

データ駆動型材料開発のためのアルゴリズム開発： 金属錯体を対象として

漆原 誠 山口 健志

Algorithm Development for Data-driven Materials Research on Metal-ligand Complexes

Makoto URUSHIHARA Kenji YAMAGUCHI

Abstract

We have applied data sciences to accelerate the development of materials research. In this paper, we show our research on machine learning-based analysis of overall stability constants of metal(M)-ligand(L) complexes. In this study, two Gaussian process regression models are developed to predict the first overall stability constant and the n-th ($n > 1$) overall stability constants. Furthermore, the feature relevance is quantitatively evaluated via sensitivity analysis. As a result, the electronegativities of both metal and ligand are found to be the most important factor for predicting the first overall stability constant. Interestingly, the predicted value of the first overall stability constant shows the highest correlation with the n-th overall stability constant of the corresponding M-L pair. Finally, the number of features is optimized using validation data where the ligands are not included in the training data, which indicates high generalizability. This study provides valuable insights and may help accelerate molecular screening and design for various applications.

キーワード：データ駆動型材料開発，マテリアルズ・インフォマティクス，機械学習，金属錯体，錯化剤，ガウス過程回帰

1. はじめに

新規材料開発や材料製造プロセスの開発に対して、データ科学・情報科学を適用するマテリアルズ・インフォマティクス (MI)¹⁾ やプロセス・インフォマティクス²⁾ が近年盛んに取り組まれている。当社においても材料開発からプロセスの開発に至るまでデータ科学の適用に幅広く取り組んでおり、その一端は、電磁波遮蔽樹脂の開発における適用事例として R&D レビューでも既に紹介した³⁾。本報では、材料の特性予測を行うために出来るだけ汎化性能（未知のデータに対する予測能力）の高いモデル化を行うための検証データ抽出の方法や特徴量の選定など、個々の材料開発に応じたアルゴリズム開発の中で、金属錯体に関して取り組んだ我々の事例⁴⁾ について紹介する。

2. 金属錯体

金属錯体は、金属イオンに配位子が結合した化合物である。めっきプロセス、金属の分離プロセス、分析化学等、当社でもなじみのある様々なプロセスで用いられているが、用途に応じて適切な金属・配位子の組合せの選定が重要となる。例えば、ある金属に注目した場合、可

能性のある配位子は数千にも上り、全ての候補の特性を評価することは現実的ではない。また、現状、用途に応じた最適な配位子の選定は経験的な方法に頼っており、未知の材料を含む幅広い選択肢の中から最適な材料を効率的に選定する方法が望まれている。仮に金属錯体の構造を指定すると特性が予測できるモデルが構築できれば、構築したモデルと各種最適化手法とを組み合わせることで、未評価の材料を含めた候補の中から最適な金属錯体や配位子を効率的に選定できることが期待される。

金属錯体の特性を特徴づけるさまざまな指標が存在するが、重要な指標の一つが水溶液中での錯安定度定数である。錯安定度定数は、金属錯体を水に溶かした時に配位子が解離せずに金属錯体として存在する度合いを示すものである。例えば金属 M と配位子 L が 1 : 1 で形成する錯体 $M-L$ が水溶液中で平衡状態にある時、平衡定数 K_1 は下記で表される。

$$M + L \leftrightarrow M-L, \quad K_1 = \frac{[M-L]}{[M][L]} \quad (1)$$

この時、錯安定度定数 β_1 は下記と定義される。

$$\beta_1 = \log K_1 \quad (2)$$

1つの金属イオンに対して配位子が複数存在する場合は、配位子1つが解離する反応を考える。例えば $M:L=1:n$ ($n>1$) の錯体に対して、 n 番目の配位子の平衡状態に対する平衡定数 K_n は下記で表される。

$$M-L_{n-1}+L \leftrightarrow M-L_n, \quad K_n = \frac{[M-L_n]}{[M-L_{n-1}][L]} \quad (3)$$

このような金属錯体において、 n 個の配位子が解離する際の錯安定度定数を、 n 次反応における全錯安定度定数 β_n として、 β_1 と同様に下記のように定義する。

$$\beta_n = \log K_1 \times \dots \times \log K_{n-1} \times \log K_n = \log \frac{[M-L_n]}{[M][L]^n} \quad (4)$$

これ以降、本報内では $n=1$ は β_1 、 $n>1$ は β_n と表記する。上式より、 β が小さい金属錯体ほど、溶液中で金属イオンと配位子の状態に解離している状態であるといえる。例えば、ある湿式プロセスにおいて、金属イオンを含む溶液中で金属を錯体状態で極力安定的に存在させたい場合は、 β がより大きくなる配位子・金属錯体を選定する。

金属錯体の錯安定度定数を予測するための機械学習モデルについて、いくつかの報告⁵⁻⁷⁾ がなされているが、それらの既報においては金属元素の種類が限定的である、 β_1 のみの予測モデルである、といった課題がある。また、錯安定度定数の予測モデルから優れた金属錯体・配位子を選定する場合、予測モデルを利用するだけではなく最適化手法と組み合わせることが考えられる。材料開発で用いられるブラックボックス最適化⁸⁾ の中でも、ベイズ最適化は効率的な手法の一つである⁹⁻¹¹⁾ が、ベイズ最適化を行うためには、ベイズ線形回帰やガウス過程回帰によるモデル化が必要である。しかしながら我々の調査の範囲内では錯安定度定数に対するそのような報告は見つけれなかった。そこで我々は、最適化と組み合わせることを視野に入れた錯安定度定数の予測モデルとするため、ガウス過程回帰モデルを構築することとした。

機械学習モデルを作成するにあたって、学習の元とな

る錯安定度定数のデータが必要となる。今回はアメリカ国立標準技術研究所 (NIST) が作成したデータベース NIST SRD 46. Critically Selected Stability Constants of Metal Complexes¹²⁾ 等から $M:L=1:m$ ($m=1\sim6$) の金属錯体に対して全安定度定数の実験値とそれらの実験条件を抽出した。超ウラン元素など一部の重金属 (Am, Cm 等) を含む錯体、また配位子内に Te, Se, As, Mn などを含む錯体については、モデル作成時に必要となる説明変数の生成の難しさから除外した。また、実験条件として 25℃、イオン強度が 0.1 となるデータを中心に収集した。なお、重複したデータが存在した場合は β が大きいものを採用した。最終的には 19810 の金属錯体を含むデータセットを作成し、モデルの学習に利用した。図 1 に金属カチオン種毎のデータ数を示す。なお、金属カチオンは 57 種類、配位子は 2706 種類と幅広い組合せのデータセットとなっている。

3. 方 法

3.1 特徴量候補

錯安定度定数を予測する機械学習モデルを作成するためには、それを説明するための特徴量が必要となる。特徴量の候補として、カチオンや配位子由来の特徴量、実験条件である温度とイオン強度を用いた。カチオンの特徴量は、カチオンの電荷、原子番号、融点、熱容量、イオン半径、分極率、電子親和力、Pauling の電気陰性度、s, p, d, f 軌道における空の電子軌道数^{13, 14)} を用意した。配位子の特徴量の算出には特徴量計算ソフト Mordred¹⁵⁾ を用いた。さらに配位子の特徴量として水溶液中の配位子の分子電荷も考慮した。Null 値や単一値である特徴量を除去し、特徴量同士のピアソン相関係数の絶対値が 0.7 以上となる場合は一方の特徴量を除外した。更に特徴量間の多重共線性を避けるために、分散拡大係数 (variance inflation factor, VIF) を計算し、VIF が最も大きくなる特徴量を削除し、再度 VIF を評価することを VIF が 4 以下になるまで繰り返した。 β_n の機械学習モデル作成時には、

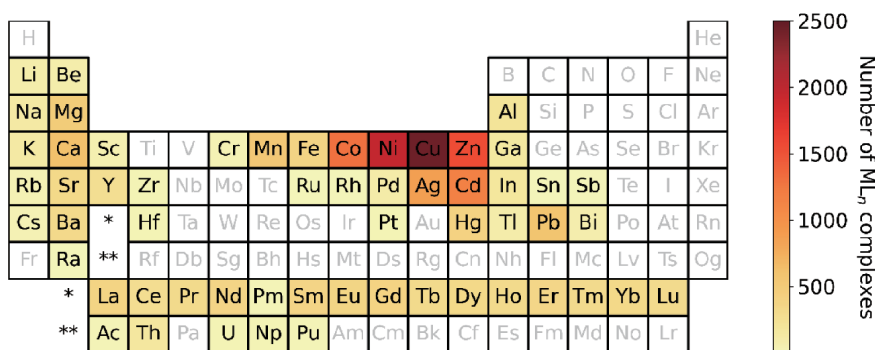


図 1 抽出された金属錯体の実験データ数。57 種の金属カチオン、2706 種の配位子を含む

Extracted total experimental data of M-L complex, which is composed of 57 metal cations and 2706 ligands

上記に加えて機械学習モデルにより予測された β_1 値とその分散、さらに錯体の電荷（カチオンとアニオンの電荷の和）も説明変数として用いた。最終的に得られた特徴量のデータセットのサイズは、 β_1 用が13559錯体データ $\times$ 118特徴量、 β_n 用が6251錯体データ $\times$ 60特徴量となった。

3.2 機械学習モデル

安定度定数を予測するための機械学習モデルとして、ガウス過程回帰（Gaussian Process Regressor, GPR）を用いた。データ間の類似性を表現するカーネルとしてはMatern kernel ($\nu=3/2$)を基に関連度自動決定（automatic relevance determination, ARD）構造に修正した下記の式で表現されるカーネルを用いた。

$$k(x_i, x_j) = \sigma^2 (1 + \sqrt{3}r_{ARD}) \exp(-\sqrt{3}r_{ARD}), \quad (5)$$

$$r_{ARD} = \sqrt{\sum_{m=1}^d \frac{(x_{im} - x_{jm})^2}{l_m^2}} \quad (6)$$

ここで σ 、 l_m 、 r_{ARD} は、それぞれ、カーネルの重み、特徴量間の相関性を表す長さスケールの指標、2つのデータ x_i と x_j のユークリッド距離を表すハイパーパラメータである。 d は特徴量の次元（数）である。GPRのモデル化では、PyTorch¹⁶⁾とGPytorch¹⁷⁾を用いた。

3.3 汎化性能

GPRのモデル化において、過学習を避け高い汎化性能を実現するために特徴量の選択を実施した。特徴量の選択を行う場合、各特徴量がモデルに与える影響を評価して影響の大きいものを選択することが行われる。GPRではカーネル中の長さスケールの逆数 l_m を使って特徴量の影響度を評価することが行われるが、この方法がうまく行かないケースが知られている¹⁸⁾。そこで本研究では、

各特徴量に対してKullback-Leibler情報量を用いた感度解析を行い、特徴量の影響度を評価した。より具体的には、特徴量候補全てを使ったGPRモデルを作成した上で、作成されたGPRとKullback-Leibler情報量を用いて、各特徴量のモデルに対する影響度を感度解析により評価した。次に影響度の高い特徴量から順にモデルの特徴量に加えたGPRモデルを作成し、作成したGPRモデルで検証データに対する予測を行い、予測値と実測値を比較することで検証を行った上で、最も検証データの再現性に優れた特徴量を採用することで最終的なGPRモデルを作成した。

モデルの汎化性能を評価する際の検証データの選択についても工夫を行った。一般に汎化性能を評価する方法として交差検証（cross-validation, CV）法がよく用いられる。しかしながら本課題において収集したデータに対してCV法を用いると、データ中に多く含まれる一般的な配位子が訓練データ・検証データのいずれにも含まれてしまい、過学習が避けられない可能性がある。そこで、検証データの抽出に対して、①訓練データ内に含まれない配位子を選定し、②学習モデルの適用範囲（Applicability Domain, AD）の端に位置するデータを抽出する、という方針を立てた。ここで②について補足する。モデルのADは学習に用いたデータセットに強く影響を受けるため、学習用に用いるデータセットを可視化した上で、その可視化した領域の端のデータを選ぶことによって、モデルのADの端のデータを検証データとして採用できたこととなり、モデルの汎化性能を確保しやすいであろう、という方針である。②を満たすデータを選定するために、UMAP（uniform manifold approximation and projection）¹⁹⁾による次元削減手法を活用して可視化した。 β_1 の実験データセットに対して、UMAPで次元削減した β_1 に対する散布図を図2(a)に示す。散布図のマップの中心からのマハラノビス距離、すなわち中心から離れるほど中心の

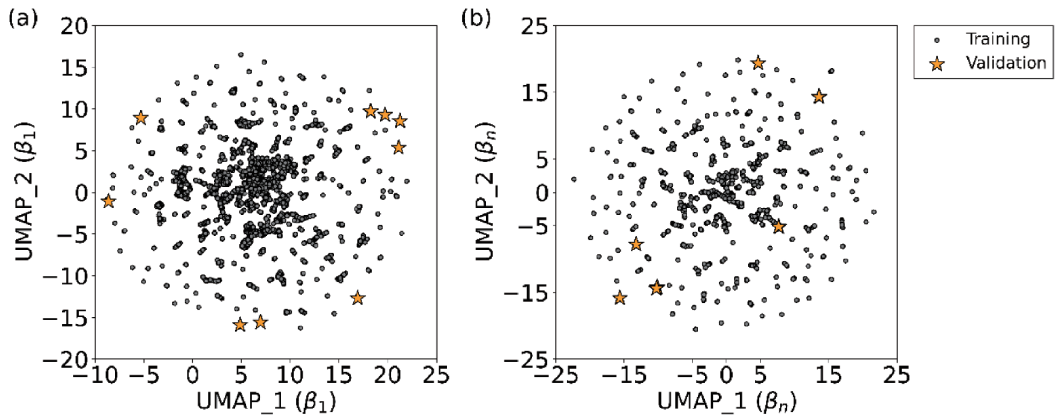


図2 UMAPを使った次元削減後の (a) β_1 と (b) β_n の散布図。灰色の丸、黄色の星はそれぞれ訓練データと検証データを示す

Visualization of the dataset for (a) β_1 and (b) β_n using UMAP. Data represented by gray circles and yellow stars were used for the training and valid data for feature optimization, respectively

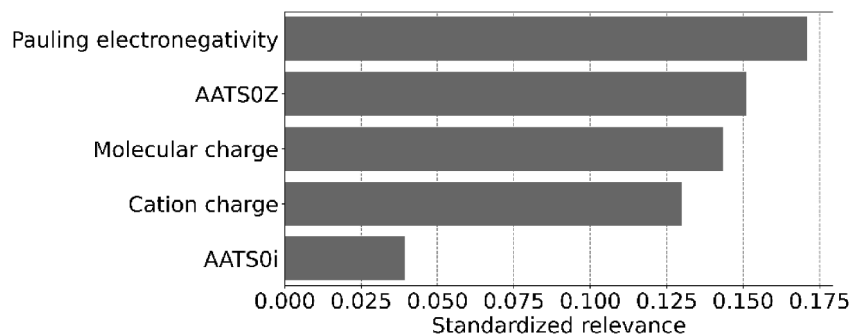


図3 Kullback-Leibler 情報量を用いた感度解析による β_1 予測モデルにおける特徴量の影響度 (上位 5)

The top 5 highest ranked features through importance analysis using a Kullback-Leibler divergence as a measure for predicting β_1

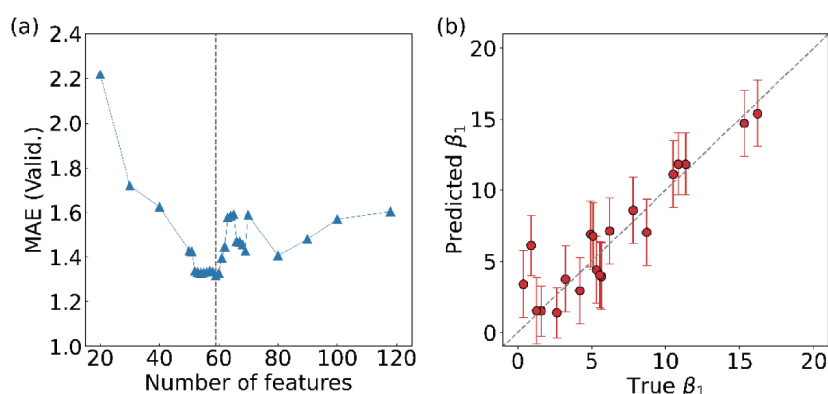


図4 (a) モデル作成時に利用する特徴量の数を変量した場合のモデルの予測性能の変化。点線は特徴量の数が 59 の場合。(b) モデルと検証データに対して 59 の特徴量を用いたモデルで予測した β_1 と実験値により得られた β_1 。エラーバーは 1σ を表す

(a) Predictive performance for the validation samples as a function of the number of features. Features are arranged in descending order of relevance. The black dashed line corresponds to the top 59 features. (b) Parity plot between true and predicted β_1 values of the validation data using the best GPR model with the top 59 features. Error bars indicate 1σ uncertainty of the predicted values

データとの類似性が低下することを表す距離、を計算して上述①②の二つの条件に合う 20 種の金属錯体のデータを、 β_1 モデル用の検証データとして選定した (図 2(a) の星印)。 β_n にも類似の方法を適用して検証データを抽出 (図 2(b) の星印) し、5 種の金属錯体のデータを検証データとして選定した。

4. 結果と考察

4.1 β_1 のモデリング

3.1 で述べた 118 の特徴量と予測用 β_1 のデータを用いて GPR モデルを作成し、作成したモデルを基に、特徴量の影響度の評価を 3.3 の方法で実施した。特徴量の影響度が大きい上位 5 つを図 3 に示した。なお、全特徴量の影響度に対して、上位 5 つの特徴量が占める影響度の割

合は 0.63 であり、この 5 つで 60% 以上の寄与を持っている。図 3 を見ると電気陰性度、分子および金属の電荷が重要な因子となっている。また、配位子の幾何学的な特徴を示す AATS0Z, AATS0i が含まれている。両者は分子の形状を表現する分子グラフに基づいて計算され、さらに配位子に含まれる元素の原子番号とイオン化ポテンシャルに依存する特徴量である。

特徴量の影響度の結果を元にして、GPR モデルで用いる特徴量選択を実施した。特徴量の影響度が高い順に特徴量を増やしてモデル化し、それぞれのモデルを使って 3.3 で説明した方法を用いて検証用データの予測を行い、予測値と実験値の平均絶対値誤差 (Mean Absolute Error, MAE) を評価した。特徴量の数と MAE の関係を図 4(a) に示す。説明変数として 59 種を用いた場合が最も精度が

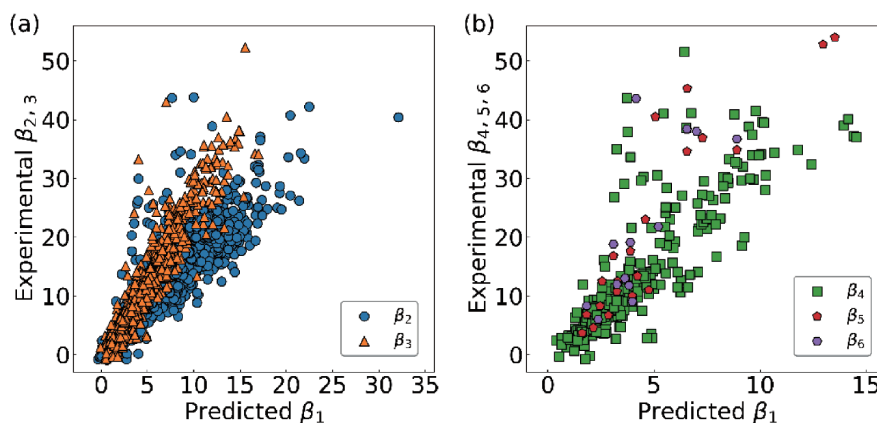


図5 実験値から得られた n 次の全安定度定数と予測された β_1 の関係。(a) β_2, β_3
(b) $\beta_4, \beta_5, \beta_6$
Relationships between experimental multi-order β_n and predicted β_1 . (a) β_2, β_3 (b) $\beta_4, \beta_5, \beta_6$

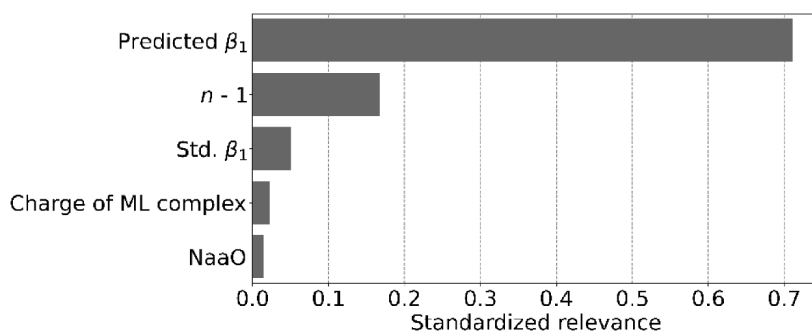


図6 Kullback-Leibler 情報量を用いた β_n 予測モデルにおける特徴量の影響度 (上位 5)
The top 5 highest ranked features through importance analysis using a Kullback-Leibler divergence as a measure for predicting β_n

よい結果となった。59 種の特徴量を用いた際の GPR モデルと検証データを使って β_1 を予測した際の予測値と実測値の関係を図 4(b) に示す。学習時に含まれていない検証データに対しても、高い予測性能を有することが確認できた。

4.2 β_n モデリング

β_n についても GPR でモデルを作成した。ここで Co^{2+} , Ni^{2+} , Cu^{2+} の金属錯体について、 β_1 と β_2 の間に線形関係があることが知られている²⁰⁾。そこで今回のデータセットで β_n と β_1 のような関係があるかを確認した (図 5)。なお、 β_n のすべての $M-L$ ペアが β_1 のデータセットに含まれるわけではない。そのため β_1 は 4.1 で作成した GPR モデルでの予測値を用いている。この結果を見ると β_n は β_1 と線形関係があることが見て取れる。つまり β_n のモデルにおいて β_1 はよい特徴量となる可能性が高いといえる。

上記の結果を元に、 β_1 と同様の方法で β_n の GPR モデルを作成した。まずは β_1 を含む 60 種全ての特徴量を用いて β_n を予測するモデルを作成し、特徴量の影響度を

Kullback-Leibler 情報量を用いて評価した。図 6 に特徴量の影響度が大きい上位 5 つを示す。なお、上位 5 つで特徴量全体の影響度の 0.97 を占めている。図 6 を見ると、 β_1 が極めて高い影響を持つ特徴量であることが分かる。その他にも金属錯体の電荷や NaaO が含まれている。なお、NaaO は Mordred により生成された特徴量で、芳香族環に挿入された O のような分子構造に起因する特徴量である。

特徴量の影響度の結果を用いて、 β_1 と同様に、影響度の高い順に特徴量を増やしてモデル化し、検証用データの予測における MAE を評価した。特徴量の数と MAE の関係を図 7(a) に示す。説明変数として 25 種を用いた場合が最も精度がよい結果となった。その時の検証データに対する予測値と実測値の関係を図 7(b) に示す。 $\beta_2 \sim \beta_6$ のいずれにおいても、学習に含まれていない検証データに対して高い予測性を持つモデルが得られたことが確認できた。

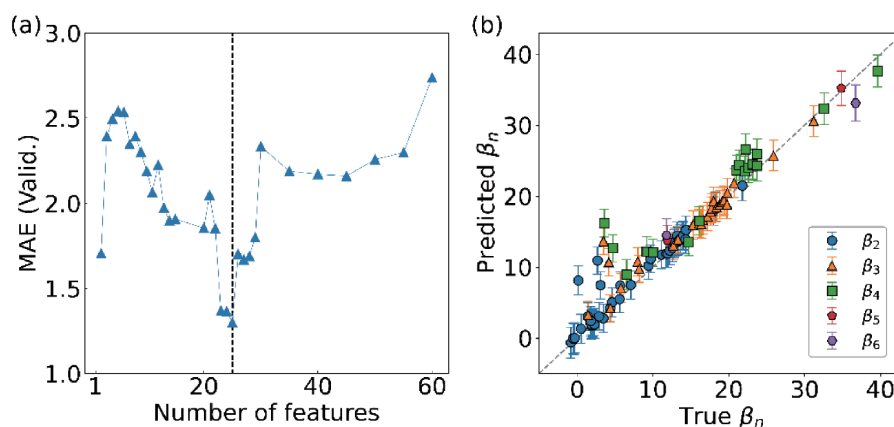


図7 (a) モデル作成時に利用する特徴量の数を変量した場合のモデルの予測性能の変化。点線は特徴量の数が25の場合。(b) 検証データに対して25の特徴量を用いたモデルで予測した β_n と実験値により得られた β_n 。エラーバーは 1σ を表す

(a) Predictive performance for the validation samples as a function of the number of features. Features are arranged in descending order of relevance. The black dashed line corresponds to the top 25 features. (b) Parity plot between true and predicted β_n values of the validation data using the best GPR model with the top 25 features. Error bars indicate 1σ uncertainty of the predicted values

4.3 選定された特徴量について

図3より、 β_1 のモデルでは、電気陰性度と電荷が影響度の大きい特徴量であることが示された。金属と配位子の間の結合は、イオン結合的なものから共有結合的なものまで含まれる。電荷や電気陰性度はそれらの結合の強弱の理解に用いられることがしばしばある。その点から β_1 の予測モデルにこれらの特徴量が高い影響度を示したことは、極めて合理的である。また、この結果は従来の報告⁵⁻⁷⁾とは異なり、様々な金属カチオンを含んだデータセットを用いたことで明らかになったと考えられる。

β_n のモデルでは、 β_1 が大きな影響度を持つ特徴量であることが示された。 β_1 の予測モデルの特徴量から、 β_1 が金属カチオンと配位子間の結合力を反映していると考えられる。 β_n と β_1 の相関を有していることは妥当であると考えられる。また、 β_n の影響度の高い特徴量としてNaOが入っていたが、この特徴量は分子構造の中の酸素に起因する特徴量である。配位子中の-O-が金属カチオンに対する配位部になることがあるが、このような構造に起因する特徴により、その影響度が高くなったものと考えられる。

上述のように、作成された $\beta_1 \cdot \beta_n$ のGPRモデルは、化学的観点からみても妥当な特徴量が高い影響度を持つモデルであることが確認できた。

5. おわりに

本報では、データセットから出来るだけ汎化性の高いモデル化を行うために、モデルの適用範囲の端のデータを検証用のデータとして抽出する方法や特徴量選定など、

個々の材料開発に応じたアルゴリズム開発に関する我々の取組例として、金属錯体の錯安定度定数を予測するモデル開発を行った事例を紹介した。この取組では、幅広い金属と配位子を含んだデータセットを用いて、金属錯体の錯安定度定数を予測するガウス過程回帰モデルを構築した。検証データの抽出やKullback-Leibler情報量を用いた特徴量の影響度の評価を通して、高い汎化性能を有したモデルを構築することができた。またモデル中で高い影響度を持った特徴量は、化学的な観点からも妥当な特徴量であることが確認され、この点からも汎化性能を有したモデルが構築できたと考えている。本報で紹介したモデルはガウス過程回帰モデルであり、材料開発でも広く用いられるベイズ最適化に適用でき得ることから、種々の用途に応じた金属錯体や配位子を幅広い候補の中から効率的に選定するために用いることが可能であると考えられる。

近年、材料開発やプロセス開発には、従来に増して高度な性能・機能・制御が要求されている。このような要求に答えるためには、従来探索されてこなかった領域や組み合わせの中から最適な材料やプロセス条件を、迅速・効率的に見出していくことが求められる。そのためには従来の開発手法に加えて、データ科学に基づくMIやプロセス・インフォマティクスの活用が必須である。著者らが所属しているCAE領域では、従来からシミュレーションを用いた研究・開発を進めているが、今後もシミュレーションとの連携も含めたデータ駆動型の研究開発を材料とプロセスの双方に対して進めていき、材料開発の迅速化・効率化に貢献していきたい。

謝 辞

本報文で紹介した金属錯体の錯安定度定数のモデル開発における金橋魁利研究員（現東京大学大学院工学系研究科助教）の多大なる貢献に対し感謝の意を表します。また、実験データ整理の遂行・助言を頂きました小田恭宏副主任研究員（現イノベーションセンター企画部 DX 推進グループ長補佐）にも感謝いたします。

文 献

- 1) 茂本 勇 等, “マテリアルズ・インフォマティクス Q A 集—解析実務と応用事例”, 情報機構 (2020).
- 2) 曾根純一, 福井弘行, “科学技術未来戦略ワークショップ報告書, 材料創製技術を確認するプロセス・インフォマティクス”, 研究開発戦略センター, (2021) 1 章, <https://www.jst.go.jp/crds/pdf/2020/WR/CRDS-FY2020-WR-12.pdf>
- 3) 倉田悠都, 三菱マテリアル中央研究所 R&D レビュー, **9**, 39-41 (2020).
- 4) K. Kanahashi, M. Urushihara and K. Yamaguchi, *Sci. Rep.*, **12**, 11159 (2022).
- 5) S. Chaube, S. Goverapet Srinivasan and B. Rai, *Sci. Rep.*, **10**, 14322 (2020).
- 6) V. Solov'ev, N. Kireeva, S. Ovchinnikova and A. Tsivadze, *J. Incl. Phenom. Macrocycl. Chem.*, **83**, 89-101 (2015).
- 7) I. V. Tetko, V. P. Solovev and A. V. Antonov, *J. Chem. Inf. Model.*, **46**, 808-819 (2006).
- 8) K. Terayama, M. Sumita, R. Tamura and K. Tsuda, *Acc. Chem. Res.*, **54** (6), 1334-1346 (2021).
- 9) V. L. Deringer, A. P. Bartok, N. Bernstein, D. M. Wilkins, M. Ceriotti and G. Csanyi, *Chem. Rev.*, **121**, 10073-10141 (2021).
- 10) T. Yamashita, N. Sato, H. Kino, T. Miyake, K. Tsuda, T. Oguchi, *Phys. Rev. Materials.*, **2**, 013803 (2018).
- 11) R. Tamura, T. Osada, K. Minagawa, T. Kohata, M. Hirose, K. Tsuda and K. Kawagishi, *Mater. Des.*, **198**, 109290 (2021).
- 12) R. M. Smith and A. E. Martell, NIST Critically Selected Stability Constants of Metal Complexes Database (NIST Standard Reference Database 46). version 8.0, (National Institute of Standards and Technology, Gaithersburg, MD, 2004), <https://www.nist.gov/srd/nist46>. Accessed 1 March 2022.
- 13) H. Yamada, C. Liu, S. Wu, Y. Koyama, S. Ju, J. Shiomi, J. Morikawa and R. Yoshida, *ACS Cent. Sci.*, **5**(10), 1717-1730 (2019).
- 14) R. D. Shannon, *Acta Cryst. A*, **32**, 751-767 (1976).
- 15) H. Moriwaki, Y.-S. Tian, N. Kawashita and T. Takagi, *J. Cheminformatics*, **10**, 1-14 (2018).
- 16) A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimselshin, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, *Adv. Neural Inf. Process. Syst.*, **32**, 8024-8035 (2019).
- 17) J. R. Gardner, G. Pleiss, K. Q. Weinberger, D. Bindel and A. G. Wilson, *Adv. Neural Inf. Process. Syst.*, **31**, 7576-7586 (2018).
- 18) H. Zhang, *J. Am. Stat. Assoc.*, **99**, 250-261 (2004).
- 19) L. McInnes, J. Healy and J. Melville, Preprint at <https://arxiv.org/abs/1802.03426>
- 20) N. M. Quang, N. T. A. Nhung and P. V. Tat, *Vietnam J. Chem.*, **57**, 500-506 (2019).



漆原 誠 Makoto URUSHIHARA
CAE 領域
主任研究員



山口 健志 Kenji YAMAGUCHI
CAE 領域
サブテクニカルマネージャー／主任研究員
博士 (理学)