jh231009

物性予測のためのグラフニューラルネットワークベース汎用 Pre-trained モデルの構築

研究代表者氏名 華井雅俊（所属機関　東京大学）

概要

本研究で材料分野におけるグラフニューラルネットワークベースの汎用 Pre-trained モデルの構築の研究および、本研究に代表される機械学習と材料科学分野における活動をサポートするモダンなデータシステムの構築を実施した。Pre-trained モデル構築テーマに関しては Pre-training を用いた Semi-supervised なアプローチによって物性予測での外挿問題における精度劣化の是正する手法を開発した。更に問題を順問題（構造から物性値を予測）から逆問題（物性値から構造を予測）に発展させるべく、生成モデルの研究を実施している。また、システム面に関して材料科学分野と機械学習分野の学際的研究を強力にサポートする Jupyter ベースのデータシステムを mdx 上に構築、さらに材料実験データの効率的な収集基盤を開発した。将来的には大規模な実験データを用いて Pretrained モデルの研究を強化する。本研究から生まれたプロトタイプシステムは、東京大学の共用材料実験施設における基盤システム”ARIM-mdx データシステム”として展開され学内外や企業ユーザー約 300 名が利用する一般利用サービスとなった。

1. 共同研究に関する情報

⑴ 共同利用・共同研究を実施している拠点名

mdx

(2) 課題分野

データ科学・データ利活用課題分野

(3) 共同研究分野(HPCI 資源利用課題のみ)

無し

(4) 参加研究者の役割分担

代表者 華井雅俊 研究総括

副代表者 河村光晶 物理学分野の補佐

共同研究者 鈴村豊太郎 機械学習分野の補佐

2. 研究の目的と意義

グラフニューラルネットワーク （Graph Neural Network, GNN） を用いた**物性値予測**の研究が盛んであり、電池、半導体、触媒、医薬品など広範囲な材料開発の基礎技術として利用される。ここでの物性予測とは、物質の基本情報である分子構造・結晶構造データ （グラフ） からより複雑な物性値を計算・予測する問題であり、特に第一原理計算などの物理シミュレーションによる計算と機械学習によるデータ学習を組み合わせることで問題の解決を目指す。物理シミュレーションによる物性値計算は一般的に多くの計算リソースを必要とし候補材料全てに対して逐一求めるの

は現実的ではなく、一部の計算結果を用いてデータ学習を行い、GNN による代理モデルを構築することで候補材料全てに対する物性値取得を実現する。

さらに、近年では**逆問題**、つまりある物性値を与えて、その値を持つような材料を見つける問題の研究も増えており、これは機械学習での生成モデルが爆発的な盛り上がりを見せたことが要因であると考えられる。一方、自然言語や画像の認識モデルや生成モデルにおいて、大規模なデータをつかった Pre-training モデルを構築するアプローチが主である。本研究では、物性予測や逆問題において、そのような Pre-training モデルを構築することを目指す。

加えて、本研究ではシステム面での貢献も目指す。自由なシステム構築な可能な mdx をベースにし、機械学習と材料科学の融合研究を効果的に進めるためのモダンなデータシステムを研究活動通じて構築する。

3. 当拠点の公募型共同研究として実施した意義

本研究は,機械学習分野と材料分野の学際的分野であるため、公募型共同研究として実施の意義があった。また、後述するように柔軟なシステム構築が可能である mdx を利用し,研究のためのデータプラットフォームの整備を進めた結果、本課題でのプロトタイプシステムをベースとして大規模な材料用データシステムの構築およびサービスローンチを行うことができた。

4. 前年度までに得られた研究成果の概要

昨年度実施の課題において、外挿予測、つまり学習データに現れない物性値の予測問 題 （例えば、全学習データより大きな値や小さな値の予測） とマルチタスク学習、つまり異なる種類 の物性値を組み合わせ、効果的に学習する手法に関して研究を行った。昨年度は主に構造からの物性値予測という順問題に関しての研究であったが、今年度は逆問題(物質を生成)へと研究を発展させた。

5. 今年度の研究成果の詳細

本研究の成果は大きく 2 点ある。１つは本研究の最終的なゴールである Pre-trained model 構築に向けた機械学習分野での成果、もう 1 点は材料研究用のデータシステム構築に関しての成果である。

(ア)Pretrain Model 構築に向けた研究

本研究の目標である汎用 Pre-trained モデル構築に際して、Pre-training を利用した Semi-supervised なアプローチによって物性予測に関する重要問題である Out-Of-Distribution 問題、つまり、データ分布の偏り（物性値の多くは典型的なデータ区間に集まる）が引き起こす性能劣化問題を是正することに成功した [2]。

また、Pre-traing を利用してデータの不均衡問題に着手した[3]。これは、例えばエネルギー値を予測する物性値予測において、典型的値にほとんどのデータが偏りモデルの予測精度も典型的値に近いほど高精度となる。一方ハズレ値のデータは非常に少なく精度を出すのが非常に難しい。しかしながら実用上の重要度は逆でありハズレ値を持つ物質のほうが重要となり高精度な予測が求められる。

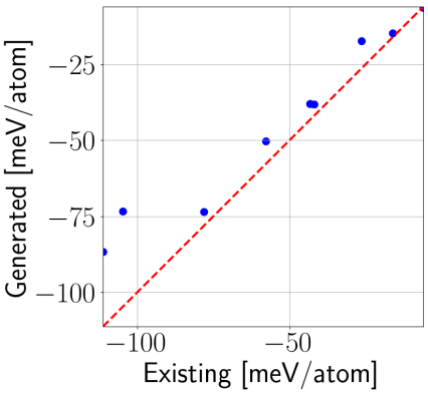
研究の発展として、よりチャレンジングな問題である生成モデルの構築に関して取り組んでいる。特に結晶構造における、繰り返し構造の取り扱いについての課題に着手している。

結晶構造において最小の繰り返し単位はユニットセルと言われ情報の過不足がないため、通常モデル化のベースとして利用される。しかしながらユニットセルを用

いるとデータ学習において問題が生じる。特に我々はユニットセルによるモデル化によってデータの連続性が著しく欠落する問題に着目し研究を行っている。ここでの連続性の問題とは、本来の構造（無限の繰り返しによる構造）において、非常に近似している構造であっても、繰り返しの単位が異なる場合、ユニットセルのモデル化によってデータ表現大きく異なる問題である。例えば、非常に近似した構造であってもユニットセルのサイズが 2 倍になる場合、従来手法を利用すると全く別の構造として処理される。本問題は物性値予測などの順問題においてはデータの前処理を十分に行うことで大きく問題にならないが、逆問題（生成モデル）においては非常に問題になることは明らかであろう。

そこで本研究では大きく 2 つのアプローチを用いた方法に着手している。1 つ目は従来手法で多く用いられる絶対座標に変わって、相対座標を用いる点である。これによってユニットセルが大きく変わっても各分子の距離は大きく変化しないので連続性が保たれると考えられる。2 つ目は生成プロセスにユニットセルも含む方法である。従来手法においてユニットセルは生成プロセスの前予測として単純な多次元 Regression 問題として計算され、生成時には固定値として扱う。本アプローチではユニットセル自体も生成プロセスの一部として含むことでユニットセルの変化に対してロバストなモデルを目指している。

グラフに本研究にて実施中の評価の一部を示す。本生成モデルにおいて、生成された構造と、その同一組成比をもつ実データとの Formation Energy 値の比較である。エネルギー値は生成モデルの学習において明示的に考慮されていないにもかかわらず、生成された物質は実際の物質のエネルギー値と概ね近い値を示している。引き続き評価および発展をすすめている。



（イ）材料データシステムの構築

本研究では、機械学習と材料分野の学際的研究をサポートするデータシステムの構築にも注力した。

本研究を通じて実際の材料開発や実験による生データへの応用に発展させるべく実験材料研究のチームとの協業を開始、その中で実験データの大規模取得のシステム分野のテーマが生まれ、成果 [1,5] において実験装置からの効率的なデータ収集をする IoT デバイスを開発した。本研究の利用目的で構築した K8s ベースの JupyterHub と統合し、東京大学の共用実験施設におけるデータシステム ”ARIM-mdx データシステム” [4] として一般サービスに発展した。

現在、50 台の IoT デバイス（ラズベリーパイベース）が東京大学内の各材料実験装置につながっており、学内外・企業ユーザー300 名以上が利用するまでに至っている。

6. 進捗状況の自己評価と今後の展望

当初の研究計画にあげた Pre-trained モデルの構築に関して、物性予測問題に関しては着実な進歩があり、今後も進めていきたい。また逆問題・生成問題に関しては非常

にチャレンジングなトピックであるが本質的な問題点が判明しつつある。また、当初の計画では注力する予定のなかったシステム面に関して、特に実験分野の研究者との協業の中で成果を上げることができた。全体の評価としては 80%として今後につなげていきたい。

7. 研究業績

(1) 学術論文 （査読あり）

(2) 国際会議プロシーディングス （査読あり）

[1] Masatoshi Hanai, Mitsuaki Kawamura, Ryo Ishikawa, Toyotaro Suzumura, and Kenjiro Taura. 2023. Cloud Data Acquisition from Shared-Use Facilities in A University-Scale Laboratory Information Management System. In 2023 IEEE/ACM 16th International Conference on Utility and Cloud Computing (UCC '23), December 4-7, 2023, Taormina (Messina), Italy. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3603166.3632147

(3) 国際会議発表（査読なし）

(4) 国内会議発表（査読あり）

[2] S. Takashige, M Hanai, T Suzumura, L Wang, K Taura "Is Self-Supervised Pretraining Good for Extrapolation in Molecular Property Prediction?" xSIG 2023 https://arxiv.org/pdf/2308.08129

[3] L Wang, M Hanai, T Suzumura, S Takashige, K Taura "On Data Imbalance in Molecular Property Prediction with Pre-training." xSIG 2023 https://arxiv.org/abs/2308.08934

(5) 公開したライブラリなど

[4] ARIM-mdx Data System: https://lcnet.t.u-tokyo.ac.jp/data_system/

(6) その他（特許，プレスリリース，著書等）

[5] 特願 2023-156343 "IoT デバイス、データ転送システムおよびデータ転送方法"（東大 TLO より）