研　究　報　告　書

「大規模データに基づく電子物性予測のための深層学習技術の創出」

研究タイプ：通常型
研究期間：　2015 年 12 月～2019 年 3 月
研 究 者：　瀧川　一学

1.　研究のねらい

物質・材料の計測や量子化学計算による電子物性の大規模なデータの蓄積とそのデータベース整備により、帰納的な知識・法則性の発見やデータに基づく予測へのこれらのデータの利活用の期待が高まっている。そこで本研究課題では、ここ数年、劇的な技術発展を遂げ多様な応用分野の標準を塗り替えている機械学習の枠組みである「深層学習」の最新知見に基づき、電子物性を高速・高精度に予測するデータ駆動型の計算技術の確立を目的とする。本研究により、対象個々に必要となる高コストな量子化学計算を、即時に結果が得られるデータ駆動型予測で置き換え、幅広い種類の可能な候補物質・材料の電子物性の網羅的自動探索のための高速な予測計算基盤技術を提供すると共に、大規模データから獲得された機械学習モデルの逆過程の解析による逆設計支援や、物性を支配する要因・法則性の帰納的理解を得る方法論の技術的検討を通して、計算科学とデータ科学を融合させた新物質・材料探索のための新たな道筋の創出を目指す。

本研究で答えを出したい科学的問いとは「深層学習技術を駆使すれば、大規模な事例データの背後に潜む量子化学的な法則性やパターンを帰納的に模倣（近似）できるのだろうか？」というものである。本課題では有機低分子の電子物性予測をターゲットとして、三次元の電子密度表現に基づくVolumetric 深層学習技術および分子グラフ表現に基づく深層学習技術を用いた予測技術を創出し、電子物性の機械学習近似性について定量的理解を得ることを目指す。またマテリアルズインフォマティクス関連領域での連携として機械学習技術を用いた共同研究も積極的に推進する。

2.　研究成果

（1）概要

近年、計算コストの高い第一原理計算による電子物性予測の代理モデルとして深層学習をはじめとする機械学習によるデータ駆動予測が着目され、様々な研究が行われている。本課題では、このような大規模データに基づくデータ駆動型の電子物性予測として深層学習技術を主とする機械学習技術に着目し、大きく分けて次の3項目について研究を行った。

研究テーマ A「電子密度表現に基づく Volumetric 深層学習技術の創出」
研究テーマ B「分子グラフ表現に基づく深層学習技術の解析と特徴付け」
研究テーマ C「マテリアルズインフォマティクス分野での機械学習の利活用の推進」

（2）詳細

研究テーマ A「電子密度表現に基づく Volumetric 深層学習技術の創出」

最も広く利用されている第一原理計算手法として密度汎関数法がある。その理論的背景である Hohenberg-Kohn 定理によれば電子物性は全電子密度の汎関数として表現でき、各電子の波動関数を経由せずとも全電子密度で決まる。この汎関数が明示的に構成可能ならば電子物性予測は可能であるが、この主張は存在定理であって陽にその汎関数が分かるわけではない。

そこで電子物性予測の機械学習による近似可能性を評価する上で、全電子密度から電子物性の汎関数を直接機械学習によって近似する手法を構築した。まず、オープンデータとして公開されているHと7原子以下のC,O,N,Fからなる網羅的な有機低分子の電子物性データ(QM9データセット, 133,885 分子)の機械学習による分析と、3D 全電子密度データの計算とデータベース整備を行った。この際、データが大規模となるため格納方式や実装についても技術研究を行なった。こうして得られた大規模データを元に、3D 全電子密度を入力としデータ中から学習できる 13 物性を出力とする Volumetric 深層学習モデルの設計と評価を行った。特に 3D 畳み込みに基づく計算ユニットに基づく各種構造を画像(2D)から拡張し、Skip 接続、並列畳み込みユニット、Squeeze-and-Excitation ユニットなどの技術を用いて、物理的に GPU メモリに格納できる制約のもとでの効果的モデル設計の研究を行った。また、拡張として、各原子の全電子密度を重畳しただけの3D 表現や人工的なポテンシャルによる 3D 表現からアップサンプリングやデコンボリューションを用いて 3D to 3D の Volumetric 深層学習技術を設計し全電子密度の予測を行った。後者は近年先行研究で Hohenberg-Kohn mapping と呼ばれる写像の一例となっており、これを初めて 3D 深層学習によって end-to-end・data-driven で構成することができた。この結果、大規模データに基づく電子物性予測の機械学習近似性が確認できた。

3D の畳み込み深層学習は画像ベースの場合、パラメタが増える上、3D 情報を 2D に射影し膨大な既存データが利用できる 2D の学習済み深層学習モデルをうまく組合せることで 3D より高い予測が得られる場合がほとんどであり、有用性が不確かであった。一方、電子密度の場合は明らかに 2D 射影では不十分でありまた学習済みの 2D モデルも利用できないことから 3D 深層学習による直接的な機械学習は一つの有力なアプローチであると考えられる。本研究ではリアルタイムなランダムデータ拡張を学習過程に導入して回転や並進普遍性自体もデータから獲得する end-to-end なアプローチであり今後さらなる精密化にも取り組みたい。

研究テーマ B「分子グラフ表現に基づく深層学習技術の解析と特徴付け」

さきがけ課題の研究開始頃より、Google や DeepMind の一連の論文を契機とし、分子構造のグラフ表現を入力とする深層学習モデルによって電子物性予測を行う手法が機械学習分野で急速に注目をあびるようになった。そこで上記の電子密度からの学習と並行し、分子グラフ

を入力とする機械学習・深層学習モデルについて研究を行った。その結果、実際の予測精度は入力の分子グラフにおいて各原子にどのような原子特徴ベクトルを用いるかに依存しており、分子グラフの深層学習は画像や音声と比べ end-to-end な学習が依然難しいことが明らかになった。例えば、入力グラフ表現が複数可能な場合、その入力を Attention 機構で選択する深層学習モデルを用いれば精度向上が見られた。

この原子特徴ベクトルはケモインフォマティクス分野の原子不変量に由来するものであり、有機分子をバーチャルスクリーニングなどの統計的予測モデリングで扱う際の「表現」の問題である。分子グラフを入力とする深層学習は主に分子を構成する各原子の近傍部分構造による特徴づけに基づくものがほとんどである。この各原子近傍の部分構造特徴を用いる経験則は SchNet 法など他の分子物性予測の深層学習法や、創薬分野での機械学習アプローチ(ECFP 法や WL カーネル法など)でも共通しており特に生物活性の統計的予測モデリング(定量的構造活性・構造物性分析)において一定の地位を確立したヒューリスティクスである。一方、特に制約しないで任意の部分構造による特徴づけも考えられる。「グラフ」が離散的な表現であり可能な部分構造が組合せ的・指数的な種類となることを考えると、どのような部分構造に着目して予測が行うのが良いかは事前には全く非自明である。そこでグラフ表現に基づく機械学習の基礎研究として、あらゆる部分構造から選択的に予測に必要となるものを同定しながら線形予測を行う手法（研究成果の論文 2）、および、同様の設定で適応的部分構造選択を伴う決定木アンサンブル法により非線形予測を行う手法(研究成果のグラフの機械学習に関する国際会議発表)を開発した。またグラフを各原子近傍の部分構造の生起カウントとみなせば、これは自然言語処理の bag of words に類似した bag of subgraph 表現となる。この表現は各グラフについて独立に非常に高速なアルゴリズムで得られるため、教師付きの確率トピック分析法(sLDA, DiscLDA, MedLDA)などによる分子データ分析も行なっている。また化学反応ネットワークや化合物対ネットワークなどグラフのグラフに対する深層学習法の研究も共同研究として行った。

研究テーマ C「マテリアルズインフォマティクス分野での機械学習利活用の推進」

上記 A と B の主テーマと並行して、機械学習によるデータ駆動予測をマテリアルズインフォマティクスで活用する共同研究を推進した。研究代表者は研究開始以前には、材料科学に関わる研究を行っていなかったため、さきがけ研究開始後に繋がりができた共同研究者が持つ問題やデータに対して機械学習研究を行った。

特に、研究開始時に科研費のナノ構造情報に関する新学術領域に参画していた北海道大学・触媒科学研究所の研究グループ(清水研一・高草木達)と共に不均一触媒の研究に対するデータ駆動型研究を開始した。開始時に最初に取り組んだ課題が、伝統的に触媒の活性を見積もるのに使われている d-バンド中心を電子状態計算なしで機械学習によって予測するものである。この研究では、二元合金の d-バンド中心は簡単な元素に関する特徴量から高精度で予測できることを示した(研究成果の論文 1)。この結果は英・王立化学会（RSC）のオンラ

インマガジン Chemistry World に取材を受けた記事「Machine-learning accelerates catalytic trend spotting」として取り上げられ、また、不均一触媒に関する機械学習活用のレビュー(Nature Catalysis 1, 2018 や AIChE 64(7), 2018)でも取り上げられた。また、米・化学会(ACS)の年会のセッション Machine Learning for Catalysis Research にも招待され発表・討論を行った。その後、清水グループの課題は JST CREST のマテリアルズインフォマティクス領域に採択され引き続き、電子状態計算結果からの機械学習(研究成果の論文 3,4)、文献データ・実験データからの機械学習(論文準備中・企業共同研究など)の研究を進めている。他にも研究成果の論文 5 や北海道大学・理学部化学科(領域のさきがけ研究者 2 名が在籍)をはじめとして共同研究で機械学習手法を利活用する研究を推進しており、有機分子に関する研究は北海道大学の世界トップレベル拠点・化学反応創成研究拠点(WPI-ICReDD)採択(研究代表者は PI の一人)や科研費・挑戦的研究(萌芽)等、次の課題へも繋げることができた。

3. 今後の展開

本研究課題で対象とした分子構造からの電子物性予測の精度向上はこの順問題の逆プロセスである望ましい電子物性を持つ分子構造の探索の前提である。開発した深層学習技術に基づいて高精度な予測が可能になれば、この逆問題も近年の関連技術の進展(代理モデル最適化、逐次実験計画、ベイズ最適化など)を活用してアプローチすることが期待できる。

また現状ではすでに大規模な計算結果データが利用可能であった有機低分子のデータセットに基づく評価・分析を行ってきているが、原理上はポテンシャルや電子密度表現は固体へも拡張可能である。したがって、様々な対象について本課題で得た手法を用いて、こうした実際の材料探索へ活用する手法を確立することが次なる課題と言える。

また、物質材料科学分野全般でのデータ駆動型のアプローチの研究も今後さらに進めていく予定である。特に、不均一系触媒や化学反応の探索・設計については、密に連携できる共同研究グループができつつあり、物質材料研究を加速化するための方法論として、専門家とともに最も効果的なデータ駆動型予測の利活用の研究を進めたい。

4. 自己評価

研究代表者は本さきがけ研究への参画以前には全く物質材料科学には関わっていなかったことを考えると、本課題で得られた「全て」は共同研究連携も含めて、本さきがけ課題により始めて可能となったものである。もし採択されていなければ研究代表者が物質材料科学という新分野に関わることも、本課題で得られた知見や技術も確実に生まれていないという意味で非常に有益な研究活動であったと考える。本さきがけ課題終了後も、引き続き物質材料科学におけるデータ駆動型研究について関わっていく技術的知見や共同研究グループも色々と得ることもできた。

また、主関心であった深層学習技術は手法的にも計算機環境的にもさきがけ実施期間に日進月歩で進展したため、こうした各種の知見をリアルタイムで効果的に「深層学習技術の創出」という主課題の技術研究を進めるのに役立てることができ、さきがけ終了後も長く技術的土台となる様々な知見や技術を得ることができた。研究費についても主課題である Volumetric データの深層学習はとにかくメモリ・ストレージ・GPU などが揃わないとなかなか実データを処理できないが、本課題支援のもとサーバー・ワークステーションの効果的な調達により研究を進める貴重な環境を構築することができた。

一方で、境界領域特有の問題により、専門的知見や知識習得や研究実施については困難も多くあった。物質材料科学に関してはゼロからのスタートであったため、より具体的にテーマを研磨していく際にはどのような方向が良いのか下調べや技術的検討に非常に時間がかかった。採択直後には、予定していた共同研究者との連携ができなくなり、分野の専門的知見や基礎的な知識を得るだけでも大変苦労した。また主領域が情報分野のため、学生の参画も1名を除いて得られず、実質は調査・構想からデータ整備や実装まで、通常の情報科学の研究や学生指導と並行して、実施する形となった。主テーマに関しては各種研究成果とデータは得られているが、現在のところまだ論文化にまで至っておらず、どのような筋で発表するか、どのような媒体に発表するか、などは引き続き最終年度内で試行錯誤し検討予定である。

新規分野へのチャレンジとして新しい問題に対する興奮や新しい知識や技術や発見、新たなネットワークによる共同研究、境界分野ならではのたくさんの試行錯誤や苦労、は考えてみると、すべてが新しい分野である物質材料科学のデータ駆動型アプローチを成立させるのに欠かせない要素であり、実施期間中に生まれた様々なシード研究を含めて、潜在的に多大な波及効果をもつと考えられ、そうした出口を見据えて今後も取り組んでいきたい。

成果還元の点では不満の残る進捗とはなったものの、総合的・科学的には非常に有益な研究活動ができたと考える。本さきがけ領域への参画は研究代表者の研究キャリアの上でも分水嶺のようであり、様々な新たな技術課題とチャレンジ、様々な領域の研究者との人脈、様々な新しい問題と発見、などこれからの研究人生の上でも貴重な研究活動となった。

5. 主な研究成果リスト

(1)論文(原著論文)発表

1. Takigawa I, Shimizu K, Tsuda K, Takakusagi S. Machine-learning prediction of d-band center for metals and bimetals. RSC Advances. 2016; 6: 52587-52595.
2. Takigawa I, Mamitsuka H. Generalized sparse learning of linear models over the complete subgraph feature set. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017; 39(3): 617-624.
3. Toyao T, Suzuki K, Kikuchi S, Takakusagi S, Shimizu K, Takigawa I. Toward effective utilization of methane: machine learning prediction of adsorption energies on metal alloys. The Journal of Physical Chemistry C. 2018; 122(15): 8315-8326.
4. Hinuma Y, Toyao T, Kamachi T, Maeno Z, Takakusagi S, Furukawa S, Takigawa I, Shimizu K. Density functional theory calculations of oxygen-vacancy formation and subsequent molecular adsorption on oxide surfaces. The Journal of Physical Chemistry C. 2018 (in press)
5. Pham T L, Kino H, Terakura K, Miyake T, Tsuda K, Takigawa I, Dam H C. Machine learning reveals orbital interaction in materials. Science and Technology of Advanced Materials. 2017; 18(1): 756-765.

(2)特許出願

研究期間累積件数:0 件(公開前の出願件名については件数のみ記載)

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

(Book Chapter) Takigawa I, Shimizu K, Tsuda K, Takakusagi S, Machine Learning Predictions of Factors Affecting the Activity of Heterogeneous Metal Catalysts. Nanoinformatics. 2018;45-64.

(Invited Talk) Takigawa I, Machine learning predictions of factors affecting the activity of heterogeneous metal catalysts, "CATL: Machine Learning for Catalysis Research", The 255th ACS (American Chemical Society) National Meeting, New Orleans, LA, March 18-22, 2018.

（国際会議発表） Shirakawa R, Yokoyama Y, Okazaki F, Takigawa I. Jointly Learning Relevant Subgraph Patterns and Nnonlinear Models of Their Indicators. The 14th International Conference on Mining and Learning with Graphs (MLG 2018) (KDD'18 Workshop), London, U.K., August 20, 2018

（招待講演）瀧川一学，分子のグラフ表現と機械学習，第 79 回応用物理学会秋季学術講演会 特別シンポジウム「インフォマティクスへの招待」～機械学習・インフォマティクスは応用物理をどう変えるか？～，2018 年 9 月 18 日，名古屋国際会議場.

（招待講演）瀧川一学，機械学習は真の発見に寄与できるのか？ MI2I・JAIST 合同シンポジウム（情報統合型物質・材料開発イニシアティブ・北陸先端科学技術大学院大学） データ科学における予測と理解の両立を目指して－分かるとは何か？－，2018 年 5 月 21 日，JST 東京本部別館 1 階ホール.

（サイエンスカフェ）瀧川一学，第 97 回サイエンス・カフェ札幌，見えるものを見る AI 見たいものを見る人間～機械に「正しく」学習させるには～．紀伊國屋書店札幌本店 1F インナーガーデン，2017 年 10 月 1 日