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　４ 年　 ６ 月 １６ 日現在

機関番号：３２６８９
研究種目：研究活動スタート支援
研究期間：2020～2021
課題番号：２０Ｋ２２５３９

研究課題名（和文）機械学習を用いた分子構造探索手法と自動的なパラメータ構築手法の開発

研究課題名（英文）Development of Molecular Structure Search Method and Automated Parameter Construction Scheme Based on Machine Learning

研究代表者
藤波　美起登（Fujinami, Mikito）

早稲田大学・理工学術院・助教

研究者番号：５０８７５３９１

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：本研究では、分子構造の迅速な探索のために、分子中の原子に働くフォースを機械学習を用いて高速に予測する手法の開発に取り組んだ。有機分子の構造最適化および有機金属錯体反応に関するフォースのデータベースを構築し、種々の機械学習手法を適用することでその予測精度を検証した。フォースの予測に必要なデータベース、記述子および機械学習手法に関する知見が得られた。また、構築したデータベースは本研究に限らず計算化学研究に有用な情報を内包する。

研究成果の学術的意義や社会的意義
本研究は、機械学習を用いたポテンシャルの予測において今日広く用いらている手法と異なり、原子のフォースを直接予測する点で特異である。これに有効な記述子や機械学習手法を検証した点は学術的な意義がある。また、本手法の精度をさらに向上させることで分子構造の迅速な探索が可能となれば、新規化合物の設計など、計算化学分野で広く行われている研究課題に対して貢献することも期待される。

研究成果の概要（英文）：In computational chemistry, the fast prediction of atom forces in molecules is essential in the rapid exploration of molecular structures, including chemical reactions. This study aimed to develop a fast prediction method of atom forces using machine learning. In this study, a database of atom forces related to geometry optimization of organic molecules and chemical reactions of the organometallic complex was constructed. The prediction accuracy was assessed by applying various machine learning methods. The knowledge about databases, descriptors, and machine learning methods for predicting atom forces was obtained. The constructed database contains information about a large number of non-equilibrium molecular structures. It is usable for a wide range of computational chemistry research.

研究分野：理論化学、ケモインフォマティクス

キーワード：量子化学計算　機械学習　構造最適化

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

量子化学計算による分子の平衡構造・遷移状態・反応経路の構造探索と各構造のエネルギー評価は、実験化学者にも広く用いられ、化学研究において重要な役割を果たしている。構造探索においては、エネルギーの各原子に対する核座標微分（フォース）を計算するが、これはエネルギー計算に比べてコストが高い。フォースの計算コストを削減することができれば、構造探索の高速化につながり、より大きな分子の計算が平易となる。

量子化学計算は、本来経験的なパラメータを必要としないが、計算コストを低減するために多くの手法で経験的なパラメータを用いている。任意の系を高精度に計算するパラメータの構築は事実上困難であり、ユーザーが計算対象とする系に最適化されたパラメータを用いる必要がある。ユーザーの計算対象に適したパラメータが存在しない場合、計算精度は悪化する。また、パラメータの決定は、それ自体が計算化学において困難な課題のひとつである。これは熟練の計算化学者の経験に基づいて行われており、しばしば職人技とも形容される。これらの状況は量子化学計算の適用範囲を限定している。

機械学習は、2010 年代に飛躍的に深化した、多数のデータ中に存在する法則を自動的に関数として定式化する技術である。十分な量のデータに対して機械学習を適用すれば、与えられたデータの範囲において最適なパラメータを有する任意の関数を構築することができる。すなわち、分子の構造の多様性を考慮したデータを作成し、フォースの予測に機械学習を用いれば、最適化されたパラメータの自動構築が可能と考えられる。計算化学の観点では、これまでに汎用的に利用可能なパラメータの構築や、特定の系で精度の高い計算を目指すパラメータの構築は行われてきたが、ユーザーが望む系に対してパラメータの自動的な構築が可能となれば、広い分野のユーザーに貢献することができる。理論化学の観点からは、パラメータ構築に機械学習を用い、職人技の世界から脱却することができれば、量子化学計算の適用範囲を大きく拡張することが期待される。

２．研究の目的

本研究は、分子の構造探索を高速に行うために、与えられた分子構造に対して、原子に働くフォースを機械学習により高速に予測する手法を開発するのが目的である。分子の構造探索に用いるフォースを学習するために、適当な分子構造およびそれに対応したフォースのデータベースを構築する。本研究では、計算化学のターゲットとして広く研究されている有機分子の構造最適化および有機金属錯体反応の過程に対するフォースのデータベースを構築する。 次に、原子のフォースを予測するための原子の環境の表現方法と機械学習手法について検証する。特に、近年の機械学習に基づくポテンシャルの開発に見られる方法とは異なり、フォースの値そのものを予測する手法の開発を試みる。構築したデータベースに含まれる有機分子および有機金属錯体反応におけるフォースについて、その予測精度を評価する。

３．研究の方法

機械学習の元となる、分子の構造とその構造における原子のフォースを含むデータベースを構築した。フォースの参照値は密度汎関数理論計算によって得た。多様な有機分子を含むデータベースとして、9 原子以下の炭素、窒素、酸素、フッ素を骨格元素として含む任意の有機分子の量子化学計算結果を内包した QM9 データベースが挙げられる。ただし、この計算結果は最安定構造のみを含むもので、分子中の原子のフォースが存在しない。そのため、QM9 データベースの分子に対して、適当な初期構造から出発した構造最適化を実行し、この過程で生じるフォースの計算結果をデータベース化した。構造の多様性をさらに持たせるために、異なる配座を初期構造とする構造最適化の過程も含めた。構築したデータベースの情報を機械学習に適用して問題がないか、データベースの性質を調査した。有機金属錯体反応については、酸化的付加反応の過程におけるフォースの予測を検証した。そのために、遷移状態計算の過程における分子の構造とフォースの情報を収集した。

与えられた分子構造において各原子が置かれた環境を表現する記述子を算出した。記述子には weighted atom-centered symmetry function (wACSF) を用いた。フォースの予測に直接的な情報を与えると期待される wACSF の微分の値を記述子に加えた。原子間の多体の効果を記述子に含めるため、フォースを予測する原子の近傍の原子の環境も記述子とした。フォースおよび wACSF の微分には軸に対する任意性があるため、正負および回転に対する座標依存性を検証した。フォースと並行な微分値と鉛直方向の微分値の寄与を検証した。さらに、値のスケーリングについても検討した。

次に機械学習を用いたフォースの予測手法を検討した。機械学習手法には、勾配ブースティング決定木、ニューラルネットワークを主として、種々の機械学習手法を検討した。ニューラルネットワークについては、全結合ニューラルネットワークの他にも異なる構造を用いた。特に、原子に働くフォースは局所的な環境によって決定される性質を考慮して、畳み込みニューラルネットワークに類似した部分的に結合された構造も検証した。さらに、異なる機械学習手法の予測結果を用いたアンサンブル学習についても検証した。

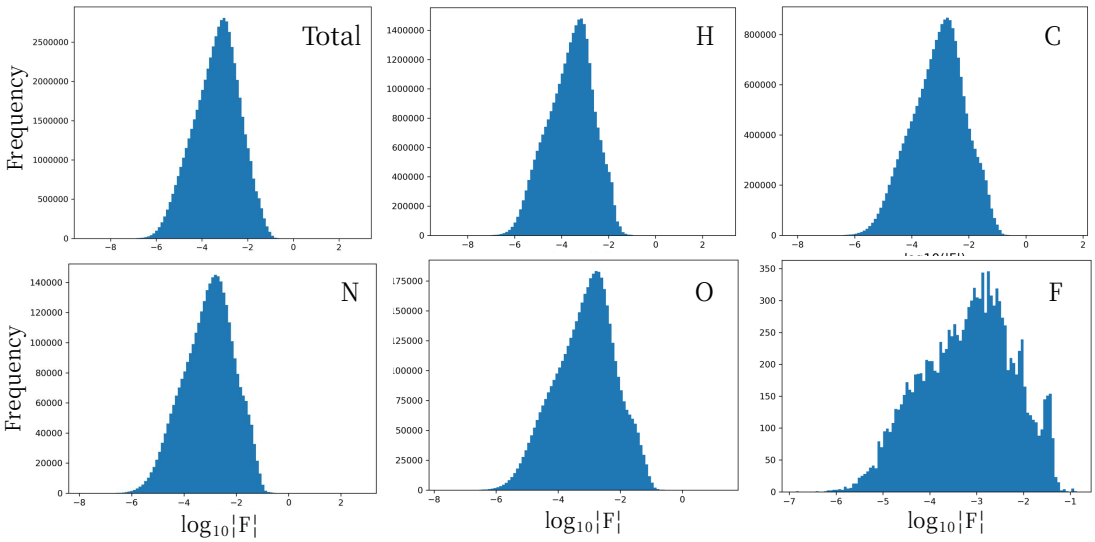

図 1．構築したデータベース中のフォースの分布

4．研究成果

有機分子の構造とそれに対応するフォースに関する大規模なデータベースを構築した。表 1 にデータベースが内包する分子、配座、構造、原子の数を示す。11 万を超える分子種から、27 万種を超える配座を生成し、500 万の構造、1 億超の原子数についてフォースの情報を収集した。それぞれのフォースの値のヒストグラムを図 1 に示す。横軸はフォースの絶対値に対して常用対数を取った値、縦軸はその頻度である。各元素の値と全ての元素の結果をまとめたヒストグラムを示している。概ねひとつのピークからなるフォースが収集できており、機械学習に応用するのに適したデータセットが構築できた。原子数が相対的に少ないフッ素において分布にばらつきが見られるが、他元素の結果よりデータ数の増加によって解消されると考えられる。

有機分子のデータベースに対するフォースの予測精度を表 2 に示す。窒素、酸素、フッ素原子に対して 0.2 mhartree/bohr 以下の精度でフォースを予測した。さらにフッ素原子に対して、フォースの局所性を利用したニューラルネットワーク構造や、異なる機械学習手法から得られた学習器によるアンサンブル学習の導入など、機械学習手法を精査したところ、平均絶対誤差 0.079 mhartree/bohr，決定係数 0.94 とさらに予測精度を改善した。

有機金属錯体反応の過程における炭素原子に対するフォースの予測結果を図 2 に示す。横軸は参照値となるフォースの値、縦軸は機械学習によるフォースの予測値である。予測値は参照値の傾向を再現し、決定係数は 0.9988 と高い値を示した。また、フォースの平均絶対誤差も 0.017 mhartree/bohr であった。

本研究で得られた結果は、本手法のさらなる改善により実用的に、ユーザー自身が所望の系に対する計算を実行するために必要なパラメータを構築できる可能性を示唆する。今後は継続してデータベースの拡充を行うとともに、励起状態における分子構造探索への応用を目指して分子の励起状態に関するデータベースの構築も進める。

表 1．データベースの内容

分子種の数	118,478
配座の数	275,075
構造の数	5,594,672
総原子数	113,895,412
H原子の数	64,810,396
C原子の数	34,832,055
N原子の数	5,731,549
O原子の数	8,427,044
F原子の数	94,368

表 2．有機分子中の原子に対するフォースの予測精度

元素	平均絶対誤差 mhartree/bohr	決定係数
N	0.20	0.80
O	0.14	0.88
F	0.10	0.89

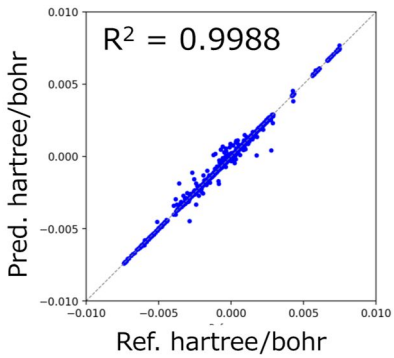

図 2．有機金属錯体反応における炭素原子に対するフォースの予測精度

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　0件）

〔学会発表〕　計5件（うち招待講演　2件／うち国際学会　2件）

１．発表者名 藤波美起登
２．発表標題 機械学習の基礎と実践のためのヒント
３．学会等名 第11回量子化学スクール（招待講演）
４．発表年 2021年

１．発表者名 藤波美起登
２．発表標題 運動エネルギー汎関数の開発、反応予測、反応条件最適化に対する量子化学計算と機械学習の応用
３．学会等名 計算科学研究センター・ナノテクノロジープラットフォーム事業合同ワークショップ（招待講演）
４．発表年 2021年

〔図書〕　計2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

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