

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。2022年度は以下の2つのテーマに関して研究を行った。(1) スペクトル情報と量子化学計算に基づく化合物の自動同定システムの開発、(2) 機械学習による周辺環境を考慮した分類と量子化学計算手法の精度推定。

2. 主な研究成果

2-1 スペクトル情報と量子化学計算に基づく化合物の自動同定システムの開発

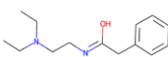
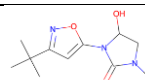
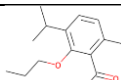
近年における自動実験の急速な発展に伴い、合成した化合物を自動的に同定できるシステムの開発が進んでいる。膨大な化合物空間の中で既知の化合物はごく僅かであり、既存の化合物を基軸とするデータベースから独立した、化合物の自動同定手法の開発は重要である。本研究では、各種スペクトルを利用した自動構造決定・同定システムを、量子化学計算とインフォマティクスの融合により確立することを目指す。具体的には次の3つを実施する。1. 量子化学計算に基づく NMR スペクトルピークおよびスペクトルを表現するための電子的情報のデータベース化、2. スペクトルピークの予測モデルの開発、3. 複数のスペクトル情報から化合物の候補を与える手法の開発。

2022年度は3を中心に研究を行った。本システムの流れを以下に記す。①赤外吸収スペクトルからフラグメントを抽出する。次に②フラグメント同士を組み合わせて分子を生成し、分子量が質量スペクトルの結果と一致する分子を候補として選択する。さらに③各候補分子の安定構造を探索し、④算出した化学シフト値を参照スペクトルの値と比較する。最後に⑤ピークの一貫性に基づき候補分子をランク付けし、分子の同定結果を出力する。本研究では特に、スペクトルの測定誤差や分析精度の影響を除外し、本システム全体の流れの妥当性を検証するため、分子の構造情報を実験スペクトルではなくデータベースから取得することで代替した。

表1に3種類の化合物(ID1~3)における、本システムによる各ステップの候補数と同定の可否を示す。各種スペクトルに対応する情報を用いることで、目的となる分子の構造を正しく予測できた。本システムを40種類以上の分子に適用し、同様の結果が得られることも確認した。

表1. 候補数と同定の可否の結果

ID	Fragment	# of generated molecule	# of candidate molecule	Correct?
1	5	21	6	Yes
2	3	131	2	Yes
3	6	491	2	Yes



ID1

ID2

ID3

2-2 機械学習による周辺環境を考慮した分類と量子化学計算手法の精度推定

密度汎関数理論 (DFT) の汎関数や機械学習ポテンシャルを利用する場合、所望の精度を満たす計算の可否の判断は参照データや対象分子に依存し、一般に簡単ではない。そのため、これまでに様々なデータセットを用いて計算精度が統計的に検証されてきた。このような検証において、官能基などの人間が解釈可能な基準を用いて計算精度を統一的に評価・推定できれば、計算精度の事前判断に加えて、既存の汎関数や機械学習ポテンシャルの改善へのヒントとなる。そこで本研究では、原子の周辺環境に基づいて、分子の構成原子を人間が解釈可能なレベルで複数のクラスターに自動的に分類した。そしてクラスター毎にエネルギーや分子物性の計算精度を推定することで、所望の分子の計算精度を事前に推定可能な手法を開発した。

2022 年度は有機分子を構成する炭素原子のみに着目した。対象原子は、データセット QM9 からランダムに選択した 5,000 分子を構成する 30,782 原子である。記述子は、周辺原子との距離・角度やクーロン相互作用を表す構造的記述子と DFT 計算に基づく原子の電子状態を表す電子的記述子から、特徴量削減により 4 個の記述子を用いた。クラスタリング手法として、ノイズを自動検出し、密度でクラスターを決める DBSCAN を用いた。その結果、ヘテロ原子の隣接の有無などの化学的性質を反映した 36 のクラスターを得た。

68 の汎関数における NMR 化学シフト計算の精度を、上記のクラスターに基づいて比較した。基底関数は 6-31G(d)を用い、参照値は MP2/cc-pVTZ により算出した。図 1 は平均平方二乗誤差 (RMSE) が小さい 10 種の汎関数による計算結果をクラスター毎に示す。この結果、同じ汎関数でもクラスター毎に誤差が異なることを確認した。この結果を基準として、計算対象に含まれる炭素原子の属するクラスターの計算精度が高い汎関数を選択することで、期待する精度を満たす手法を適切に選択した NMR 化学シフト計算が実行可能となる。

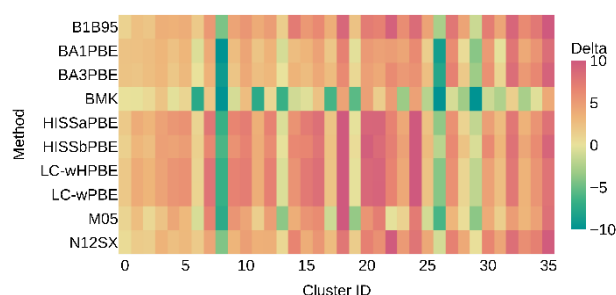


図 1. NMR 化学シフトにおける参照値からの誤差のヒートマップ

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)

中嶋 裕也 (理工学術院総合研究所・客員次席研究員)

速水 雅生 (理工学術院総合研究所・嘱託)

4. 研究業績

4.1 総説・フォーラム

- (1) H. Nakai, M. Kobayashi, T. Yoshikawa, J. Seino, Y. Ikabata, Y. Nishimura, Divide-and-Conquer Linear-Scaling Quantum Chemical Computations, *J. Phys. Chem. A*, **127**(3), 589 (2023). **(Feature Article)** DOI: 10.1021/acs.jpca.2c06965
- (2) 清野淳司, “理論化学と人工知能技術”, *IQCE-NEWS*, **106**, 1 (2022).

4.2 招待講演

- (1) “「実験化学・理論／計算化学・AI 技術」の融合による DX～ケム・インフォマティクスの基礎知識と化学法則の自動構築・材料探索・反応予測・自動解析技術・実験条件最適化への応用～”、清野淳司、情報機構オンラインセミナー、2022 年 4 月、オンライン
- (2) “ケム・インフォマティクスによる理論化学の DX”、清野淳司、第 24 回理論化学討論会、2022 年 5 月、金沢商工会議所
- (3) “マテリアルズインフォマティクス／化学原理構築から材料設計まで”、清野淳司、計算材料科学連続セミナー 化学材料第 1 シリーズ、2022 年 6 月、オンライン
- (4) “化学におけるマテリアルズインフォマティクスを実践するための基礎と応用事例”、清野淳司、サイエンス&テクノロジーオンラインセミナー、2022 年 7 月、オンライン
- (5) “量子化学計算と人工知能技術の融合”、清野淳司、スーパーコンピュータワークショップ 2022 「複雑電子状態の理論・計算科学」、2023 年 1 月、計算科学研究センター

4.3 受賞・表彰

- (1) 第 12 回 CSJ 化学フェスタ 2022、優秀ポスター発表賞、磯田拓哉、“シンボリック回帰に基づいた化学における明示的な初等関数/微分方程式の自動導出”

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 幹事、2021 年－現在
- (3) 日本化学会関東支部 幹事、2022 年－現在
- (4) 第 12 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、複数のスペクトル情報を用いた自動構造決定・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。特に未知分子の同定には量子化学計算の利用が重要であり、その精度を把握することが鍵を握る。そのため 2－2 の研究と融合させることで、あらゆる分子を同定可能なシステムの完成に大きく近づくことになる。