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　５ 年　 ６ 月 ２３ 日現在

機関番号：１４３０１
研究種目：研究活動スタート支援
研究期間：2020～2022
課題番号：２０Ｋ２２４６６

研究課題名（和文）実験プロセスや試料構造の影響を考慮して物性値を予測する次世代MIの開発

研究課題名（英文）Development of Next-Generation Machine Intelligence for Predicting Material Properties, Considering the Influence of Experimental Processes and Sample Structures

研究代表者

熊谷　将也（Kumagai, Masaya）

京都大学・複合原子力科学研究所・特定助教

研究者番号：００８８１０５４

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：本研究では、プロセス・構造・物性の各要素間の関係性の分析、独自の大規模データセットの作成を行った。本研究におけるプロセス情報の収集は、論文PDFからのテキスト抽出により実施した。構造情報と物性値との関係性分析は、X線回折パターンを入力、結晶系、体積、密度、体積弾性率を学習対象とする機械学習モデルを構築し、X線回折パターンと学習対象との関係性を明らかにした。本研究期間で作成した独自の大規模データセットは、Figshare上に公開した。また本研究内容は、国内外の学会や論文誌への投稿など、様々な形で外部に報告した。

研究成果の学術的意義や社会的意義
プロセス情報を含めた物性の予測を可能にすることは、新規材料の発見のみならず、製造プロセスの改善に貢献することができるため大きな意義がある。また、XRDと物性との関係性を大規模なデータを利用して明らかにできたことは、これまで結晶構造と物性の関係性を紐解く上で学術的に意義がある。さらに、本研究期間に作成した大規模実験データは、これからの実験MIを推進する基盤データとなると考えている。

研究成果の概要（英文）：In this study, we analyzed the relationships among process, structure, and physical properties and created our own large dataset. The process information in this study was collected by extracting text from PDFs of research papers. To analyze the relationship between structural information and physical properties, a machine learning model was constructed using X-ray diffraction patterns as input and crystal systems, volume, density, and volume modulus as learning targets. The original large dataset created during this research period has been publicly released on Figshare. The findings of this research were also disseminated externally through various means, including submissions to domestic and international conferences and journals.

研究分野：マテリアルズ・インフォマティクス

キーワード：マテリアルズ・インフォマティクス　プロセス・インフォマティクス　機械学習　材料工学

## １．研究開始当初の背景

新しい材料の発見は、現代の地球温暖化やエネルギー問題などの社会問題を解決する鍵を握っている[1, 2]。ところが材料設計は、元素の組合せや結晶相等の複雑な多次元最適化問題であるため、経験や勘、実験の試行錯誤に依存した現在の手法では、最適解の導出が困難である。そこで近年、大規模な材料データとデータ科学によって解決を図る Materials Informatics（MI）が注目されている。

現在の MI は、第一原理計算データを利用した研究が主流である[2, 3]。ところが、理想的な条件下計算されたデータであるため、予測結果が実験結果と異なることがしばしば発生する。一方、実験的な材料開発は、図１に示すような４つの要素（プロセス、構造、物性、性能）とそれらをつなぐ論理構造で表現できる[1]。各プロセスが様々な構造（結晶構造や電子構造、微細構造や試料形状など）の形成に影響を与え、得られた構造が物性を決め、その物性を組み合わせて応用に必要な性能が評価される。そこで本研究では、この論理構造を紐解くことで、既存の MI とは異なり実験で得られる物性値を高精度に予測できると考えた。

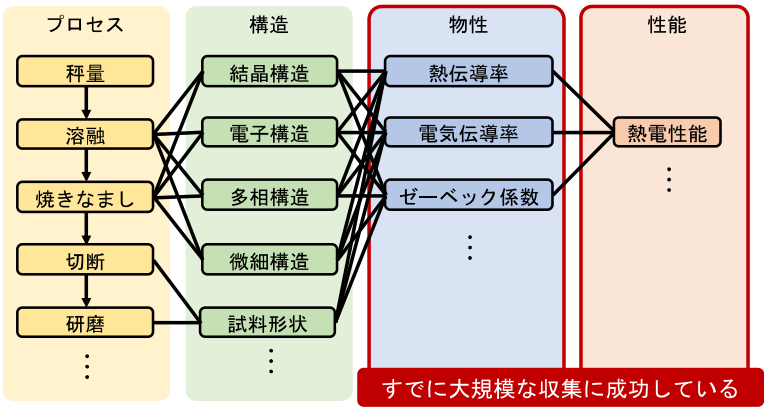


図１ プロセス、構造、特性、性能の相互関係の例

## ２．研究の目的

本研究では、論文中から多様な実験的要素（実験プロセスや試料構造）を収集するとともに、解釈性の高い機械学習手法を利用することで実験的物性値との関係性を解明し、その関係性を考慮した独自の大規模実験データを作成する。それにより、各実験プロセスや試料構造の影響を考慮した高精度な物性値予測ができる次世代 MI を開発することを目的とする。

本研究では、論文から実験的物性値を抽出する Web システム Starrydata2[4]をすでに開発し、世界最大規模の実験的物性値の収集に成功している（図 2）。特に本研究では、すでに独自性のある大規模な実験的物性値を基軸に実験的要素を紐づけた独自のデータを作成する。

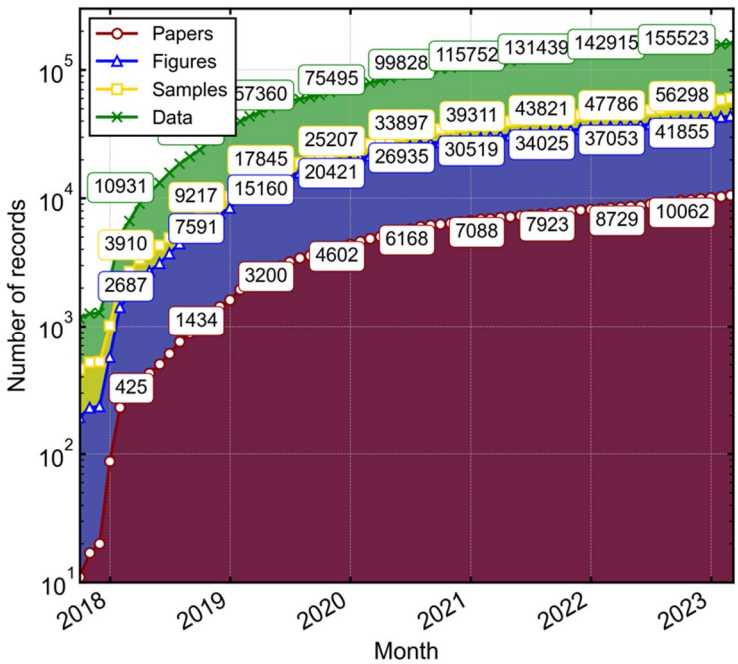


図２ Starrydata2[2]によって収集したデータ数(2023 年 3 月時点)

## ３．研究の方法

### (1)実験プロセスの収集と独自データセットの作成

Starydata2 で物性値の収集がすでに完了している論文に対して、フルテキストを取得し、テキストマイニングによるプロセス抽出および実験的物性値との紐付けたデータセットの作成を行った。論文からプロセスを抽出する手法については、関連研究[5]を利用した。

### (2)構造による物性予測と関係性解明

本研究では、構造情報として X 線回折パターンに着目した。材料開発では、必ず X 線回折パターンを測定する。そのため X 線回折パターンは、実験から得られる構造情報として最も取得しやすいデータである。しかし、論文等から大規模に収集するには多くの時間を要する。そこで、まず Materials Project[2]が提供している X 線回折パターンと物性値を利用し、X 線回折パターンを基軸に物性値を予測する機械学習モデルを作成した。また、X 線回折パターンと物性値との関係性を可視化によって明らかにした。

## ４．研究成果

### (1)実験プロセスの収集と独自データセットの作成

Starrydata2 に収録されている論文のうち約 5,000 本の論文 PDF の取得およびテキストの抽出を行い、Starrydata2 の実験的物性値と紐付けたデータセットを作成した。また、論文のテキストからプロセス情報（合成条件等）を取得するプログラムを先行研究に基づいて実装し、プロセス情報と実験的物性値を紐付けた。図 3 は、実験的物性値に対してテキストを紐づけた結果の一例である。論文から抽出したテキストは、それぞれ論文の製造方法や組成などの特徴が現れていることが確認できた。

独自のデータセットを作成する場合、そのデータセットが持つバイアスの影響を示すことが重要となる。本研究では、定量的な構造活性相関研究で用いられる適用範囲（AD）という概念を利用したバイアスの影響の可視化を行った。図 4 は、本研究で構築した機械学習モデルによる予測値と実際の実験値との関係性を示している。ここで機械学習には、MI 分野で一般に利用されるランダムフォレストを利用した。AD 外における予測精度は、AD 内と比べて予測精度が低いことが確認できた。また、AD 内においても、既知ノードの数が多いほど予測精度が高いことが確認できた。この結果は、学習に使用した既知データのバイアスを受けることを意味している。本研究で作成した Starrydata2 をベースとしたデータセットは、熱電変換材料の収録数が多いことがわかっており、化学組成や合成方法も熱電変換材料分野のバイアスを含んでいることに注意が必要である。

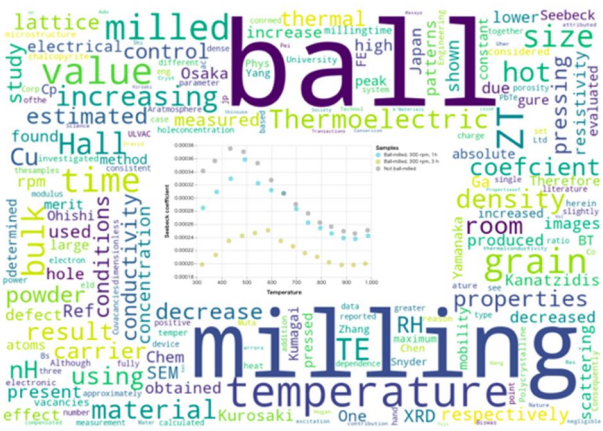


図 3　$CuGaTe_2$ に関する論文 [6]のワードクラウド

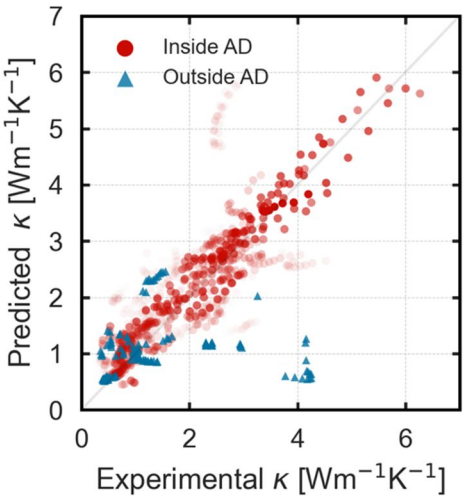


図 4　熱伝導率の予測値と実験値[7]．色の濃さは、既知ノードの数が多いほど濃くなっている．

### (2)構造による物性予測と関係性解明

構造情報からの物性値予測に関しては、学習に利用するデータとして第一原理計算データベース Materials Project から X 線回折パターンを含む結晶構造情報を約 63,000 レコード取得した。取得した X 線回折パターンを入力とし、結晶構造情報や体積弾性率などの基礎物性を学習・予測対象とする機械学習モデルを構築した。入力である特徴量ベクトルには、次の 3 種類を用意した：①ピーク数とピーク角度を利用した NA ベクトル、②ピーク数とピーク強度利用した NI ベクトル、③ピーク数とピーク角度とピーク強度を利用した NAI ベクトル。目的変数には、結晶系、体積、密度、体積弾性率の 4 つである。

図 5 は、4 つの目的変数に対する 3 つのベクトルの予測精度を比較した結果である。ここで、結晶系の予測には、評価指標として F 値を使用しており、他の 3 つには決定係数 $R^2$ を用いてい

る。結晶系と体積の予測精度は、角度情報を有している NA ベクトルと NAI ベクトルを利用した機械学習モデルが高い値を示した。このことは、物質の格子面間隔に依存する角度情報により、結晶構造の形状が高い精度で推測できたからであると考えられる。一方、密度と体積弾性率の予測精度は、強度情報を有している NI ベクトルと NAI ベクトルを利用した機械学習モデルが高い値を示した。密度と体積弾性率は、形状のみならず、構成元素の重さや結合状態が関わっている。このことから、原子種に依存する強度情報により、構成元素の重さや結合状態の影響をある程度推測できたと考えられる。また、結晶系や体積、密度や体積弾性率の予測結果から、結晶構造の形状や物性の予測には、ピーク角度とピーク強度の両方の情報が必要であることがわかった。

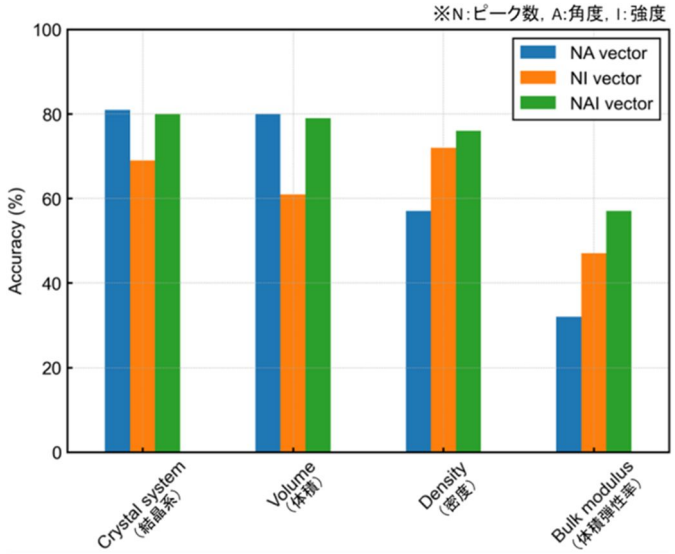


図 5 NA、NI、NAI ベクトルの４つの目的変数の予測精度比較

引用文献

[1] G. B. Olson, Science 288 (2000) 993-998.
[2] A. Jain, K. A. Persson and G. Ceder, APL Mater. 4 (2016) 053102.
[3] S. Curtarolo, W. Setyawan, G. L. W. Hart et al., Comput. Mater. Sci. 58 (2012) 218-226.
[4] Y. Katsura, M. Kumagai et al., Sci. Technol. Adv. Mater. 20 (2019) 511-520.
[5] T. Onishi, T. Kadohira and I. Watanabe, Sci. Technol. Adv. Mater. 19 (2018) 649-659.
[6] M. Kumagai, K. Kurosaki, Y. Ohishi, H. Muta, and S. Yamanaka, Materials Transactions 55 (2014)1215-1218,.
[7] M. Kumagai, Y. Ando, A. Tanaka, K. Tsuda, Y. Katsura, K. Kurosaki, STAM Methods 2 (2022) 302-309.

５．主な発表論文等

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

| | |
|---|---|
| １．著者名<br>Kumagai Masaya、Ando Yuki、Tanaka Atsumi、Tsuda Koji、Katsura Yukari、Kurosaki Ken | ４．巻<br>2 |
| ２．論文標題<br>Effects of data bias on machine-learning-based material discovery using experimental property data | ５．発行年<br>2022年 |
| ３．雑誌名<br>Science and Technology of Advanced Materials: Methods | ６．最初と最後の頁<br>302～309 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1080/27660400.2022.2109447 | 査読の有無<br>無 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

〔学会発表〕　計6件（うち招待講演　1件／うち国際学会　3件）

| |
|---|
| １．発表者名<br>熊谷 将也 |
| ２．発表標題<br>既知材料との類似性に基づいた熱電特性予測モデルの適用範囲 |
| ３．学会等名<br>日本熱電学会 |
| ４．発表年<br>2021年 |

| |
|---|
| １．発表者名<br>Masaya Kumagai |
| ２．発表標題<br>Applicability domain for prediction models of thermoelectric properties based on similarity to known materials |
| ３．学会等名<br>TMS2022（国際学会） |
| ４．発表年<br>2022年 |

| |
|---|
| １．発表者名<br>波頭 直輝 |
| ２．発表標題<br>機械的特性予測のためのX線回折パターンに基づく特徴量の設計 |
| ３．学会等名<br>日本金属学会 |
| ４．発表年<br>2021年 |

１．発表者名
Naoki Hato

２．発表標題
Design of Features Based on X-ray Diffraction Patterns for Prediction of Mechanical Properties

３．学会等名
MRS2021（国際学会）

４．発表年
2021年

１．発表者名
Naoki Hato

２．発表標題
Direct prediction of mechanical properties from X-ray diffraction patterns using machine learning

３．学会等名
TMS2022（国際学会）

４．発表年
2022年

１．発表者名
熊谷 将也

２．発表標題
マテリアルズ・インフォマティクス― 大規模な実験データ収集Webシステムの開発と応用 ―

３．学会等名
複合原子力化学研究所第56回学術講演会（招待講演）

４．発表年
2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（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 |
| --- | --- |