氏　　名　草場 穫

学位(専攻分野)　博士(統計科学)

学 位 記 番 号　総研大甲第 2315 号

学位授与の日付　2022 年 3 月 24 日

学位授与の要件　複合科学研究科　統計科学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Machine Learning for Chemical Elements and Crystal Structures

論 文 審 査 委 員　主　査　日野 英逸
統計科学専攻　教授
吉田 亮
統計科学専攻　教授
Wu Stephen
統計科学専攻　准教授
持橋 大地
統計科学専攻　准教授
三宅 隆
産業技術総合研究所　材料・化学領域
研究チーム長

（様式３）

博士論文の要旨

氏　　名　草場　穫

論文題目 Machine Learning for Chemical Elements and Crystal Structures

Materials informatics is a technique which aim to improving the efficiency of development of new and innovative materials with a way of informatics. Traditionally, the discovery of new inorganic compounds has been done by the human intuition and experiment. However, thanks to the progress of *ab initio* computations in the density functional theory (DFT) framework, computing power, and memory storage, computational study of materials has significantly developed for the past decades. It's accuracy and computational efficiency of DFT calculations enabled computational studies of a large number of compounds, leading to the rapid expansion of the computational crystal structure databases (detailed in the section 2.2). The accumulation of these online databases has greatly facilitated the application of machine learning in this study area. In recent years, more efficient materials development methods, which combine accurate but time consuming DFT calculations (deductive) and fast machine learning methods (inductive), has been proposed (detailed in the section 2.1 to 2.6). We have conducted two different studies which related to the field of materials informatics, and they are reported in the section 3, and 4 respectively.

In the section 3, we report the study, titled "Recreation of the Periodic Table with an Unsupervised Machine Learning Algorithm". In 1869, the first draft of the periodic table was published by Russian chemist Dmitri Mendeleev. In terms of data science, his achievement can be viewed as a successful example of feature embedding based on human cognition: chemical properties of all known elements at that time were compressed onto the two-dimensional grid system for a tabular display. In this study, we seek to answer the question of whether machine learning can reproduce or recreate the periodic table by using observed physicochemical properties of the elements. To achieve this goal, we developed a periodic table generator (PTG). The PTG is an unsupervised machine learning algorithm based on the generative topographic mapping (GTM), which can automate the translation of high-dimensional data into a tabular form with varying layouts on-demand. The PTG autonomously produced various arrangements of chemical symbols, which organized a two-dimensional array such as Mendeleev's periodic table or three-dimensional spiral table according to the underlying periodicity in the given data. We further showed what the PTG learned from the element data and how the element features, such as melting point and electronegativity, are compressed to the lower-dimensional latent spaces.

In the section 4, we report the study, titled "Data-driven crystal structure prediction

using structure similarity". Prediction of the stable structure of a given chemical composition is a basic and prerequisite task for the discovery of new materials. The major solution to this problem is based on an optimization problem of the free energy which requires a significant computational cost. In this study, we propose a method which makes crystal structure prediction by selecting crystal structures that are predicted to be similar to the stable structure of a given chemical composition from the existing crystal structures in the database. The prediction of crystal structure similarity is performed by a machine learning model built using prior information about crystal structure similarities in the database. Our method does not require the computationally expensive density functional theory framework, except for the validation part of the suggested structures. The effectiveness and characteristics of our method were demonstrated on a benchmark set.

Both of the above two studies can be said to be applications of machine learning to materials data, but they contrast in the following points. The study in the section 3 dealt with small data in unsupervised manner. The study in the section 4 dealt with relatively large data, mainly in supervised manner. Furthermore, the latter is a practical study in terms of materials science, while the former is not.

Additionally, in the section 2, we review the brief history and recent developments of materials informatics on inorganic materials. Also, we describe the position of our research, in the overall flow of the materials informatic researches. Historical and recent developments of crystal structure databases are summarized in the section 2.2. The topics about crystal structure prediction and visualization of materials data and chemical elements are particularly detailed in the section 2.6, 2.7, and 2.8 respectively. Due to strong relationship between PTG and GTM, GTM is particularly detailed in the section 2.9 and 2.10. Finally, in the section 5, we review the above two studies together and state our conclusions.

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis defense

博士論文審査結果

Name in Full
氏 名　草場　穫

Title
論文題目　Machine Learning for Chemical Elements and Crystal Structures

[論文の概要]

申請論文は 5 章 65 頁からなる．論文の構成は，物質の表現と学習に関するマテリアルズインフォマティクスの二つの研究をまとめた形となっている．

一つ目の研究は,「教師なし学習による元素周期表の自動設計」である. 周期表の原型は, 19 世紀後半にロシアの化学者ドミトリ・メンデレーエフによって発明された．メンデレーエフは，当時見つかっていた 50 個ほどの元素の特性がある周期的な振る舞いを示すことに着目し，そのパターンを表形式にまとめることで現在の周期表の原型を発明した．この発明に至る過程をデータ科学の視点から解釈すれば，メンデレーエフは，元素の多次元データをそのパターンに応じ二次元座標上の格子点（表）に配置するという「データの次元圧縮・可視化」を行ったといえる．本研究は，機械学習で元素のデータから周期表を自動設計できるかという問いから出発している. 問題は, 高次元データの「表形式の次元削減」に帰着する．申請者は，Generative Topographic Mapping という手法を拡張して高次元データを表形式に縮約する教師なし学習の手法を開発し，メンデレーエフの周期表とほぼ同等の表現を得ることに成功した．さらに，提案手法を用いて 3 次元円錐螺旋型の周期表を構築した．この周期表から元素の新しい分類基準を示唆する興味深いルールが見出された．

二つ目の研究は,「距離学習による化学組成からの結晶構造予測」である．材料の化学組成の情報のみから原子や分子の集合系が形成する結晶構造を予測できるかというのが本研究の問いである．従来の物理化学的なアプローチでは，多体電子系電子状態密度の第一原理計算と進化計算を組み合わせてエネルギー最小化問題を解き，安定な結晶構造を予測する．したがって，遺伝的操作で生成した候補構造に対して通常は数千回以上の第一原理計算を繰り返し実行するため, 膨大な計算時間を要することになる. 特に結晶の単位胞が 30–40 個以上の原子を含む系においては，現在の計算機の演算能力ではこの問題を解けない．本研究は，既知の結晶構造の化学組成のデータに距離学習を適用することで，第一原理計算をほぼ実行せずに，組成情報のみから結晶構造を非常に高い精度で測できることを明らかにした．これまでに合成されてきた様々な結晶系に提案手法を適用した結果，機械学習の予測モデルは第一原理に基づく従来法を大きく上回る予測能力を示すことが分かった．

各章の概要は以下の通りである．

1 章：イントロダクション

2 章：無機物質のマテリアルズインフォマティクスに関する先行研究やデータ解析手法のレビュー，第 3 章と第 4 章の研究成果の位置付け

3 章：教師なし学習による元素周期表の自動設計

4 章：距離学習による化学組成からの結晶構造の予測

5 章：まとめと今後の課題

[論文の評価]

物質科学の基本問題に対し，データ科学の独自の視点から新しい研究手法を見出したことが，二つの研究成果の特筆すべき点である．申請者は，周期表の設計というタスクを統計的次元圧縮・可視化の問題として定式化した．また，結晶構造予測のタスクが距離学習の問題形式に帰着できることを見出した．これらの着眼点は，非常に斬新である．また，データ科学的観点における学術的新規性も認められる．データ科学の分野では，様々な次元圧縮の方法論が研究されてきたが，表形式の次元圧縮の方法については，先行研究がほとんどない．周期表の研究は，このようなデータ分析の基本タスクに対する新しい方法論を提案したといえる．結晶構造予測については，従来の第一原理的なアプローチでは全く解けないような結晶構造を高精度で予測できる手法が提案された．本研究成果の物質科学における学術的貢献は十分に認められる．

[その他]

第 3 章の内容をまとめた論文は，査読付きジャーナル Scientific Reports 誌（第一著者）に掲載されている．第 4 章の研究については，学術誌への投稿が受理され，現在は査読中という状況である（ArXiv プレプリントは公開済み）．