

物性予測タスク訓練済みモデルの包括的ライブラリ XenonPy.MDL を公開

～転移学習で材料インフォマティクスのスモールデータの壁を乗り越える～

概要

統計数理研究所と物質・材料研究機構の共同研究グループは、低分子、高分子、無機材料の45種類の特性を対象に約140,000個の機械学習の予測モデルを開発し、訓練済みモデルライブラリ XenonPy.MDL^{※1}を公開しました。XenonPy^{※2}は、統計数理研究所ものづくりデータ科学研究センターのチーム（Chang Liu 特任助教、野口瑤 特任研究員、Stephen Wu 助教、山田寛尚 特任研究員、吉田亮 同センター長）が物質・材料研究機構と共同開発しているマテリアルズインフォマティクス（MI）のオープンソースプラットフォームです。XenonPyにはMIの様々なタスクを実行する機械学習アルゴリズムが実装されており、ユーザーはAPI^{※3}経由でXenonPy.MDLの訓練済みモデルを再利用し、材料設計の様々なワークフローを構築できます。今回アメリカ化学会の論文誌 ACS Central Science に掲載された論文では、XenonPy.MDLのファーストリリースを発表するとともに物質・材料研究の様々なタスク（スモールデータに基づく低分子・高分子・無機結晶の物性予測）において転移学習^{※4}が持つ潜在的な予測性能を炙り出すことに成功しました。

本研究の一部は、科学技術振興機構（JST）のイノベーションハブ構築支援事業「情報統合型物質・材料開発イニシアティブ（MI²I: “Materials research by Information Integration” Initiative）」（拠点：物質・材料研究機構、プロジェクト実施期間：2015-2019年度）の支援により実施されました。

研究の背景

機械学習の他の応用領域に比べると材料研究のデータ量は圧倒的に少なく、これによりMIの適用領域が限定されています。原因として、次の三点が挙げられます：(a)データ取得の高コスト性、(b)研究者のニーズの多様性によるコモンデータベース創出の難しさ、(c)競合相手に対する情報秘匿の意識が高く、研究者にデータ公開に対するインセンティブが働きにくい。したがって、オープンデータベースの開発が中々進まず、さらに、先端領域に近づくにつれてスモールデータの傾向はより顕著になります。また、コミュニティ全体でコモンデータを創出しようという動向も極めて低調であり、少なくとも短中期的には大学のラボや一企業で生産可能なデータセットがMIの標準になることが予想されます。

研究内容と成果

同グループは、転移学習の活用がMIの限界を突破する鍵になると考えてきました。転移学習は、あるタスクで訓練されたモデルを他のタスクに転用する解析手法です。例えば、大量の画像データを用いて動物の種類を判定するニューラルネットワークを訓練し、少数の花の画像データを用いて訓練済みモデルを改変することで、花の種類のカテゴリを構築します。動物のカテゴリは、訓練の過程で汎用的な画像特徴量を獲得していることが期待され、その一部は花のカテゴリにも転用できる可能性があります。その場合、花のカテゴリを一から学習するのではなく、少数のデータで動物のカテゴリを修正すれば十分かもしれません。ヒトの脳には、少ない経験でも合理的に予測を行うメカニズムが備わっています。例えば、小さい頃からピアノを学んでいた人は、音楽に関する一般的な知識を獲得しているため、他の楽器の演奏技術を比較的容易に習得できます。このような推論メカニズムを模倣したものが転移学習です。脳の情報

プレスリリース

処理との対比で言えば、多様且つ包括的な訓練済みモデル群を実装することは、多くの経験から記憶の集合体を獲得することに相当します。モデルの多様性が増すほど強力な転移学習を実現できる可能性が高まると同グループは考えています。

今回アメリカ化学会の論文誌 ACS Central Science に掲載された論文では、XenonPy.MDL のファーストリリースをアナウンスするとともに材料研究の様々なタスク（スモールデータに基づく低分子・高分子・無機結晶の物性予測）で転移学習が持つ驚異的な性能を実証しました。その中で最も特筆すべき点は、転移学習による外挿性の獲得の可能性が示されたことです。一般に革新的な物質の周辺にはデータは存在しません。機械学習は訓練データとテストデータの類似性に基づいて予測を行うため、周辺に訓練データが全く存在しない外挿領域では予測能力を失います。しかしながら、転移学習を巧みに適用することで、極めて少数の訓練データでも時には超外挿的ともいえる予測モデルを構築できる可能性が示されました（図 1）。関連するタスクで広大な物質空間を包含する大量のデータを用いて事前学習を行います。すると、訓練されたモデルは広大な物質空間に適用可能な「特徴表現」を獲得します。このように訓練済みモデルが獲得した汎用的な特徴抽出器を目標ドメインに転移することで、目標タスクの訓練データが極めて少数だったとしても、転移されたモデルには既存データとは大きく異なるドメインの予測能力が備わっていることが明らかになりました。

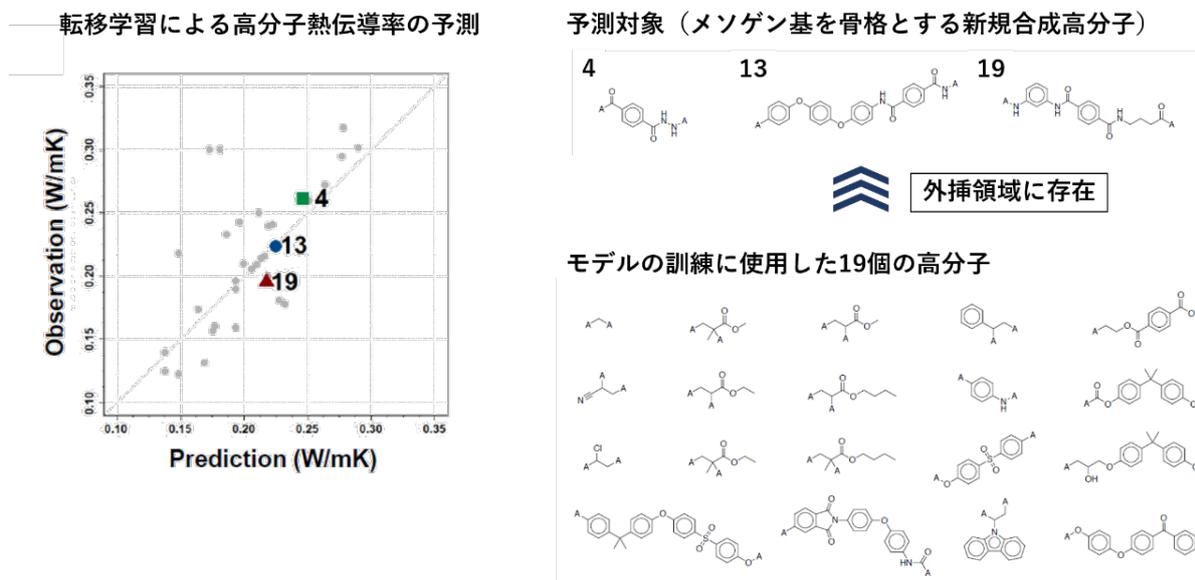


図 1: 転移学習による高分子物性（熱伝導率）の予測。高分子熱伝導率の 19 件の訓練データを用いて、XenonPy.MDL の高分子のガラス転移温度、低分子化合物の比熱容量等の訓練済みモデルを転移し、外挿性を持つ予測モデルを構築することに成功（Yamada and Liu et al. ACS Cent Sci 2019）

今後の展開

同グループは、転移学習の導入と戦略的活用が MI の究極の目標である「外挿的予測と発見」の実現に向けた大きな一歩になると考えています。材料デバイスの開発では、転移学習で個々の構成素子の物性予測モデルを統合することで、デバイス特性の少数データを用いて高精度なデバイス特性のモデルを構築できる可能性があります。また、高分子・無機結晶の熱伝導率、化学合成の反応予測等、転移学習は MI の様々なタスクで驚異的な性能を発揮することが分かってきました。優れた研究者が過去の経験から大量且つ多様な知識体系を獲得しているのと同様に、転移学習の成功の鍵は超包括的で高品質な訓練済みモデルライブラリを実装することです。同グループは今後も継続的に XenonPy.MDL の開発を推進していきます。

プレスリリース

掲載論文

題目 : Predicting materials properties with little data using shotgun transfer learning

著者 : Hironao Yamada, Chang Liu, Stephen Wu, Yukinori Koyama, Shenghong Ju, Junichiro Shiomi, Junko Morikawa, Ryo Yoshida

雑誌 : ACS Central Science

DOI: 10.1021/acscentsci.9b00804

掲載日時 : 2019 年 9 月 30 日午後 9 時 (9 月 30 日午前 8 時 米国東部標準時 (EST))

用語説明

※1 物性予測訓練済みモデルライブラリ XenonPy.MDL:

https://xenonpy.readthedocs.io/en/latest/tutorials/5-pre-trained_model_library.html

※2 MI オープンソースプラットフォーム XenonPy: <https://github.com/yoshida-lab/XenonPy>

※3 API: Application Programming Interface

※4 転移学習 : あるタスクの学習モデルを他のタスクのモデルに改変するための方法論。通常の機械学習に比べると、転移学習はより少ないデータでモデルの予測性能を引き上げることができると言われている。

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 ものづくりデータ科学研究センター長、教授
- 国立研究開発法人 物質・材料研究機構 統合型材料開発・情報基盤部門 情報統合型物質・材料研究拠点 物質・材料記述基盤グループ長

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
〒190-8562 東京都立川市緑町 10-3

TEL: 050-5533-8580 FAX: 041-526-4348

E-mail: ask-ura@ism.ac.jp

国立研究開発法人 物質・材料研究機構 経営企画部門 広報室

〒305-0047 茨城県つくば市千現 1-2-1

TEL: 029-859-2026, FAX: 029-859-2017

E-mail: pressrelease@ml.nims.go.jp