様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 １４ 日現在

機関番号：１２６０８
研究種目：若手研究
研究期間：2019～2020
課題番号：１９Ｋ１５２７５

研究課題名（和文）高精度第一原理計算と能動学習を用いた汎用的物性値予測モデルの開発

研究課題名（英文）Construction of general models for predicting material properties from first-principle calculations and active learning

研究代表者
高橋　亮（Takahashi, Akira）

東京工業大学・科学技術創成研究院・助教

研究者番号：８０８２２３１１

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：系統的な第一原理計算のための自動化プログラムを構築し、開発したプログラムを用いて多様な結晶構造型を持つ1,266の酸化物の誘電定数の系統的な計算を行った。また、誘電率を電子系・格子系の寄与に分け、それぞれについて機械学習により予測モデルを構築し、また誘電率の支配因子を抽出した。
一方で能動学習の手法について、既存データベースにuncertainty samplingを適用することで、未知データに対する予測モデルの予測精度が、少数のデータの追加によって大きく減少させることができることが分かった。さらに、所望の範囲に物性値が収まるデータを優先的に計算対象として探索を進める手法を開発した。

研究成果の学術的意義や社会的意義
本研究により、多数の物質の多様な物性を効率よく計算できるプログラムを開発し、大規模な計算材料データベースを構築することができた。誘電率について機械学習を行うことにより、高速・高精度な予測モデルの構築に成功し、さらに誘電率の支配因子をデータ科学的に求めることに成功した。
一方で、能動学習により未知のデータを効率よく選択しバランスの良いデータベースを構築できることを示した。
本研究により開発されたプログラム、データベース及び機械学習に構築された予測モデルと支配因子、能動学習の手法は今後のデータ科学・機械学習を用いた効率的な材料探索の基盤技術となると考えられる。

研究成果の概要（英文）：We developed software to perform a large number of first-principle calculations automatically and systematically. We calculated the dielectric constants of 1,266 oxides including various atomic frameworks and developed the prediction models and extracted determining factors of both electronic and ionic contributions to dielectric constants by machine learning technique.
On the other hands, we applied uncertainty sampling technique for an existing database and demonstrated the prediction error of machine learning can be significantly reduced by small number of additional sampling. Moreover, we developed active learning method to efficiently collect data whose material properties fitting in desirable range.

研究分野：材料科学

キーワード：第一原理計算　機械学習　能動学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

無機材料は現代社会において幅広く応用されており、その使用例はコンデンサやセンサー、LED、触媒、光触媒など枚挙にいとまがない。従って無機材料の安定性や電子・光学特性の解析や、これらの特性の制御方法を確立することは極めて重要である。

これまで材料特性の解析は実験的アプローチが主であったが、近年の計算機・アルゴリズムの発展により、基礎的な物性値については第一原理計算によって実験値と比較できる精度で予測することが可能になっている。更に最近では、数千～数万の第一原理計算結果から材料計算データベースを構築し、機械学習の手法を適用して高速な物性値予測モデルの構築や物性値の支配因子の解析を行うマテリアルズインフォマティクス(MI)が流行している。だが今までの多くの MI の研究では、試験的に機械学習の手法を導入してその適用性を確認するに留まっており、材料科学に貢献する統一的な学理の構築や未知材料の開発に成功している例はごく少数である。その理由として、従来の MI 研究に広く使われている Materials Project、Open Quantum Materials Database、AFLOW などの既存の材料計算データベースにおける次の２点の問題が挙げられる。

第一に、計算精度の問題である。一般的にデータベース構築のための第一原理計算には高速な PBE-GGA の近似が用いられるが、この枠組みでは格子定数・バンドギャップのような物性の予測値に大きな誤差を内包する。これらの問題は現在、PBEsol や non-self-consistent (nsc) hybrid のような新しい計算手法により解決できることがわかってきている。また PBEsol や nsc hybrid は比較的高速で、データベースを現実的な時間で構築できる。

第二に、データの偏りの問題である。一般に計算材料データベースは ICSD のような既報材料の結晶構造データベースを基に作られている。ICSD には重複を含めて現在 188,631 のデータが掲載されているが、このうち酸化物が 92,297 と半数近くを占めている一方、硫化物については 17,244 しかない。一般に、回帰分析により構築された予測モデルはデータの多い領域に過剰に適合してしまうため、類似データの少ない領域の予測の信頼性は低い。したがってデータベースの偏りの解消は MI を実行する上で必須である。

２．研究の目的

上述した背景から、本研究ではまず数千の既報の無機物質に PBEsol や nsc hybrid の第一原理計算手法を適用し、現実的な時間で高精度な材料計算データベースの構築を行う。次に回帰分析の手法でバンドギャップなどの物性値予測モデルの構築を行う。その後、データベースに含まれていない構造と元素の組み合わせで未知物質の計算モデルを作り、統計学的にデータの偏りを解消するように能動学習の手法を用いて、データの追加と予測モデルの検証および再構築を行う。これら一連の流れを自動化し、未知物質に対する高精度第一原理計算結果の追加を繰り返すことで精度・汎用性の高い物性値予測モデルを作ることを目的とした。

３．研究の方法

まず多数の物質に対して系統的に第一原理計算を行う必要があるが、上述したような様々な手法を用いる必要があり、また多様な電子物性の計算・解析には非常に煩雑な処理を要する。したがって、本研究ではまず、第一原理計算の自動化プログラムを構築し、既報材料の系統的な計算を行い、信頼性の高いデータベースを構築する。

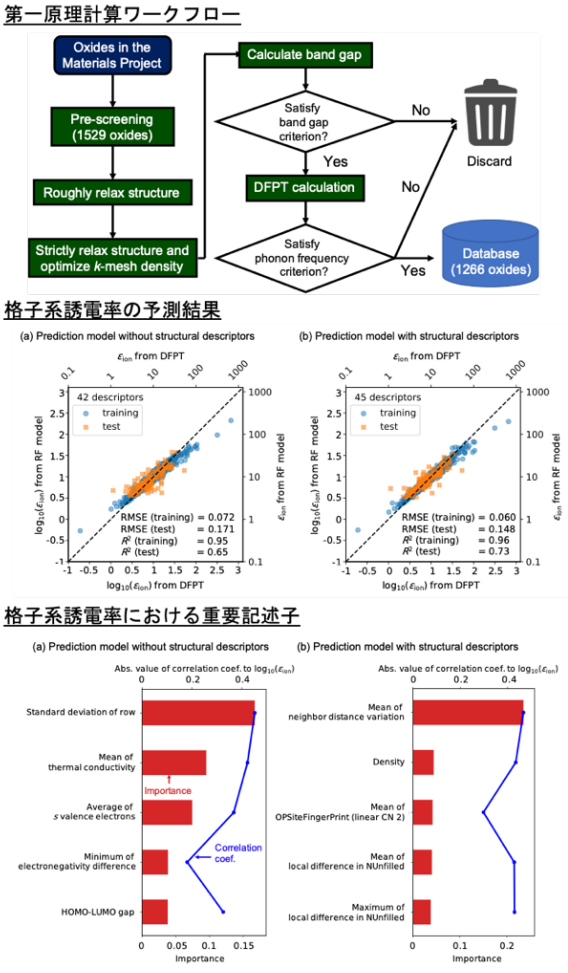

図　１　誘電率の自動化計算プログラムのワークフローと、得られたデータベースの機械学習の結果の例。[Phys. Rev. Mater. (2020)]

次に、得られたデータベースに回帰分析などの機械学習の手法を用いることにより、組成や構造情報から物性値を予測するモデルを構築する。また、各物性値と元素種・組成・構造などの情報との相関を調べて物性値の支配因子を明らかにし、データ科学の視点から材料設計指針を構築する。

最後に、ここまで得られたデータベースや機械学習モデルを基に、今まで報告例が無い物質を計算対象に加え、多種多様な物質をバランスよく含むデータベースの構築を行う。

4．研究成果

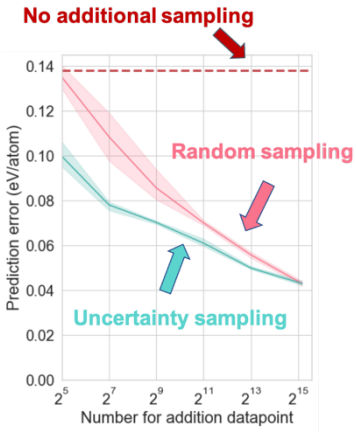

図 2 Materials project に掲載されたエネルギーのデータについて Uncertainty Sampling を適用した結果。

まず系統的な第一原理計算のための自動化プログラムを構築し、開発したプログラムを用いて Materials project に掲載されている多様な結晶構造型を持つ 1,266 の酸化物の誘電定数の系統的な計算を行った。得られた計算を実験と比較した結果、誘電率が極端に大きいものや温度の寄与が効いていると考えられるもの以外、計算値と実験値は良い一致を示した。次に、誘電率を電子系・格子系の寄与に分け、それぞれについて組成情報のみから予測を行うことができるモデルと、結晶構造の情報を用いて高精度な予測が可能なモデルの 2 つをランダムフォレスト法により構築した。学習に用いない未知データに対する予測モデルの予測精度について、電子系誘電率では構造情報なし、ありのモデルのそれぞれで決定係数が 0.87, 0.89 であり、格子系誘電率では構造情報なし、ありのそれぞれで 0.65, 0.73 であった。これらの予測精度は、high-k 等の高誘電率を持つ材料探索の際の初期スクリーニングを行う際に十分な精度であると考えられる。また、ランダムフォレストの特徴量の重要度解析を行なった結果、電子系誘電率では原子質量の平均と質量密度が、格子系誘電率では主量子数の標準偏差と隣接原子との距離のばらつきが重要記述子として抽出された。これらの記述子が誘電率の支配因子となっていることについて、データの詳細な解析や考察を行うことにより、従来の化学的・物理学的な描像と矛盾しないことを示した。これらの成果は図 1 に示すように、Phys. Rev. Mater. 誌にオープンアクセスで掲載されており、さらに github 上に得られたデータベースと予測モデルを公開している。

また現在ではプログラムの拡張を行い、HSE06 や nsc-hybrid 等の計算手法や、有効質量・光吸収係数などといった電子物性の計算に対応させている。Materials project 等に含まれない未知の物質の物性計算を行い、データベース拡張・材料探索・機械学習による解析を進めている。

一方で能動学習の手法について、図 2 に示すように、既存データベースに uncertainty sampling を適用することで、未知データに対する予測モデルの予測精度が、少数のデータの追加によって大きく減少させることができることが分かった。その一方で、データベースの拡張のために多数の未知の結晶構造を系統的に作成した場合、非常に不安定な物質や興味の対象に無い物質などを多く含むと考えられるが、実際に材料探索と能動学習を用いたデータベースの拡張を両立させる際には、これらを除外して計算する必要がある。そこで、本研究では、所望の範囲に物性値が収まるデータを優先的に計算対象として探索を進める手法を開発した。この手法を既存データベースに適用して物質探索シミュレーションを行なった結果、所望の物性値を高く評価する評価関数と EI や PI といった従来の手法を組み合わせたベイズ最適化に比べて、所望のバンドギャップを持つ物質を効率的に同定することができた。

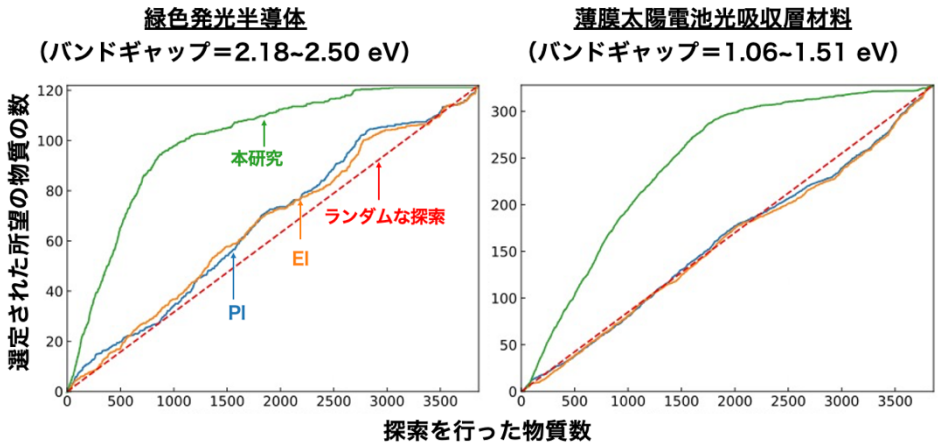

図 3　能動学習を用いた物質探索シミュレーションの結果。

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Takahashi Akira、Kumagai Yu、Miyamoto Jun、Mochizuki Yasuhide、Oba Fumiyasu	４．巻 4
２．論文標題 Machine learning models for predicting the dielectric constants of oxides based on high-throughput first-principles calculations	５．発行年 2020年
３．雑誌名 Physical Review Materials	６．最初と最後の頁 103801(1-13)
掲載論文のDOI（デジタルオブジェクト識別子） 10.1103/PhysRevMaterials.4.103801	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計4件（うち招待講演　1件／うち国際学会　2件）

１．発表者名 高橋亮，熊谷悠，宮本惇，望月泰英，大場史康
２．発表標題 酸化物誘電率計算データベースと機械学習による誘電率予測モデルの構築
３．学会等名 金属学会 2019年秋季大会
４．発表年 2019年～2020年

１．発表者名 Akira Takahashi, Yu Kumagai, Jun Miyamoto, Yasuhide Mochizuki, Fumiyasu Oba
２．発表標題 Machine Learning Model for Predicting Dielectric Constants of Oxides
３．学会等名 The 4th International Symposium on Creation of Life Innovation Materials for Interdisciplinary and International Researcher Development (iLIM-4)（国際学会）
４．発表年 2019年～2020年

１．発表者名 Akira Takahashi
２．発表標題 Prediction of Material Properties from First Principles and Machine Learning
３．学会等名 ICMASS2019 (iLIM-s session)（招待講演）（国際学会）
４．発表年 2019年～2020年

１．発表者名 高橋亮，青木宏賢，熊谷悠，大場史康
２．発表標題 計算材料データベースと機械学習を用いた緑色発光半導体の効率的探索手法の開発
３．学会等名 日本金属学会　2021年春期大会
４．発表年 2020年～2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関