機械学習を用いた第一原理吸着エネルギー面計算の高速化

Application of Machine Learning
for Fast Evaluation of Chemisorption Energy Surface
from the First Principles

物理学専攻　北原慎平
Dept. of Phys., Shimpei Kitahara

1. Introduction

準結晶分野の研究においては、単一の元素からなる準結晶を発見することが大きな目標の一つである。そのための方法として、基板を与えてその基板構造を模した単元素の準周期配列の薄膜を作成するという研究がある [1][2]。これらの研究では、Ag-In-Yb 準結晶基板上に Pb や Bi を蒸着させる事によって、基板構造を模した準周期構造の Pb や Bi を結晶成長させることに成功している。またそこでは、準結晶表面に吸着する Pb や Bi の吸着エネルギーのポテンシャル面を第一原理計算により求め、実験結果の解釈を行っている。しかしこういった準結晶研究の問題として、計算コストが膨大なことが挙げられる。周期的な結晶の場合、安定した吸着構造を得るためには通常、ポテンシャル勾配に従って原子を動かす構造緩和計算が行われる。一方、準結晶は表面平行方向に周期性を持たないため、通常のバンド計算の手法の中で扱うにはクラスターで近似する必要がある。しかし、クラスターのエッジ付近では本来の結晶中の状態と異なるため、原子に働く力を正確に把握することができない。そこで吸着エネルギーを準結晶表面上の全ての格子点において計算することで安定な吸着構造を推論することになる。

本研究では吸着エネルギーの効率的な計算を目指して機械学習の手法を導入し、既得の Ag-In-Yb 5 回表面上の Bi 吸着エネルギーデータを用いてより少ないサンプリングで吸着エネルギー面の全容を予測する可能性を検討した。そこで吸着現象を考える上で必要なモデルの精度目標を設定し、その精度を得るために必要なサンプリング回数をいかに減らすことが出来るかを本研究の目的とした。この目的に沿った機械学習手法として、我々はベイズ最適化を採用した。ベイズ最適化を用いることで、吸着構造の特に低エネルギー領域を効率的に予測することができるが、適切な設定が求められるパラメ―タが複数存在する。そこで我々は目的を達成するための最適な条件を調査した。この条件を見つけることで、他の物質への応用に加えて、今回扱った Bi の計算にもまだ使用できる。準結晶は厳密には表面上に同じ原子配置は存在しない。本研究で対象とした面は 5 回表面のほんの一部分であり、少し離れたところに未調査の異なる原子配置の面がある。また 5 回表面自体も実験的に出やすい面を扱っているに過ぎず、その面上の原子配置で確定ではない。吸着エネルギーは吸着子と表面原子間の相互作用によって決まるので、エネルギー的に不安定な原子を表面から排除した場合、再度吸着エネルギー面を計算し直さなければならない。このように Bi の計算においても多くの未調査の面が存在するが、現状は計算コストの問題で気軽に計算してみることが困難である。よって本研究で計算を高速化する方法をベイズ最適化の枠組みの中で調査した。

2. Method

ベイズ最適化は、既得データからガウス過程回帰により目的変数 $y = f(x)$ の予測値 $\mu(x)$ と分散 $\sigma^2(x)$ を求め、次にそれらを変数に持つ獲得関数が最大となるように次の候補点の x を決定する。ここで n 個のデータがある時、$n+1$ 個目の x における $\mu\left(\mathbf{x}^{(n+1)}\right)$ と $\sigma^2\left(\mathbf{x}^{(n+1)}\right)$ は以下のように表される。

$$\mu\left(\mathbf{x}^{(n+1)}\right) = \boldsymbol{\sigma}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}_{\mathrm{obs}} \tag{1}$$

$$\sigma^2\left(\mathbf{x}^{(n+1)}\right) = K_{\mathrm{GPR}}\left(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n+1)}\right) - \boldsymbol{\sigma}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\sigma} \tag{2}$$

ここで $\mathbf{y}_{\mathrm{obs}}$ は y の n 個の要素を持つベクトル、K_{GPR} はカーネル関数である。また $\boldsymbol{\sigma}^\top$、$\boldsymbol{\Sigma}$ は以下の通りである。

$$\boldsymbol{\sigma}^\top = \begin{pmatrix} K_{\mathrm{GPR}}\left(\mathbf{x}^{(1)}, \mathbf{x}^{(n+1)}\right) & \cdots & K_{\mathrm{GPR}}\left(\mathbf{x}^{(n)}, \mathbf{x}^{(n+1)}\right) \end{pmatrix} \tag{3}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} K_{\mathrm{GPR}}\left(\mathbf{x}^{(1)}, \mathbf{x}^{(1)}\right) & \cdots & K_{\mathrm{GPR}}\left(\mathbf{x}^{(1)}, \mathbf{x}^{(n)}\right) \\ \vdots & \ddots & \vdots \\ K_{\mathrm{GPR}}\left(\mathbf{x}^{(n)}, \mathbf{x}^{(1)}\right) & \cdots & K_{\mathrm{GPR}}\left(\mathbf{x}^{(n)}, \mathbf{x}^{(n)}\right) \end{pmatrix} \tag{4}$$

次に本研究で用いた２つの獲得関数の定義を示す。

- **Lower Confidence Bound (LCB)**

LCB は以下の式で定義される。

$$\mathrm{LCB} = \beta \times \sigma(x) - \mu(x) \tag{5}$$

ここで β は標準偏差 $\sigma(x)$ をどの程度重視するかを決定するパラメータで、β が大きいほど分散の大きい領域を次に探索する。

- **Expected Improvement (EI)**

EI は現在までに見つかっているサンプル点の最小値 f_{min} の改善度 $I = f_{min} - z$ $(z \sim N(\mu, \sigma^2))$ の期待値であり、正規分布の確率密度関数 $p(z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$ を用いて以下の式で定義される。

$$\mathrm{EI} = \int_{-\infty}^{f_{min}} (f_{min} - z)\, p(z) dz \tag{6}$$

また式 (6) を変形することで、最終的に以下のように表現されることも多い。

$$\mathrm{EI} = \sigma(x) \cdot \phi(Z) + (f_{min} - \mu(x)) \cdot \Phi(Z) \ (\sigma(x) > 0) \tag{7}$$

$$Z = \frac{f_{min} - \mu(x)}{\sigma(x)},\ \Phi(Z) = \int_{-\infty}^{Z} \phi(t) dt = \int_{-\infty}^{Z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} t^2\right) dt \tag{8}$$

$\Phi(Z)$ は標準正規分布の累積分布関数、$\phi(Z)$ は標準正規分布の確率密度関数である。

3. Computational Method

予測精度の計算方法

ベイズ最適化に基づいて吸着エネルギーの回帰モデルを作成、更新するたびに真分布との誤差を計算する。精度指標としては平均絶対誤差 (MAE) を使用した。MAE は以下の式で定義される。

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{9}$$

(N：データ数, y_i：真値, $\hat{y}_i$：予測値)

求めたいエネルギー構造

実際の吸着位置を推定する上で、我々が考えるべき領域は、「全ての極小領域」と、「極小領域間のエネルギー障壁」である (以下、これらを合わせて「概形」と呼ぶ)。

MAE を計算するエネルギー範囲

概形を求めるには、基本的に吸着エネルギーが負の領域に注目すればよい。しかしエネルギー障壁は必ずしも負の範囲に収まるわけではない。そこで一つの目安として「極小領域の最小値から 1eV」を考える。また、表面上に複数ある各々の極小領域から 1eV の範囲を割り出す事は困難であるため、極小値の最も小さい１つの領域のみに注目し、その極小値から 2eV の範囲を考えることで、なるべく全ての極小領域とエネルギー障壁を含むように MAE を計算した。

予測モデルの精度目標

我々は概形予測の精度目標を、以下の図 4.4 から MAE が 90meV と設定した。この図は、右上の「iter=」の数字がその段階でのサンプル点数に対応する。また青色の濃い部分が極小領域であり、黄緑色の部分がエネルギー障壁に該当する。すなわち色のついている花びら状の領域が概形といえる。ここで概形を予測するには、極小領域の位置と深さ、また花びら領域の輪郭 (赤色部分) の形状を求める必要がある。それが得られているのが、例えば iter=277 の図のモデルである。iter=501 の図ではより正確なモデルとなっているが、実用上、iter=277 のモデルで吸着位置の推定が可能である。

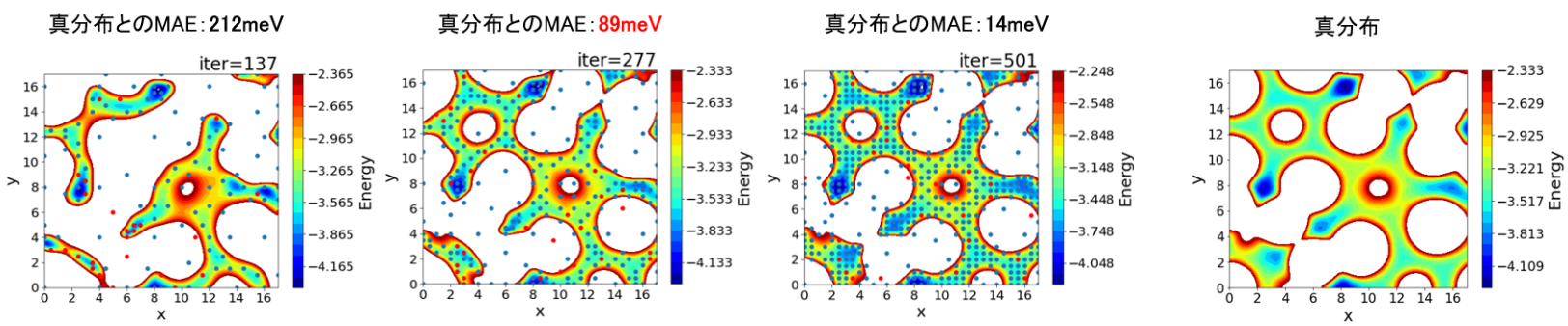

図 1　モデルの最小値から 2eV 以内の吸着エネルギー面の推移

4. Result and Discussion

右の図 2 は各獲得関数ごとに色分けし、各サンプリング終了時の MAE の値をモデルの最小値から 2eV の範囲で計算し、その 10 回平均を取ったものである。(使用したベイズ最適化実行ライブラリ「GPyOpt」の仕組みで、次のサンプル点の座標を決める際にランダム性が入るため、MAE が計算によって変化する。その傾向を見るために平均を取っている)。この 10 回の計算中 10 回で 90meV を達成したサンプル点数を以下の表 1 で示す。この表を見ると、LCB の $\beta = 3$ と EI で最小の 305 点のサンプリングを行えば目標を達成出来ることが分かった。これは全 1225 点の約 24.8% であり、サンプル点数を約 4 分の 1 に減少させることに成功した。また他の β の値でも最高で 389 点となっており、ランダムサンプリングのみ 529 点となった。以上から、ベイズ最適化が本研究で用いた吸着エネルギー面の探索において有用であることが示唆された。

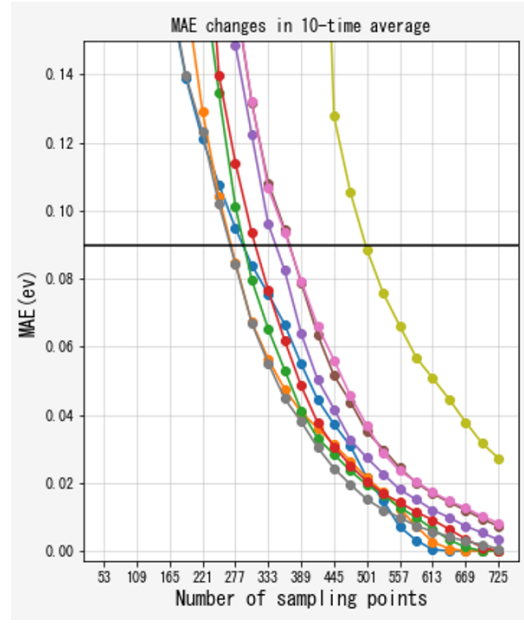

図 2　MAE の推移

表 1　90meV を達成するのに要したサンプル点数

獲得関数	LCB ($\beta = 2$)	$\beta = 3$	$\beta = 4$	$\beta = 5$
サンプル点数(個)	333	305	333	333

$\beta = 8$	$\beta = 11$	$\beta = 14$	EI	ランダム
389	389	389	305	529

5. Conclusion

獲得関数として LCB の $\beta = 3$ と EI を用いることで、最も少ない 305 点で目標を達成できた。全 1225 点の約 4 分の 1 に減少できたことが本研究での成果となる。MAE を計算する範囲やベイズ最適化のパラメータなど課題は多いが、本研究を通して獲得関数の違いによりどのようにサンプリングが進み、MAE が減少していくかを定性的、定量的に考察することができた。

参考文献

[1] Kazuki Nozawa and Yasushi Ishii 2017 J. Phys.: Conf. Ser. 809 012018.

[2] Sharma, H. R. et al. Templated three-dimensional growth of quasicrystalline lead. Nat. Commun. 4:2715 doi: 10.1038/ncomms3715 (2013).

[3] 野 澤 和 生, 石 井 靖 まてりあ 第 55 巻 第 6 号 (2016) doi:10.2320/materia.55.259. 単元素準結晶の結晶成長.

[4] Masanori Sato et al. Materials Transactions, Vol. 62, No. 3 (2021) pp. 350 to 355.