プレスリリース

ISM2024-07
2025 年 3 月 3 日
大学共同利用機関法人 情報・システム研究機構　統計数理研究所
JSR 株式会社

報道関係各位

データの外の世界を予測する方法を学ぶ AI 技術
～ データ駆動型材料研究における有効性を実証 ～

- 外挿的予測の一般的な方法を学習するメタ学習技術 “E2T アルゴリズム”とソフトウェアを開発
- 材料物性予測タスクにおいて、学習データに含まれない元素の組み合わせや構造的特徴を持つ材料に対しても高い予測精度を実現
- 大量の外挿的タスクを経験したモデルが、新しいタスクへの迅速な適応能力を獲得することを実証

概 要

統計数理研究所 JSR-ISM スマートケミストリラボ（JSR 株式会社と統計数理研究所の共同研究部門）の野田考平研究員（JSR 株式会社）および吉田亮教授（統計数理研究所）らの研究グループは、学習データの範囲外での予測を可能にする機械学習技術“E2T”アルゴリズムを開発し、その有効性を材料研究において実証しました。

材料科学の究極の目標は、データが存在しない未踏領域から新しい材料を発見することです。しかし、機械学習の予測は一般に内挿的であり、その有効性は既存のデータ分布に近い範囲に限られます。さらに、材料研究ではデータ取得コストが非常に高いため、十分な学習データを確保することが難しく、必然的にデータの範囲外、すなわち外挿的領域の探索が求められます。

この課題を解決するために、研究グループは E2T (extrapolative episodic training)という新しい機械学習アルゴリズムを開発しました。E2T では、手元のデータセットから人工的に生成した大量の外挿的タスクを用いて、メタ学習器という特殊なモデルを訓練します。その結果、モデルは外挿的な予測を実現するための汎用的な“学習方法”を自律的に習得します。

本研究では、E2T を材料特性の予測タスクに適用し、学習データには存在しない元素の組み合わせや構造的特徴を持つ材料の特性についても高精度な予測を実現できることを実証しました。また、大量の外挿的タスクを経験したモデルは、未知ドメインに少量のデータを追加するだけで迅速に予測能力を獲得できることも明らかになりました。

本研究成果は、2025 年 2 月 22 日付で Communications Materials 誌から発表されました。

研究成果

近年、機械学習を活用した新材料の発見が次々と報告されています。その基盤を支えているのが、機械学習に基づく物性予測技術です。数百万あるいは数十億の候補材料をモデルに入力することで、広大な探索空間から所望の特性を持つ候補材料を効率的に同定できるようになりました。しかしながら、多くの研究では、十分なデータを確保することが難しく、機械学習を適用できる範囲は限定的であるのが現状です。また、材料科学の究極的な目標は革新的な特性を持つ未知の材料を発見することです。しかしながら、一般に機械学習で予測可能な領域は学習データの近傍に限られており、データの外側の未知領域を探索するのは容易ではありません。たとえば、近年の大規模言語モデルなどの生成 AI も、基本的には人間が経験したタスクを模倣する内挿的な性質を持つ機械といえます。データの外の世界を予測する AI 技術の創出は、材料科学のみならず、次世代 AI 研究全体のグランドチャレンジといえます。

外挿的予測を実現するために、これまで以下のような機械学習手法が研究されてきました。

(1) ドメイン一般化 (domain generalization)：異なるデータ分布間で共通の特徴表現を学習する手法
(2) データ拡張（data augmentation）：学習データの多様性を増やし、モデル性能を向上させる手法
(3) 物理知識との統合：物理法則などの事前知識を機械学習に組み込むアプローチ (例: physics-informed neural networks)
(4) メタ学習：多数の異なるタスクをモデルに経験させることで、一般的された学習則を習得させる手法

本研究では、特に“メタ学習”に基づくアプローチに着目し、外挿的予測のための汎用的な学習則をモデルが学習できる技術を開発しました。具体的には、Matching Neural Networks という注意機構（attention mechanism）を備えたニューラルネットワークを利用し、外挿的な予測を実現するために必要な学習方法をモデルに学習させました(図 1)。具体的には、与えられたデータセットから訓練データ集合 D と外挿的な関係にある入出力の組 (x, y) をサンプリングします。ここで x は材料、y はその特性を表します。これら三つの要素の組をエピソードといいます。エピソードは任意に生成できます。そこで、人工的に生成した大量のエピソードを用いて、材料 x から特性 y を予測するメタ学習器 $y = f(x, D)$を学習します。このように訓練されたモデルは、データ集合 D と外挿的関係にある (x,y) を予測するために必要な一般的な関数 f を学習します。研究グループは、この学習アルゴリズムを E2T（extrapolative episodic training）と命名しました。

研究グループは、E2T を高分子材料や無機化合物に関する 40 種類以上の物性予測タスクに適用し、その性能を評価しました（図 2)。その結果、ほぼ全てのケースで、E2T で訓練されたモデルは通常の機械学習モデルの性能を上回る、あるいは同等以上の外挿性能を示しました。また、訓練データ近傍における予測性能についても、E2T は従来の機械学習モデルの性能と同等以上の精度を達成していることが確認されました。一方で、E2T の外挿性能は、外挿領域を含むすべてのデータセットで学習した理想的なモデル（オラクル）の性能には及びませんでした。つまり、E2T は外挿領域の予測性能を安定的に向上させるものの、“究極の外挿能力”の獲得には至りませんでした。

さらに注目すべき点として、大量の外挿的タスクで訓練されたモデルは、わずかなデータを用いたファインチューニングにより、未知の外挿タスクに迅速に適応する能力を示しました。特に、外挿領域を含むデータで学習したオラクルよりも遥かに少ないデータ量で、同等の性能に達することが実証されました。人間の迅速な適応力は、個人の生まれ持った資質だけでなく、訓練や経験を通じて強化される能力であるといわれています。本研究は、AI の学習においても同様な現象が起こりうることを明らかにしました。

今後の展望

材料研究の究極の目標は、データが存在しない未踏領域を開拓することにあります。例えば、研究者はこれまでに試したことがない元素や原料の組み合わせ、あるいは試料作製プロトコルを大きく変更した場合の材料特性に関心があります。本研究は、現在のデータセットで外挿性を獲得するための訓練を受けたモデルは外挿性や未知の環境への適応能力を獲得できるのではないかという素朴な問いから始まり、その問いに対する非常にシンプルな解法を提示しました。現時点では限られたケースにおけるエビデンスに留まっていますが、E2T の学習能力が普遍的なものであると証明されれば、そのインパクトは、材料科学を超えて AI for Science の多くの分野に波及する可能性があります。特に、基盤モデルの学習への E2T の応用が期待されます。基盤モデルは、大規模かつ汎用性の高いデータセットを用いて訓練されたモデルで、多様なタスクに適応する能力を備えていることが求められます。基盤モデルを特定のタスクやドメインに適応させるために追加学習を行うことで、必要なデータ量を削減しながら、特定のタスクにおける高精度な予測を実現します。E2T が持つ外挿性能やドメイン適応力は、基盤モデルの開発に新機軸をもたらし、科学の進展に大きく貢献することが期待されます。

発表論文

タイトル: Advancing extrapolative predictions of material properties through learning to learn using extrapolative episodic training

著者: Kohei Noda, Araki Wakiuchi, Yoshihiro Hayashi, Ryo Yoshida
掲載誌: Communications Materials
DOI: 10.1038/s43246-025-00754-x

プログラム公開ウェブサイト

https://github.com/JSR-ISM-Smart-Chemistry-Lab/E2T

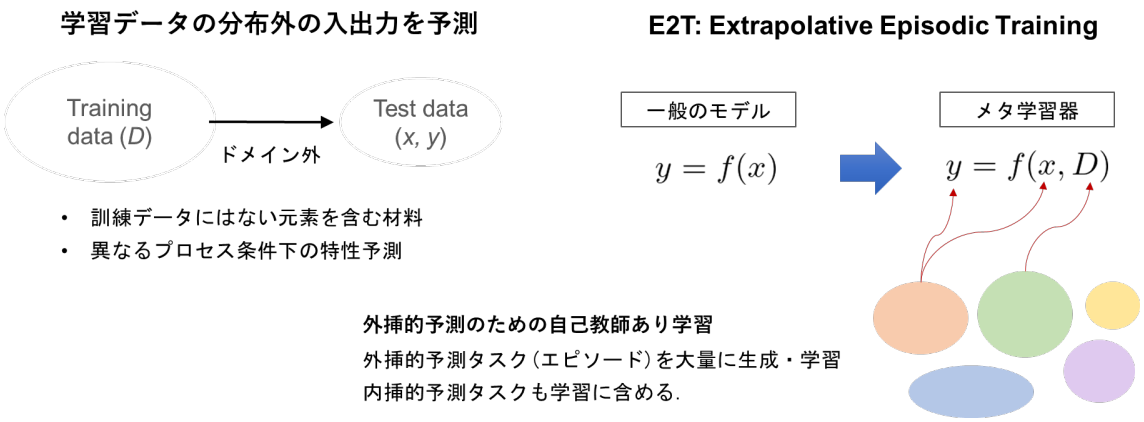

図 1: E2T アルゴリズムによる外挿的予測のための学習方法の学習

大量の外挿的予測タスクを学習したモデルは外挿性を獲得

- データ：1,345個の有機無機ハイブリッドペロブスカイト化合物のバンドギャップ
- 外挿領域の定義 {Ge, F} を含む化合物，{Pb, I}を含む化合物を訓練データ集合から除外

外挿性能の “改善”，内挿性能も低下しない.

Methods	HOIP-GeF		HOIP-PbI	
	R^2	RMSE (eV)	R^2	RMSE (eV)
MPNN-Linear	0.255 ± 0.198	0.361 ± 0.046	0.545 ± 0.064	0.207 ± 0.014
MPNN-FCNN	-0.088 ± 0.614	0.427 ± 0.106	0.508 ± 0.185	0.213 ± 0.037
ANE [40]	0.361 ± 0.105	0.336 ± 0.027	0.510 ± 0.108	0.214 ± 0.024
E2T	$\mathbf{0.486 \pm 0.095}$	$\mathbf{0.301 \pm 0.027}$	$\mathbf{0.605 \pm 0.057}$	$\mathbf{0.193 \pm 0.013}$
Oracle	0.557 ± 0.166	0.253 ± 0.042	0.766 ± 0.113	0.140 ± 0.024

外挿領域のデータでメタ学習器をファインチューニング

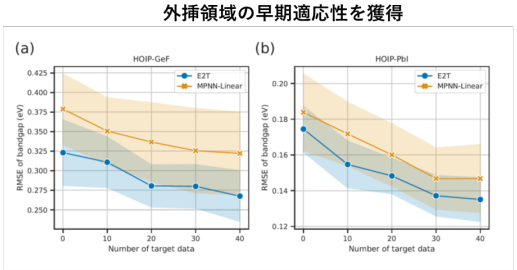

図 2: 有機・無機ハイブリッドペロブスカイトのバンドギャップ予測

本件に関するお問い合わせ先
【研究内容について】
大学共同利用機関法人 情報・システム研究機構 統計数理研究所
先端データサイエンス研究系 マテリアルズインフォマティクス研究推進センター
教授（センター長）
吉田 亮（よしだ りょう）
TEL：050-5533-8534　E-mail：yoshidar@ism.ac.jp

【報道・広報について】
大学共同利用機関法人 情報・システム研究機構 統計数理研究所
運営企画本部 企画室 URA ステーション
TEL：050-5533-8500（代表）　E-mail：ask-ura@ism.ac.jp
〒190-8562　東京都立川市緑町 10-3