jh250048

生成モデルを用いた固体材料発見手法の追求

華井雅俊（東京大学情報基盤センター）

概要

本プロジェクトでは、機械学習を用いた固体材料の発見的手法に関して研究をおこなう。特に、新しい材料構造の発見や物性の極値予測など、対象が学習データ分布の外側にあるいわゆる外挿問題に注力する。生成モデル・物性予測モデルの構築、大規模データセットの作成を総合的に実施することで問題の汎化を行うとともに、それぞれのターゲット材料・ドメインへの応用を探求する。今年度においては、主にオープンデータの収集・分析、ドメイン特化の物性データの大規模生成とそれらを使った予測モデルの構築とデータ学習を実施した。

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

東京大学 情報基盤センター

(2) 課題分野

データ科学・データ利活用課題分野

(3) 参加研究者一覧と役割分担

華井雅俊（研究総括）
河村光晶（材料分野担当）
鈴村豊太郎（機械学習分野担当）
吉見一慶（データセット計算タスク最適化）
青山龍美（データセット計算タスク最適化）
本山裕一（データセット計算タスク最適化）
李 恒宇（データセット計算タスク最適化）
二塚 俊洋（データセット作成・分析）
LI AOWEN（データセット作成・分析）
久住 太一（データセット作成・分析）

2. 研究の目的と意義

昨年のノーベル物理学賞・化学賞に代表されるように、機械学習手法の自然科学分野における応用が活発であり、材料分野においてもデータ駆動型の材料発見や物性予測の研究が盛んに行われている。機械学習や生成モデルで顕著な成功をおさめる自然言語・画像処理分野と大きく異なり、 自然科学分野、特に材料分野におけるデー タ駆動型研究開発の難しさは発見的タス クを本質的に内在することに起因することが大きい。つまり、既存のデータから未知の材料や特異な物性値を発見すること（外挿）が材料開発では強く求められるが、 今日のニューラルネットワークをベースとした 機械学習モデルは内挿能力に比べ外挿能力が非常に限定的であることが知られ困難をきわめる。材料分野における外挿能力の追求はこれまで様々行われており、例えば、既知の物性値から未知の物性値への Transfer Learning・Multi-Task Learning や、外挿の学習方法を学習する Meta Learning、 Pretraining + Fine tuning を利用した方法などが挙げられる。しかしながら、これらは主に物性値予測における問題に終始しており、未知の構造を発見する生成タスクにおける外挿問題の研究は未発展である。また共通の評価ベンチマークが未確立なため個々の事例に基づく評価に限定されてしまっている。実際、結晶構造の生成問題

に関する最新研究においても、その多くが既知のデータセットからどれだけ再現できるかを定量評価の中心としており、未知の構造に関しては個々の結果に関してのアドホックな検証や定性的な評価に限定されている。そこで、本研究では結晶構造予測を中心とした生成タスクにおける外挿問題の追求を目的とする。

3. 当拠点の公募型共同研究として実施した意義

本研究はデータ科学と材料科学分野との学際領域である。申請メンバーの内、華井がデータ科学部分の統括を担当し、河村が材料科学分野の統括し研究を遂行する。また、近年のデータセット肥大化から複数の GPU ノードを用いた大規模データ学習が不可欠であり、JHPCN 資源の利用が必要であった。

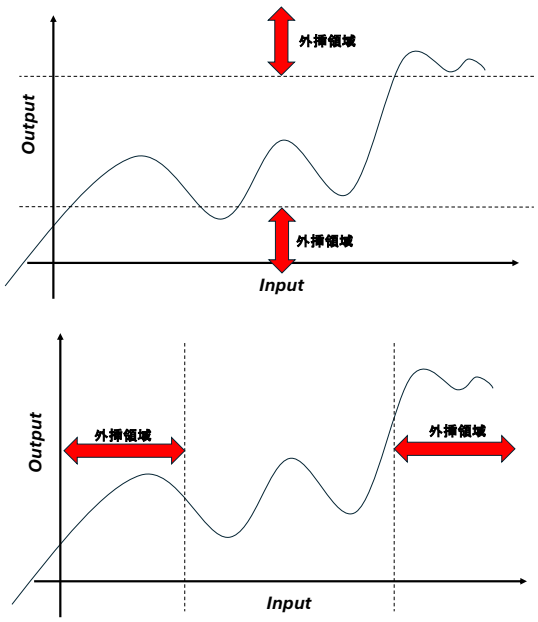
また、計算機利用の効率化を図るべく、東京大学物性研究所ソフトウェア開発・高度化チームをメンバーに加えて研究チームの強化を実施した。

4. 前年度までに得られた研究成果の概要

新規課題のため該当せず。

5. 今年度の研究成果の詳細

今年度は、主にデータセットの構築にて成果があった。まず、発見的タスクベンチーマークにおいて、固体物質の構造データセットの



うち発見の年代と論文情報を含む ICSD のデータを全取得し分析を進め、より大規模な Materials Project, Omat24, Alexandria のデータとともに外挿の分類・分析を進めた。具体的には図にあるように、Input (構造データなど)による分類と Output(エネルギー、力やその他の物性値など)による分類を行い、データ分布に関する調査を実施した。

より具体的かつ有用な外挿タスクとしてオープンデータで一般的なエネルギー値や Bandgap だけでなく熱伝導物性に取り組んでおり、こちらは DFT による大規模データセット作成に関して論文としてまとめた(https://phonix-db.org/)。 次年度以降、本データセットを使って熱伝導が極値であ

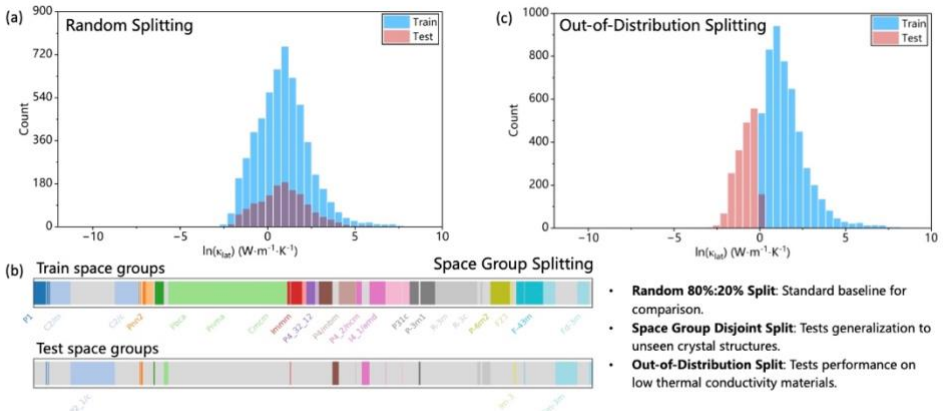


Figure 2: Dataset split for the benchmark. (a) Random 80%:20% split. (b) Space-group disjoint split. (c) Out-of-distribution split.

る材料の新規発見をターゲットとしてモデル開発を進める予定である。

熱伝導物性における各種既存モデルの外挿予測パフォーマンスに関して現状の結果を以下報告する。図は、データ学習における Train/Test スプリットに関して示している。大きく、物性値（output）による外挿（Out-of-Distribution Splitting）と、Space-group 分類（input)による外挿を定義した。比較として、単純な random split を定義し、“構造から熱伝導物性値”の予測を評価した。

図は、各モデルの分類とその結果である。（i） Pysical-Informed Feature+ML は、物理学の知識を反映した Human-craft の特徴量を利用する方法、(ii) MLIP Embedding + ML は、汎用 Machine Learning Inter-atomic Potential を経由し、力とエネルギーを予測したのち物性予測をする方法、(iii)End-to-end Deep Neural Network は、構造データから熱伝導物性値を直接予測する方法、である。

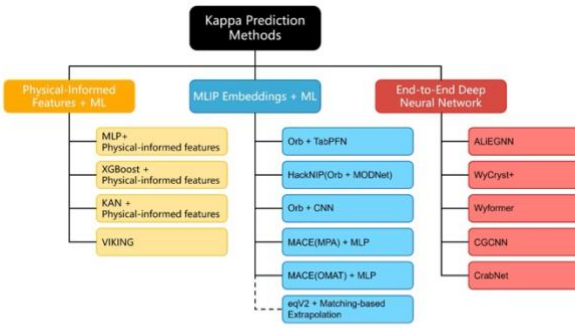


Figure 3: Overview of the surrogate models categorized into three groups, with corresponding developers indicated for each model.

個々のモデルにおいて、差異はみられるものの、Out-of-distribution においては Random Split・Space Group Split と明確に性能特性が異なっており、End-to-end な方法が比較的優位となる結果となった。外挿予測においては物理的な特徴量や汎用モデルの効果が限定的となっており、今後の研究方向性を決める発見となった。

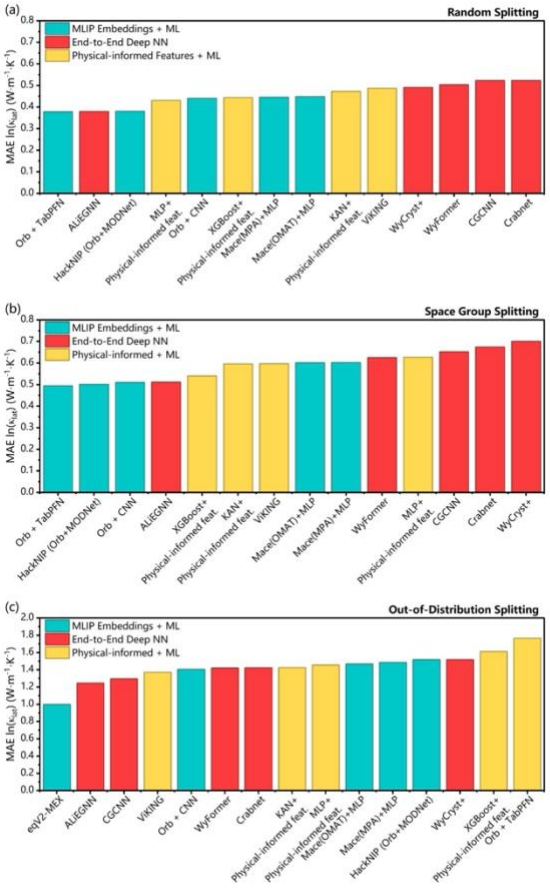
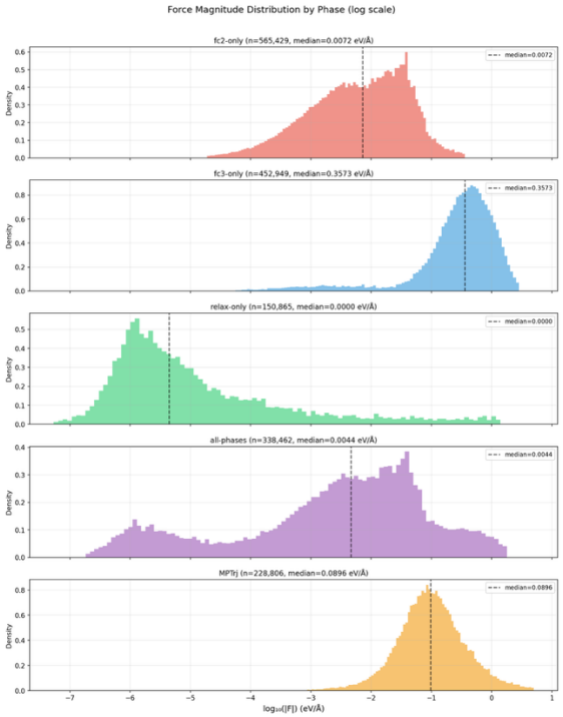


Figure 4: The performance across the 15 surrogate models for each individual data split, with color separation into the three identified categories.

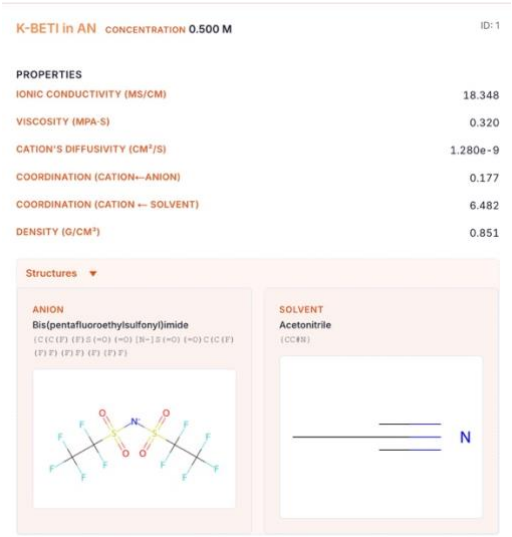
また、熱伝導物性の導出中に得られた DFT 計算の結果の MLIP への利活用に関しても研究を進めている。DFT による理論計算は他物性の中でも非常に計算コストがかかることが知られ（非調和フォノン計算を含むため)、データセット作る過程でより高密度な DFT 計算が得られる。特に、(1) Super cell を用いた広い領域での相互作用データ（Primitive cell や Conventional cell といった冗長性を省いた表現ではなく）である点、(2) ポテンシャルエネルギー曲面の FC2, FC3 をいった高次の微分係数導出に耐えうるだけの高精度データである点、に特徴がある。

図は生成の DFT データ（Phonix）と オープンデータ MPTrj における Force の比較である。特に構造最適化時（relax-only）と FC2 計算時（fc2-only）においてオーダーレベル（10-3~-2)での違いがあり、分布の違うデータを効果的に学習する必要がある。今後こ

れらデータを用いた汎用 MLIP モデルへの適応を実施予定である。



また、電池材料向けの電解液データベースの構築も別プロジェクトとの共同で進めており（ https://oedb.jp/ ）、こちらでも外挿的な応用 （秀でた性能を持つ電解液の発見）が求められる。今後、物性予測モデルの構築を進める予定である。 電解液データベースでは、1つの材料（Anion, Cation, Solvent の組み合わせ）に対して6つの物性を含んでおり、各物性のデータ分布の違いを考慮した予測モデルの構築が求められる。図は、1つの電解液に対して計算した物性の例である。



最後に計算機利用に関しての報告であるが、オープンデータの収集を通じて、実際の DFT 計算や MD 計算によってデータを増やしていくことが不可欠であることを痛感した。データ生成を更に加速するために、スパコン上での物性計算の自動 Job 生成ツールである Moller (https://github.com/issp-centerdev/Moller)の Miyabi 対応を始めた。Miyabi へのポーティングに際し、東京大学物性研究所ソフトウェア開発・高度化チームをメンバーに加えて研究チームの強化を実施した。

大規模データの収集・管理およびそれらの公開に関しては mdx 上に構築した web service である ARIM-mdx Data System（ https://arim.mdx.jp/ ）を利用した。熱伝導データに関しては合計 20TB 以上、電解液データに関しては 1 PB 以上となっており、miyabi やその他スーパコンピュータからの効率的な転送や、圧縮の方法に関して今後進めていく必要がある。

6. 進捗状況の自己評価と今後の展望

概ね良好である （80%の達成）。2025 年度は主にデータセットの構築にて成果があったが、これらデータセット作成をさらに効率化するとともに、作成したデータを使ってエネ

ルギーと力のような汎用物性だけでなく、より応用に即した外挿予測モデルの構築に研究対象を広げていく。

本課題は 2026 年度においても継続実施しており、2026 年度においては、以下を実施する。

（i） 外挿予測 Task のよりハイレベルな分類とベンチマーク化:

2025 年度は熱伝導物性と電解液のデータセット構築に関して特に注力した。これら特定の応用から得られた知見を整理・分類することで、外挿予測問題を適切に定義しそれら問題の難しさや各手法の優位性をより広範囲に分析する。また、2025 年度に収集した既存データセット(Materials Project, AFLOW, JARVIS, ICSD, Open Catalyst, OMAT, GNoME)をベースにした新しいベンチマークの設計を実施する。

(ii) 生成モデルへの適応

2025 年度は生成モデル（所望の条件:構成元素や物性値など => 出力: 条件を満たすような結晶構造） の研究実施には至っておらず物性値予測 （構造 => 物性値） の追求に終始していた。2026 年度は 2025 年度での知見を更に深め、生成モデルの研究を開始する。

(iii) スケーラブルな大規模モデルの構築

2025 年度は Meta の FAIRChem (eSEN モデル) を対象に ARM 対応、省メモリ対応、1GPU per node 構成などに対応した改良を行い、Miyabi へポーティングを実施した。2026 年度においては、これらの知見をベースに単なるプログラム改良ではなく、より本質的かつ汎用的なモデル改良を実施する。特に、近年の汎用 MLIP においてはモデルの予測精度はメモリサイズに強く律速されていることが知られており、プログラムの最適化や GPU の効率利用に関して研究の余地がまだまだ大きいことがわかっている。

※7. 研究業績はウェブ入力です