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 ２１ 日現在

機関番号：８２６２６
研究種目：基盤研究(B)（一般）
研究期間：2020～2022
課題番号：２０Ｈ０２７４７

研究課題名（和文）データ駆動に基づく記述子構築法と有機合成反応および触媒反応予測への展開

研究課題名（英文）Data-Driven Descriptor Generation and its Application to Prediction of Organic Synthetic and Catalytic Reactions

研究代表者
矢田　陽（Yada, Akira）

国立研究開発法人産業技術総合研究所・材料・化学領域・研究チーム長

研究者番号：７０６１９９６５

交付決定額（研究期間全体）：（直接経費）　　13,700,000円

研究成果の概要（和文）：本研究では、有機化学や触媒化学分野において、少数データに対して予測性能の高い機械学習モデルを構築するための新しい方法論の構築に取り組んだ。具体的には、まず、有機化合物の情報（入力）と量子化学計算値（出力）とを相関づけて、グラフニューラルネットワーク（GNN）による機械学習モデルを構築した。この機械学習モデルの最終中間層を取り出して触媒反応の収率や選択性などの予測モデルの入力として活用できることを明らかにした。また、本手法に活用可能な有機化合物の独自データベースの構築に取り組み、約6000の有機リン化合物のデータベースの構築に成功した。

研究成果の学術的意義や社会的意義
本研究成果は、コストや時間が掛かる実験を実施しなければデータが収集できない実験科学において、少数のデータでも予測性能の高い機械学習モデルを構築するための方法論を提供するものである。大量のデータが収集できない分野は多岐に渡ると考えられ、本研究技術はそのような分野にも波及して幅広く応用されることが期待され、学術的および社会的意義は大きいと考えている。また、有機合成化学・触媒化学分野においても、新たなAIの活用方法を提供することで、新しい原理・原則の発見と化学の進歩につながると同時に、本研究代表者が目指す触媒の自動発見に向けて大きく前進すると期待している。

研究成果の概要（英文）：In this research project, the development of a new methodology to build a machine learning model for predicting organic reactions and catalytic reactions with high predictive performance for a small number of data is conducted. Thus, a machine learning model using a graph neural network (GNN) which correlates the information of organic compounds (input) and the calculated values of quantum chemistry (output) is constructed. It is clarified that the final layer of this machine learning model can be extracted and used as an input for predictive models such as the reaction yield and selectivity of catalytic reactions. In addition, the construction of a unique database of organic compounds that can be used in this methodology is performed and finally succeeded in constructing a database which includes approximately 6,000 organophosphorus compounds.

研究分野：有機合成化学

キーワード：機械学習　転移学習　有機合成　触媒

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

人工知能（AI）は人間の仕事を代替して、さまざまな作業を自動で全て行えるようになることが期待されている。われわれの身近なところ、例えば自動車の自動運転や掃除ロボット、オペレーション業務、不正検知、検索エンジンなどでは、AI の導入・活用が進んで、その有用性が世間に認められつつある。一方、化学においては、AI の活用に向けた試行錯誤が始まったばかりである。化学において AI の活用が進まない理由の一つに、過去の知見に基づいた仮説提唱とその実証を繰り返しながら新しい価値を発見するという、化学特有の高度な知的作業を AI で代替することが困難であることが挙げられる。また、斬新なアイデアや偶然、セレンディピティー的な発見などを AI で行うことも極めて難しい。むしろ化学における AI は、人間の作業を代替するのではなく、人間の思考を補完するために活用することが期待される。すなわち、知的作業を実施する主体は人間であり、AI は人間には想像できなかった仮説を提唱する役割を果たさせる。今までにはない仮説が人間にさまざまな思考・考察を促して、人間の創造性を増強させることにつながり、その結果、新しい原理・原則の発見と化学の進歩をもたらすと期待できる。

また、AI を活用するためには一般的には大量のデータが必要であるが、実験化学においてはコストや時間が掛かる実験を実施しなければデータが収集できない。したがって、データの収集が困難な実験化学において AI をいかにうまく活用していくかは重要な課題である。特に、触媒開発や反応開発等の新しい分子の創成が要求される分野では、少数のデータに対していかに予測性能の高いモデルを構築できるかが極めて重要となる。

２．研究の目的

本研究は、有機化学や触媒化学分野において、少数データに対して予測性能の高い機械学習モデルを構築するための新しい方法論の構築することを目的とする。本研究により、新しい原理・原則の発見と化学の進歩につながると同時に、触媒の自動発見に向けて大きく前進させることを目標とする。

３．研究の方法

本研究では、以下の課題に取り組んだ。

（１）転移学習を活用した分子パラメーター生成のための方法論の構築

本研究課題の基盤となる技術である、転移学習による分子パラメーター生成のための方法論の構築に取り組む。転移学習の基となる有機化合物のデータセットを準備し、有機化合物の情報（入力）と量子化学計算値（出力）とを相関づけて、グラフニューラルネットワーク（GNN）と呼ばれる手法でニューラルネットワークモデルを構築する。生成されたニューラルネットワークモデルの最終中間層を取り出して、次の（２）における機械学習の入力（分子のパラメーター）として活用する。

（２）有機合成および触媒反応の収率・選択性の予測モデルの構築と反応基質や触媒の設計技術の開発

（１）のニューラルネットワークモデルの最終中間層から取り出されたパラメーターを入力、触媒反応の収率や選択性などの反応評価を出力として、触媒反応の機械学習モデルの構築に取り組み、さらに構築したモデルの予測性能の検証を実施する。機械学習法は、申請者が最近開発した触媒反応の収率予測モデルの構築に活用した LASSO を中心に検討する（*Chem. Lett.* **2018**, *47*, 284–287）。LASSO を用いると重要なパラメーターが自動で客観的に選択され、さらにそれらの重要度も可視化できる。そのため、モデルの解釈が容易になると同時に、新しい分子の設計指針が得られる利点がある。

（３）データセット拡張による独自の有機分子データベースの構築

転移学習を活用して生成できる分子パラメーターは、準備した有機化合物データセットに含まれる構成元素に限定されてしまう。有機化合物は主に炭素（C）、水素（H）、酸素（O）、窒素（N）原子で構成されるが、それ以外にもハロゲン（F、Cl、Br、I）、リン（P）、硫黄（S）、ホウ素（B）、ケイ素（Si）など、さまざまな元素によって構成される。さまざまな有機化合に対応するために、データセットを拡充して独自データベースを構築することを試みる。

４．研究成果

まず、転移学習による分子パラメーター生成のための方法論の構築に取り組んだ。具体的には、C 、H、O、N、F の 5 つの原子で構成される重原子数（C, O, N, F）が 9 以下の約 14 万個の有機分子について、量子化学計算によって最適化された 3 次元構造と、内部エネルギーや HOMO/LUMO レベルなどの 12 種類 の量子化学計算値が格納されている QM9 データセット（*Scientific Data* **2014**, *1*, 140022.）を用いて、入力（分子の 3 次元構造と原子種）と出力（量子化学計算値）とを相関づけたニューラルネットワークモデルを構築した。例えば、有機化合物

の内部エネルギーのニューラルネットモデルを GNN によって構築し、そのモデルによって QM9 データセットに含まれない分子の内部エネルギーを高精度で予測できることが確認できた（図 1）。同様に、他の量子化学計算値を予測するニューラルネットモデルの構築にも可能であることが明らかとなった。

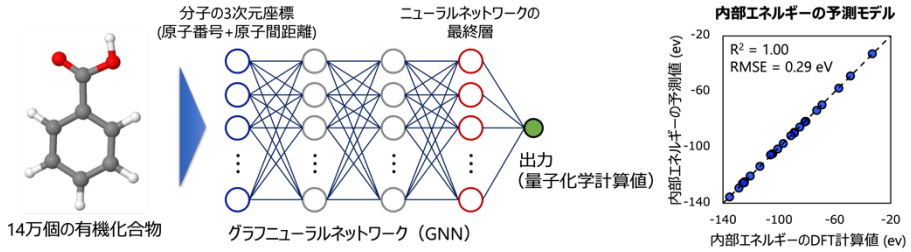

図 1. ニューラルネットワークモデルによる有機化合物の内部エネルギーの予測

そこで次に、上記の方法で構築したニューラルネットワークモデルから最終中間層を取り出して、それを触媒反応予測モデル構築におけるパラメーターとした。すなわち、この新規パラメーターを機械学習の入力、触媒反応の収率や選択性などを出力として機械学習させ、機械学習モデルの構築を行なった（図 2）。触媒反応としては論文ですでに報告されているものを対象として、本手法に実施した。その結果、既存の分子パラメーター生成法と比較して本パラメーターが収率予測モデルの構築に有用であることを確認できた。

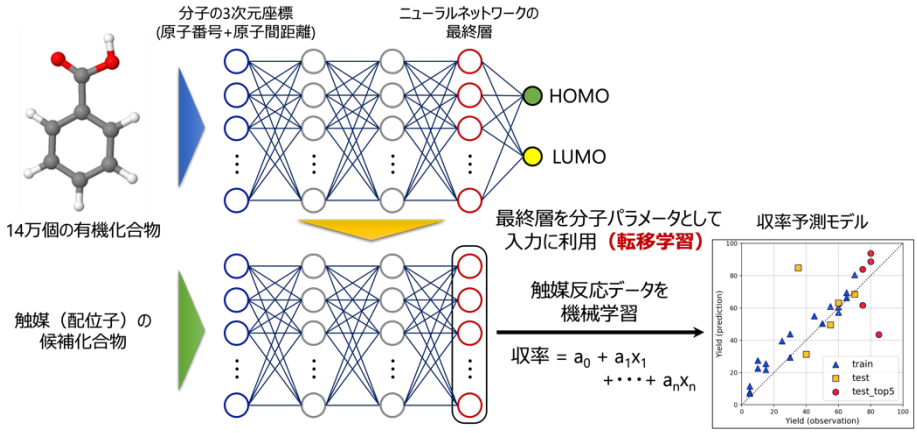

図 2. 転移学習による分子パラメータ生成と触媒反応の収率予測

さらに、転移学習に利用できる有機化合物の拡張を目指して、QM9 データセットに含まれない元素種を有する有機分子の量子化学計算を実施し、独自の有機分子データベースの構築を行なった。具体的には 3 価のリン原子を含む有機化合物で市販されているものを調査し、それらのリスト（約 1000 化合物）を作成した。また、全ての 3 価のリン原子に酸素 2 重結合または硫黄二重結合を付与して、5 価のリン化合物を人工的に発生させた。これら 3 価リン化合物、5 価リン化合物の約 3000 化合物全てについて配座探索と量子化学計算（構造最適化と振動解析）を実施した。一方、本研究を実施している途中に、3 価の有機リン化合物類似するデータベース構築に関する論文が発表された（*J. Am. Chem. Soc.* **2022**, *144*, 1205–1217）。そこで、この論文で報告されている有機リン化合物で、すでにリスト化していたものと重複しない有機リン化合物を追加して、データベースの拡張を試みた。以上のような結果、約 6000 の有機リン化合物についての独自データベースの構築が完了した。今後も上記の手法に基づいたデータベース拡張を継続していく予定である。また、構築したデータベースは転移学習のデータセットとして利用可能であり、さまざまな有機合成反応や触媒反応の収率や選択性等の予測にも応用していく予定である。

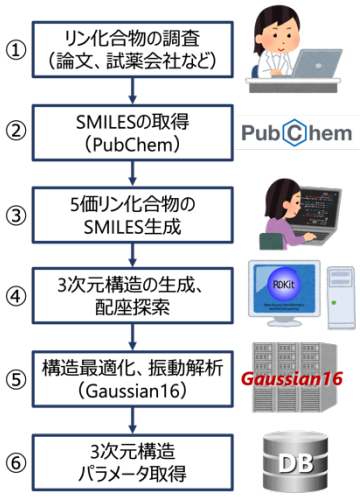

図 3. データベース構築の流れ

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Yada Akira、Sato Kazuhiko、Matsumura Tarojiro、Ando Yasunobu、Nagata Kenji、Ichinoseki Sakina	４．巻 -
２．論文標題 Ensemble Learning Approach with LASSO for Predicting Catalytic Reaction Rates	５．発行年 2020年
３．雑誌名 Synlett	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1055/a-1304-4878	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Kon Yoshihiro、Nakashima Takuya、Yada Akira、Fujitani Tadahiro、Onozawa Shun-ya、Kobayashi Shu、Sato Kazuhiko	４．巻 19
２．論文標題 Pt-Catalyzed selective oxidation of alcohols to aldehydes with hydrogen peroxide using continuous flow reactors	５．発行年 2021年
３．雑誌名 Organic & Biomolecular Chemistry	６．最初と最後の頁 1115～1121
掲載論文のDOI（デジタルオブジェクト識別子） 10.1039/d0ob02213f	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Kiyohara Shin、Tsubaki Masashi、Mizoguchi Teruyasu	４．巻 6
２．論文標題 Learning excited states from ground states by using an artificial neural network	５．発行年 2020年
３．雑誌名 npj Computational Materials	６．最初と最後の頁 1-6
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41524-020-0336-3	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 Tsubaki Masashi、Mizoguchi Teruyasu	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 76402
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、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 33 (NeurIPS 2020)	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 オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計19件（うち招待講演　4件／うち国際学会　3件）

1．発表者名 矢田陽
2．発表標題 キャタリストインフォマティクスによる触媒設計
3．学会等名 第16回ケムステVシンポ「マテリアルズインフォマティクス」
4．発表年 2021年

1．発表者名 矢田陽
2．発表標題 キャタリストインフォマティクス：触媒×人工知能による触媒設計
3．学会等名 有機合成夏期セミナー「明日の有機合成化学」
4．発表年 2021年

1．発表者名 矢田陽
2．発表標題 キャタリストインフォマティクス -触媒化学と機械学習の融合-
3．学会等名 接着・接合技術コンソーシアム 第1回データ駆動WG
4．発表年 2021年

１．発表者名
矢田陽

２．発表標題
キャタリストインフォマティクスによる触媒反応の収率予測

３．学会等名
化学反応経路探索のニューフロンティア2021（招待講演）

４．発表年
2021年

１．発表者名
Akira Yada

２．発表標題
Machine learning approach for prediction of reaction yield-aiming at the discovery of innovative catalysts

３．学会等名
The 2021 International Chemical Congress of Pacific Basin Societies (Pacifichem 2021)（国際学会）

４．発表年
2021年

１．発表者名
Akira Yada

２．発表標題
Materials and Process Informatics in Organic Synthesis and Catalyst

３．学会等名
The Material Research Meeting 2021 (MRM2021)（国際学会）

４．発表年
2021年

１．発表者名
矢田陽

２．発表標題
合成化学者が合成化学にAIを活用するためには?

３．学会等名
日本化学会 第102春季年会(2022)

４．発表年
2022年

１．発表者名
矢田　陽

２．発表標題
キャタリストインフォマティクスによる触媒反応の収率予測

３．学会等名
化学工学会第51回秋季大会（招待講演）

４．発表年
2020年

１．発表者名
矢田　陽

２．発表標題
キャタリストインフォマティクスによる触媒活性予測

３．学会等名
第14回触媒劣化セミナー

４．発表年
2021年

１．発表者名
矢田　陽

２．発表標題
分子機能予測のための人工知能技術

３．学会等名
第４回 食・触コンソーシアム シンポジウム

４．発表年
2021年

１．発表者名
矢田　陽

２．発表標題
キャタリストインフォマティクスによる触媒活性予測

３．学会等名
日本化学会「R&D懇話会214回」AI を活用した研究開発の現状と展望～超超PJ における研究事例～

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 第101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

１．発表者名
矢田　陽

２．発表標題
データ駆動型触媒設計・開発の現状と今後

３．学会等名
第2回錯体化学会フロンティアセミナー

４．発表年
2022年

１．発表者名
矢田　陽

２．発表標題
キャタリストインフォマティクス:触媒開発から医薬品合成への貢献を目指して

３．学会等名
筑波大学 トランスボーダー医学研究センター 情報医学研究部門 第四回講演会

４．発表年
2022年

１．発表者名
Akira Yada

２．発表標題
Machine Learning Prediction of Catalytic Activity toward Rapid Design of Innovative Catalyst

３．学会等名
22nd Tateshina Conference on Organic Chemistry（国際学会）

４．発表年
2022年

１．発表者名
矢田　陽

２．発表標題
触媒インフォマティクス

３．学会等名
最近の化学工学講習会71 「カーボンニュートラルに貢献する触媒・反応工学」

４．発表年
2023年

１．発表者名
矢田　陽

２．発表標題
データ駆動型触媒化学

３．学会等名
第6回 SPIRITS生物-無機-有機融合化学セミナー

４．発表年
2023年

１．発表者名
矢田　陽

２．発表標題
機械学習による触媒反応最適化の基礎

３．学会等名
日本化学会 第103春季年会(2023)

４．発表年
2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

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
研究分担者	椿　真史 (Tsubaki Masashi) (80803874)	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 (82626)	

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