データ駆動科学の立場からみた
物質科学と情報科学の接点

安藤 康伸

AIST, 機能材料コンピュテーショナルデザイン研究センター

自己紹介

名前：安藤康伸（36）

出身：愛知県西尾市

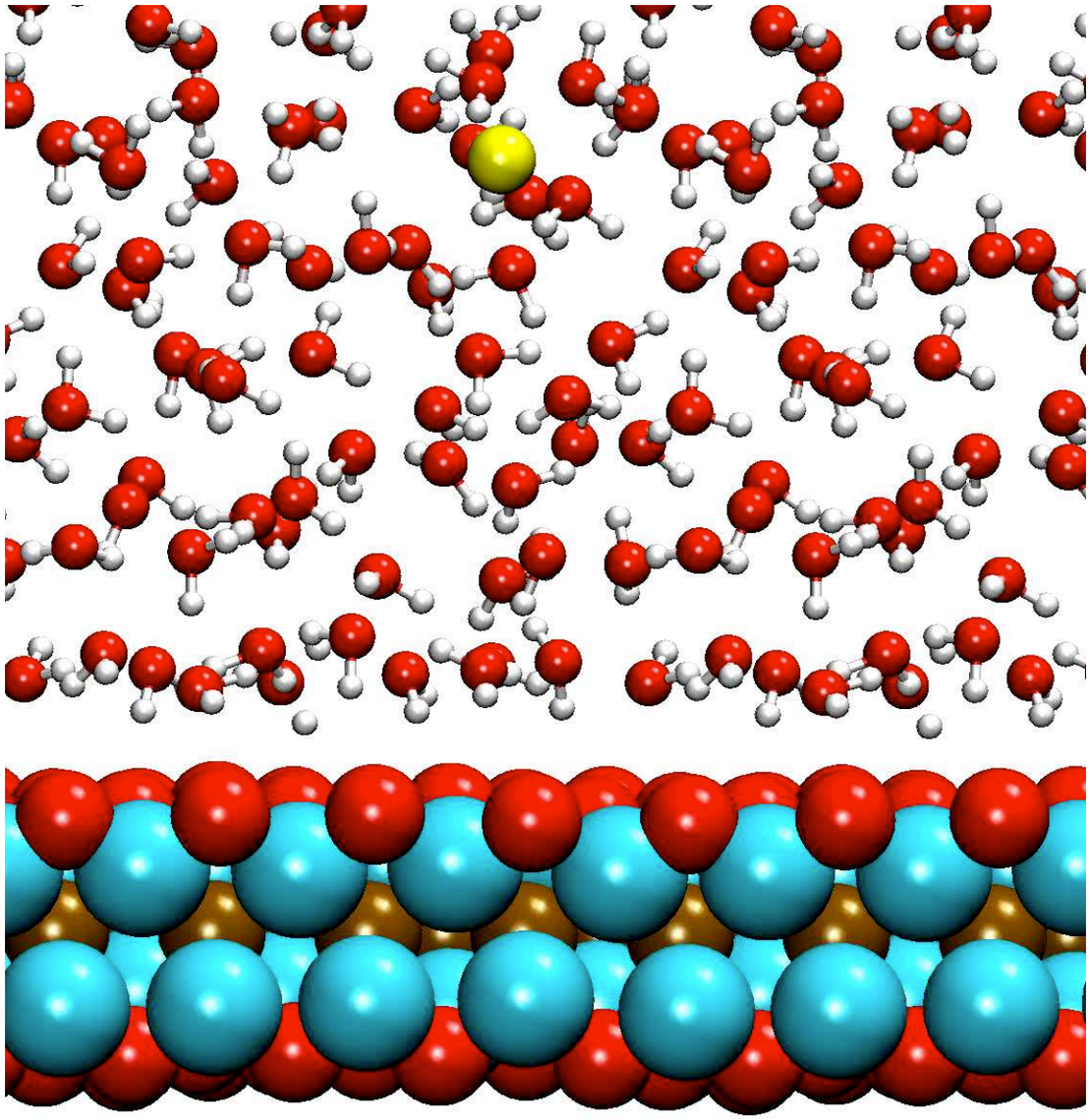
<u>略歴</u>

- ✓ 2012. 3.　博士（理学） @東大院理
- ✓ 2012. 4 ~ 2013. 4　ポスドク@AIST
- ✓ 2013. 5 ~ 2016. 3　助教 @東大マテ工
- ✓ **2016.4** ~ 研究員 @ 機能材料コンピューテショナルデザイン研究センター (CD-FMat)，AIST

Back ground

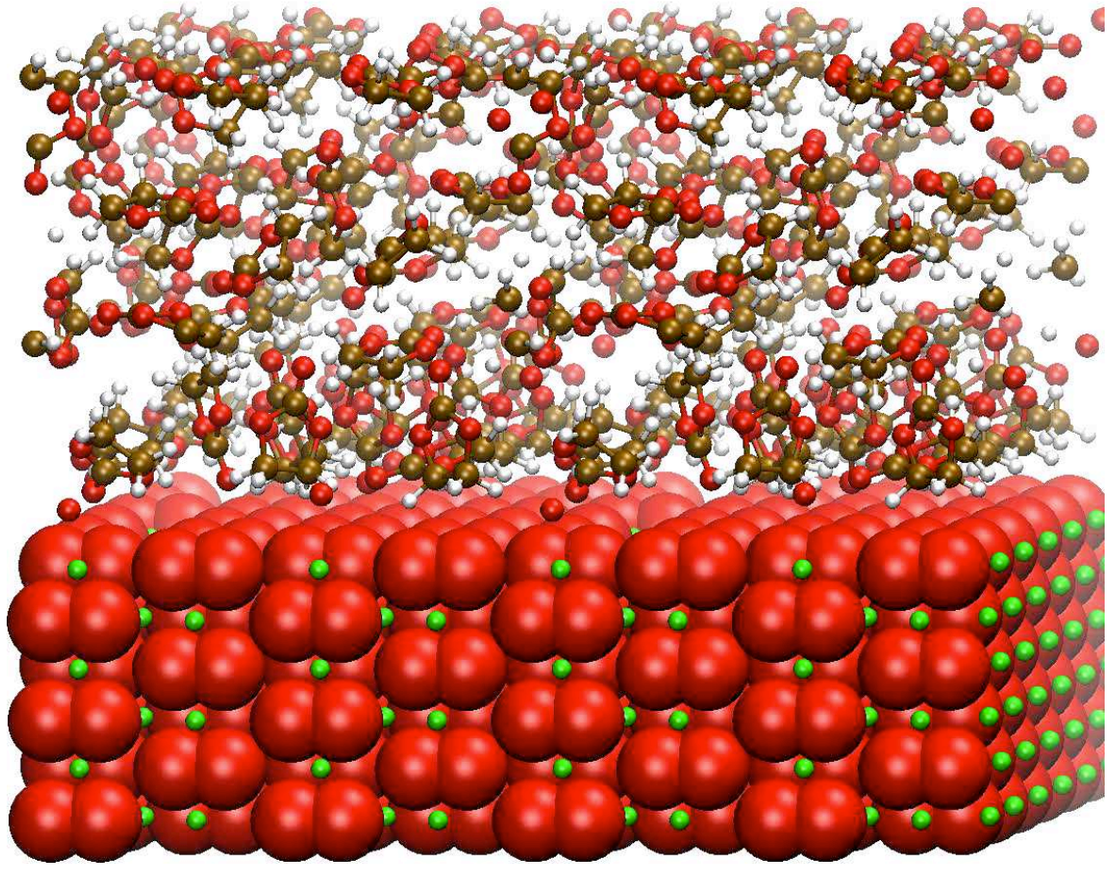
- ✓ 物性物理学
- ✓ 表面・界面科学
- ✓ 計算物質科学

情報科学×物質科学

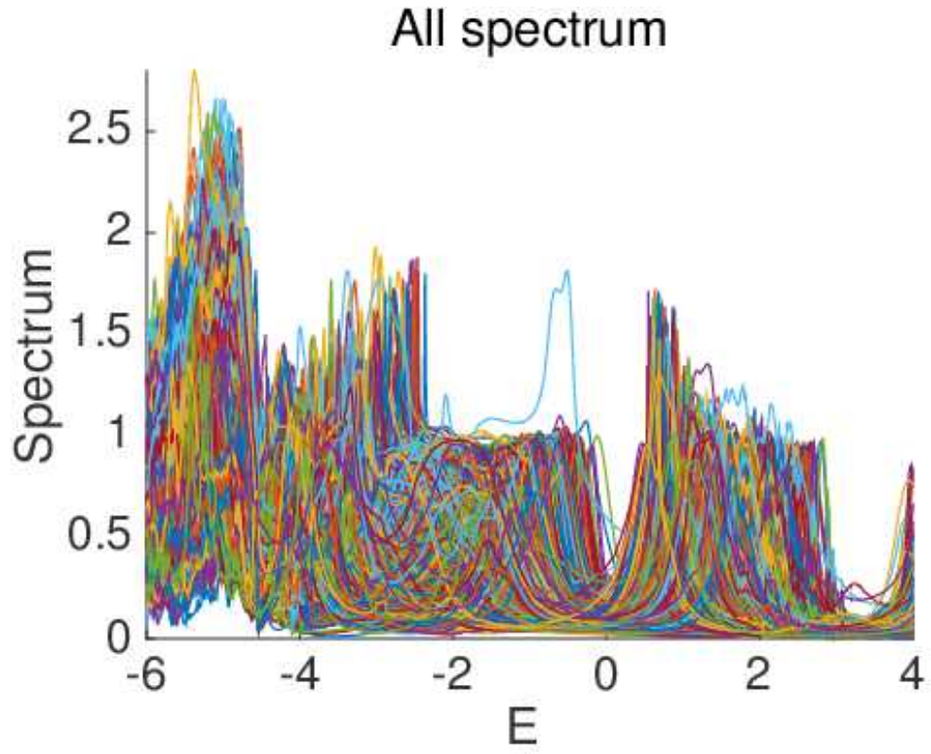


電極ー電解液界面の第一原理分子動力学

情報科学に至った経緯



MDの結果をひたすら”観察”



600本… 一体どうすれば。。

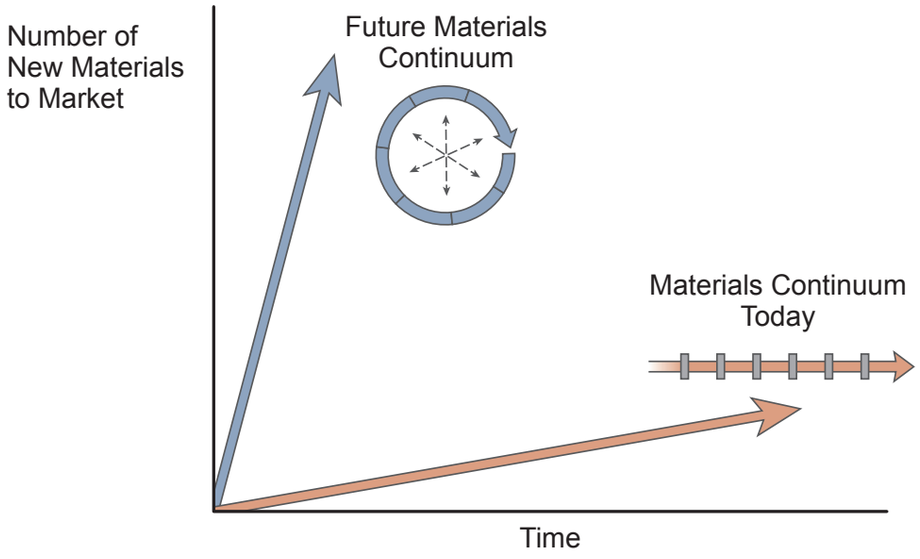
その他 約10万本のスペクトル処理など

効率的な解析技術の習得が必須

情報科学ブームの火付け役

Materials Genome Initiative（MGI from 2011）

戦略目標



1. 材料イノベーションのインフラ整備
2. 先端材料開発に関する国家目標達成
3. 次世代の研究従事者育成

→ **データベース・機械学習の活用を掲げる**

物質・材料科学？？？


Vanilla essence
Vanilla oil
Choco spray
Pastel pink spray
Color sugar vermicelli
Mix color spray
Choco chips
Choco chips white
Choco chips caramel
Minihearts strawberry
Color sugar heart

製品の”原材料”を研究する仕事

シリコン インゴット

工業製品をもっと
よくするような原材料
についての研究

電池, CPU, モニター等
工業製品の大事なパーツ

情報的な視点から見た物質科学

ビッグデータというよりスモールデータ

- ✓ 典型的な実験：高々100, ようやく1,000サンプル出せることも
- ✓ 他の実験室の結果とも直接比較が困難（環境・装置・人）

⇔ シーケンサー

汎用的な物質表現・処理方法の不在

- ✓ 対象：高分子・半導体・金属 etc.
- ✓ スケール：1 nm ～ 1 m（９桁！！）
- ✓ 構造やシステムの表現方法の多様性

⇔ ゲノム情報

物質科学の強みである「制御性（再現性）」「理論（事前知識）」を生かした中規模な情報処理スタイルの開発が必要

事例の蓄積が鍵

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

ハイスループット計測データ解析

High-Throughput Experimental (HTE) methodologies

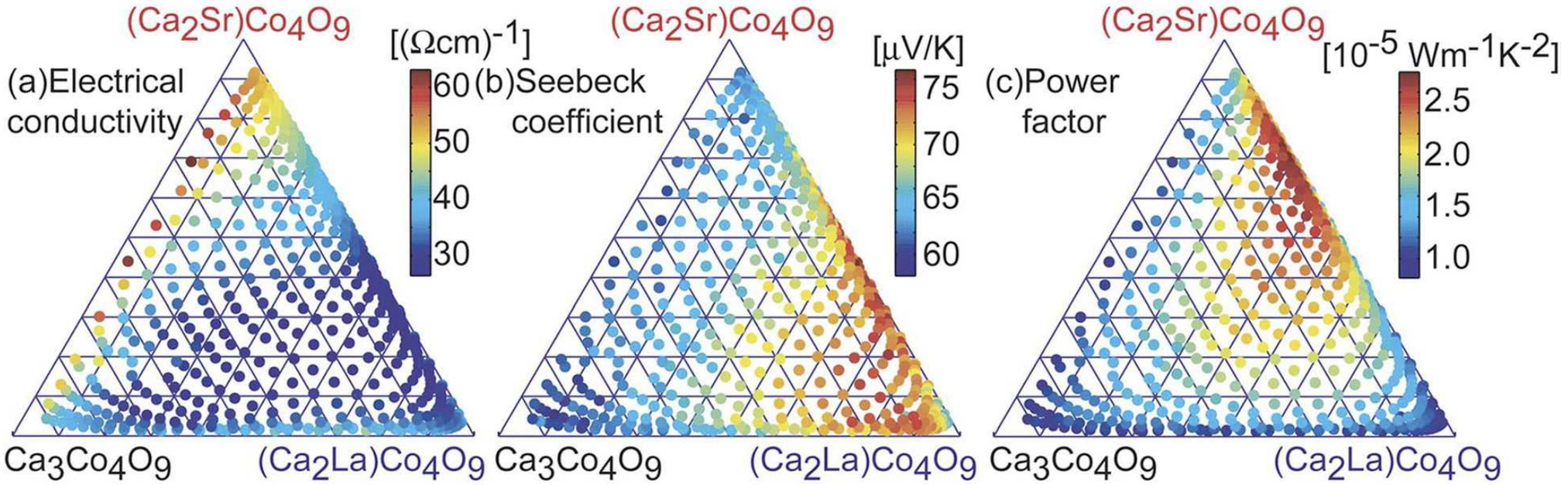


FIG. 2. (a) Electrical conductivity, (b) Seebeck coefficient, and (c) power factor of the composition-spread $(Ca_{1-x-y}Sr_xLa_y)_3Co_4O_9$ film ($0 < x < 1/3$ and $0 < y < 1/3$). Reproduced with permission from Appl. Phys. Lett. **91**, 3 (2007). Copyright 2007 AIP Publishing LLC.[56]

Materials "Library" : 単一シートに複数の組成を連続的に散布

物質空間を一気に観測、最適な組成を高速に発見する

（観測に~6 h）

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

準粒子干渉測定のダウンサンプリング

課題：**『短い観測時間でも正しい結果を得たい』**

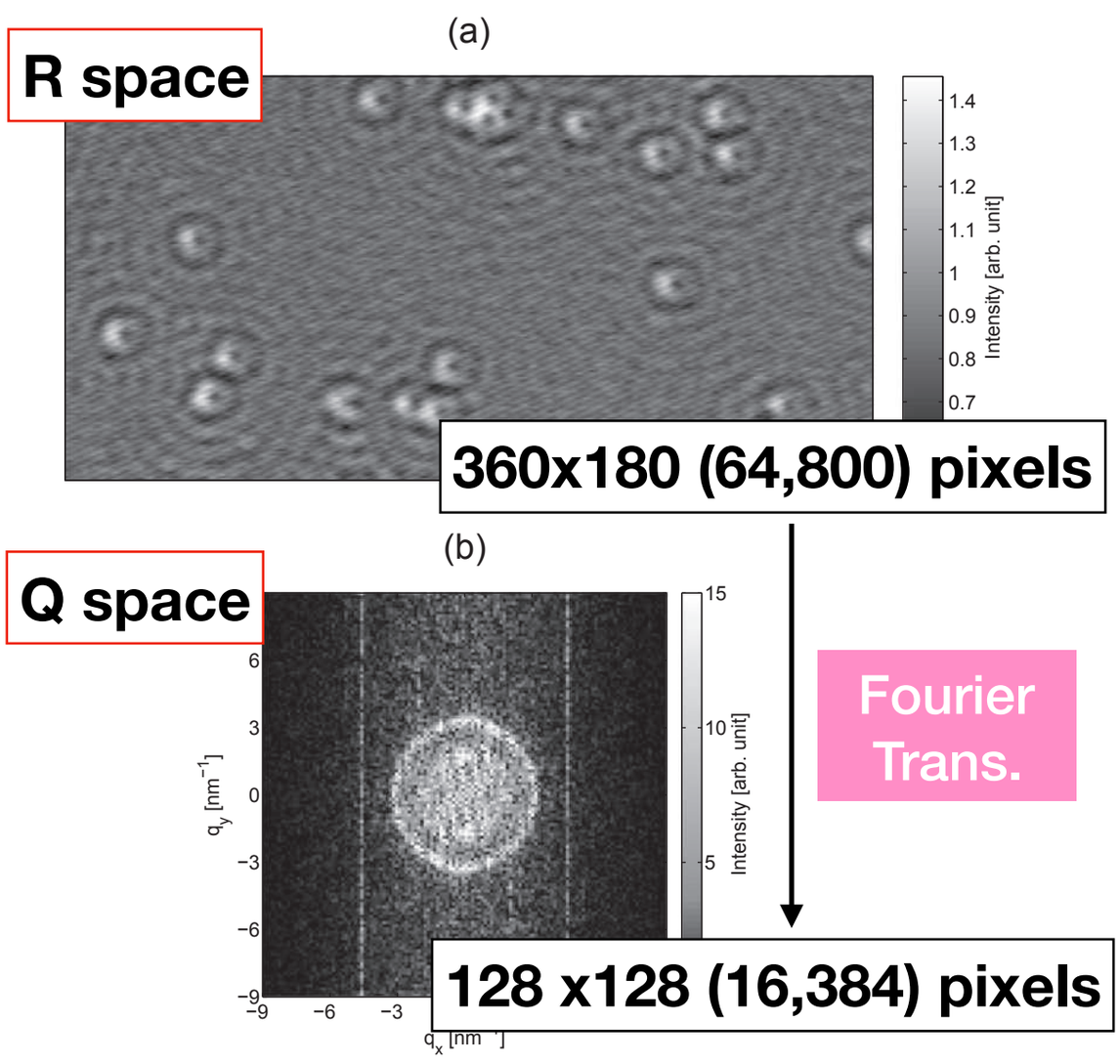
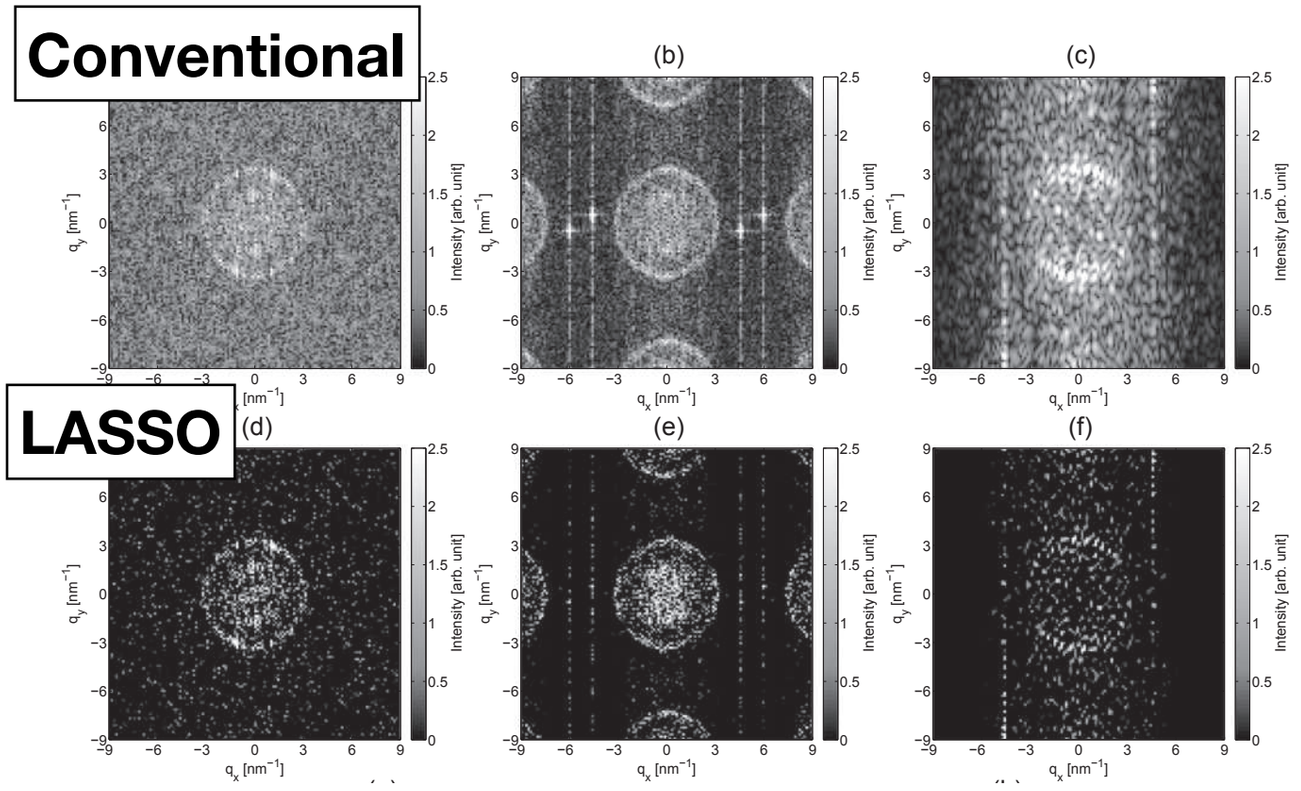


Fig. 1. (a) dI/dV map of Ag(111) surface. (b) FT of (a) obtained by conventional method.

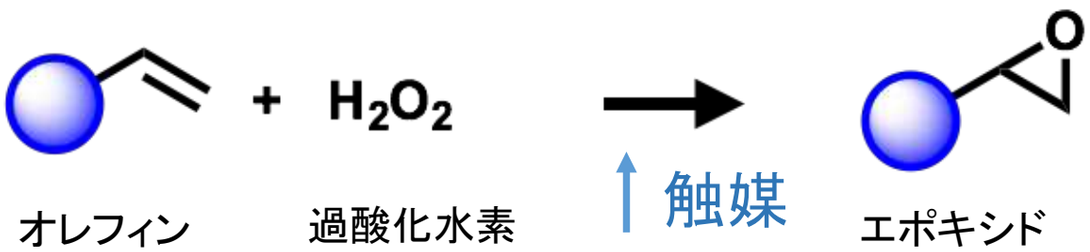


Q space is really sparse
Sampling in R space can be reduced.

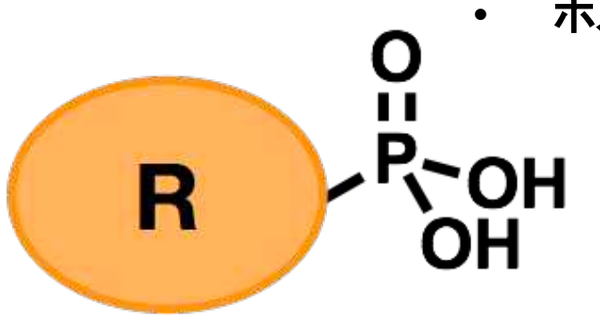
Chem. Lett. **47**, 284–287 (2018).

計測データを計算データでフィット

高活性ホスホン酸触媒をシミュレーションからデザインすることに成功 （矢田・安藤・永田）



- タングステン触媒
- 相間移動触媒
- **ホスホン酸触媒**

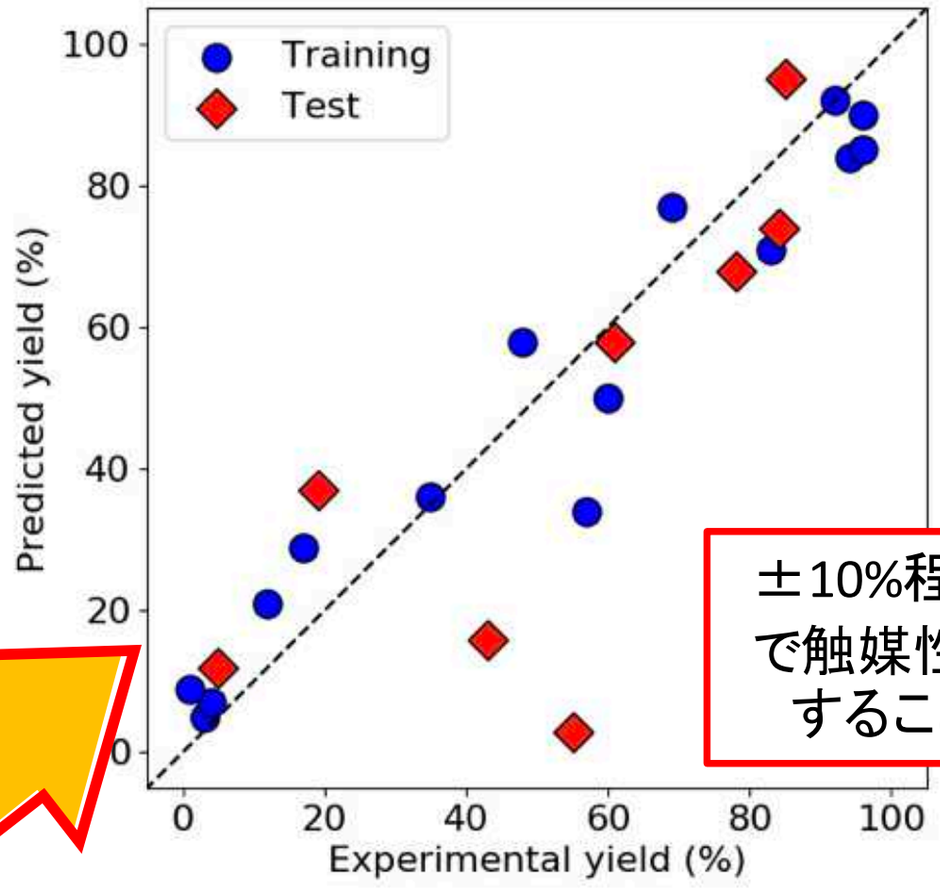


エポキシド高収率なホスホン酸の側鎖Rを予測したい

量子化学計算によって触媒を数値化

実験によって触媒性能を評価

AI技術で関連付）



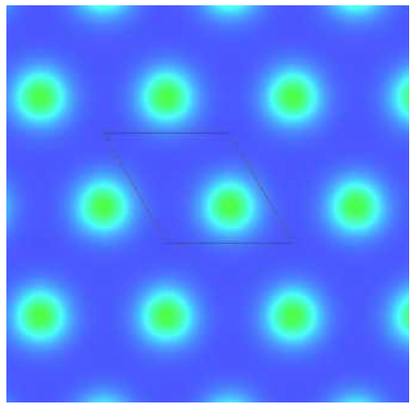
±10%程度の精度で触媒性能を予測することに成功

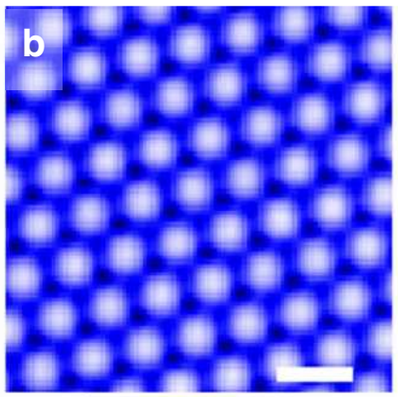
- **実験をせずに高収率触媒を予測・性能を実証**
- **専門家の勘を再現することに成功**
- **AI, 機能材料, 触媒科学3センター連携による成果**

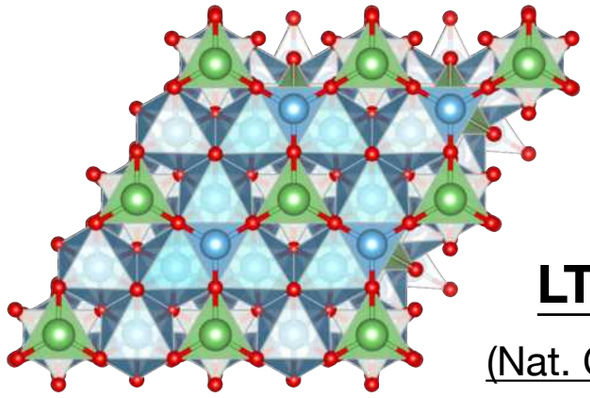
第一原理計算データを利用した機械学習ポテンシャル研究

原子・分子シミュレーション

安定構造探索





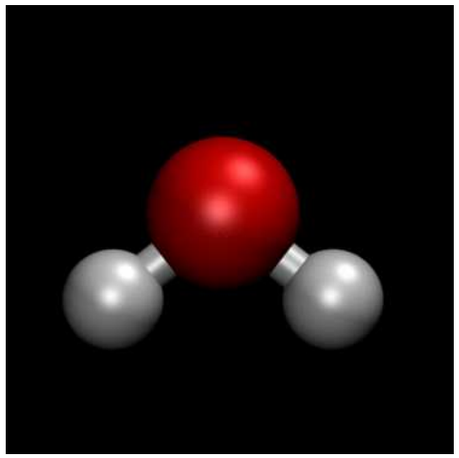


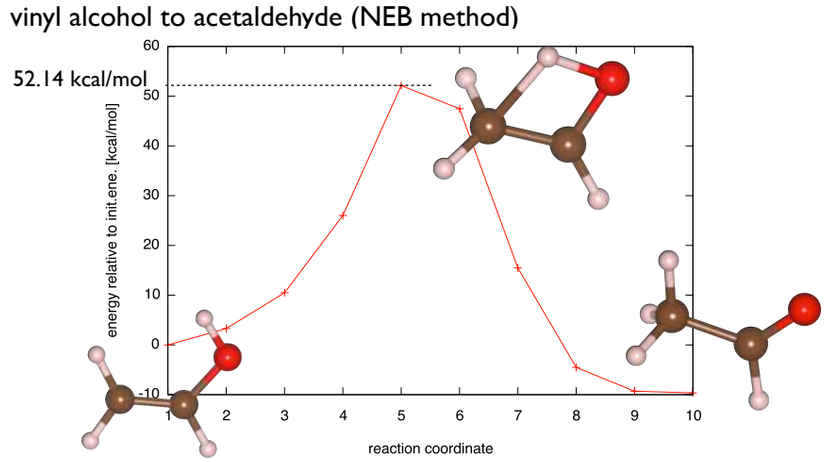
LTO表面構造の同定

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

- 静的な安定状態から動的な挙動まで
- 構造と特性・観測データを結びつけるキー
- ナノスケール物理では必須のツール

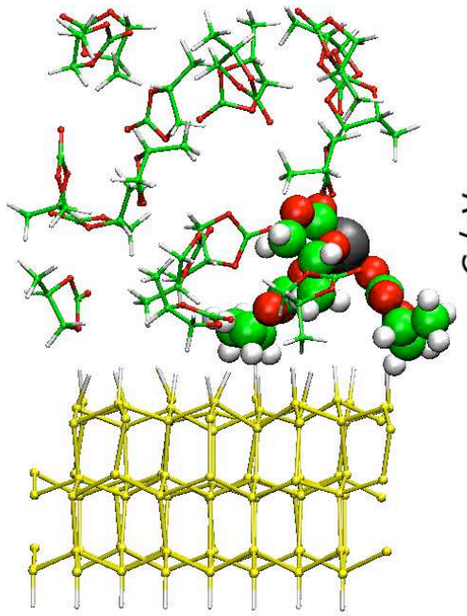
振動モード・反応座標

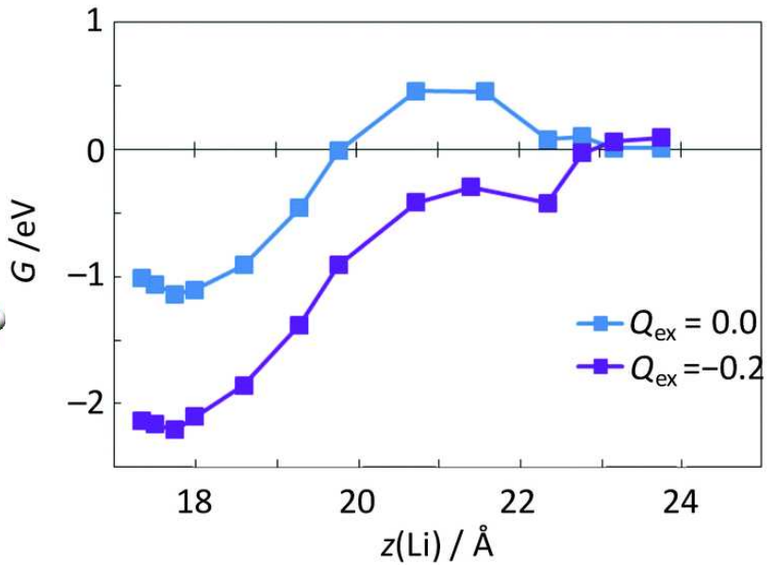




DMol3 TS search (LST/QST) 51.473 cal/mol, -10.851 cal/mol (PBE)

統計サンプリング(自由エネルギー・エントロピー)

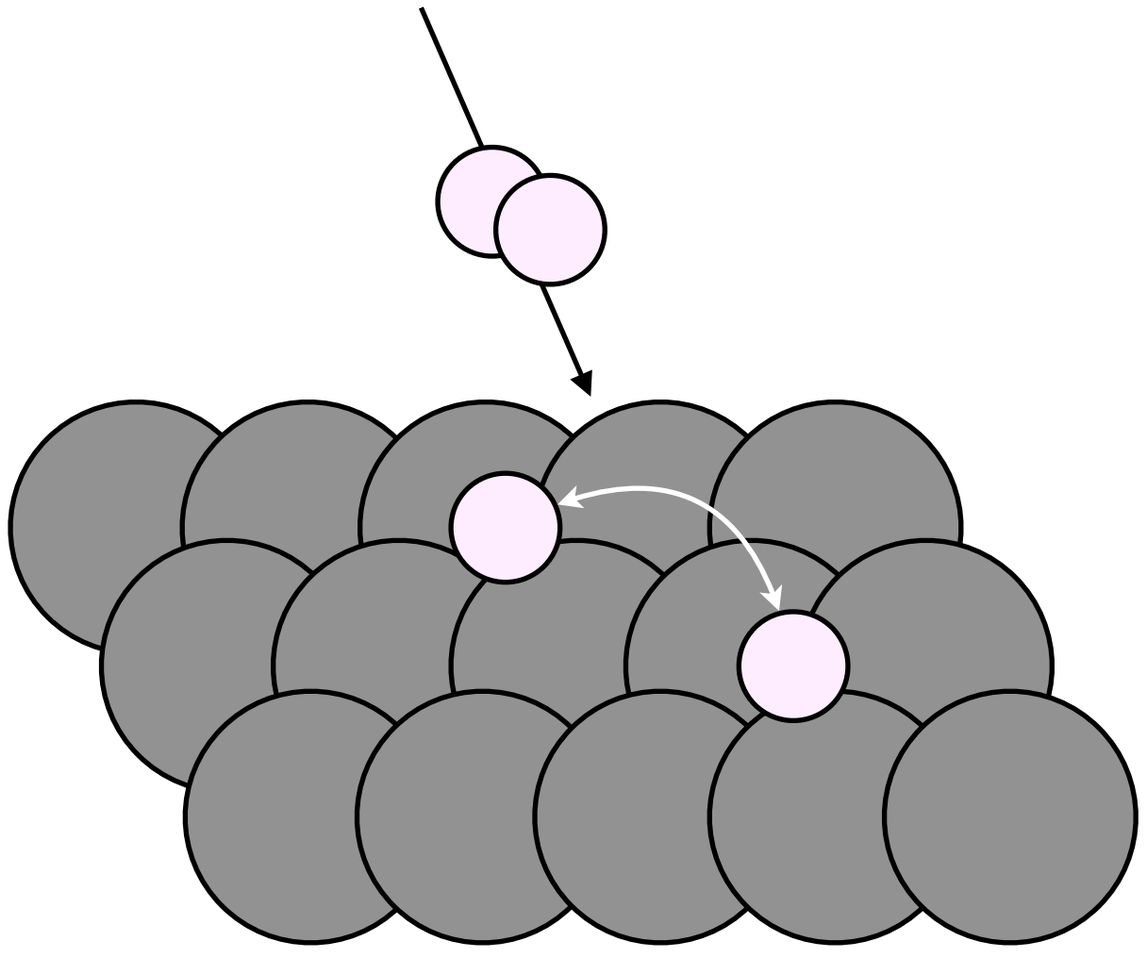




Phys. Chem. Chem. Phys., 2018, 20, 11586-11591

NNポテンシャル発展の鍵

表面化学反応のシミュレーション



Pd表面でのH_2分解反応

方法論に求められること

1.電子状態変化が取り扱えること
2.エネルギー超曲面上でのダイナミクス
3.ダイナミクスの統計性

- 電子状態変化を伴うため、DFT計算が必須
- DFTのみでは計算が大変で統計が集まらない

NNによるモデリングの試みは

1995年, 2004年にすでに報告あり

J. Chem. Phys. **103** (10), 8 (**1995**) .
Chem. Phys. Lett. 395, 210 (**2004**).

NNフィッティングの諸問題

表面構造の特徴量設計

fcc(111)の表面形状に合わせてフーリエ変換を元に構成する方法を提案[1]

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

ただし問題に合わせて複雑な関数を設計することは決して容易ではなくもっと簡便な表現が必要に。

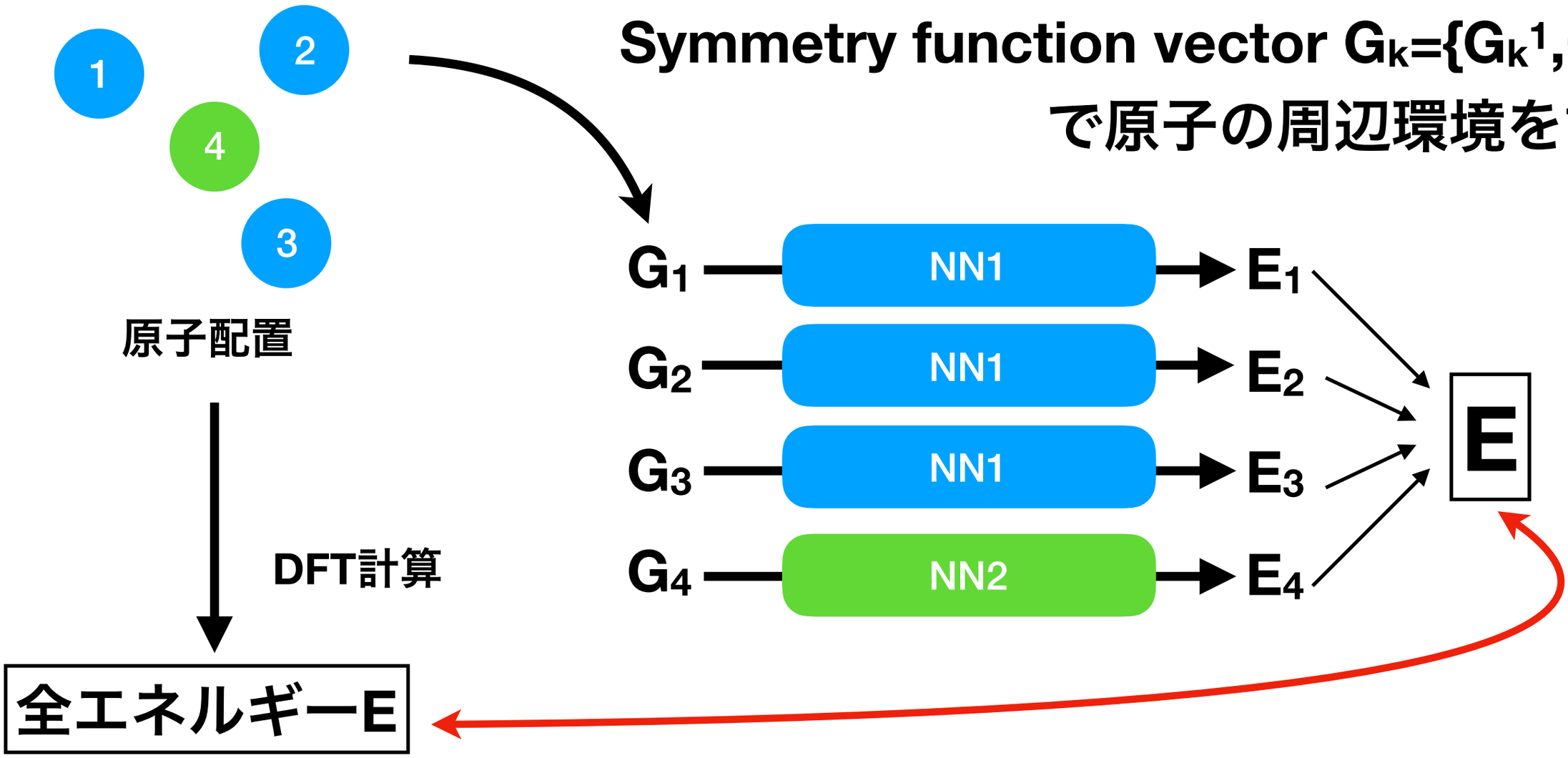
３次元配置を入力にしたNNの問題

- 入力層が原子数に依存するため、異なる系には適用できない（汎用性）
- 同種粒子の入れ替えに対する対称性が破れてしまう
- 同様にポテンシャルが満たすべき対称性が保証されない

入力表現・汎用性・対称性がキーワード

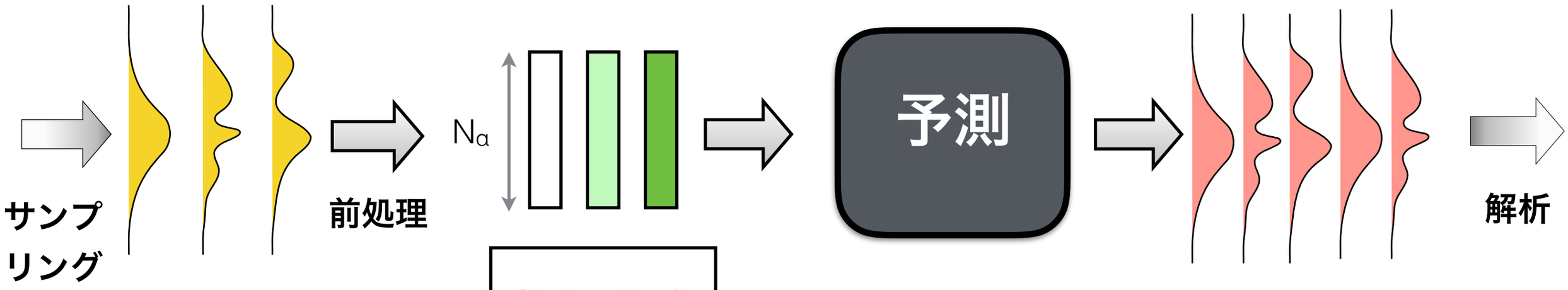
Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).

Behler - Parrinelloの方法



- 正味のINPUTは、サブネットの入力 x 原子数
- 学習すべきNNのパラメータは、原子種ごとに共通。
- また入力ベクトルGの次元は固定できるので原子数を増やすことは容易

機械学習ポテンシャル作成の構成技術



生データ

構造の記述

モデリング

データ予測

DFT計算結果

（説明変数）原子の配置

（目的変数）エネルギー

（またはforce）

説明変数の低次元化

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

回帰モデル

Neural network
多項式展開

ガウス過程回帰

Kernel ridge回帰

High-Throuput MD

- 安定構造探索
- 材料スクリーニング
- MDからの拡散係数計算
- 網羅的NEB計算

機械学習ポテンシャルは**（1）構造の記述（2）モデリング**の違いで大別

使用するメリット・目的で手法を使い分ける必要がある

アモルファス中のイオン安定配置の全探索

大問題：高精度に計算できるDFTではコストがかかりすぎる

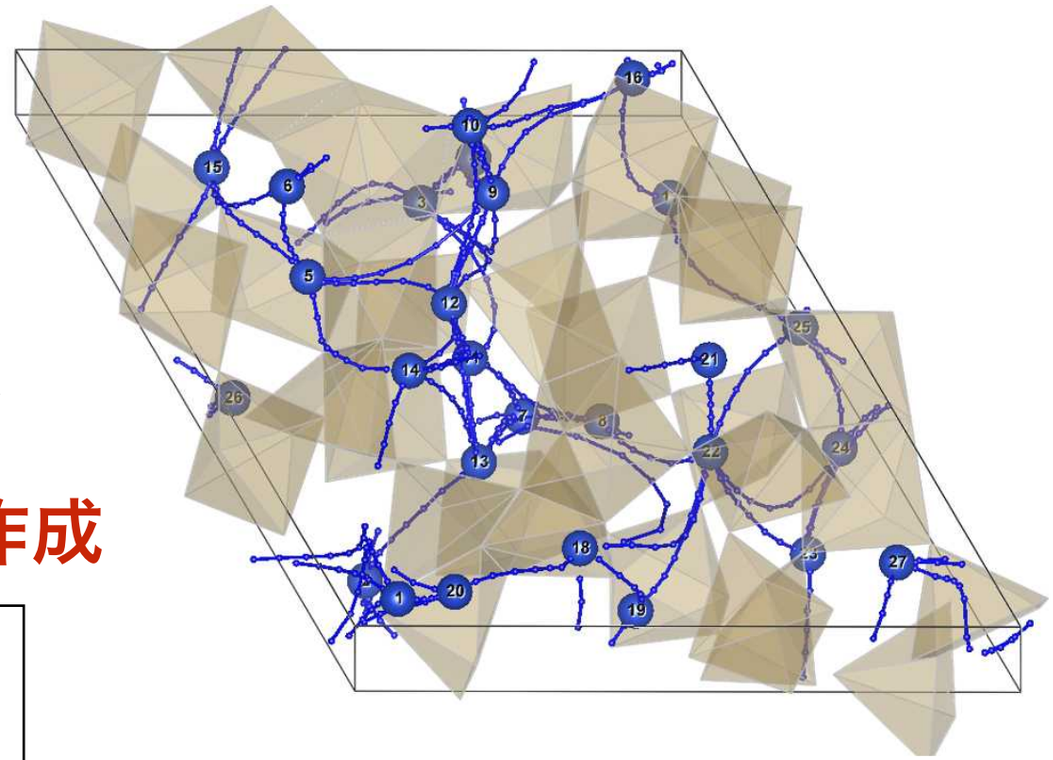
(N^3 回の構造最適化 x 計算時間 T)　N~50, T~1 hour以上 > 10 year

さらに拡散障壁の計算も必要…

$$E_{\mathrm{tot}} = E_{\mathrm{amorph}} + \Delta E_{\mathrm{Cu}} + \Delta E_{\mathrm{opt}}$$

Cu挿入による エネルギー変化

構造緩和による エネルギー変化



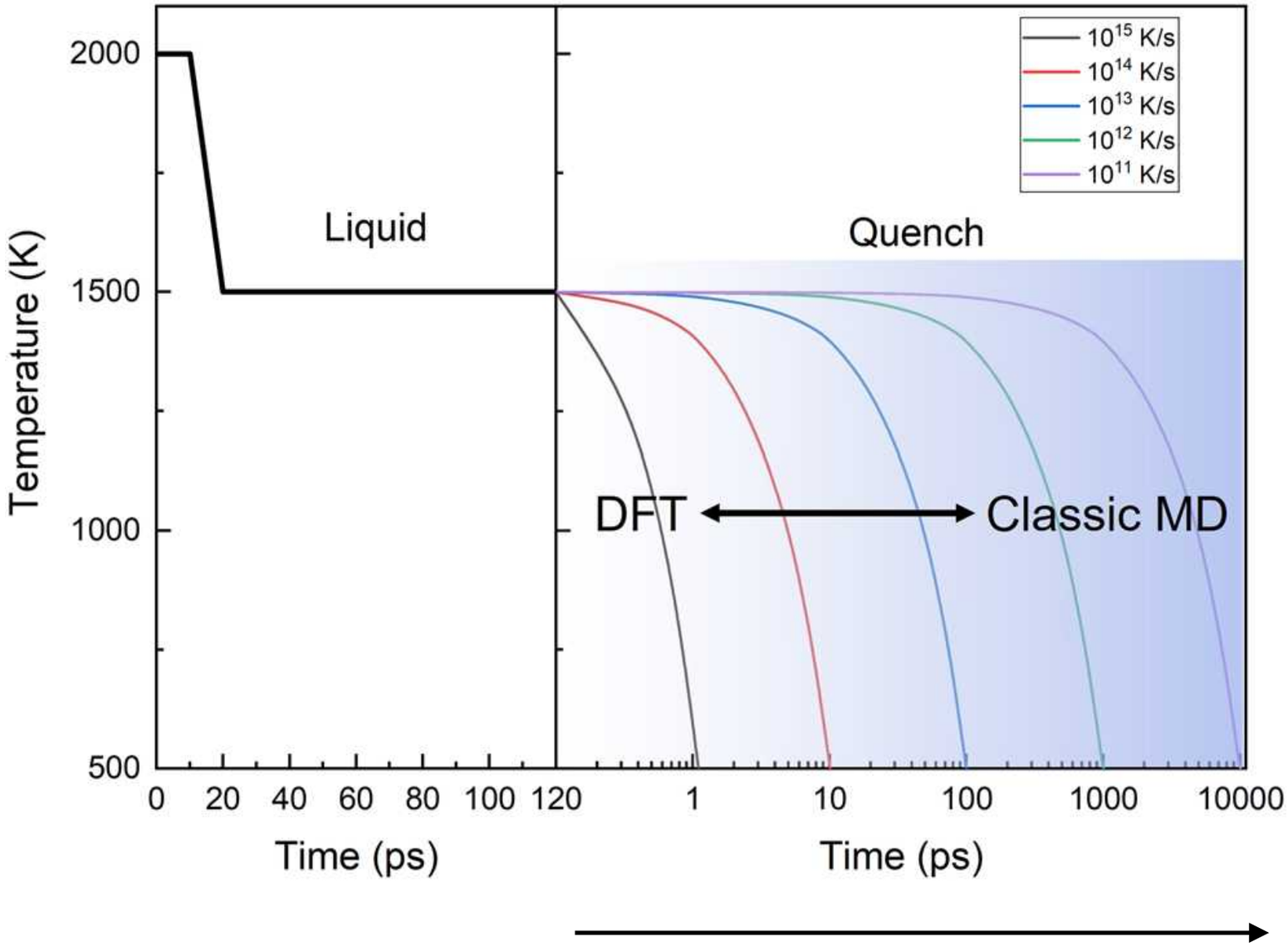
拡散ネットワークの全体図

Behler-Parrinelloの方法でポテンシャル作成

- ✓ アモルファス母体を「場」として取り扱う
- ✓ 学習対象は拡散粒子（Cu）のみ
- ✓ ３元系でありながら小規模なネットワークで記述可能
- ✓ 事実上１粒子の計算なので、エネルギー評価も容易
- ✓ アモルファスは動かせないのでMDはできない

W. Li, and Y. Ando, J. Chem. Phys. **151**, 114101 (2019).

焼きなまし法の冷却速度依存性



計算コストが10倍ずつ増加

- **DFTで実現できる範囲では冷却が早すぎる**
- **Classical MDのポテンシャルが妥当か怪しい**

機械学習ポテンシャルで克服

NN configuration	RMSE (meV/atom) Training	Testing
48-10-10-10-1	6.6	6.5
48-20-20-20-1	6.3	6.4
48-30-30-30-1	6.1	6.2
48-40-40-40-1	**5.9**	**6.2**
48-50-50-50-1	5.8	6.2

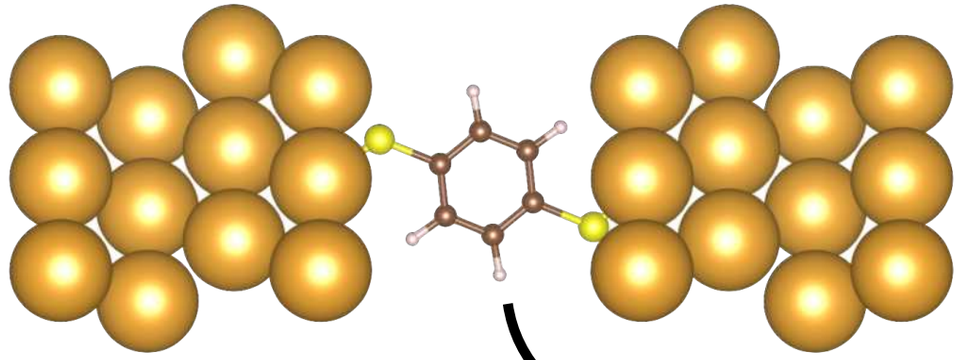
計算・計測データの自動解析技術

安藤, 藤掛, 渡邉, 表面科学 36, 515-520 (2015).

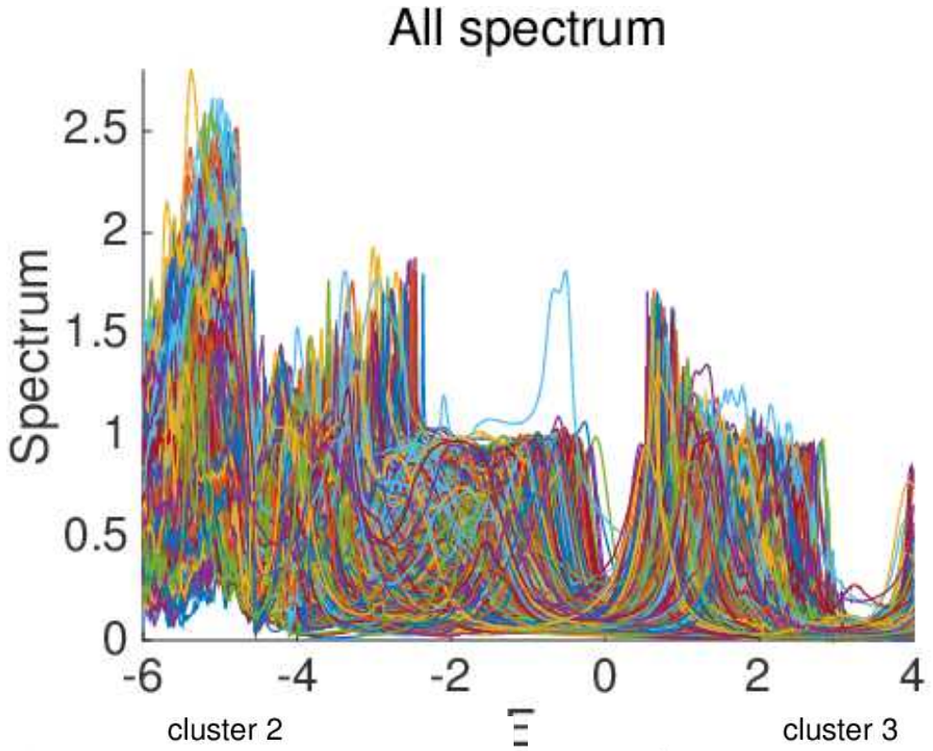
自明なマップ上の類似度による分類

課題：**『たくさんあるデータを自動で分類したい』**

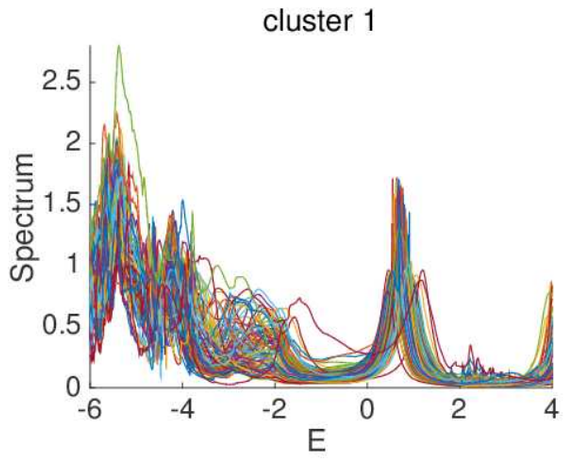
単分子電気伝導スペクトル

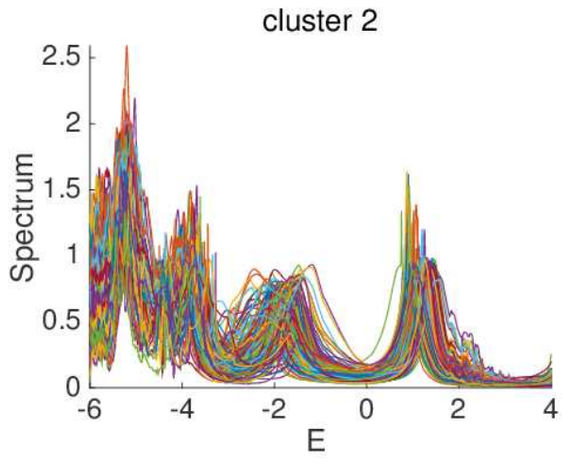


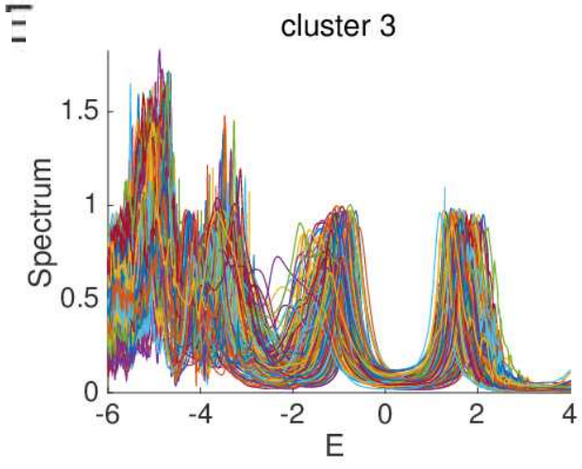
挟み込む分子の
配置で形状が非自明に変化



クラスター解析

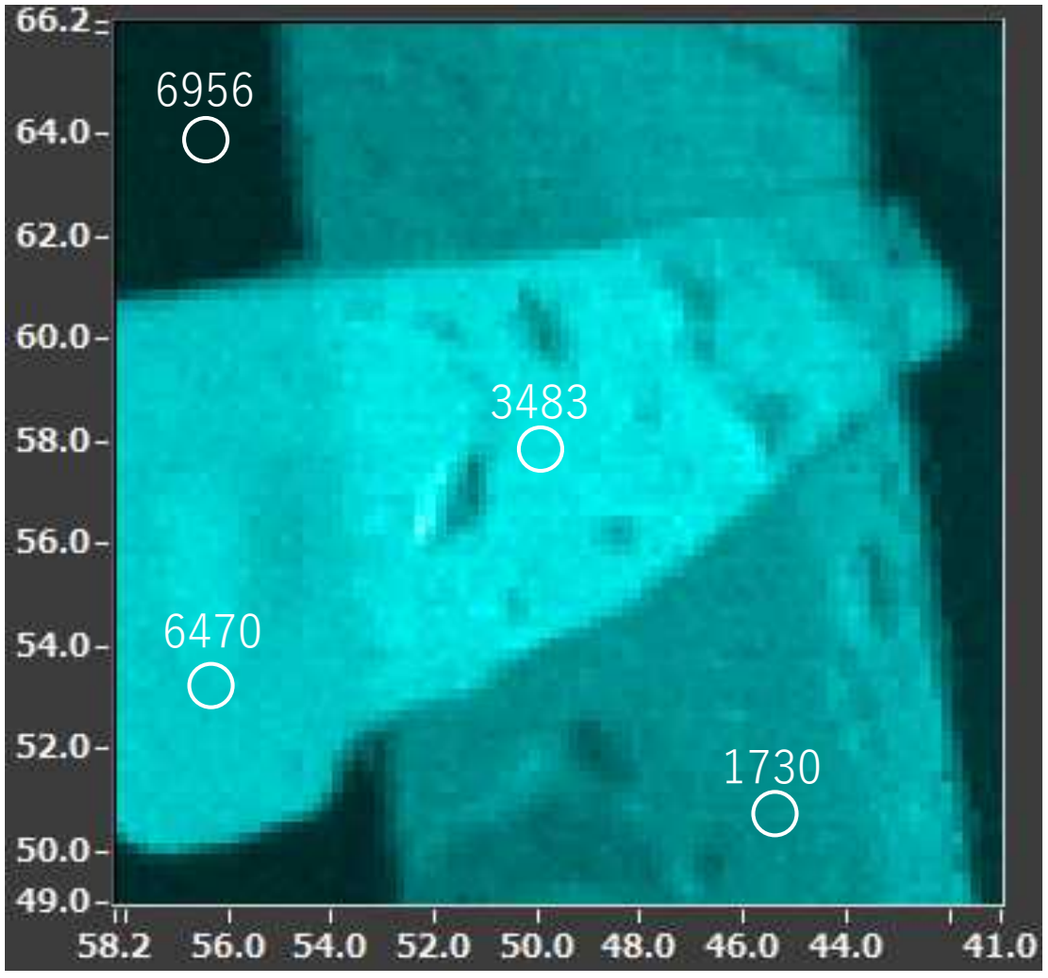






類似スペクトルが持つ共通構造を解析

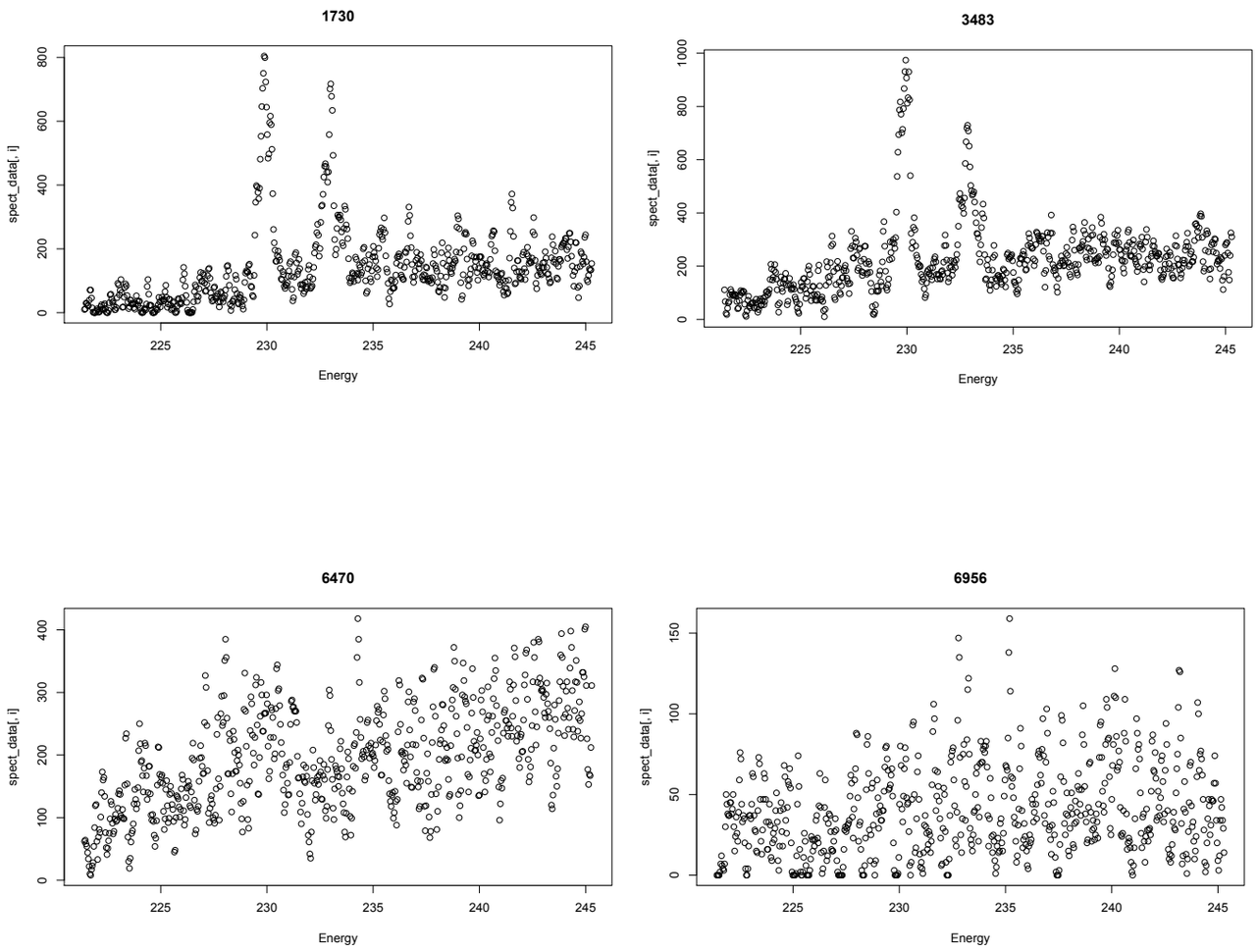
他の「スペクトルが多すぎる」問題



白丸の上の数字が右のスペクトルデータの数字に対応

白丸が各データの得られた大体の位置に対応

(86 x 86 pixels) 7,396

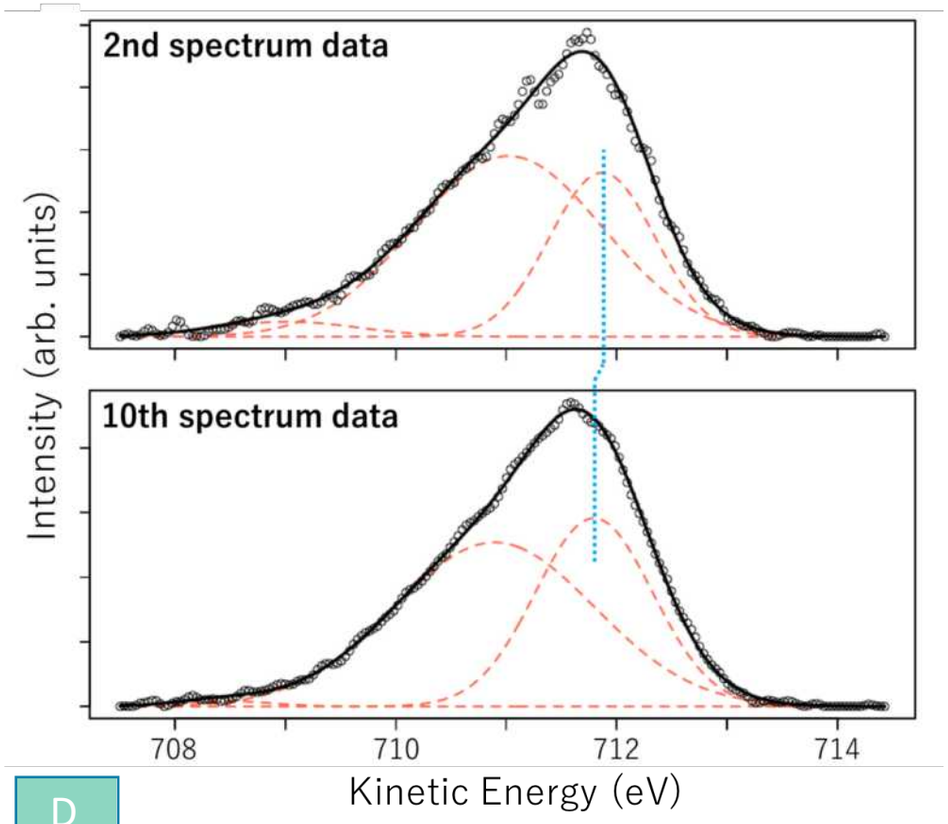


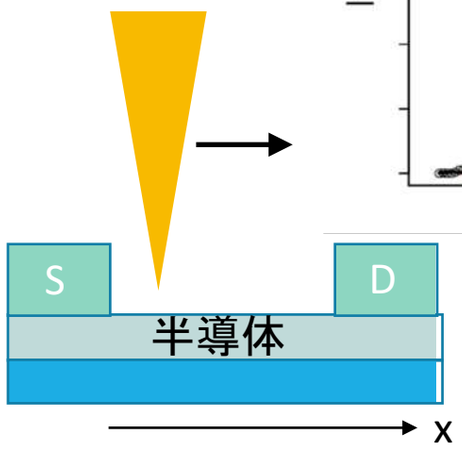
二本のピークが綺麗に見える領域(上二つ)と見えない領域(下二つ)があった

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

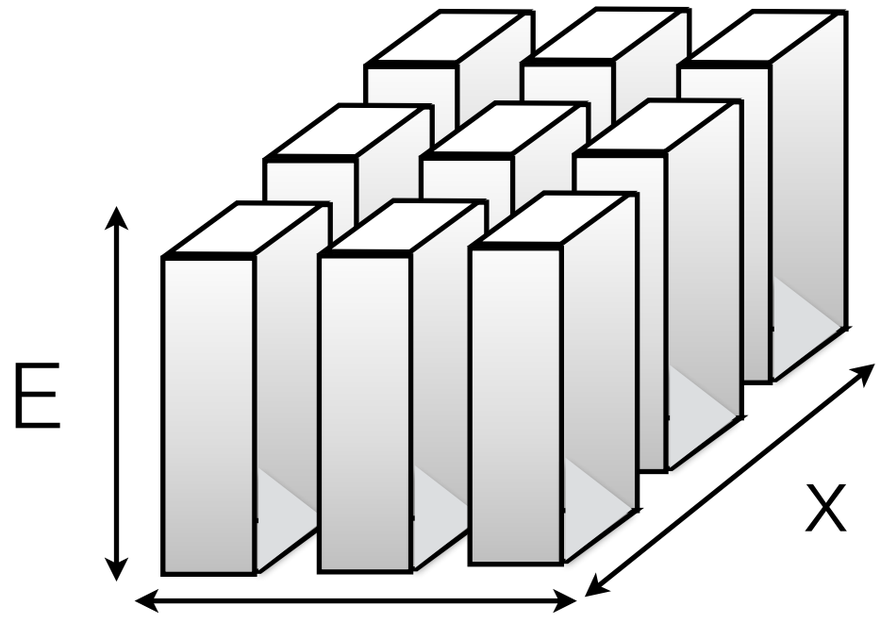
ピーク位置の自動推定

課題：**『ピークの位置を大まかにでも自動抽出したい』**





X線を使った局所分光実験



特徴画像化

全てのデータからピーク位置を抽出する必要あり

複数のガウス関数でフィット？

実験研究者の悩み

非線形最小二乗フィットはしんどい

非線形最小二乗フィット = いい感じの初期値探し

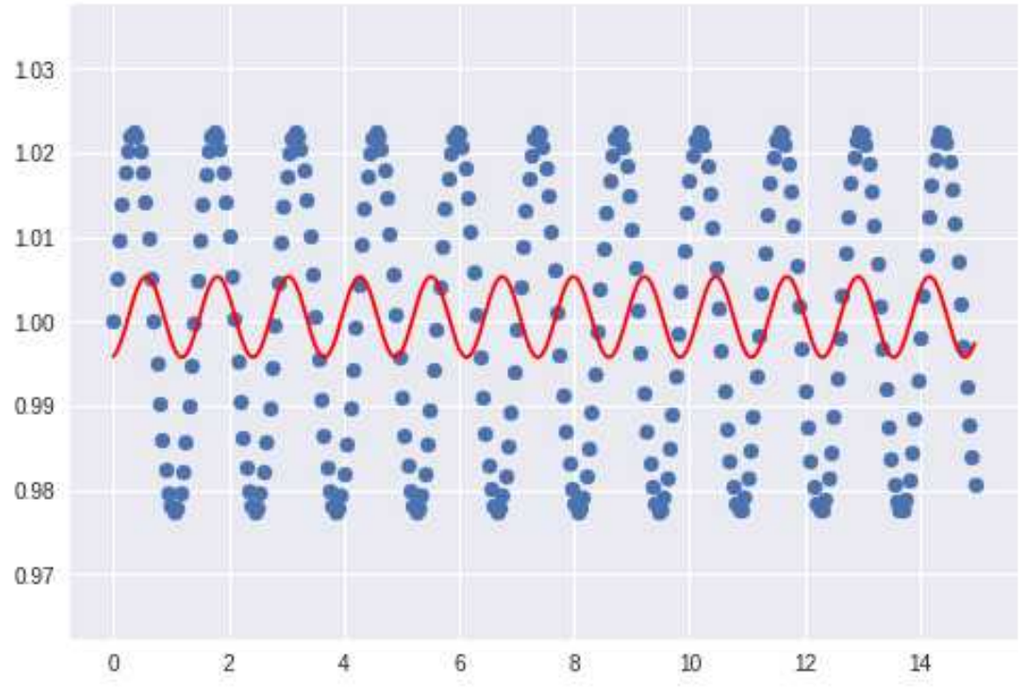
データを見て、”経験と勘”でいい感じの初期値を探す

データそれぞれを研究者が見て処理する必要あり

取得データ数が増えると、とてもこなせない！！！

解決策

パラメータを効率的に改善（EMアルゴリズム）
試行錯誤を自動化（モンテカルロ法）



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

ノイズなしデータでも、初期値が悪いと
明らかにフィットしてない

EMアルゴリズムの適用

混合ガウスモデル推定

推定対象

1. 分布の平均 μ_k と分散σ^2_k, 混合率π_k
2. データ点n の潜在変数 r_{nk}

E (Expectation)**-step**：（1）をもとに（2）を推定

データx_nに関する潜在変数r_{nk}の期待値（負担率 $\gamma(z_{nk})$）

$$E[z_{nk}] = \frac{\pi_k N(x_n, \theta_k)}{\sum_j \pi_k N(x_n \mid \theta_k)} = \gamma(z_{nk})$$ ← (1)とx_nに依存

M (Maximization)**-step**：（2）をもとに（1）を推定

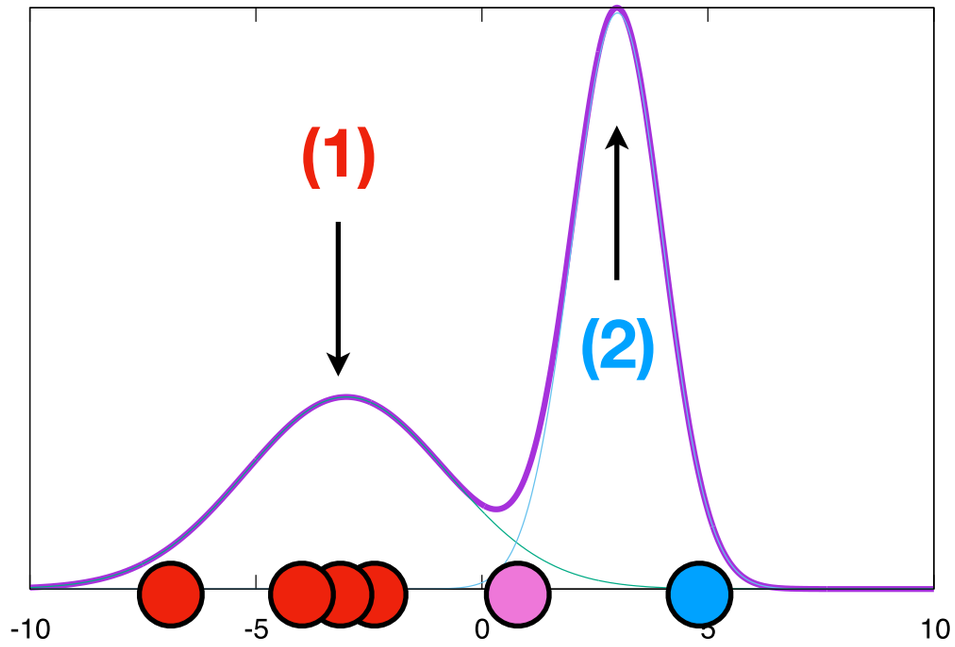


図３－１ 混合ガウス分布の例

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

完全データ（データx_nと潜在変数r_{nk} の組）から対数尤度の期待値を計算して最大化

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

繰り返すと尤度が単調増加！

実データへの適用上の課題

放射光スペクトルはイベント数が膨大（~10^7 events）

混合ガウス分布のモデルと対数尤度

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

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

データ数に関する和

データ数が少ないと精度がでないが、
データ数が多いと計算が重くなる

放射光スペクトルのイベント数を適切に生かし
かつ高速に動作するように改良が必要

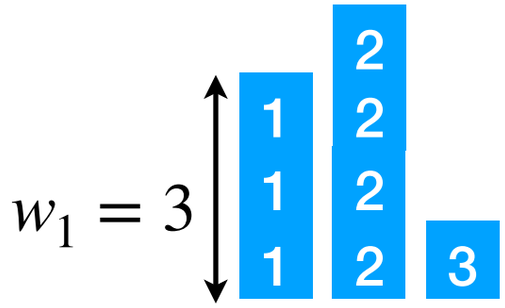
スペクトル解析に適したEMアルゴリズム

通常のEMアルゴリズムの入力: $X_{1D} = \{x_1, x_2, x_3, \cdots, x_N\}$

イベント列 1 2 3 1 2 2 2 1

改良型EMアルゴリズムの入力: $X_{2D} = \begin{Bmatrix} \hat{x}_1, & \hat{x}_2, & \hat{x}_3, & \cdots, & \hat{x}_M \\ w_1, & w_2, & w_3, & \cdots, & w_M \end{Bmatrix}$

ヒストグラム形式



改良型EMアルゴリズムの対数尤度

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

$$= \sum_{n=1}^{M} w_n \ln \left(\sum_{k=1}^{K} \pi_k N(\hat{x}_n \mid \mu_k, \sigma_k) \right)$$

（ポイント: N >> M）

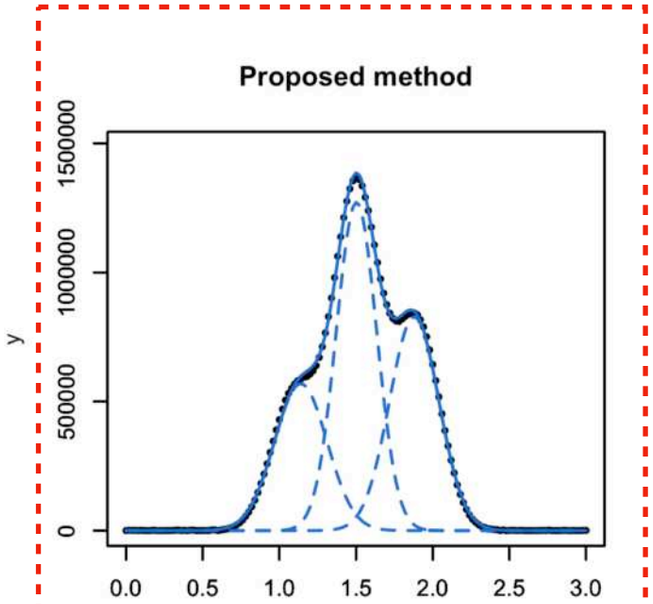
ヒストグラムのビン数に関する和

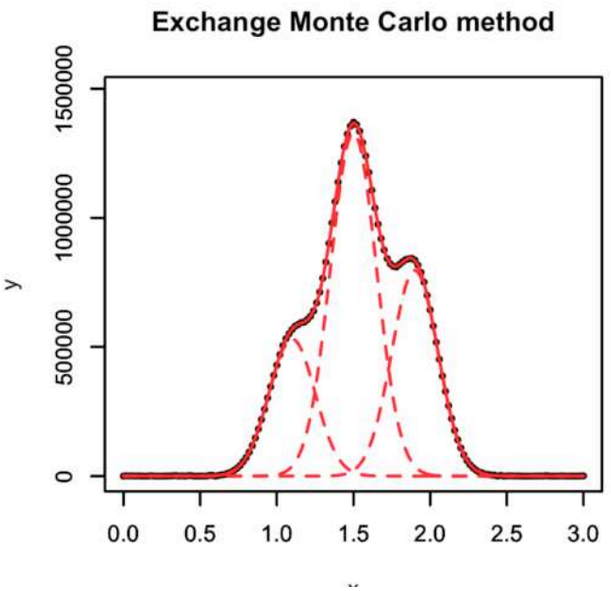
- ビン数はイベント総数より圧倒的に小さい
- イベント数が増えても演算回数が変わらない

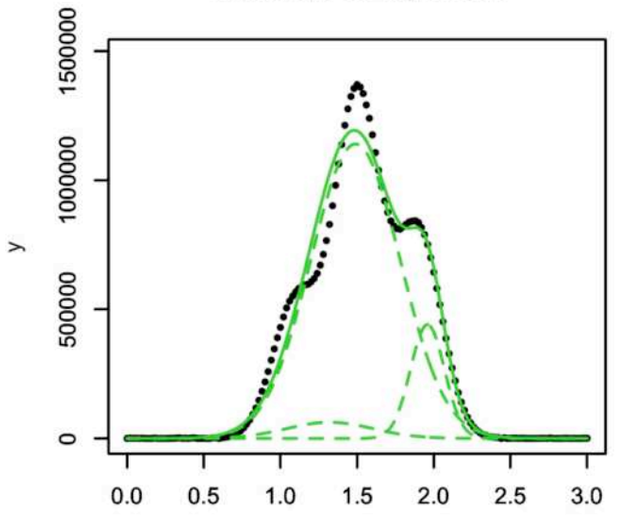
T. Matsumura, N. Nagamura, S. Akaho, K. Nagata, and Y. Ando, Sci. Tech. Adv. Mater. **20**, 733 (2019).

計算速度と精度に関する性能評価

通常のEMアルゴリズム：計算負荷が測定イベント数に依存　⇦　改善！



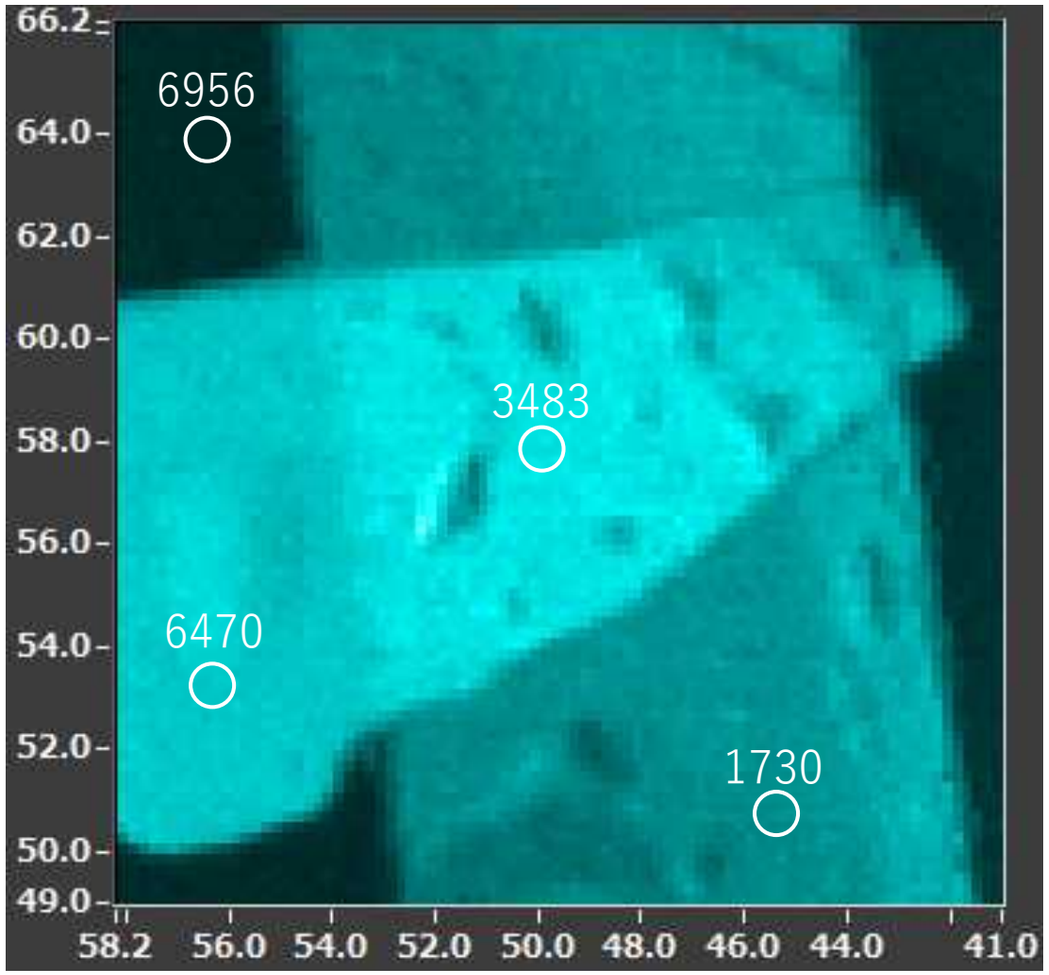




	Proposed method	Exchange Monte Carlo method	Ordinary EM algorithm
時間[s]	**0.8**	481.0	19748.5
RMSE	8707.4	933.8	66086.0

- ランダムな初期値でも安定して動作、極めて低コストでモデル推定（ピーク抽出）が可能
- 初期値を振って最良のモデルを探せばさらに精度をあげられる
- 高ノイズデータ・埋もれたピークに関してはモデリングが困難（事前知識が必要）

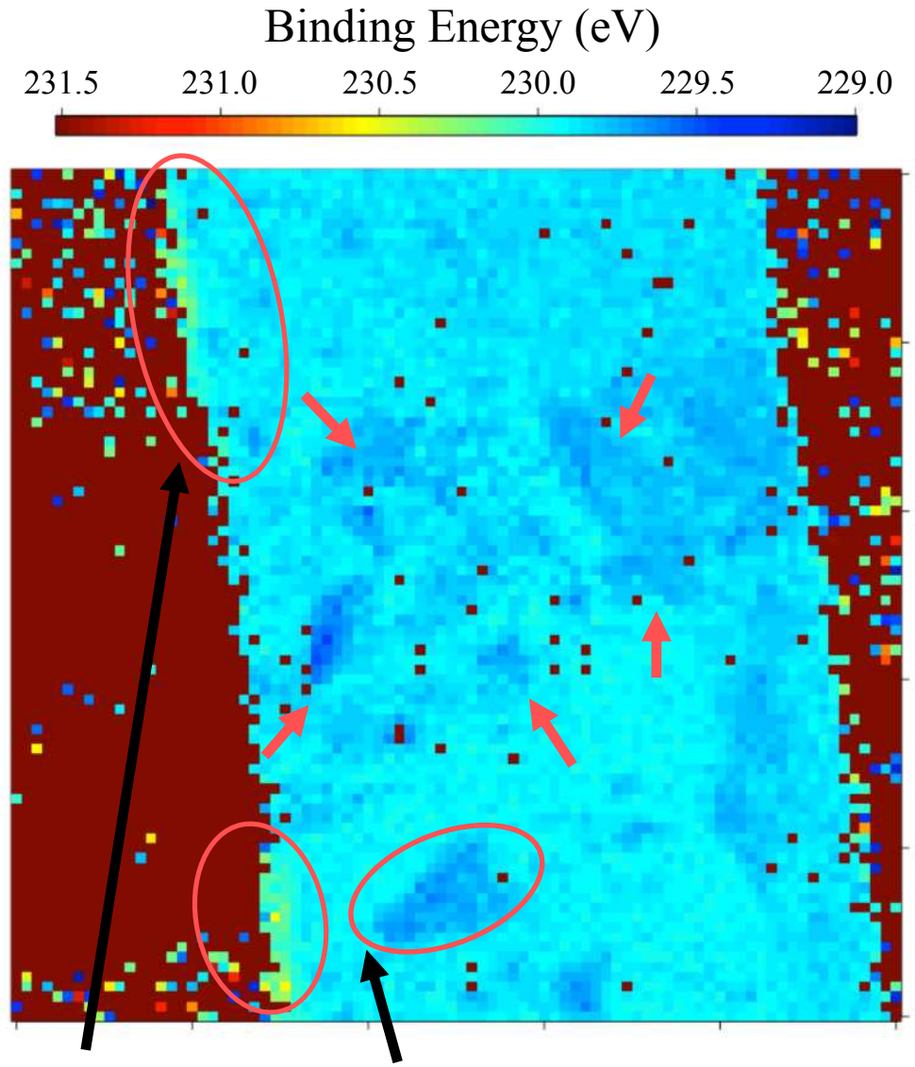
やりたかったこと



白丸の上の数字が右のスペクトルデータの数字に対応

白丸が各データの得られた大体の位置に対応

光学顕微鏡の観測結果



この辺りのピーク位置の違いは光学顕微鏡像に現れていない

EMアルゴリズムによるピーク位置推定結果

（~12 h程度）

まとめ：今後、物質科学者として目指したいこと