jh221005

グラフニューラルネットワークとマルチタスク学習による汎用的物性予測モデルの構築

研究代表者氏名 華井雅俊（東京大学）

概要

電池、半導体、触媒、医薬品などの材料開発全般において、膨大にある候補材料のさまざまな物性を解析することが必要であり、近年、Graph Neural Network(GNN)を利用した分子構造からの物性予測が注目される。本研究では、GNN での物性予測において、さまざまな種類の物性データを利活用し、マルチタスク学習ベースの汎用的物性予測モデルの構築を目指す。一般にマルチタスク学習では、最終的な目的タスクに対し、関連した別の問題（擬似タスクや別ラベル)を学習することでの精度向上をはかるが、物性予測問題のマルチタスク学習においては、特に、データからの別タスクの設計と学習時のデータ分布不均衡性が顕著な問題となる。本研究では、汎用的な予測モデル構築へ向けた取り組みとして、ノードとエッジ情報の相互的マルチタスク学習、不均衡データにおける外挿学習・事前学習について研究を行った。

1. 共同研究に関する情報

(1) 共同利用・共同研究を実施している拠点名（該当するものを残す）

mdx

(2) 課題分野（該当するものを残す）

データ科学・データ利活用課題分野

(3) 共同研究分野（HPCI 資源を利用している研究課題のみ，該当するものを残す）

(4) 参加研究者の役割分担

代表者 華井雅俊 研究統括・実装・評価
副代表者 大西正人 物理学分野での補佐
共同研究者 鈴村豊太郎 機械学習分野での補佐
共同研究者 塩見淳一郎 物理学分野での補佐

2. 研究の目的と意義

研究の目的：電池、半導体、触媒、医薬品などの材料開発全般において、膨大にある候補材料のさまざまな物質的性質（物性）を比較解析することが不可欠であるが、それら候補全てを実際に作り検証することは現実的でない。そのため分子構造などの比較的簡単に得られる物質情報から物性を予測・計算することが重要である。

近年、既知の物性値データと機械学習を利用した物性値予測モデルの研究が盛んである。ある物性値が広範囲な材料群に対し既知である場合、予測モデルを構築することが可能となるが、しかし一方で、多くの物性値においては既知である材料が少数であり学習データが十分に取得できないため、高精度の予測モデルを構築することは難しい。

本研究は、グラフニューラルネットワークとマルチタスク学習を利用し、様々な物性を高精度に予測可能な汎用機械学習モデルを構築することを目的とする。物質の分子構造（グラフ）を入力とし、グラフニューラルネットワークを用いてその物性値を予測する。また、マルチタスク学習を利用することで、広範囲な材料群で明らかになって

いる一部の物性値だけでなく、少数のデータのみ取得できる他の多くの物性値データにも適応可能な予測モデルの構築をめざす。

研究の意義: 本研究の意義は、物理学的側面と情報科学的側面で大きく 2 つある。

まず、汎用かつ高精度な物性値の予測は材料開発全般において実用面での意義が大きい。高精度予測のためには、単一の物性値に対してパラメーター条件や実験条件が共通するまとまった大規模学習データが不可欠であるが、そのようなデータの取得は一部の物性に限られる。本研究では、少数のみ取得できるような物性値データへも適応可能な汎用的予測モデルを構築することを目指す。

3. 当拠点の公募型研究として実施した意義

本研究の公募型研究としての意義は、学際性にあり、特に物理学 (物性予測) の問題に情報科学(機械学習) の手法を適応する点である。分野の境界領域にて、これまでの機械学習の一般的な問題(画像認識や自然言語処理) では見られない課題が浮き彫りになり、着手することができた。

4. 前年度までに得られた研究成果の概要

5. 今年度の研究成果の詳細

マルチタスク学習を用いた汎用的な物性予測モデルの構築に関連して、今年度は外挿予測・Out of Distribution 問題、学習データの不均衡問題の研究を行った。さらに応用として、ガラスのダイナミクス予測におけるマルチタスク学習の研究を行った。

外挿予測・Out of Distribution(OOD)問題: 本研究では、外挿予測や Out of Distribution と呼ばれる、データ分布が学習データ-予測データ、事前学習データ-Downstream タスク間、で異なる場合に起きる精度劣化に関する問題に着手した。外挿予測・OOD は我々の目指すマルチタスクベースの汎用的モデルの構築に際して中心となる問題の 1 つである。例えば、異なるタスク間において、物性値やラベルのデータ分布の違いは大きくなることが多く、単純な相互学習をすると性能劣化することがわかっている。

本研究では特に学習データと予測データの間で分布が異なる場合について注力し、事前学習を用いて外挿予測 (予測データが学習データに含まれない問題) の精度がどのように改善するかを評価した。事前学習には擬似タスクを用い、(1) 分子中の元素を一部 masking し再構築、(2) 分子に含まれる結合構造のパターン(グラフパターン)を予測、(3) 分子に含まれる元素と結合構造のパターン(グラフ Motif) を予測、を利用した。

ターゲットの物性として、Homo-lumo energy gap を利用し、学習データと予測データを閾値によって 2 分割することで外挿問題を設計した (閾値以下で学習し、閾値以上を予測)。事前学習によってターゲットタスクでのデータ分布とは独立した分布を学習でき、ターゲット物性のラベルがない入力データを効果的に学習に利用できるため、外挿予測の精度が大きく向上した。また、擬似タスクは (1), (2), (3)の順で難しくなるが、予測精度も擬似タスクの難易度に従って向上することがわかった。(業績[2,4])

学習データの不均衡問題: 本研究では、データ内の要素 (例えば、分子中の元素の頻度) の不均衡が物性予測にどのように影響するかを調査し、学習時に不均衡を是正することで精度にどのような影響があるかを研究した。物性予測において、データ内の要素は非常に偏っている。例えば、有機物質のほとんどが炭素から構成されるため、データは炭素に偏る。しかし、物性を特徴づけるのはしばしば炭素以外の元素であり、それらの少数元素は学習時に過少評価されてしまう。また、物性や予測タスクが異なると注目すべき元素が変わってくるため、データ内の要素の不均衡性はマルチタスク学習でのモデル汎用化において中心的な問題となる。

本研究では特に、ターゲット物性値予測の別タスクとして、分子中の元素を一部 masking し復元予測する擬似タスク (node masking) を用いた。

Node masking はシンプルながらも BERT など多くの既存研究で用いられており高い効果が実証されている。また、データの不均衡性がラベルに直接的に現れるため本研究では用いた。

本研究では、有機物質の分子データとその Homo-lumo energy gap のデータをターゲット物性として利用した。有機物質においては構成元素のほとんどが炭素であるため、Node masking の擬似タスクを行うと、その正解ラベルのほとんどが炭素元素となり過大評価される。学習時に元素の頻度に基づいた不均衡是正を行うことで、精度改善が見られた。(業績[3])

ガラスのダイナミクス予測への応用：本研究では実際にガラス構造の時間経過を予測する問題に着手した。本問題では、予測対象が時間経過後の構造全体でありどの情報を学習するか（つまりラベルデータの設計と学習方法)、 実データをグラフによってどのようにモデル化するかが注力する点となる。既存研究では、実データの原子（グラフにおけるノード）に着目し GNN モデルにて学習をおこなっていたが、本研究では原子に加え原子間の相互位置（グラフにおけるエッジ）に着目し、ノードとエッジ情報のマルチタスク学習を行うことで精度の向上を実現した。(業績[1])

6. 進捗状況の自己評価と今後の展望

当初の予定では具体的な物性に関して様々な適応を目指すことになっていたが、物性データ生成の部分での遅れがあり、ベンチマークやオープンデータを使っての基礎的な問題の研究に方針を変更したため、自己評価は 75%とする。次年度は、準備中の様々な物性値データ（特に、固体系でのフォノン由来の熱伝導率、Tc、DOS 等）を実際に活用し応用面での貢献に注力したい。

画像認識や自然言語処理にて十分に実証され確立された手法であっても、物性予測の問題に応用すると申請者の予想を遥かに超えた難しさが多く発見された、本研究分野への貢献を引き続き進めていきたい。

7. 研究業績

(1) 学術論文 （査読あり）

[1] Hayato Shiba, Masatoshi Hanai, Toyotaro Suzumura, and Takashi Shimokawabe, ”BOTAN: BOnd TArgeting Network for prediction of slow glassy dynamics by machine learning relative motion.” The Journal of Chemical Physics 158, no. 8, 084503, 2022.

(2) 国際会議プロシーディングス （査読あり）

(3) 国際会議発表 （査読なし）

(4) 国内会議発表 （査読あり）

[2] Shun Takashige, Masatoshi Hanai, Toyotaro Suzumura, Limin Wang, Kenjiro Taura, ”Is SelfSupervised Pretraining Good for Extrapolation in Molecular Property Prediction?” xSIG 2023 (cross-disciplinary workshop on computing Systems, Infrastructures, and programminG)

[3] Limin Wang, Masatoshi Hanai, Toyotaro Suzumura, Shun Takashige, Kenjiro Taura, ”On Data Imbalance in Molecular Property Prediction with Pre-training” xSIG 2023 (cross-disciplinary workshop on computing Systems, Infrastructures, and programminG)

(5) 公開したライブラリなど

(6) その他（特許，プレスリリース，著書等）

[4] Shun Takashige, Masatoshi Hanai, Toyotaro Suzumura, Limin Wang, Kenjiro Taura, ”Is SelfSupervised Pretraining Good for Extrapolation in Molecular Property Prediction?” (Under review on NeurIPS 2023)