$ 

### **Reproduce the RDF and Diffusion constant estimated by DFT**

## Vibration property depending on structure





1728 atoms

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

## Automatic estimation of peak position

#### **Experimental data-sets of spectrum**



**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"**



## 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**)



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

## **Maximum likelihood approach**

ML estimation For Gaussian distribution

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

**Likelihood: probability of obtaining observed data**

**Maximizing it** 

ML estimation for Gaussian mixture model

$$
p(x) = \sum_{k=1}^{K} \pi_k N(x | \mu_k, \sigma_k)
$$
  
 
$$
\text{Equation are in log. } \boxed{\sum_{n=1}^{N} \ln \left( \sum_{k=1}^{K} \pi_k N(x_n | \mu_k, \sigma_k) \right)}
$$
  
 
$$
\text{Different to solve analytically}
$$

## **Expectation-Minimization algorithm**

#### **Gaussian mixture model**

**Estimation** 

 $E[z_{nk}] =$ 

1. mean  $\mu_k$ , variance  $\sigma^2$ <sub>k</sub>, mixing ration  $\pi_k$ 

2. Latent variable  $r_{nk}$  for data n

 $\pi_k N(x_n, \theta_k)$ 

 $\sum_{j} \pi_{k} N(x_{n} | \theta_{k})$ 

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

**expectation of r<sub>nk</sub>** (: responsibility  $γ(z<sub>nk</sub>)$ )



図3-1 混合ガウス分布の例



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

**Maximizing expectation of log. Likelihood from complete data (pairs of x<sub>n</sub> and r<sub>nk</sub>)** 

 $=\gamma(z_{nk})$   $\longleftarrow$  Depend on (1) and x<sub>n</sub>

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

#### **Likelihood monotonically increase!**

## **Spectrum modeling of EM algorithm**

**improved** ! Conventional EM algorithm: the cost depends total data number  $\equiv$ 



- 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"**



## 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".

### **NO universal representation**

- ✓ object: polymer・semiconductor・metal etc.
- $\sqrt{\phantom{a}}$  Scale: 1 nm  $\sim$  1 m (9 order!!!)
- ✓ Diversity of representation for systems and scales

**Key aspects of Materials sciences** **"controllability (reproducibility)" and "theory (prior knowledge)"**

Prior cases are important

Genomics

Sequencer

## 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.
- $\checkmark$  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.

