様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月　 ８ 日現在

機関番号：８２６２６
研究種目：若手研究
研究期間：2020～2022
課題番号：２０Ｋ１９８７６

研究課題名（和文）データ駆動科学における量子物理・化学的に解釈可能な深層学習手法の開発とその検証

研究課題名（英文）Development and validation of quantum physics and chemistry-interpretable deep learning methods in data-driven science

研究代表者
椿　真史（Tsubaki, Masashi）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・研究員

研究者番号：８０８０３８７４

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：主な研究成果は、以下の三つである。まず、上記の新たな深層学習モデルを実装し、分子のエネルギーをある誤差の範囲内で外挿予測できることを示した。また、それを実装する過程で、従来の深層学習モデルが量子化学計算における波動関数の重ね合わせと等価な計算を行っていることを、数学的に示すことにも成功した。さらに、新たな深層学習モデルを簡単な低分子で学習し、その学習済みモデルをより複雑な高分子の物性予測へと転移させた。それぞれの研究成果について、一本ずつ合計三本の論文が国際学術誌に掲載された。

研究成果の学術的意義や社会的意義
学術的意義は、物理化学と機械学習という二つの分野を適切に融合できたことである。分子データを扱う際には、どちらかの分野の理論やアプローチのみに偏ることなく、それぞれの分野の良い部分をうまくミックスさせることが必要不可欠であり、それを達成することができた。また社会的意義は、分子データは製薬企業や材料企業のすべてが密接に関わるデータであり、その分野の研究者や技術者にとっての基礎技術を開発できたことである。学習モデルは、製薬企業や材料企業が独自に持つデータでも再学習可能であり、広く使われることが期待できる。

研究成果の概要（英文）：The three main research results are as follows. First, we implemented the new deep learning model described above and showed that it can extrapolate and predict molecular energies within a certain error range. In the process of implementing this model, we also succeeded in mathematically demonstrating that the conventional deep learning model is equivalent to the superposition of wave functions in quantum chemical calculations. Furthermore, the new deep learning model was trained on a simple small molecule, and the learned model was transferred to the prediction of properties of more complex polymers. A total of three papers, one for each of these research results, were published in international journals.

研究分野：機械学習

キーワード：深層学習　密度汎関数理論

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

材料開発や創薬の分野では、化合物のさまざまな物性値（物質のエネルギー、触媒の反応収率、発電材料の効率、薬剤の活性など）の計算・予測が必要不可欠である。その計算・予測には、量子物理学に基づく理論計算・シミュレーションが有用と認識されてきた一方で、膨大な計算コストという問題がある。これを解決するために近年、人工知能技術の一つである深層学習が用いられるようになってきた。しかしながら、理論計算・シミュレーションとは異なり、深層学習の計算の中身はブラックボックスなので、材料開発や創薬の現場で重要な解釈性・信頼性が低いことが大きな問題となっている。さらに人工知能技術は一般に、存在するデータから答えを導く内挿は得意だが、存在しないデータを推測して答えを導く外挿は不得意であり、性能が著しく悪化することが多い。例えば、物性値予測での外挿とは、学習用のデータと分子構造が大きく異なる化合物の物性予測などである。なお、内挿とはその逆で、分子構造がほぼ同じ化合物の物性予測である。外挿予測は新規の材料や薬剤の開発に極めて重要である。

研究開発当初の背景としては、深層学習が様々な研究分野とデータに適用される時代が到来しており、物理化学分野における材料データに対してもその波は来ていた。材料開発や創薬の分野では、化合物のさまざまな物性値の計算が必要不可欠であるが、深層学習技術を用いて物性値を予測することで、計算量を抑えられることが知られている。特に、分子のエネルギーを予測する深層学習モデルが登場し、第一原理計算あるいは量子化学計算と遜色ない精度を達成したことが大きな話題となっていた。一方でそれは、偏ったベンチマークデータにオーバーフィットした予測精度でしかないことが、多くの研究者から指摘され続けているのも現状であった。例えば、非常に簡単な分子（水素分子）や原子が一個分増えた分子などについては、学習データにない場合は予測精度が著しく悪化することが、他の論文で既に示されていた。このように、データと機械学習のみに頼るアプローチでは、このような致命的な予測の失敗が起き、これは第一原理計算あるいは量子化学計算では絶対に起きないことであり、問題となっていた。

２．研究の目的

上記の研究背景を踏まえて、未知分子に対しても適切に外挿予測できるような深層学習モデルの開発を目的とした。この目的を達成することで、限られたデータからでも広範囲に実用可能な深層学習アプリケーションを構築することができる。またこれは、物理化学と機械学習という二つの分野を適切に融合するという学術的な意義もあり、これも目的とした。分子データを扱う際には、どちらかの分野の理論やアプローチのみに偏ることなく、それぞれの分野の良い部分をうまくミックスさせることが必要不可欠であり、それを達成することを目的とする。

３．研究の方法

上記の目的を達成するための方法として、データや深層学習に過度に頼らない、物理化学に基づくアプローチを採用した。特に、第一原理計算あるいは量子化学における電子の軌道（波動関数）や密度の概念を踏襲し、分子のエネルギーを予測する新たな深層学習モデルを実装した。実装したモデルを、小さな分子で学習し大きな分子を予測する、つまり外挿予測で精度を評価した。今回開発した技術では、深層学習モデルの内部に、波動関数と電子密度という量子物理的に最も基本的な情報を顕わに表現することによって、現在深層学習で大きな問題となっている予測結果の解釈性・信頼性の問題を解決する。また、波動関数と電子密度という、データの偏りに影響されない普遍的な情報に基づくことで、学習データとは分子構造が大きく異なる未知化合物の物性を外挿予測できる。これによって、材料開発や創薬の分野における大規模な有用物質探索への貢献が期待される。今回開発した技術では、まず化合物Ｍの原子配置の情報を、理論計算・シミュレーションで用いられる原子の波動関数φに変換して、量子物理的に正しい計算の出発点を得る。次に、波動関数の重ね合わせの原理に従い、このφから分子の波動関数ψを計算する。そして、このψから物性値Ｅを学習する。加えて、分子の波動関数ψから得られる電子密度ρと原子配置から計算できるポテンシャルＶ　とが一対一対応するというホーヘンベルグ・コーンの定理を、モデル全体への物理制約として課す。これらはすべて、密度汎関数理論の枠組みに基づいていることが重要な点である。このモデルを、化合物の原子配置（入力）と物性値（出力）に関する大規模データベースを用いて学習させることで、波動関数と電子密度を経由した物性値の予測が可能になる。量子物理的に最も基本的な情報であるψやρを経由して物性値が導かれ

るため学習データの偏りに影響されない化合物の本質を捉えることができ、物性値の外挿予測が可能になる。具体的には、分子の波動関数ψから物性値Eの予測を行うニューラルネットワークと、電子密度ρにポテンシャルVの制約を課すニューラルネットワークという、二つのニューラルネットワークを交互に学習する。ψとEを繋ぐ関数と、ρとVを繋ぐ関数は、どちらも正確な形がわかっていない複雑な関数であり、これらを大規模データベースから学習する。

４．研究成果

理論計算で得られる値は実験で得られる値を1～2kcal/molの誤差で予測できる一方で、今回の技術はその理論計算値を1～3kcal/molの誤差で予測できる。つまり、実験値を2～5kcal/molの誤差で外挿予測できることになり、これは従来技術よりも高い精度であり充分実用に耐えうる精度と言える。さらに、理論計算は1種類の分子に数十分から数時間かかるが、今回の技術は数分で1万種類の分子を予測できる。このように、実用に耐えうる外挿精度を保ちながら理論計算を10万倍以上高速化した今回の技術は、新規の材料や薬を大規模に探索し効率的に発見・開発するという実応用では重要となると考えられる。

主な研究成果は、以下の三つである。まず、上記の新たな深層学習モデルを実装し、分子のエネルギーをある誤差の範囲内で外挿予測できることを示した。また、それを実装する過程で、従来の深層学習モデルが量子化学計算における波動関数の重ね合わせと等価な計算を行っていることを、数学的に示すことにも成功した。さらに、新たな深層学習モデルを簡単な低分子で学習し、その学習済みモデルをより複雑な高分子の物性予測へと転移させた。それぞれの研究成果について、一本ずつ合計三本の論文が国際学術誌に掲載された。また、この成果の学術的意義は、物理化学と機械学習という二つの分野を適切に融合できたことである。分子データを扱う際には、どちらかの分野の理論やアプローチのみに偏ることなく、それぞれの分野の良い部分をうまくミックスさせることが必要不可欠であり、それを達成することができた。また社会的意義は、分子データは製薬企業や材料企業のすべてが密接に関わるデータであり、その分野の研究者や技術者にとっての基礎技術を開発できたことである。学習モデルは、製薬企業や材料企業が独自に持つデータでも再学習可能であり、広く使われることが期待できる。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　1件／うちオープンアクセス　3件）

１．著者名 Masashi Tsubaki and Teruyasu Mizoguchi	４．巻 17
２．論文標題 Quantum deep descriptor: Physically informed transfer learning from small molecules to polymers	５．発行年 2021年
３．雑誌名 Journal of Chemical Theory and Computation	６．最初と最後の頁 7814-7821
掲載論文のDOI（デジタルオブジェクト識別子） 10.1021/acs.jctc.1c00568	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Masashi Tsubaki and Teruyasu Mizoguchi	４．巻 125
２．論文標題 Quantum Deep Field: Data-Driven Wave Function, Electron Density Generation, and Atomization Energy Prediction and Extrapolation with Machine Learning	５．発行年 2020年
３．雑誌名 Physical Review Letters	６．最初と最後の頁 206401
掲載論文のDOI（デジタルオブジェクト識別子） 10.1103/PhysRevLett.125.206401	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Masashi Tsubaki and Teruyasu Mizoguchi	４．巻 33
２．論文標題 On the equivalence of molecular graph convolution and molecular wave function with poor basis set	５．発行年 2020年
３．雑誌名 Advances in Neural Information Processing Systems	６．最初と最後の頁 1982--1993
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計4件（うち招待講演　4件／うち国際学会　0件）

１．発表者名 椿真史
２．発表標題 量子化学計算のための深層学習技術の基礎と応用
３．学会等名 顕微鏡計測インフォマティックス研究部会（招待講演）
４．発表年 2021年

１．発表者名 椿真史
２．発表標題 深層学習に基づく波動関数・電子構造の記述子表現と転移学習への応用依頼講演
３．学会等名 日本化学会 第101回春季大会（招待講演）
４．発表年 2021年

１．発表者名 椿真史
２．発表標題 深層学習に基づく波動関数・電子構造の記述子表現と転移学習への応用
３．学会等名 日本化学会春季年会 2021（招待講演）
４．発表年 2021年

１．発表者名 椿真史
２．発表標題 創薬と新材料開発のための人工知能
３．学会等名 情報処理学会全国大会 2021（招待講演）
４．発表年 2021年

〔図書〕　計0件

〔出願〕　計1件

産業財産権の名称	発明者	権利者
物性予測方法及び物性予測装置	椿真史	国立研究開発法人産業技術総合研究所
産業財産権の種類、番号	出願年	国内・外国の別
特許、2020-090714	2020年	国内

〔取得〕　計0件

〔その他〕

開発したソフトウエアの公開ページ https://github.com/masashitsubaki

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関