分子構造からのNMRスペクトル予測：トポロジー記述子とQSPRモデルの応用

関連製品： 核磁気共鳴装置(NMR)

本資料では、分子構造からNMRスペクトルを予測する「順解析」の手法として、QSPR（Quantitative Structure–Property Relationship）モデルの応用可能性を紹介します。
本手法は、JEOL製のNMRデータ解析ソフトウェアJASON[1]と、そのPython®[2]インターフェースであるBeautifulJASON[3]を活用することで、分子構造の記述子抽出からスペクトル予測までを一貫して実施可能です。
少数データ環境下でも、分子の電子的・構造的特徴を数値化し、ppm値の予測精度とモデルの解釈性を両立することを目指します。
本稿では、以下の2点を中心に検討を進めます：

- 少数データ環境における高精度なNMRスペクトル予測モデルの構築
- 記述子の寄与分析による説明性の高いモデル設計と、クラスタリングによる構造傾向の可視化

これらの取り組みは、NMR解析の自動化とマテリアルズ・インフォマティクス（MI）の融合を通じて、分子探索・設計の効率化と信頼性向上に寄与することを目的としています。
なお、NMRスペクトルから分子構造を推定する「逆解析」については、アプリケーションノート NM250008「NMRスペクトルからの分子構造推定：QSPRモデルによる逆解析の実践」も併せてご参照ください。

NMRスペクトルの定量性の課題とQSPRモデル

NMRスペクトルは、分子構造解析において極めて有用な情報源である一方で、測定シーケンスやその条件（溶媒、濃度、温度など）に強く依存するため、ピーク位置や積分値が変動しやすく、定量的な特徴量抽出には課題が残ります。
特に「順解析」においては、実測スペクトルとの整合性を保ちながら、構造的特徴を定量的に捉える手法が求められます。
従来の量子化学計算によるスペクトル予測は高精度であるものの、計算負荷が高く、現場での即時活用には限界があります。
これに対し、分子構造の電子的・トポロジー的特徴を数値化したQSPRモデルは、軽量かつ高速な予測を可能にし、さまざまなスペクトルとの複合的な照合や構造探索において有効な手法として注目されています。
本検討では、記述子を活用することで、少数データ環境下でも安定したppm予測を実現し、構造–スペクトル関係の定量的理解と可視化を目指します。

解析アプローチ

順解析について

分子構造からNMRスペクトルを予測する「順解析」は、構造情報とスペクトル特性の関係性を定量的に捉えるための重要なアプローチです。
本資料では、QSPRモデルを用いた順解析に焦点を当て、分子構造の電子的・トポロジー的特徴を数値化することで、ppm値の予測を試みます。

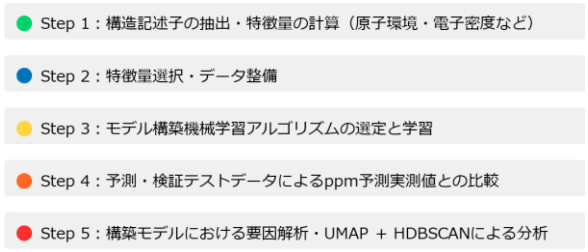


図1. 順解析ワークフロー例

図1に示す順解析ワークフローでは、電子的・構造的・トポロジー的記述子を組み合わせ、非線形回帰モデルによるppm予測とクラスタリングによる構造傾向の可視化を行います：

- 局所記述子の設計と選定（電子密度、混成軌道、グラフ中心性など）
- 非線形回帰モデル（Gradient Boostingなど）によるppm予測
- クラスタリング（UMAP+HDBSCAN）による構造空間の可視化と誤差傾向の分析

これらの手法により、分子構造の特徴がNMRスペクトルに与える影響を定量的に把握し、構造–スペクトル関係の理解を深めることが可能となります。

表1. 本検討で選定した主な記述子群

記述子カテゴリ	代表例	内容・意味
電子的特徴	Gasteiger 電荷、共役性	電子密度やπ結合の影響
構造的特徴	芳香環、立体中心、環の有無	空間配置や結合性に影響
トポロジー指標	グラフ中心性（closeness等）	分子内での原子の位置関係や重要度
SMARTSベースの分類	アミド、エステル、カルボニル等	官能基に対応した化学的分類

特徴量設計

表1に示すように、選定された記述子群を一覧化しています。
これらの記述子は、ppmシフトに直接関係する物理・化学的要因を数値的に捉えたものです。

日本電子株式会社
本社・昭島製作所
〒196-8558 東京都昭島市武蔵野3−1−2 TEL：(042) 543−1111(大代表) FAX：(042) 546−3353
www.jeol.co.jp ISO 9001・ISO 14001 認証取得

検討した分子リスト

表2に示すように、ppm予測モデルの構築にあたり、芳香族性、官能基の多様性、立体化学的特徴を有する複数の分子を選定しました。
これらの分子は、モデルの汎化性能を検証する目的で選定され、SMILES表記をもとに構造記述子を抽出し、機械学習モデルの入力データとして使用しました。図2に示すように、各分子の構造は以下の通りです。

表2. 学習に使用した主な分子について

分子名	SMILES表記	主な特徴
cinnamic acid	C1=CC=C(C=C1)/C=C/C(=O)O	芳香族カルボン酸、共役二重結合
cahe	CC/C=C¥¥CCOC(=O)/C=C/C1=CC=CC=C1	不飽和脂肪酸エステル、芳香環含有
reserpine	CO[C@H]1[C@@H](C[C@@H]2CN3CCC4=C([C@H]3C[C@@H]2[C@@H]1C(=O)OC)NC5=C4C=CC(=C5)OC)OC(=O)C6=CC(=C(C(=C6)OC)OC)OC	複雑な立体構造、インドールアルカロイド
cinchonidine	C=C[C@H]1CN2CC[C@H]1C[C@H]2[C@@H](C3=CC=NC4=CC=CC=C34)O	キノリン骨格、天然由来アルカロイド
riboflavin	CC1=CC2=C(C=C1C)N(C3=NC(=O)NC(=O)C3=N2)C[C@@H]([C@@H]([C@@H](CO)O)O)O	ビタミンB2、多官能基・多環構造
diflubenzuron	C1=CC(=C(C(=C1)F)C(=O)NC(=O)NC2=CC=C(C=C2)Cl)F	農薬成分、ハロゲン・尿素基含有

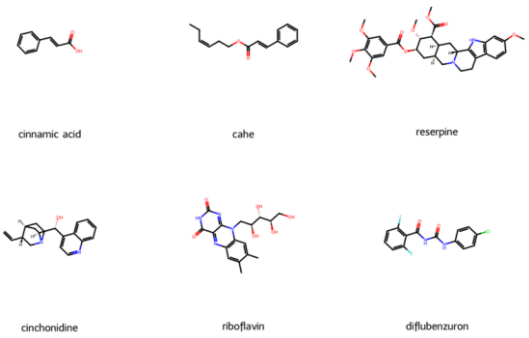


図2. 学習に使用した主な分子の分子構造

モデル構築と評価

回帰モデルの比較結果

ppm値の予測精度向上を目的として、表3に示すように、複数の回帰アルゴリズムを比較しました。評価指標には、RMSE（Root Mean Squared Error）および決定係数（R^2）を用いました。
表3に示すように、Gradient Boostingは他の手法と比較して最も低い誤差および高い決定係数を示しました。
この結果から、本検討におけるppm値の予測において高い精度と汎化性能を有するモデルであることが確認されました。

表3. ppm予測における学習モデル比較結果

Model	RMSE	R^2
GradientBoosting	2.054239	0.997909
RandomForest	3.899582	0.992464
Linear	8.649999	0.962921
BayesianRidge	8.777687	0.961818
Ridge	9.235034	0.957736

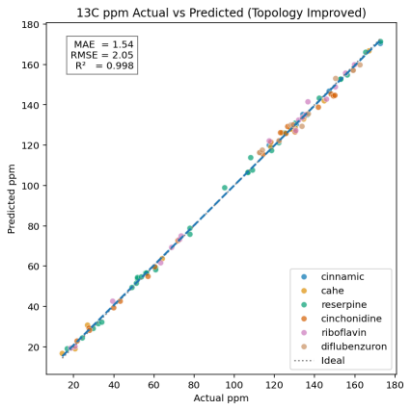


図3. 学習モデルによる^{13}C ppm 予測

表4. 分子誤差統計結果

Molecule	RMSE	Mean Error	Max Error
cahe	1.71	1.34	3.79
cinchonidine	2.57	2.02	5.45
cinnamic	1.24	0.87	2.53
diflubenzuron	2.50	2.12	4.43
reserpine	1.69	1.26	5.59
riboflavin	2.24	1.62	4.80

- **Gradient Boosting**：勾配ブースティングは、弱学習器を逐次的に構築することで予測精度を高めるアンサンブル学習手法です。
  本検討では、非線形性を捉えた安定した予測が可能であり、誤差分布も良好であることが確認されました。
- **Bayesian Ridge回帰**：ベイズ的アプローチを採用した線形回帰手法で、正則化項により過学習を抑制します。
  ただし、本検討のデータにおいては非線形性の表現が不十分で、実測値との乖離が大きい結果となりました。

本検討モデルによるppm予測結果

図3に示すように、Gradient Boostingモデルによるppm予測結果を示した散布図から、予測結果は実測値と良好に一致しています。
実測値（Actual ppm）と予測値（Predicted ppm）の関係を可視化しており、各データ点は分子ごとに色分けされています。
理想線（x = y）および線形回帰線を併記することで、予測精度の視覚的な評価が可能です。
図3左上は、主要な評価指標であるMAE（Mean Absolute Error）、RMSE、R^2であり、モデルの性能を定量的に把握できる指標が表示されています：

- **MAE**：誤差の絶対値の平均。外れ値に強く、平均的なズレを把握できます。
- **RMSE**：誤差の二乗平均の平方根。大きな誤差を強調します。
- **R^2**：予測がどれだけ実測値を説明できているかの指標です。1に近いほど良好です。

本モデルでは、RMSEが2.054、決定係数（R^2）が0.998と高い精度を示しており、ppmスケールにおいて十分に小さい誤差で安定した回帰が実現されています。
この結果から、少数データ環境下でも、分子記述子を活用することで高精度なppm予測が可能であることが確認されました。
分子ごとの誤差統計については表4に示します。
なお、図3の結果は、トポロジー記述子を改良したモデルに基づいており、その効果については後述の記述子傾向分析にて詳しく検討します。

JEOL 日本電子株式会社
本社・昭島製作所
〒196-8558 東京都昭島市武蔵野3−1−2 TEL：(042) 543−1111(大代表) FAX：(042) 546−3353
www.jeol.co.jp ISO 9001・ISO 14001 認証取得

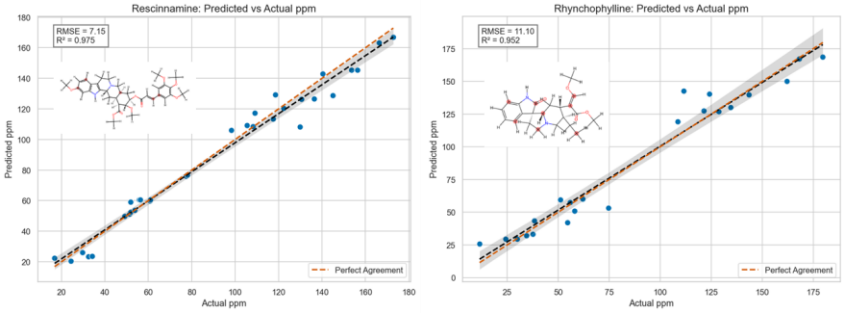


図4. 未知分子による^{13}C ppm値予測: Rescinnamine and Rhynchophylline

未知分子に対する予測性能の検証

図4には、構築したモデルを用いて未知分子RescinnamineおよびRhynchophyllineの^{13}C ppm値を予測した結果を示しています。

少数データによるモデル学習の段階でありながら、両分子に対して良好な予測結果が得られ、モデルの汎用性と応用可能性が示唆されました。

今後、学習データの拡充や記述子設計の最適化を進めることで、さらなる精度向上が期待されます。

記述子の寄与分析

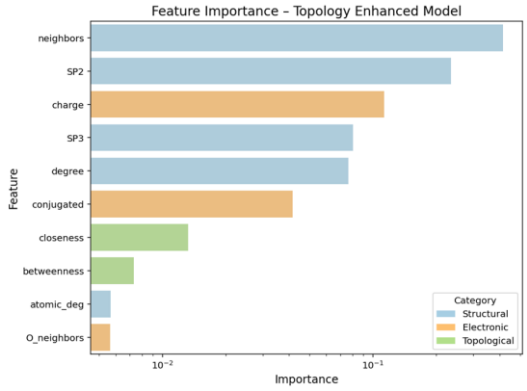
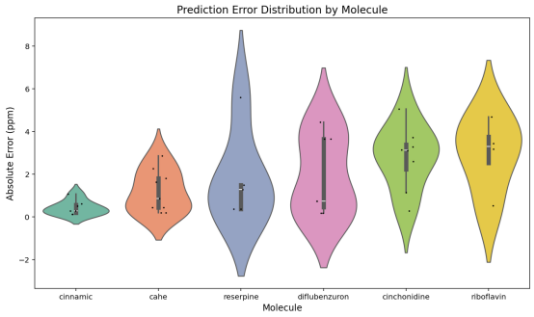


図5. 記述子の寄与度



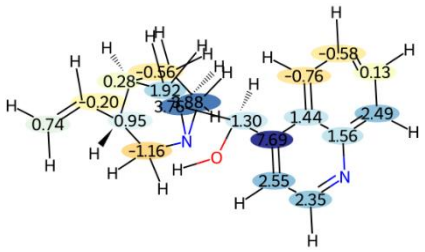


図6. トポロジー記述子による効果: Violin Plot and Error Reduction Map (Cinchonidine)

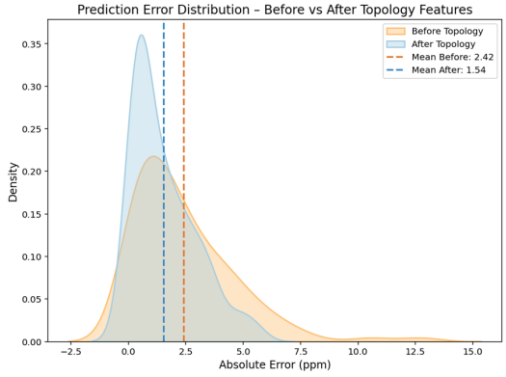


図7.予測誤差分布のKDE Plot:トポロジー記述子導入前後の比較

特徴量重要度(Feature Importance)

図5に示すように、ppm予測モデルにおける特徴量(記述子)の重要度が定量的に評価されています。

特徴量重要度とは、各記述子がモデルの予測精度向上にどれだけ寄与しているかを定量的に評価する指標です。

この重要度は、決定木モデルにおける分割時に、各特徴量が目標変数の不純度(例:平均二乗誤差)をどれだけ削減したかに基づいて算出されます。

アンサンブル全体の貢献度を反映するため、信頼性の高い指標とされています。

図5の結果から、ppm予測においては構造的特徴(例:原子の近接環境や混成状態)と電子的特徴(例:部分電荷や共役性)の両方が重要であることが示されました。

特徴量「num_neighbors」は、対象原子の近接原子数を示す指標であり、ppm予測における主要な寄与因子として評価されました(図5参照)。

分子内の局所構造がppm予測に強く影響する主要因であることが示唆されます。

さらに、トポロジー記述子の活用により、構造的要素と電子的要素の関係性を補完することで、モデル全体の性能向上に寄与していると考えられます。

トポロジー情報の追加における誤差傾向と構造的改善効果

図6に示すように、分子ごとの予測誤差改善の分布と、Cinchonidineに対してトポロジー記述子を追加した際の改善箇所を差分として可視化しています。

トポロジー記述子とは、原子の「つながり方」や「構造内での位置関係」をグラフ理論に基づいて定量化したものであり、分子構造の抽象的な特徴を捉えるのに有効です。

分子構造上のマップにおける濃淡は、誤差改善の度合いを表しており、青色が濃いほど改善効果が大きい元素であることを示しています。

図6のViolinプロットから、特に構造が複雑な分子において、トポロジー記述子の追加による予測精度の改善が顕著であることが確認できます。

例えばCinchonidineについては、図6の差分プロットより、トポロジー情報の追加によって分子全体の予測傾向が改善していることが示されました。

図7に示すように、トポロジー記述子の追加によるモデル性能の予測誤差傾向をKDEプロットで比較しています。

具体的には:

- トポロジー記述子の追加により、RMSEは3.26から2.05へと約37%低下し、予測誤差が大幅に改善されました。
- 決定係数(R^2)も0.995から0.998と向上し、モデルの説明力が強化されました。

これらの結果は、トポロジー記述子がppm予測モデルにおいて有効な情報源であることを示しており、構造的・電子的特徴の補完的役割を果たしていると考えられます。

このように、分子構造をトポロジー的観点から捉える手法は、分子設計やスペクトル予測などの応用分野において、今後さらに実用性の高い改善手法として発展していくことが期待されます。

JEOL 日本電子株式会社
本社・昭島製作所
〒196-8558 東京都昭島市武蔵野3−1−2 TEL:(042) 543−1111(大代表) FAX:(042) 546−3353
www.jeol.co.jp ISO 9001・ISO 14001 認証取得

クラスタ分析による構造傾向と誤差改善の可視化

UMAP+HDBSCANによるクラスタ分析結果

図8に示すように、分子構造空間におけるクラスタ分布と、各クラスタにおける予測誤差の改善傾向を比較しました。

高次元の分子記述子データを効率的に可視化するため、非線形構造の保持に優れたUMAP（Uniform Manifold Approximation and Projection）による次元削減を実施しました。

さらに、データの密度構造を反映するクラスタリング手法であるHDBSCAN（Hierarchical Density-Based Spatial Clustering）を用いて、構造空間における自然なクラスタ分布を抽出しました。

クラスタごとの誤差改善量を分析した結果、クラスタ4・2・3においては、予測誤差の改善量の中央値が高く、安定した改善傾向が確認されました。

特にクラスタ2はデータ件数が多く、モデル改善において中心的な構造領域であると推察されます。この結果は、構造空間における特定の領域がモデルの予測性能向上に寄与していることを示しており、今後の記述子設計やデータ選定において重要な指針となると考えられます。

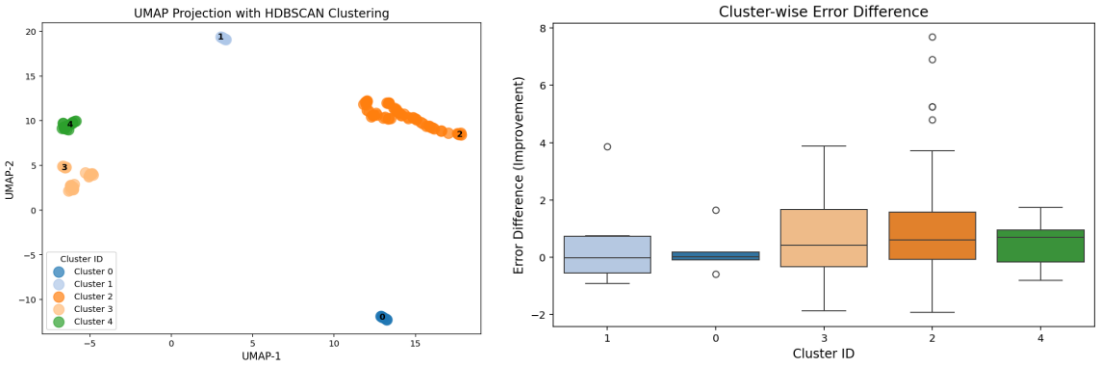


図8. UMAPおよびHDBSCANによる構造空間におけるクラスタ分布とクラスタ毎の予測改善傾向

誤差改善に寄与する記述子とクラスタ構造の関係

分子構造空間における特徴量の意味を理解するには、クラスタごとの記述子傾向と誤差改善との相関分析という、異なる視点からのアプローチが有効です。

例えばSpearman相関分析は、誤差改善との関係性を定量的に評価する手法であり、モデルがどの特徴量に対して高い感度を持つかを明らかにします。これにより、モデルの予測性能に寄与する記述子を特定することが可能です。

図9に示すように、誤差改善に関するSpearman相関分析の結果から、記述子と誤差改善の相関関係を明らかになりました。

分析の結果、is_aromaticおよびis_in_ringといった記述子が、予測誤差の改善に寄与していることが示されました。

これらの記述子は、分子内の芳香族性や環構造の有無を示すものであり、モデルの予測精度向上において重要な因子であると考えられます。

図10では、各クラスタにおけるis_aromaticおよびis_in_ringの平均値をプロットし、記述子と誤差改善の傾向をクラスタから示しています。

クラスタ2は、芳香族性や極性原子を多く含む分子群であり、予測誤差の改善傾向が安定していることから、今後の記述子設計における重要なターゲット領域と考えられます（図10参照）。

分子構造およびトポロジー情報を取り入れた誤差改善分析により、芳香族原子や環構造に属する原子に対しては、予測誤差が改善される傾向が観察されました。

一方で、結合次数や周囲の原子数といった局所的な構造量に対しては、改善傾向が弱く、モデルの反応性は分子グラフ全体の特異性よりも、共鳴性や拘束構造といった分子全体の構造的制約に影響されている可能性が示されました。

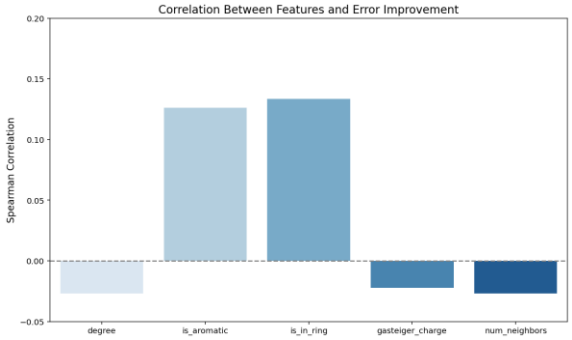


図9. 誤差改善に関する記述子のSpearman相関分析結果

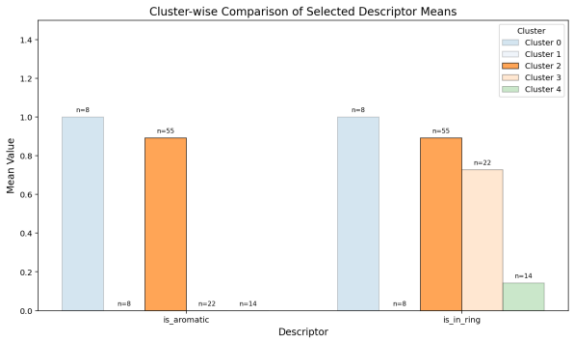


図10. 芳香族性および環構造に関する記述子のクラスタ平均

改善原子の空間的広がりと構造空間の関係

Spread Scoreによる改善原子の空間的広がりの分析

モデル改善の要因を探るため、改善された原子の周辺環境に着目し、改善された炭素の空間的広がりを評価しました。

具体的には、改善された炭素原子を中心に一定の半径内に存在する隣接原子の数をカウントし、炭素数に距離に応じた重みを加えて算出した指標を「Spread Score」として評価しました。

JEOL 日本電子株式会社
本社・昭島製作所
〒196-8558　東京都昭島市武蔵野3－1－2　TEL：(042) 543－1111(大代表)　FAX：(042) 546－3353
www.jeol.co.jp　ISO 9001・ISO 14001　認証取得

図11. クラスタ毎のSpread Score分布

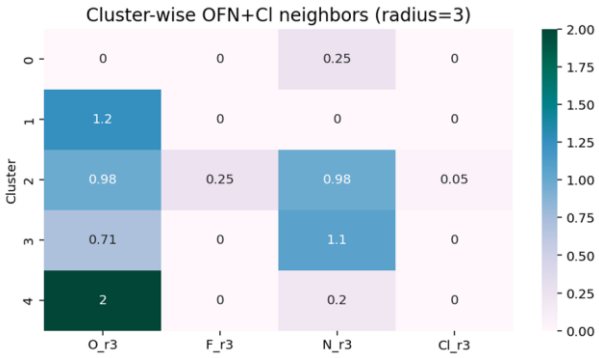
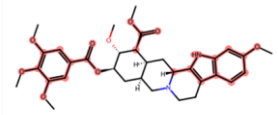
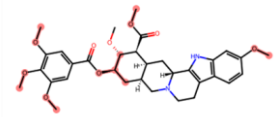


図12. 改善原子周辺のヘテロ原子の広がり

表5. 改善原子のサブグラフ抽出

クラスタ	改善原子のサブグラフ抽出
2	
4	

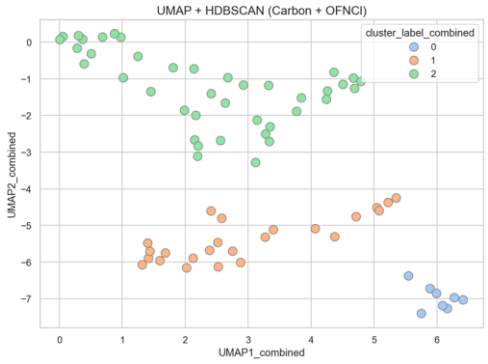


図13. 改善原子周辺環境における構造空間分類

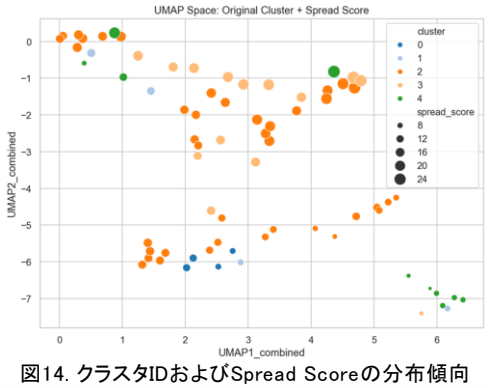


図14. クラスタIDおよびSpread Scoreの分布傾向

図11は、クラスタごとのSpread ScoreをViolinプロットで示し、Spread Scoreのクラスタにおける分布傾向を可視化しています。

この結果から、クラスタ2および3は高いSpread Scoreを示し、構造的に広範囲な改善が得られていることが確認されました。

またクラスタ0および1では、Spread Scoreが低く局所的な改善傾向となっていることがわかります。

一方、クラスタ4は特異的な分布を示し、限定的な構造環境において高いSpread Scoreを示していることがわかりました。

ヘテロ原子特徴のクラスタ別傾向

改善された原子の周辺における構造的多様性を評価するため、ヘテロ原子（N, O, F, Cl）の分布を分析しました。

図12に示すように、改善原子周辺に存在するヘテロ原子の分布と環境の可視化から、改善に寄与するヘテロ原子の構造空間における関係が明らかになりました。

クラスタ2では、ヘテロ原子の分布が広範囲に及んでおり、構造的多様性が高いことが示されました。

一方でクラスタ0および1は、炭素・ヘテロ原子ともに分布が限定的であり改善が局所的かつ不安定であることがわかります。

またクラスタ4では、酸素（O）に関する構造的特徴が特異であり、Spread Scoreの分布も他クラスタとは異なる傾向を示しました。

サブグラフ抽出による構造的特徴の確認

表5では、クラスタ2およびクラスタ4に属する改善原子に対して抽出されたサブグラフの構造的特徴を示しています。

図12で示唆されたクラスタごとの構造的特徴が、サブグラフから明確に示されました。

炭素・ヘテロ原子記述子による構造空間と改善傾向の分布

図13に示すように、改善原子の周辺環境に関する分析結果を統合し、炭素およびヘテロ原子（N, O, F, Cl）に関する記述子を用いて、UMAPとHDBSCANによる構造空間の分類を実施しました。

図14に示すように、図8におけるクラスタIDとSpread Scoreを図13に重ね合わせて、炭素およびヘテロ原子記述子に基づく構造空間の分類結果を示しました。

データポイントの大きさはSpread Scoreに対応し、構造的改善傾向の空間的な広がりを示しています。

クラスタ2領域に、主に図8におけるクラスタ2・3・4のデータが集積し、高いSpread Scoreを持つ構造が分布していることから、芳香族性や極性原子の多様性が改善に寄与している可能性が示唆されました。

図15に示すように、炭素・ヘテロ原子記述子に基づく分類におけるSpread Scoreの分布と統計結果から、図13のクラスタ2が改善に大きく寄与するグループであることが定量的に裏付けられました。

これらの領域は、モデルにとって「改善しやすい」構造群であると判断されていることが示唆されます。

設計した記述子において芳香族性や極性原子の電子的・構造的な多様性がモデルの改善に寄与していることが示されました。

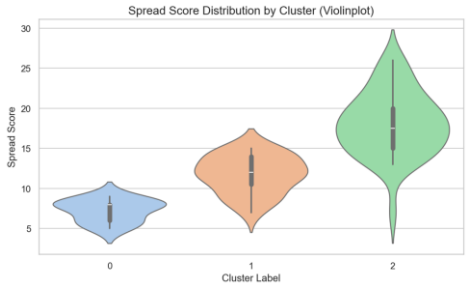


図15. 改善原子周辺におけるSpread Score分布と統計結果

JEOL 日本電子株式会社
本社・昭島製作所
〒196-8558 東京都昭島市武蔵野3−1−2 TEL：(042) 543−1111(大代表) FAX：(042) 546−3353
www.jeol.co.jp ISO 9001・ISO 14001 認証取得

まとめ

本検討では、NMRスペクトルに基づくppm予測モデルの構築と、分子構造に対する誤差改善傾向の解析を通じて、少数データ環境下でも記述子を活用することで高精度な分子評価が可能であることを示しました。
Gradient Boostingを中心とした機械学習モデルにより、非線形性を捉えた安定した予測が実現され、未知分子に対する汎化性能も確認されました。
記述子の重要度分析では、芳香族性や環構造に関連する特徴量が誤差改善に寄与することが明らかとなり、特にクラスタ2および3に属する分子群がモデル改善の中心的領域であることが示されました。
改善された炭素原子の局所環境を定量評価する空間的指標を用いた解析により、周辺環境の広がりが予測精度向上に寄与していることが定量的に裏付けられました。
さらに、UMAPおよびHDBSCANによる構造空間の分類と、ヘテロ原子（O, F, N, Cl）を含む記述子の統合により、従来の分類では捉えきれなかった改善傾向を可視化することができました。
この結果から、設計した記述子に含まれる芳香族性や極性原子の電子的・構造的なバラエティが、トポロジー特徴と組み合わせることで包括的に評価され、モデル改善に寄与していることが示されました。

今後の展望

本手法は、NMRスペクトルのさらなる活用に向けて、分子の物理化学的特性を記述子で定量的に評価し、構造的傾向を抽出することで、分子設計に資する指針を導出する有効なアプローチです。
既存のマテリアルズ・インフォマティクスツールとの親和性も高く、今後の実験設計や分子探索への応用が期待されます。
本検討で構築したスペクトル予測モデルは、構造アノテーション支援や未知分子のスクリーニング、逆設計などへの応用が期待されます。
また、MSやXRFなど他スペクトルとの統合解析により、より多角的な構造理解が可能となります。
今後は、分子生成AIとの連携やモデルの説明性向上を通じて、材料設計支援や標準化への貢献が期待されます。
また、クラウド型解析ツールとの統合により、実験・解析・設計の循環的な強化も可能となります。

参考：使用環境とライブラリー覧

本検討は、Python 3.13環境にて、分子構造解析・機械学習・統計解析・可視化を統合的に実施しました。
使用した主要ライブラリは以下の通りです：

- RDKit[4]：化学構造の操作、記述子計算、構造描画
- pandas / numpy[5][6]：データ処理と数値演算
- Networkx[7]：分子グラフ構造の解析
- scikit-learn[8]：機械学習モデルの構築と評価（線形回帰、アンサンブル学習、SVM、ガウス過程回帰など）
- scikit-learn内モジュール：主成分分析（PCA）、特徴量の標準化、ラベルエンコード

これらのライブラリを組み合わせることで、記述子の設計からモデル構築、構造空間の可視化を実施しました。

参照
[1] JEOL Analytical Software Network: https://www.jeoljason.com/
[2] Pythonは、Python Software Foundationの商標または登録商標です。公式ドキュメント（日本語）: https://docs.python.org/ja/
[3] BeautifulJASON ドキュメント: https://www.jeoljason.com/beautifuljason/docs/
[4] RDKit: オープンソース化学情報ツールキット（Version 2025.03.6） https://www.rdkit.org
[5] pandas：https://pandas.pydata.org/docs/
[6] numpy：https://numpy.org/doc/
[7] networkx：https://networkx.org/
[8] scikit-learn：https://scikit-learn.org/stable/

JEOL 日本電子株式会社
本社・昭島製作所
〒196-8558 東京都昭島市武蔵野3－1－2 TEL：(042) 543－1111(大代表) FAX：(042) 546－3353
www.jeol.co.jp ISO 9001・ISO 14001 認証取得