実験・理論・計算が融合したケム・インフォマティクス研究

研究代表者　清野　淳司
（先進理工学研究科 化学・生命化学専攻 准教授（任期付））

１．研究課題

理論・計算・実験化学と機械学習をはじめとするインフォマティクスの融合を図ることにより、計算化学と実験化学の間のギャップを埋め、計算化学の社会実装を加速させる。さらに、これまでの研究手法では困難であった化学分野の諸問題を解決へと導き、新たな研究領域ケム・インフォマティクスを定着させる。2021 年度は以下の 2 つのテーマに関して研究を行った。(1) NMR スペクトルの自動構造決定・同定システムの開発、(2) シンボリック回帰に基づく化学原理自動抽出手法の開発。

２．主な研究成果

２－１　NMR スペクトルの自動構造決定・同定システムの開発

核磁気共鳴（NMR）スペクトルは、有機化合物、高分子材料、生体物質などの化合物同定や分子構造決定のなどのために有用な情報を与える。NMR スペクトルのピークは対象となる原子核周辺の環境に直接的に影響されるため、特に低分子有機化合物において、経験則や簡便なアルゴリズムから確度の高い化合物同定が可能である。一方、周期表の下方にある高周期元素を包含した化合物や、タンパク質などの生体分子においては、データが不足している、またはスペクトルが複雑であることから、ピークの経験的な帰属は低分子有機化合物ほど容易ではない。さらに近年の自律的に化学実験を行うロボティクスの発展により、人の手を介さないスペクトル自動同定技術の開発は不可避である。

本研究では、NMR スペクトルの自動構造決定・同定システムを、量子化学計算とインフォマティクスの融合により確立することを目指す。このシステムは、次の二つの手法から構成される。一つ目は NMR スペクトルから、インフォマティクスにおける画像認識技術を応用して情報を抽出し、化合物候補をリストアップする手法である。しかし、ここで予測される化合物は複数提案される可能性がある。そこで候補化合物を絞り込むために、二つ目では反対に、量子化学計算データを利用して、化合物の構造から NMR スペクトルの各ピークを高精度に予測する手法を開発する。また予測手法の開発過程において、高精度にスペクトルを再現するために重要な構造的・電子的情報の選定や、それぞれの因子の寄与度を評価する。ここで得られた知見は既存の経験則とは異なる視点からの情報を与える可能性を有するため、任意の NMR スペクトルに対する新たな指導原理の構築を目指す。具体的には次の 3 つを実施する。1. 量子化学計算に基づく NMR スペクトルピークおよびスペクトルを表現するための電子的情報のデータベース化、2. スペクトルピークの予測モデルの開発、3. 複数の原子核に対する NMR スペクトル画像から化合物の候補を与える手法の開発。

2021 年度は 1.と 2.を中心に研究を行った。1.において、C, O, N, F, H 元素の内の数個の組み合わせから構成される QM9 データセット内の約 13 万の化合物に対して、B3LYP/6-31G**レベルの量子化学計算を行うことにより、約 235 万個の H 原子に対する核磁気遮蔽定数データベースを構築した。

2.において、分子の構造的/電子的記述子を用いて、量子化学計算により得られた ^{1}H 原子に対する核磁気遮蔽定数の予測モデルを機械学習（XGBoost）により構築した．その結果、高精度に予測可能であること、核磁気遮蔽定数の局所性を考慮することで、小さな分子で構成されるデータセットに対する学習から、タンパク質といった大きな分子に対する予測が可能であることが示された（**図 1**）。

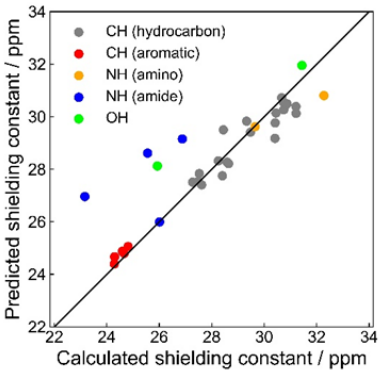

図 1. ペンタペプチドの ^{1}H 核磁気遮蔽定数に対する予測と実際のプロット

２－２　シンボリック回帰に基づく化学原理自動抽出手法の開発

今日の化学において、様々な観点からのビッグデータが生まれている。例えば、多くの実験・計算化学者が生成したデータを寄託し、集約したデータベースが作られている。また、ロボティクスと融合して膨大なデータを生成するハイスループット実験が行われている。一方、このビッグデータを利用してどのよう原理や知識を得るのかは、現在の化学における大きな課題の一つである。このような原理や知識を獲得するためには、どれくらいの明瞭さまたは精密さで化学モデルを構築するかが非常に重要である。近年多くの研究で用いられる機械学習は、予測能の高い化学モデルを構築できる。しかし、このようなモデルの多くは複雑な関数で表現されるため、人間が解釈することが困難である。そのため近年ではインフォマティクスの分野において、説明可能な技術の重要性に関する議論が進められている。

このような説明可能なインフォマティクス技術の一つとして、シンボリック回帰手法が近年注目されている。この手法として、四則演算、記述子/変数、定数パラメータに対する組み合わせ最適化手法である遺伝的プログラミング（GP）や、対称性・分割可能性・簡略化・次元解析を活用して、できるだけ簡易的な関数形を探索する AI-Feynman などが提案されている。本研究では、シンボリック回帰手法を化学の諸問題に適用できるようにカスタマイズすることで、膨大なデータから化学原理・法則を自動的に抽出するための基盤技術とすることを目指す。シンボリック回帰により構築された化学モデルは、学習したデータの範囲内での基本原理・法則を表す。さらに各化学モデルに対して類似性の観点から解析し分類する、または学習するデータ空間を拡大してより一般的な化学モデルを構築することで、より包括的な原理・法則を導くことができる。

2021 年度は、GP や機械学習などを利用した既存のシンボリック回帰（RLS, AI-Feynman）を、20 世紀前半以前に提案されたシンプルな化学における数理モデルに適用し、現状における表現能力を検証した（**表 1**）。この結果から、複雑な数理モデルを高速に構築するためには、より大きな関数空間の中から如何に効率的に最適な関数を探索するかが重要であることがわかった。

表 1. 化学法則の導出に要する計算時間（秒）

Level	式	GP	RLS	AI-Feynman
0	ヘンリーの法則	2.6	9.5	1243.8
1	ラウールの法則	2.6	7.5	701.8
2	2 次反応の反応速度式	×	18.1	3599.6
3	ファンデルワールス方程式	×	113.0	4729.0

３．　共同研究者

中野　匡彦（理工学術院総合研究所・招聘研究員）
速水　雅生（理工学術院総合研究所・嘱託）

４． 研究業績

4.1 学術論文

(1) “Database-assisted local unitary transformation method for two-electron integrals in two-component relativistic calculations”, C. Takashima, J. Seino, H. Nakai, *Chem. Phys. Lett.* **777**, 138691 (2021). **(Editor’s Choice)**, DOI: 10.1016/j.cplett.2021.138691.

4.2 総説・著書

(1) “人工知能技術と融合した量子化学理論”、清野淳司、*化学と教育*、ヘッドライン「*AI* *が開く新たな化学領域*」、**70**, 118 (2022).

4.3 招待講演

(1) “化学研究と人工知能技術の融合に関する基礎と応用事例”、清野淳司、CMC リサーチウェビナー、2021 年 8 月、オンライン

(2) “相対論的量子化学計算とインフォマティクスからみた化学反応”、清野淳司、シンポジウム「化学反応経路探索のニューフロンティア 2021」、2021 年 9 月、オンライン

(3) “AI-Assisted Orbital-Free Density Functional Theory Calculation”, J. Seino, M. Fujinami, Y. Ikabata, H. Nakai, The International Chemical Congress of Pacific Basin Societies 2021 (Pacifichem2021), December, 2021, Online.

4.4 学会および社会的活動

(1) 文部科学省科学研究費助成金 基盤研究（C） 「精度保証を考慮したオンライン機械学習型軌道非依存密度汎関数理論の開発」、清野淳司（研究代表、2021 年度－2023 年度）

(2) “理論化学会・将来構想委員会活動報告書”、藪下聡、鷹野景子、波田雅彦、内田雅人、庄司光男、清野淳司、中田彩子、村岡梓、横川大輔、*理論化学会誌フロンティア*、 **3**, 182 (2021).

(3) 理論化学会 将来構想委員会 委員、2020 年－2021 年

(4) 理論化学会 幹事、2021 年－現在

(5) 日本化学会関東支部 幹事、2022 年－現在

(6) 第 11 回量子化学スクール 世話人

５． 研究活動の課題と展望

本研究で実施してきた、NMR スペクトルの自動構造決定・同定システムを完成させ、近年の発展が目覚ましいロボティクスと融合させることで、人の手を介さない全自動の化合物合成技術が確立される。また、シンボリック回帰手法における関数の探索能力を向上することにより、将来的には、化学反応予測や実験条件最適化などの現実の問題への適用や、密度汎関数理論（DFT）における厳密に近く、かつ物理的に意味のある汎関数の構築も可能な技術となる。