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。2023 年度は以下のテーマに関して研究を行った。(1) 機械学習と量子化学計算を利用した混合物スペクトル分解手法の開発、(2) 大規模言語モデルに基づく対話型情報検索システムの開発、(3) 機械学習を用いたヒトがん細胞に対する抗がん活性予測システムの開発。

２．主な研究成果

２－１　機械学習と量子化学計算を利用した混合物スペクトル分解手法の開発

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。我々は 2022 年度までに単一の化合物のスペクトルに対して、機械学習と量子化学計算を利用して化合物の自動同定を行うシステムを開発してきた。2023 年度は混合物スペクトルに対して適用できるように、機械学習と量子化化学計算を利用して、混合物スペクトルを個々の化合物の純成分スペクトルに分解する手法の開発を行った。

本システムは、学習段階と予測段階から構成される。学習段階では、①データベース内の化合物のスペクトル情報を量子化学計算により取得し、②そのスペクトルの組み合わせから仮想的な混合物スペクトルを生成する。次に③混合物スペクトルと純成分スペクトルを記述子として、混合物スペクトルを構成する化合物の有無を判断する識別器を構築する。予測段階では、④分解したい混合物スペクトルと純成分スペクトルの情報を記述子として、学習段階で構築した識別器を用いて化合物の有無を予測する。混合物スペクトルと純成分スペクトルが一致しているかを機械学習によって判断するため、学習フェーズで使用するスペクトルは実験や計算に関係なく使用することができる。

実験で測定された純成分スペクトルを組み合わせて、仮想的な混合物スペクトルを作成し、精度検証を行った。SDBS データベースから分子量 22 から 85 までの 300 個の有機化合物の ^{1}H NMR スペクトルを取得した。また学習フェーズと同様に、二種類の化合物からなる混合物スペクトルを生成した。テストしたデータ数は 89,000 個である。表 1 に予測における混同行列を示す。この結果、正解率が約 89%となり、当該化合物の存在を高精度で予測できることが確認された。

表 1. 該当する化合物が含まれるか否かの予測に関する混同行列

		Predicted	
		Positive	Negative
Actual	Positive	36,873	7,977
	Negative	2,213	42,637

２－２　大規模言語モデルに基づく対話型情報検索システムの開発

従来の紙あるいは電子実験ノートでは、情報管理や共有に関する限界が存在する。本研究では、最近の化学分野でも応用研究が盛んな大規模言語モデル（LLM）に基づき、人工知能時代の実験ノートのあり方を探求し、研究者の日々の業務を強力にサポートする役割にしたいと考える。特に本研究では、実験ノートから関連情報を検索し、その内容をもとに LLM を用いて対話的に文章生成するシステムを開発した。

本研究では､外部の知識ソースを参照することで LLM の回答の質を向上させる retrieval augmented generation（RAG）のフレームワークに基づき、システムを構築した。本システムでは予め、読み込ませた文書を embedding 変換し、 ベクトルデータベースに保存する。次にユーザーが入力した実験ノートに関する質問を embedding 変換し、データベースから類似文書を検索する。最終的に、データベース内の関連文書に基づき作成したプロンプトから、GPT-4 により回答を生成する。

本研究では、構築したシステムの性能検証のため、実験ノートのサンプルおよび化学物質に関する災害事例の文書をシステムに予め読み込ませた上で、関連する質問を入力し、出力された回答内容をマニュアルで評価した。生成した回答を評価するために、研究が先行する医療分野の特化型 LLM の評価基準を参考に､化学分野向けの RAG システムに関する評価基準を定義した。評価基準は、14 種の評価軸から構成され、3 つのグループ（質問への応答品質、情報ソースの活用精度、専門的な適応性）に大別できる。スコアリングには評価軸ごとの基準を定め、1~5 点の 5 段階で評価した。

評価基準に沿った質問の平均スコアを図 1 に示す。言葉遣いの丁寧さや簡潔性など言語的な観点での応答品質が高いこと、および記載されている実験情報と化学物質の安全性を問う基本的な質問に対して適切に回答可能であることが確認された。

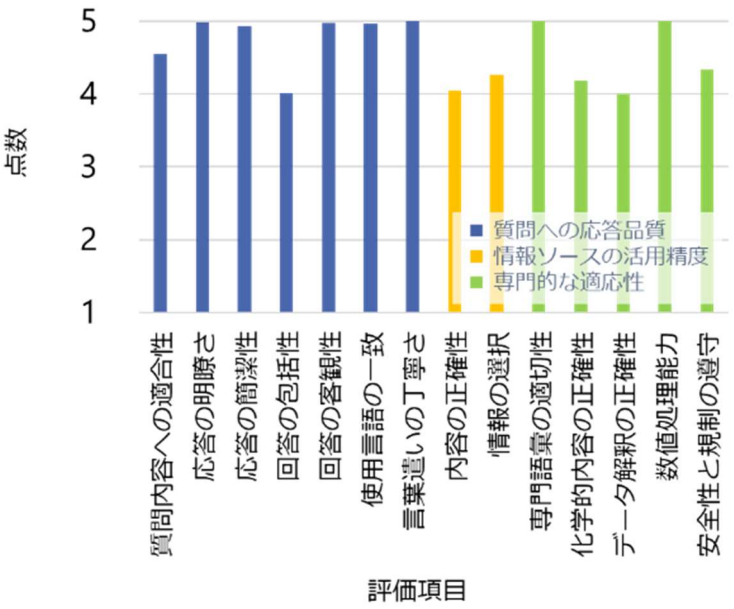

図 1. 14 種の評価軸における質問の平均点数

２－３　機械学習を用いたヒトがん細胞に対する抗がん活性予測システムの開発

創薬における天然化合物の探索研究では、新規化合物の発見だけでなく、既知の化合物について新しい生物活性を見出すことも同様に重要である。しかし、潜在的な生物活性の探索には実験による生物活性の評価が必要となるため、膨大な金銭的・時間的コストが不可避である。本研究では化学構造から生物活性を機械学習により予測し、既存の抗がん剤（基準化合物）との類似度判定により、基準化合物に類する抗がん活性を有する候補を提示するシステムを開発した。

アメリカ国立がん研究所で確立されたヒトがん細胞株 60 種に対する抗がん剤スクリーニング用パネル内の薬剤感受性データをデータベースとして活用した。本研究で予測する活性として、50%増殖阻害濃度値（GI50）を採用した。約 25,000 個の化合物を用いて各細胞株に対して予測モデルを構

築した。分子フィンガープリントや 2 次元/3 次元の構造を表現する 7,922 個の記述子を使用し、ランダムフォレスト回帰により予測した。

図 2 に細胞株 60 種に対する決定係数（R^2）を示す。ほぼすべての細胞株で R^2 が 0.8 程度となり、部位の違いによらず高い精度で予測できることが確認された。

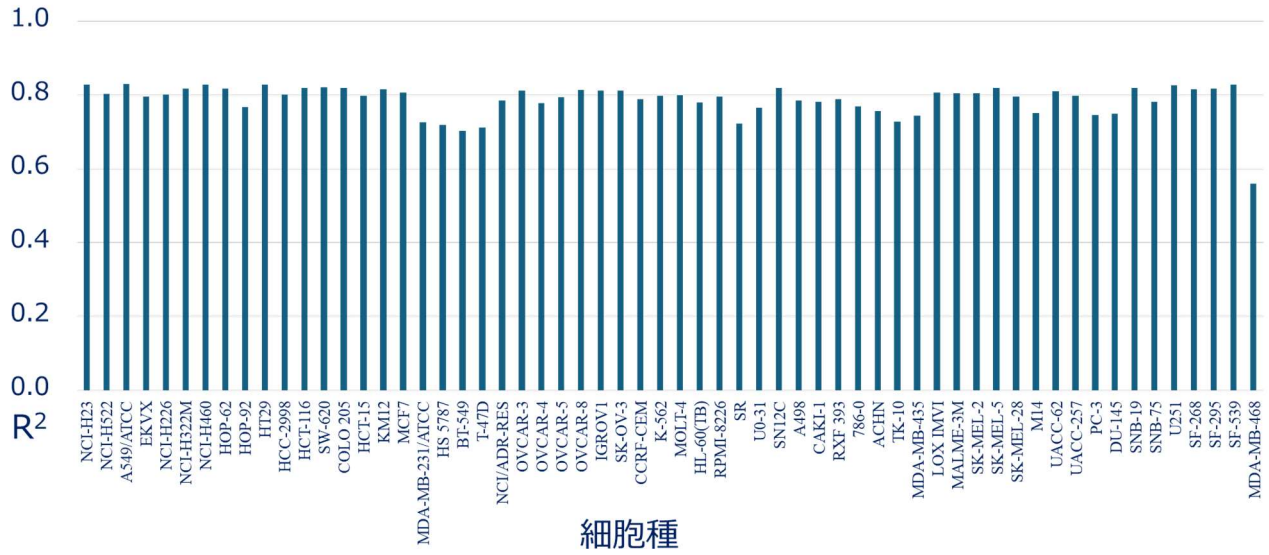

図 2. 60 種の細胞における予測精度

３． 共同研究者

中尾　洋一（先進理工学部・化学・生命化学科・教授）
町田　光史（先進理工学部・化学・生命化学科・講師（任期付））
神平　梨絵（先進理工学部・化学・生命化学科・助教）
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 学術論文

(1) T. Kumagai, Y. Nakajima, J. Seino, Automatic molecular identification system based on spectral information and quantum chemical calculation, *J. Comput. Chem. Jpn.* **22**, 12 (2023). DOI: 10.2477/jccj.2023-0029.

(2) T. Isoda, S. Takahashi, M. Nakano, Y. Nakajima, J. Seino, Validation of extrapolation in symbolic regression and its application to perovskite catalysts, *J. Comput. Chem. Jpn.* **22**, 37 (2023). DOI: 10.2477/jccj.2023-0028.

(3) Y. Nakajima, T. Ohmura, J. Seino, Using atomic clustering based on structural and electronic descriptors that consider surrounding environment to evaluate local properties of DFT functionals, *J. Comput. Chem.* in press. DOI: 10.1002/jcc.27375.

4.2 総説・フォーラム

(1) 清野淳司、“シンボリック回帰による化学データの解釈と外挿的な材料探索の可能性”、 フロンティア、受理

4.3 招待講演

(1) “ケム・インフォマティクスに基づく材料開発・化学反応速度の自動解析”、 清野淳司、 反応駆動学公開シンポジウム「原子・分子ダイナミクスと反応駆動」、2023 年 6 月、東京工業大学大岡山キャンパス

(2) “ケム・インフォマティクスの実験・計算・理論化学への展開”、 清野淳司、 セミナー「化学産業を改革する最新 DX（AI・IoT）技術」～製造、開発、研究、それぞれの工程における DX 取り組み事例～、2023 年 6 月、大阪科学技術センター

(3) “Construction of Orbital-Free DFT Scheme and Its Evaluation by ML”, J. Seino, The International Symposium on Machine Learning in Quantum Chemistry (SMLQC), November 2023, Uppsala (Sweden).

4.4 受賞・表彰

(1) 日本コンピュータ化学会 2023 年春季年会、奨学賞、磯田拓哉、“シンボリック回帰の化学の諸問題への適用：外挿性の検証と反応速度論への応用”

(2) 第 46 回ケモインフォマティクス討論会、優秀ポスター賞、熊谷拓海、“複数のスペクトル情報と量子化学計算を利用した機械学習による部分構造予測”

4.5 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－2023 年度

(4) 第 13 回量子化学スクール 世話人

(5) 講演会「マテリアルズインフォマティクスの最先端～化学産業への展開～」世話人

(6) イベント「君たちの将来と化学の未来～早大で過ごす化学な週末 2023～」世話人

５． 研究活動の課題と展望

本研究の２－１で実施してきた、化合物自動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。このためには量子化学計算と実験値のギャップを埋める手法の開発が不可欠である。２－２の対話型情報検索システムにおいては、実用化に向けて実際に実験室にて使用されている実験ノートへ適用する。２－３の抗がん活性予測システムでは、新規の抗がん剤を提案できるようにシステムを拡張し、実験における評価を通してその有用性を議論する予定である。