

機械学習を用いた無機材料の バンドギャップ予測に関する研究

五十嵐 広樹¹ 安尾 信明² 関嶋 政和^{1,2}

概要：ペロブスカイト太陽電池は、従来のシリコン型太陽電池の課題である、小型化・薄型化が可能となる新しい太陽電池として近年注目されている。しかし、現在は人体・環境に有害な鉛が含まれているものが主流になっており、鉛を用いないペロブスカイト太陽電池の開発が求められている。また、材料開発と情報技術・計算科学を組み合わせたマテリアルズ・インフォマティクスという分野が近年活発になってきており、更なる高速化・高度化を目指し、機械学習手法を取り入れた材料開発に関する研究が活発になっている。本稿では、機械学習手法を用いて、未知の鉛フリーペロブスカイト材料の、物性の一つであるバンドギャップを予測することを目的とした。ある1元素に注目し、学習データにその元素を含まないときの予測モデルを構築し、評価実験を行った。

Research on prediction of band gap of inorganic material using machine learning

1. はじめに

従来の材料開発では、研究者の知識や経験を基に、実験と評価を繰り返し行う材料探索を要しており、実用化までには20年以上といった長い期間を要する場合がある[1]。この材料開発のコストを削減するべく、材料開発に情報技術や計算科学の手法と組み合わせた、マテリアルズ・インフォマティクスという分野が近年盛んになっている。マテリアルズ・インフォマティクスでは、材料候補を合成してから物性を調べるのではなく、必要な物性を持つ材料候補をシミュレーション・予測し、合成する材料を決めることにより、材料開発における属人性を減少させ、効率良く材料を探索することが可能とされている。特に、材料分野では、物性の近似計算の主流な手法である量子化学計算によって算出された値は、実験値と近い値で求めることが可能となっている。しかし、その計算量は化合物によっては大きく増大することがあるため、仮想化合物も考慮した膨大な化合物を網羅的に計算することは難しい。そのため、最近では、材料データベース[2][3][4][5]内のデータを用いて機械学習手法を適用した材料スクリーニングや、物性予

測に関する研究が増えている[6]。

材料開発において、太陽電池の材料として、近年では従来のシリコン型に代わり、ペロブスカイト型の普及への期待が高まっている[7]。ペロブスカイト太陽電池は、シリコン型に比べて低コストで製造され、薄型化・小型化も可能であり、設置場所を制限されにくいという利点がある。太陽電池の性能は、物性の一つであるバンドギャップに依存しており、高性能な太陽電池の開発には、適切なバンドギャップを持つ材料を選ぶことが必要である。しかし、現在高性能とされているペロブスカイト太陽電池には、毒性のある鉛(Pb)を含んでおり、今後、ペロブスカイト太陽電池の普及のためには、毒性のない鉛フリーペロブスカイト太陽電池材料の探索が課題となっている。

2. 評価実験

2.1 研究目的

ペロブスカイト太陽電池の候補材料は、単原子イオン、多原子イオンすべてを考慮すると、その組み合わせの候補数から、仮想化合物を含めても膨大であり、量子化学計算では、全てを網羅して計算することは現状現実的ではない。また、機械学習を用いて新たな材料を効率よく探索するためには、未だ量子化学計算結果がない化合物や、未実験の

¹ 東京工業大学 情報理工学院 情報工学系

² 東京工業大学 科学技術創成研究院 スマート創成研究ユニット

表 1 先行研究のデータセット

ラベル	
distance	A(B) サイトと X サイトとの原子間距離
eleneg	各サイトの電気陰性度
hoe	各サイトの最高占有原子エネルギー
ionenergy	各サイトのイオン化エネルギー
luep	各サイトの最低非占有原子エネルギー
rs	各サイトの s 軌道の半径
rp	各サイトの p 軌道の半径
rd	各サイトの d 軌道の半径
cubic	空間群
ind_gap	バンドギャップ (eV)

化合物の物性を予測可能である必要がある。よって、本研究では、量子化学計算では網羅しきれないような未知の(データセット外の)ペロブスカイト化合物のバンドギャップを予測できるような機械学習モデルを構築することを目的とした。

2.2 データセット

本研究では、先行研究 [6] のデータセットを用いている。このデータセットでは、 $A_2B^{1+}B^{3+}X_6$ で組成されている、鉛フリーのハロゲン化合物ダブルペロブスカイト (ハライドダブルペロブスカイト) の電子構造に関するデータが含まれている。データベース内のハライドダブルペロブスカイトを構成している元素は、次のようになっている。A サイトはカリウム (K), ルビジウム (Rb), セシウム (Cs)。B¹⁺ サイトは、銅 (Cu), 銀 (Ag), インジウム (In), 金 (Au), タリウム (Tl)。B³⁺ サイトは、アルミニウム (Al), ガリウム (Ga), インジウム (In), ヒ素 (As), アンチモン (Sb), ビスマス (Bi)。X サイトは、塩素 (Cl), 臭素 (Br), ヨウ素 (I) がそれぞれ含まれている。データセット内の各特徴量のラベルと詳細を表 1 に示す。合計 32 個の特徴量を元に、各特徴量同士の相関係数を求め、特徴次元を 20 まで減らした。 $\bar{x}$ と $\bar{y}$ を、それぞれ n 個の化合物に対する i 番目の特徴 x_i および y_i の平均とすると、相関係数 R_{xy} は式 1 で計算される。

$$R_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

この操作によって、選ばれた 20 個の特徴量を表 2 に示す。本研究では、これらの特徴量を用いてバンドギャップ予測モデルを構築した。以下では、便宜上、B¹⁺ サイトを B1(b1) サイトと呼び、B³⁺ サイトを B2(b2) サイトと呼称する。

2.3 手法

本節では、本研究における提案手法である、1 元素に注目したバンドギャップ予測のためのデータ分割について述べる。

表 2 選ばれた特徴量

cubic	distance_a	distance_b1	distance_b2
eleneg_a	eleneg_b1	eleneg_b2	eleneg_x
hoe_a	hoe_b1	hoe_b2	luep_a
luep_b1	luep_b2	rs_a	rs_b1
rs_b2	rd_a	rd_b1	rd_b2

表 3 各サイトにおける本手法での学習・テストデータ数

サイト	学習データ	テストデータ
A	360	180
B1	432	108
B2	450	90
X	360	180

実際に実験室で合成する化合物や構成元素を含ませるかを定める際に、機械学習に基づいたマテリアルズ・インフォマティクス手法を用いれば、求める物性を持つ化合物を(そしてそれを構成する元素も)効率よく決めることが可能になる。このとき、所有しているデータにない化合物や構成元素の物性を正確に予測することができれば、今まで想定していなかった材料を新たに発見できるかもしれない。このような新材料探索を実現させるために、本研究では学習データにない元素を持つ化合物の物性を予測する。具体的には、1 元素の置換のためのデータセット分割手法を提案する。

データ分割の例を図 1 に示す。ここでは、A サイトのカリウム (K) のバンドギャップ予測のためのデータ分割手順について述べる。データセット内の化合物は基本的に 4 種類の元素 (元素イオン) によって構成されている。K を含む化合物のバンドギャップを予測するときには、データセットを学習データとテストデータに分ける際、A サイトに注目し、テストデータには K を含む化合物データ、学習データのは同じ A サイトであるルビジウム (Rb) もしくはセシウム (Cs) を含む化合物データを選ぶ。このとき、学習データ数は 360、テストデータ数は 180 になる。これを B1 サイト (5 種類), B2 サイト (6 種類), X サイト (3 種類) に対しても同様の操作を行い、それぞれについてバンドギャップ予測を行う。

この手法を用いて、合計で 17 元素分の予測を行った。また、A サイト, B1 サイト, B2 サイト, X サイトそれぞれについての学習データとテストデータの割合は異なっているため、表 3 にそれぞれのデータ数を示す。

本研究では、先行研究同様、決定木に基づいた機械学習モデルである XGBoost[8] を用いて、化合物のバンドギャップの予測を行った。

2.4 評価方法

本研究では、モデル評価に決定係数 R^2 と、Root Mean Squared Error (RMSE) を用いた。

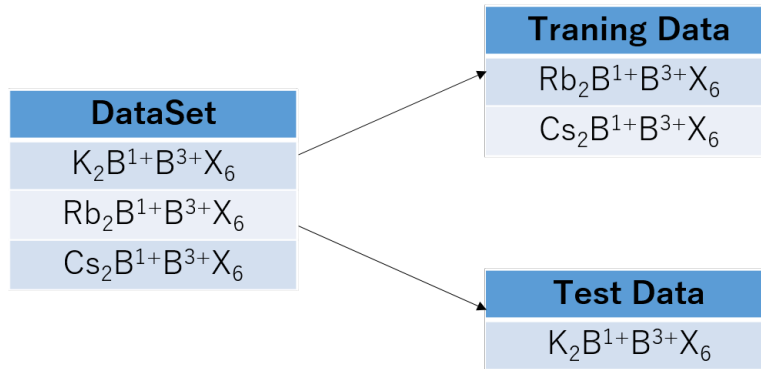


図 1 データの分割手法の例：カリウム (K) を含んだ化合物の予測の場合

表 4 XGBoost による精度比較

サイト	元素	R^2	$RMSE$
A	K	0.935	0.212
	Rb	0.912	0.245
	Cs	0.865	0.288
B1	Cu	-2.22	0.755
	Ag	0.0581	0.564
	In	0.682	0.448
	Au	0.367	0.722
	Tl	0.436	0.655
B2	Al	0.289	0.919
	Ga	0.760	0.364
	As	0.643	0.354
	In	0.785	0.364
	Sb	0.919	0.173
X	Bi	0.837	0.237
	Cl	0.515	0.650
	Br	0.632	0.447
	I	0.118	0.450

観測値を y_i ，モデルから計算された予測値を $\hat{y}_i$ ，観測値の平均を $\bar{y}$ とすると，決定係数 (R^2) は，式 2 によって表される。

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2)$$

観測値を y_i ，モデルから計算された予測値を $\hat{y}_i$ とすると， $RMSE$ は式 3 によって表される。

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

3. 結果

本節では，本研究で行った実験の結果を述べる．データセット内の特徴量 20 個を用いて，各元素における 1 置換を考慮して構築されたバンドギャップ予測モデルの R^2 と $RMSE$ で精度比較を行う．表 4 にそれぞれの予測結果を示す。

この結果から，A サイトにおける精度は，3 つ全てにおいて他サイトに比べて高精度となっている．その他のサイ

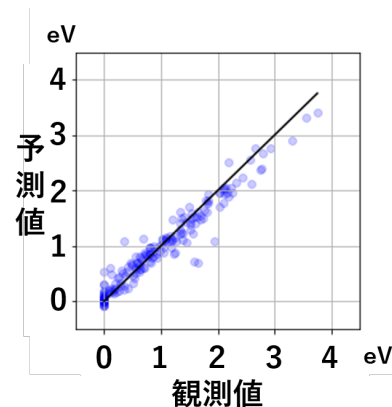


図 2 カリウム (K) におけるバンドギャップ予測の散布図

トに関しては，全く精度が出ていないものや，かなり高く精度が出ているもの，その中間ほどの精度になっているものが，予想対象とした元素によって違いが大きく表れている．この違いを示すために，いくつかの元素をピックアップして比較する．比較的高精度だった A サイトのカリウム (K) のバンドギャップ予測値と観測値の散布図を図 2 に示す．また比較的精度悪かった B2 サイトのアルミニウム (Al) と X サイトのヨウ素 (I) における散布図を図 3 に示す．これらを見ると，精度が悪くなっているものの傾向として，全体にバンドギャップ値が過大 (過小) 評価されていることが分かる．他の元素に関しても，程度の差はあれ，同様の傾向が見られた。

4. 考察

3 でも示したように，図 3，を見ると，Al においては，バンドギャップが全体的に過小評価されており，I においては，全体的に過大評価されている．その原因を探るために，B2 サイトと X サイトにおいて，各元素に対するバンドギャップ値の Violin plot を図 4 に示す．横軸は対象元素を含む化合物を表し，縦軸はデータセット内のバンドギャップ値 (eV) を表す．バンドギャップ値は，全て 0 以上の値をとっている。

，A サイトに関しては各元素によるバンドギャップの分

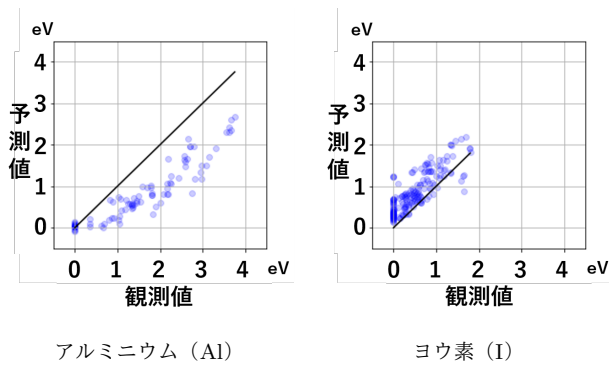


図 3 バンドギャップの予測値と観測値の散布図

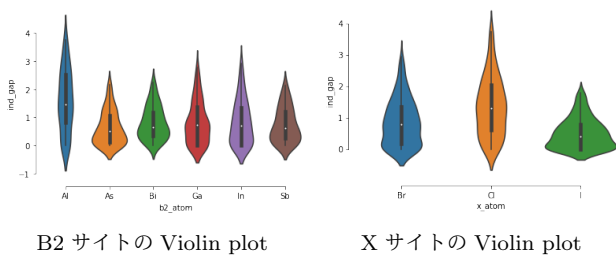


図 4 Violin plot

布に大きな差はないと考えられるが、B2、X サイトでは、元素によってバンドギャップの取り得る範囲が大きく異なっているものがある。B2 サイトの中で Al 含有化合物だけがバンドギャップの取り得る範囲が広く、X サイトの中で I 含有化合物が特にバンドギャップの取り得る範囲が狭い。このことから、1 元素置換のための学習データとテストデータの分け方として、どれだけ双方のバンドギャップ値の範囲に差があるのかが、大きく精度の違いとして表れていることがわかる。しかし、低精度なものにおいて、ある一定の傾向が見られることが明らかになったため、全体的に過大予測・過小予測されている元素に対して、本研究におけるバンドギャップ予測モデルを利用したまま、誤差を修正する手法を構築することが可能かもしれない。

5. おわりに

本研究では、鉛フリーハライドダブルペロブスカイト太陽電池の材料探索、そして、未知の、または手元のデータにはない新たな材料を素早く・正確に探索することを目指し、機械学習を用いて、1 元素置換に対応するための鉛フリーハライドダブルペロブスカイト化合物のバンドギャップ予測モデル構築した。1 種類の元素に注目し、含む/含まない化合物をテスト・学習データとして分け、データセット内 17 元素に対してバンドギャップ予測を行った。

実験の結果、選んだ元素により既存手法の精度を超えるものや、大きく下回るもの等、様々な傾向の結果が出た。原因として、その元素が含まれている化合物のバンドギャップの分布と、学習データとして扱った化合物のバンドギャップの分布の関係が考えられる。このことから、過

小予測・過大予測の修正を組み込んだ予測モデルへの拡張が挙げられる。

また、今回予測に用いたデータセットでは、 $A_2B^{1+}B^{3+}X_6$ で構成されるハライドダブルペロブスカイトしか用いられていない。しかし既に、各サイトが部分的に置換されているようなペロブスカイトも合成可能になっている。以下に例を挙げる。

- $(CH_3NH_3)Pb(Br_{1-x}Cl_x)_3$ [9]
- $Li_xLa_{(1-x)/3}NbO_3$, $(Li_{0.25}La_{0.25})_{1-x}Sr_{0.5x}NbO_3$ [10]
- $CsPb_{1-x}Sn_xIBr_2$ [11]

よって、まだ実験・計算されていない様々なペロブスカイト化合物のバンドギャップを予測するためには、より細かな置換を含むペロブスカイトを考慮した予測モデルを構築することが今後の課題として挙げられる。

謝辞 圧倒的感謝

参考文献

- [1] WWW.MGI.GOV: *Materials Genome Initiative*, <https://www.mgi.gov/>.
- [2] Jain, Anubhav and Ong, Shyue Ping and Hautier, Geoffroy and Chen, Wei and Richards, William Davidson and Dacek, Stephen and Cholia, Shreyas and Gunter, Dan and Skinner, David and Ceder, Gerbrand and Persson, Kristin a. : *The Materials Project: A materials genome approach to accelerating materials innovation*, APL Materials, Vol.1, No.1, pp. 011002, (2013).
- [3] Stefano Curtarolo, Wahyu Setyawan, Gus L.W. Hart, Michal Jahnatek, Roman V. Chepulskii, Richard H. Taylor, Shidong Wang, Junkai Xue, Kesong Yang, Ohad Levy, Michael J. Mehl, Harold T. Stokes, Denis O. Demchenko, Dane Morgan. : *AFLow: An automatic framework for high-throughput materials discovery*, Computational Materials Science, Vol.58, No.1, pp. 218-226, (2013).
- [4] James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig and C. Wolverton. : *The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies*, npj Computational Materials, Vol.1, pp. 6741-6747, (2015).
- [5] Scott Kirklin, James E Saal, Bryce Meredig, Alex Thompson, Jeff W Doak, Muratahan Aykol, Stephan Rühl and Chris Wolverton. : *Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD)*, The Journal of The Minerals, Metals and Materials Society, Vol.65, No.11, pp. 1501-1509, (2013).
- [6] Im, Jino and Lee, Seongwon and Ko, Tae-Wook and Kim, Hyun Woo and Hyon, YunKyong and Chang, Hyunju. : *Identifying Pb-free perovskites for solar cells by machine learning*, npj Computational Materials, Vol.5, No.1, pp. 1-8, (2019).
- [7] Kojima, Akihiro and Teshima, Kenjiro and Shirai, Yasuo and Miyasaka, Tsutomu. : *Organometal Halide Perovskites as Visible-Light Sensitizers for Photovoltaic Cells*, Journal of the American Chemical Society, Vol.131, No.17, pp. 6050-6051, (2013).
- [8] Tianqi Chen and Carlos Guestrin. : *XGBoost: A Scalable Tree Boosting System*, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery

- and data mining, pp.785–794, (2016).
- [9] Zhang, Taiyang and Yang, Mengjin and Benson, Eric E and Li, Zijian and van de Lagemaat, Jao and Luther, Joseph M and Yan, Yanfa and Zhu, Kai and Zhao, Yixin.: *A facile solvothermal growth of single crystal mixed halide perovskite* $(\text{CH}_3\text{NH}_3)\text{Pb}(\text{Br}_{1-x}\text{Cl}_x)_3$, Chemical Communications Vol.51, No.37, pp.7820–7823 (2015).
- [10] Kawakami, Yoshihiro and Ikuta, Hiromasa and Wakihara, Masataka.: *Ionic conduction of lithium for Perovskite-type compounds, $\text{Li}_x\text{La}_{(1-x)/3}\text{NbO}_3$ and $(\text{Li}_{0.25}\text{La}_{0.25})_{1-x}\text{Sr}_{0.5x}\text{NbO}_3$* , Journal of Solid State Electrochemistry Vol.2, No.4, pp.206–201 (1998).
- [11] Li, Nan and Zhu, Zonglong and Li, Jiangwei and Jen, Alex K-Y and Wang, Liduo.: *Inorganic $\text{CsPb}_{1-x}\text{Sn}_x\text{IBr}_2$ for Efficient Wide-Bandgap Perovskite Solar Cells*, Advanced energy materials Vol.8, No.22, (2018).