

令和 6 年度化学物質規制対策

「大学・公的研究機関と連携した化学物質管理高度化推進事業」

AI 技術を用いた魚類における
「代謝されにくさ」を決定づける化学物質の特徴探索

調査報告書

令和 7 年 2 月
国立大学法人 九州工業大学

目次

はじめに.....	3
実施体制.....	4
1. 調査の目的.....	5
2. 調査方法.....	5
2.1 データ収集.....	5
2.2 データ統合.....	5
2.3 カテゴリー分類.....	5
2.4 分子記述子の生成.....	6
2.5 モデル構築のワークフロー.....	6
2.6 特徴量選択.....	7
2.7 機械学習モデル.....	7
2.8 精度評価指標.....	8
3. 調査結果.....	10
3.1 各分子記述子と BCF の関係性.....	10
3.2 特徴量選択.....	10
3.3 回帰モデルの構築と評価.....	11
3.4 SHAP による特徴量の解釈.....	12
3.5 予測誤差が大きい化学物質の構造.....	12
4. 考察.....	13
4.1 調査結果の意義.....	13
4.2 研究の限界と今後の展望.....	13
5. 付録（データ・図表）.....	15
参考文献.....	24
調査結果の発表状況.....	25

はじめに

化学物質の安全性評価において、生物蓄積性は重要な指標の一つである。特に、化学物質の「代謝されにくさ」は、その体内残留時間に影響を与え、結果として蓄積性や毒性リスクを決定づける要因となる。従来、生物濃縮係数（BCF）を指標とした蓄積性の評価が行われてきたが、その実測試験には膨大な時間・コスト・労力が必要であり、また動物実験に関する倫理的な問題も指摘されている。このような背景から、AI 技術を活用した化学物質の蓄積性評価手法の開発が求められている。

近年、機械学習や深層学習を用いた BCF 予測モデルが報告されている。従来のモデルは予測精度の向上を重視する一方で、どの特徴が「代謝されにくさ」に寄与するかの解釈が困難であるという課題があった。そこで本研究では、AI を用いた予測モデルを構築し、魚類における「代謝されにくさ」を決定づける化学物質の特徴を探索することを目的とする。本研究では、この課題を解決するため、**カテゴリーアプローチと説明可能な AI（XAI）を導入した「代謝されにくさ」予測モデルを構築する**。具体的には、以下のアプローチを採用する。

カテゴリーアプローチによる予測精度の向上

化学物質を物理化学的特性（ファンデルワールス力、イオン性など）に基づき分類し、それぞれに最適化された AI モデルを構築する。

説明可能な AI（XAI）手法の導入

SHAP（Shapley Additive Explanations）を用いて、各特徴量が予測に与える影響を定量的に評価し、「代謝されにくさ」に寄与する化学物質の特徴を明確化する。

実験的検証

AI モデルの予測精度を検証するため、魚類肝脾 S9 を用いた *in vitro* 代謝試験を実施する。

本研究の成果により、魚類における化学物質の代謝特性をより正確に理解し、新たな蓄積性評価の基準を提供することが期待される。さらに、本手法は生物濃縮試験の代替手法として活用され、環境リスク評価や化学物質管理の高度化に貢献できる可能性がある。

実施体制

研究統括：

飯田 緑（国立大学法人 九州工業大学）

A I 技術を用いたモデル構築のためのワークフロー作成：

前田 和勲（国立大学法人 九州工業大学）

コイ肝脾 S9 を用いた代謝実験：

田上 瑠美（国立大学法人 愛媛大学）

コイ肝脾 S9 を用いた代謝実験に供する試料の養成：

仲山 慶（国立大学法人 愛媛大学）

1. 調査の目的

本研究の最終的な目標は、生物蓄積性を決定づける化学物質の特徴を同定し、実証検証を行うことで、新たな蓄積性の判断基準となりうる指標を提示することである。

2. 調査方法

2.1 データ収集

本研究では、製品評価技術基盤機構（NITE）によって公開されている蓄積性データを利用した。このデータは、生物濃縮性を TG305 などの国際評価基準に基づいて評価している信頼性の高い情報源である。公開されているデータは、測定年代の違いによって二種類ある（000064113.xlsx、000134734.xlsx）。それぞれ、1569 化学物質、1185 化学物質の情報が登録されている。

2.2 データ統合

本研究では、000064113.xlsx、000134734.xlsx の 2 つのファイルに存在する化学物質のうち、評価されている生物濃縮性が BCF のものだけを使用した。また、化学構造を簡略化して文字列で表現する形式である SMILES（Simplified Molecular Input Line Entry System）があり、混合物ではないものを抽出した。さらに、複数濃度区に対しそれぞれ BCF が計測されている場合、BCF 値を平均して、1 つの SMILES に対して 1 つの BCF となるようにデータを統合した。この結果、1508 物質に関する BCF 値を取得した（データセット 1：本研究で使用した物質のリスト）。

2.3 カテゴリー分類

BCF の予測精度の向上には、カテゴリー分類が有用であることが報告されている（製品評価技術基盤機構, 2009）。そこで本研究では、1508 物質を 2 つの方法で分類した。1 つ目は、化学物質を物理化学的特性であるイオン化するかどうかでカテゴライズする方法である。1508 物質のうち、イオン化する物質は 176 物質、イオン化しない物質は 1332 物質であった。このため、以降の解析にはイオン化しない物質の BCF のみを使用した（データセット 2：本研究で使用したイオン化しない物質のリスト）。

2 つ目は、化学物質を CERI で提案されている 5 つのカテゴリへ分割する方法（CERI 分類法）である（製品評価技術基盤機構, 2009）。この方法では、化学物質を以下の 5 つのグループに分類した：「脂肪族・芳香族炭化水素およびそのハロゲン化物」「エーテル、ケトン、アミド、アルコール」「カルボン酸、アミン」「チオール」「エステル、リン酸エステル」。この分類では、3.4 に示した分子記述子の計算方法を用いて取得された分子記述子に基づき分類を行った：「脂肪族・芳香族炭化水素およびそのハロゲン化物」

fr_alkyl_halide, fr_halogen, fr_benzene, fr_alkyl_carbamate、「エーテル、ケトン、アミド、アルコール」fr_ether, fr_ketone, fr_amide, fr_Al_OH, fr_ketone_Topliiss, fr_priamide、「カルボン酸、アミン」fr_COO, fr_NH2, fr_Al_COO, fr_Ar_COO, fr_COO2、「チオール」fr_thiol、「エステル、リン酸エステル」fr_ester, fr_phosphate。表 1 には、この方法で分類した各化学物質のカテゴリに含まれるデータ数を示している。今回のデータセットには、チオールとエステル・リン酸エステルは含まれていなかったため、結果として 3 カテゴリとなった。CERI 分類法による BCF 予測モデルの構築の詳細については、大槻卒業論文（資料 1：令和 6 年度九州工業大学卒業論文 大槻竜裕）に記載している。

2.4 分子記述子の生成

分子記述子とは、化学構造を数値化し、分子の物理化学的性質や構造的特徴を定量的に表現する指標である。本研究では、SMILES を使用して化学構造を入力し、RDKit (ver. 2024.03.6) を用いて各化合物の分子記述子を生成した。431 個の分子記述子が計算された。このうち、“AUTOCORR2D_”から始まる分子記述子は、2 つ以上の分子記述子を用いて合成された値のため解釈性が低い。そこで、“AUTOCORR2D_”から始まる分子記述子を除去し、分子記述子数は 201 個となった（データセット 2：本研究で使用したイオン化しない物質のリスト）。

CERI 分類法による BCF 予測モデルの構築（資料 1）では、RDKit (ver.2020.09.01) を用いて、211 個の分子記述子を計算している。

2.5 モデル構築のワークフロー

本研究では、これまで我々が提案してきた「PPAR α との結合に重要な PFAS の特性を解明する説明可能な機械学習アプローチ」(Maeda et al., 2024) を応用した。提案手法のステップは Maeda et al., 2024 の Figure 1 に示されており、図 1 に転記した。ここでは、簡単にワークフローについて記載する。

まず、製品評価技術基盤機構（NITE）のデータベースから取得したイオン化しない 1332 物質について 3.4 「分子記述子の生成」セクションに示した通り分子記述子を計算した。その後、データをランダムにトレーニングセットとテストセットに分割した。このとき、トレーニングセットとテストデータの割合は 8：2 となるようにした（データセット 2：本研究で使用したイオン化しない物質のリスト）。それぞれのデータを次に、「特徴選択」セクションで説明するように、体系的かつ客観的な特徴選択を実施した。選択・ランキングされた分子記述子を用いて複数の機械学習モデルを構築し、高い予測性能を持ち、かつ使用する分子記述子が少ない最適なモデルを選択した。モデルの開発および選択はトレーニングデータセットのみを使用して行った。最後に、選択されたモデルを SHAP を用いて解析し、BCF を予測する上で重要な特性を明らかにした。

2.6 特徴量選択

機械学習モデルのトレーニング前に、以下の3つの理由から特徴選択を行った。本研究の特徴選択手法は、Comesana らによって提案された体系的な特徴選択手法(Comesana et al., 2022)を修正したものである。

- ・分子記述子の数を減らすことで、より解釈しやすい機械学習モデルを開発できる。
- ・少数の分子記述子を使用したシンプルなモデルにすることで、過学習 (overfitting) を回避できる。
- ・分子記述子の数を減らすことで、特徴の多重共線性を低減できる。

まず、95%以上の分子で同じ値を持つ分子記述子を削除した。

次に、相関のある記述子の削除をした。このため、記述子間の相関を評価するために、スピアマン順位相関係数 (Spearman's rank correlation coefficient) を用いて記述子の相関行列を構築した。次に、絶対値が 0.7 を超える相関係数を持つ記述子の数をカウントし、最も多くの記述子と相関している記述子を選択し、それと相関のある記述子をすべて削除した。その後、相関行列を再生成し、このプロセスをすべての記述子を選択または削除されるまで繰り返した。

最後に、再帰的特徴消去 (RFE) による特徴ランキングを行った。選択された記述子を再帰的特徴消去 (Recursive Feature Elimination, RFE) を用いてランキングした。まず、選択されたすべての記述子を用いてランダムフォレストモデルをトレーニングし、不純物

(impurity) に基づく特徴の重要度を計算した。その後、最も重要度の低い記述子を削除し、この手順をすべての記述子なくなるまで繰り返した。最終的に、記述子は削除された順番とは逆に昇順でランク付けされた。

テストデータセットを特徴選択プロセスに含めると、データリーク (data leakage) が発生し、モデルの性能が過大評価される可能性がある。そのため、本研究では、特徴選択のための分子記述子の選択にはトレーニングデータセットのみを使用した。RFE には Python (3.9.16) と scikit-learn (1.3.0) を使用した。

2.7 機械学習モデル

アンサンブル学習は、トレーニング中に複数の決定木を組み合わせて予測を行う学習手法である。各決定木はランダムに選ばれた特徴とデータのサブセットを用いて構築されるため、過学習を抑制し、頑健なモデルを構築できる。

本研究では、高い汎化性能と計算効率のバランスを考慮し、機械学習モデルとしてランダムフォレスト (Random Forest) と勾配ブースティング決定木 (Gradient-Boosting Decision Trees) の改良版である XGBoost を採用した。具体的には、ランダムフォレスト回帰と XGBoost 回帰を使用した。

2.8 精度評価指標

機械学習モデルの予測性能を評価するために、以下の 2 つの指標を使用した。

- ・ 決定係数 (R^2)
- ・ 二乗平均平方根誤差 (RMSE: Root Mean Squared Error)

R^2 は式 (1) によって定義される。

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$$

ここで n はサンプル数 (すなわち化学物質数)、 y_i は i 番目のサンプルの BCF、 $\hat{y}_i$ は y_i の予測値を示す。また、

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

である。 R^2 (決定係数) の値が大きいほどモデルの性能が高いことを示し、もし完全に正確な予測 (すなわち $y_i = \hat{y}_i$ がすべての i で成り立つ) であれば、 R^2 の最大値は 1 となる。

また、モデル評価指標として 二乗平均平方根誤差 (RMSE: Root Mean Squared Error) も使用した (式 (2))。

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

RMSE の範囲は 0 から無限大であり、値が小さいほどモデルの性能が高いことを示す。もし完全に正確な予測ができた場合、RMSE は最小値 0 となる。

2.9 SHAP (SHapley Additive exPlanations)

SHAP (Shapley Additive exPlanations) 値は、ゲーム理論に基づいた特徴重要度の指標である。モデルの予測値は、SHAP 値の合計として表される (式 (3))。

$$f^{(i)} = \phi_0 + \sum_{j=1}^m \phi_j^{(i)} \quad (3)$$

ここで、

m は特徴量の数

$f^{(i)}$ は i 番目のサンプル (化学物質) のモデル予測値 (BCF)

ϕ_0 は トレーニングデータセットにおけるモデル予測値の期待値

$\phi_j^{(i)}$ は i 番目のサンプルにおける j 番目の特徴量 (分子記述子) の SHAP 値 を示す

また、

$\phi_j^{(i)} > 0$ の場合、j 番目の特徴量は i 番目のサンプルのモデル予測値を増加させる

$\phi_j^{(i)} < 0$ の場合、j 番目の特徴量は i 番目のサンプルのモデル予測値を減少させる

SHAP 値は、各特徴量がモデルに追加される際の影響を、すべての特徴量の導入順序にわたって平均化した重み付き和として計算される (式 (4))。

$$\phi_j^{(i)} = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{j\}}(x_{S \cup \{j\}}^{(i)}) - f_S(x_S^{(i)}) \right] \quad (4)$$

ここで、

$\phi_j^{(i)}$ は i 番目のサンプルにおける j 番目の特徴量の SHAP 値

F は全特徴量の集合

$F \setminus \{j\}$ は j 番目の特徴量を除いた全特徴量の集合

S は $F \setminus \{j\}$ のあらゆる部分集合

$f_{S \cup \{j\}}(x_{S \cup \{j\}}^{(i)})$ は j 番目の特徴量を含む状態でのモデル予測値

$f_S(x_S^{(i)})$ は j 番目の特徴量を含まない状態でのモデル予測値

SHAP の特徴重要度は、各記述子の全体的な重要性を示す指標であり、SHAP 値の絶対値の平均として定義される (式 (5))。

$$I_j = \frac{1}{n} \sum_{i=1}^n |\phi_j^{(i)}| \quad (5)$$

ここで、n はサンプル数を示す。SHAP 値と SHAP 特徴重要度の計算には、R の future ライブラリ(1.34.0)を使用した。

3. 調査結果

3.1 各分子記述子と BCF の関係性

製品評価技術基盤機構 (NITE) のデータベースから 1,332 種類の化学物質の SMILES データと BCF 値を取得し、RDKit を用いて 201 種類の分子記述子を計算した。分子記述子と BCF データはデータセット 2 (本研究で利用したイオン化しない物質のリスト) に、BCF の頻度分布と統計値は図 2 (本研究で利用したイオン化しない物質の BCF 値の頻度分布) にまとめられている。

頻度分布解析の結果、102 種類の化学物質が 1000 倍 ($\log_{10} 3$) より高い BCF を示した。BCF 値が 1000 倍以上であると高蓄積性であると判断される。本研究データでは、全体の 1 割弱程度が高蓄積性であることが示された。

説明可能な機械学習アプローチを実施する前に、分子記述子と BCF の間に識別可能な関係があるかどうかを確認するため、散布図を用いた予備解析を行った。図 3 に示すように、疎水性を示す指標である MolLogP を含む 3 つの分子記述子と BCF の散布図を作成した。この結果、MolLogP と BCF の間に明確な相関が見られた。この結果は、これまで報告されている結果と同様である。また、BCF は酸素と窒素の数の指標である NOcount と負の相関、分子の電荷状態をしめす MinAbsEStateIndex と正の相関を示した。この結果は、BCF の予測が、複数の特徴量の線形な相互作用によって決定される可能性があることを示唆している。このため、資料 1 では、BCF と線形な相関を持つ特徴量に絞って回帰モデルを構築したが、モデルの精度は 0~0.6 とあまり良くなかった。したがって、本研究では、非線形な関係性も考慮にいれ、説明可能な機械学習アプローチを適用する必要があると判断した。

3.2 特徴量選択

本研究では、BCF のデータセットを 8:2 の割合でトレーニングデータとテストデータに分割した。トレーニングデータセットは、特徴選択、モデル開発、およびモデル選択に使用した。一方、テストデータセットは、開発した機械学習モデルの性能評価のみに使用した。

少ない記述子を使用することで、機械学習モデルの解釈性が向上する。そのため、説明可能な機械学習アプローチの最初のステップとして、特徴選択 (Feature Selection) を実施した。まず、多くの化学物質において同じ値を持つ分子記述子は情報量が少ないため、95% 以上の化学物質で値が一致する分子記述子を削除した。この処理の後、151 種類の分子記述子が残った。それらのリストと相関行列をデータセット 3 (トレーニングデータの相関行列) に示した。次に、スピアマン相関係数 (Spearman's correlation coefficient) の絶対値が 0.7 を超える相関のある分子記述子を削除した。この段階で、記述子の数は 75 種類まで削減された。最後に、RFE (Recursive Feature Elimination, 再帰的特徴消去) を用いて、

これら 75 種類の分子記述子をランキングした。選択された 75 種類の分子記述子は データセット 4 (トレーニングデータの分子記述子の重要度ランキング) に示されている。

3.3 回帰モデルの構築と評価

次に、選択した分子記述子を用いて機械学習モデルを開発した。具体的には、1、2、3、4、5、10、20、50、75、および 201 種類の分子記述子を使用した 10 種類の機械学習モデルを構築した。第 1 モデルでは、特徴選択で 1 位にランク付けされた分子記述子のみを使用した。第 2 モデルでは、1 位と 2 位の分子記述子を使用した。第 3 モデルから第 9 モデルも同様に、それぞれ異なる数の上位ランクの分子記述子を用いて構築した。第 10 モデル (201 種類の分子記述子) は、RDKit によって計算された “AUTOCORR2D” 以外のすべての分子記述子の特徴選択なしで使用した。したがって、このモデルはベースラインモデルとみなすことができる。

機械学習アルゴリズムには Randomforest 回帰を採用した。これは、XGBoost 回帰 (XGBoost) モデルよりも性能が良かったためである。また、モデルの性能評価にはトレーニングデータセットを用いた 5 分割交差検証 (5-fold cross-validation) を実施した。図 4 に、開発した各モデルの予測性能を示している。4 つ以上の分子記述子を使用したモデルとベースモデル間では、統計的に有意な性能差は見られなかった (t 検定、 $p > 0.05$)。一方で、3 つの記述子を用いたモデルの R^2 値は、ベースモデルと比較して有意に低かった (t 検定、 $p = 0.007$)。一般に、予測性能 R^2 と解釈性 (使用する記述子の数) にはトレードオフが存在するため、最適なバランスを見極めることが重要である。本研究では、十分な予測性能を持ち、かつ使用する分子記述子の数が少ないモデルとして、4 つの分子記述子を用いたモデルを選択した。この選択されたモデルを、後続の解析 (downstream analysis) に使用した。

図 5 は、選択された機械学習モデルによって予測されたテストデータセットの BCF と、NITE によって測定された実際の BCF の散布図を示している。各データポイントは化学物質を表している。データポイントが対角線付近に分布しており、モデルの予測は大まかにできていることを示している。決定係数 (R^2 値) は 0.39 と予測モデルとしては、精度はあまり良くない。一方で、RMSE (二乗平均平方根誤差) は 0.71 と 1 よりも小さかった。同じデータセットを EPI Suite で予測したところ (データセット 5 EPI Suite による予測)、予測の誤差を示す RSME は 1.04 となり、本モデルの方が予測の誤差は小さいことが示唆された。

これらのことから、本研究で構築したモデルは、予測の精度は中程度だが、BCF の予測誤差を一桁以内 (10 倍以内) に抑えられており、大きな予測誤差はあまりないモデルであることが示唆された。

3.4 SHAP による特徴量の解釈

開発した機械学習モデルにおいて使用した 4 つの分子記述子の重要度を評価するため、SHAP 重要度を計算した。その結果、MolLogP（疎水性）が最も重要な特徴であると予測された（図 6）。これは、既報とも一致する（Arnot and Gobas, 2006; Kobayashi and Yoshida, 2021; Lunghini et al., 2019）。

2 番目に重要な記述子は NOcount であり、これは窒素原子と酸素原子の数であり、分子量などの物理化学的特性とも相関している。

3 番目に重要な記述子は BCUT2D_CHGHI であり、分子内の電荷分布の特徴を数値化したものである。

4 番目に重要な記述子は MinAbsEStateIndex であり、分子の電子的特性を数値化したものである。

総合すると、1332 種類の化学物質を用いた機械学習に基づく統計解析により、MolLogP、NOcount、BCUT2D_CHGHI および MinAbsEStateIndex の 4 つの記述子が重要であることが明らかになった。この結果は、疎水性、分子サイズおよび分子内の電気的な特性が化学物質の蓄積性予測において重要であるというこれまでの QSAR モデルとも一致する（Kobayashi and Yoshida, 2021; Pore et al., 2024）。

次に、分子記述子が BCF 予測に与える影響を調査した。図 7 に SHAP 値と分子記述子の値の散布図を示す。

MolLogP（疎水性）が 2.5 以下または 7.5 以上の時は、BCF に負の影響を及ぼす。一方で、2.5～7.5 までは、BCF 予測に正の影響を及ぼしていることが明らかとなった。また、5 付近で SHAP 値はピークを示した。このことから、疎水性の値は 5 付近のものが、最も BCF が高くなる傾向が示唆された。

NOcount は 2 以上となると、BCF に負の影響を与えていた。この影響は、数値が大きくなるほど強くなっており、分子サイズが大きくなると、BCF は下がる傾向にあることが示唆された。

BCUT2D_CHGHI は 2.5 以上となると SHAP 値が正の値を取ることから、BCUT2D_CHGHI が 2.5 以上の分子は BCF が高くなる傾向があることが示唆された。

MinAbsEStateIndex は他の記述子と比較して、SHAP 値が小さいことから、通常は MinAbsEStateIndex が BCF に及ぼす影響は小さいと考えられる。また、MinAbsEStateIndex が 4 を超えると、SHAP 値が最大となっていることから、MinAbsEStateIndex は 4 以上の分子は BCF が高くなるとが示唆された。ただし、テストデータセットに含まれる 225 化学物質のうち 2 つのみが MinAbsEStateIndex > 4 となっていることから、この結果の解釈には注意が必要である。

3.5 予測誤差が大きい化学物質の構造

最後に、どのような特徴を持つ化学物質の予測が外れやすいのか、予測誤差が大きい化学物質について、その構造的特徴を調査した。図 8 に予測誤差が大きかった化学物質を色付

けした予測 BCF と実測 BCF 値の散布図を示す。図 8 を作成するためのデータは、データセット 6（予測と実測の BCF 値）に示す。

図 8 のピンクで示した化学物質は予測よりも実測値が 10 倍以上小さいことを示す。225 個のテストデータセットのうち、21 個の化学物質が予測よりも実測値が 10 倍以上小さかった。このことから、これらの化学物質は体内で代謝されやすい物質であると考えられる。図 8 の青で示した化学物質は予測よりも実測値が 10 倍以上大きいことを示す。225 個のテストデータセットのうち、21 個の化学物質が予測よりも実測値が 10 倍以上大きかった。また、このうち 2 つの化学物質については、実測値が予測値よりも 100 倍以上大きかった。このことから、これらの化学物質は体内で代謝されにくい物質であると考えられた。最も予測が大きく外れていた物質は、2-（2H-ベンゾトリアゾール-2-イル）-6-sec-ブチル-4-tert-ブチルフェノール（36437-37-3）で、実測の $\log_{10}(\text{BCF})$ が 4.16 に対して予測した $\log_{10}(\text{BCF})$ は 1.21 だった。次に予測が大きく外れていた物質は、2-(2H-ベンゾトリアゾール-2-イル)-4,6-ビス(1,1-ジメチルプロピル)フェノール（25973-55-1）で、実測の $\log_{10}(\text{BCF})$ が 3.25 に対して予測した $\log_{10}(\text{BCF})$ は 1.25 だった。これらはどちらもフェノール基を持つことから、分子サイズから予測されるよりも、疎水性が高いと考えられた。

4. 考察

4.1 調査結果の意義

本研究では、機械学習を用いた BCF（生物濃縮係数）予測モデルを構築し、化学物質の蓄積性に影響を与える重要な分子記述子を 4 つ特定した。本研究で構築したモデルは、モデルの当てはまりを示す R^2 値の観点からは、中程度の予測程度であったが、従来の EPI Suite モデルよりも予測誤差（RMSE）が低く、安定した結果を示した。さらに、BCF 予測が大きく下に外れた化学物質について詳細な解析を行い、予測誤差が大きい物質の特徴を明らかにした。この結果は、BCF 予測において「代謝の影響」や「特定の官能基による例外的な挙動」を考慮する必要があることを示している。今後の研究では、代謝を考慮したモデルの改良や、分子構造に基づく補正係数の導入が有効と考えられる。

4.2 研究の限界と今後の展望

本研究で得られた知見を基に、以下の改良が求められる：

1. データセットの拡充：より多様な化学物質を含むデータセットを用いてモデルを再構築し、予測の汎化性を向上させる。
2. 代謝影響の考慮：予測誤差が大きかった化合物の代謝経路を詳細に解析し、BCF の修正モデルを開発。

3. 異常値を考慮したモデルの最適化：代謝や特定の官能基に着目した補正手法を導入し、より精度の高い予測を目指す。
4. 予測モデルの検証：予測誤差が大きかった化合物について、魚類肝脾 S9 を用いた *in vitro* 代謝試験を実施する。本年度は、愛媛大学において、S9 代謝試験に供するコイの飼育及び、実験系の確立と手技の習熟を進めており、以降の検証実験を速やかに実施するための準備を整えた。

5. 付録（データ・図表）

図 1 モデル構築フロー

図 2 本研究で使用したイオン化しない物質の BCF 値の頻度分布

図 3 BCF と 3 つの分子記述子の散布図

図 4 BCF 予測モデルの精度評価

図 5 予測された BCF と実際の BCF の散布図

図 6 BCF 予測に重要な記述子

図 7 SHAP 値と分子記述子の値の散布図

図 8 予測誤差が大きかった化学物質の予測 BCF と実測 BCF 値の散布図

表 1 カテゴリー分類（CERI 分類法）

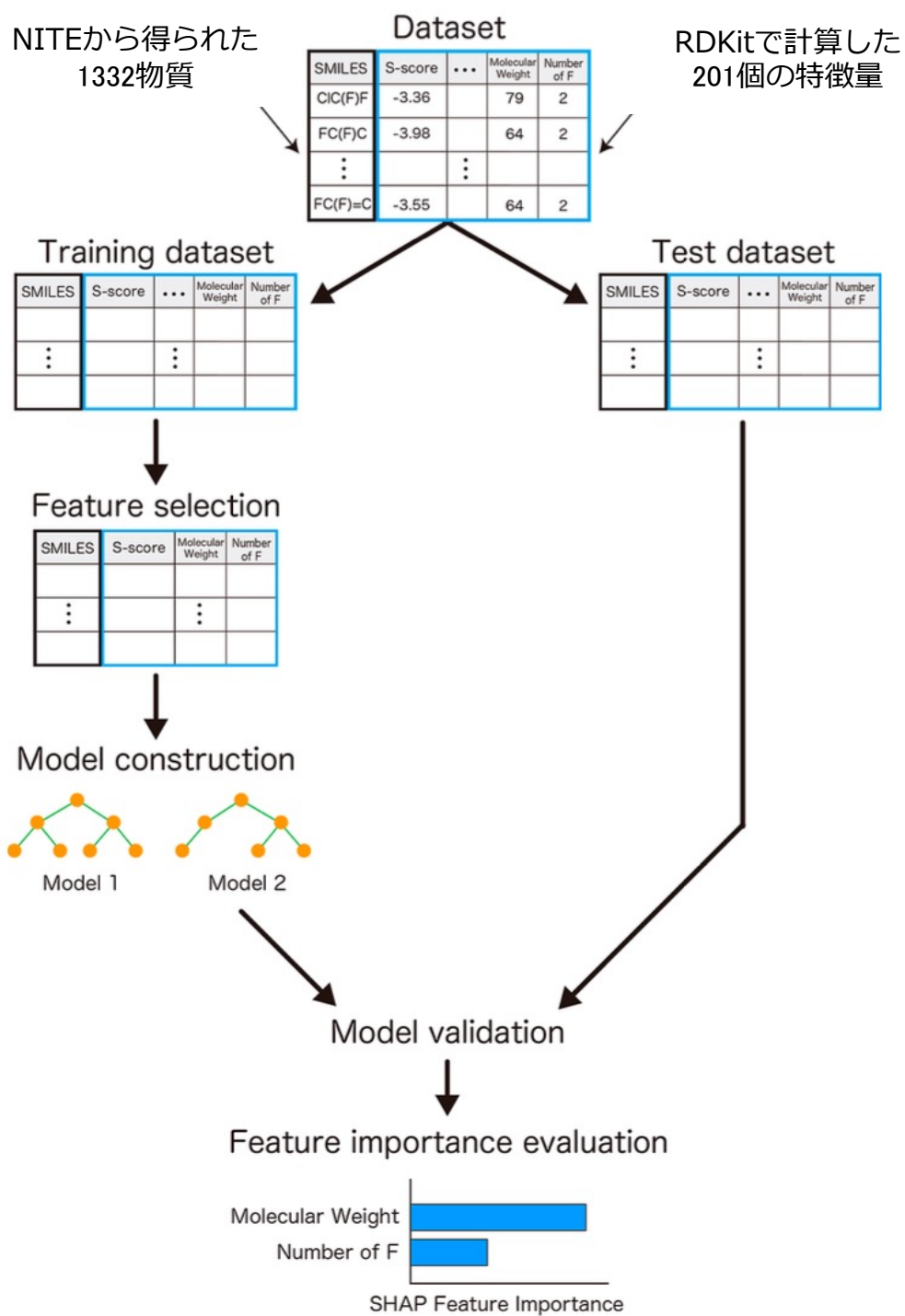


図 1 モデル構築フロー
(Maeda et al., 2024 から一部を変更し記載)

経済産業省以外の第三者の知的財産権が関与するため非公開

図 2 BCF 値の頻度分布と統計値

論文掲載予定のため非公開

図 3 BCF と 3 つの分子記述子の散布図

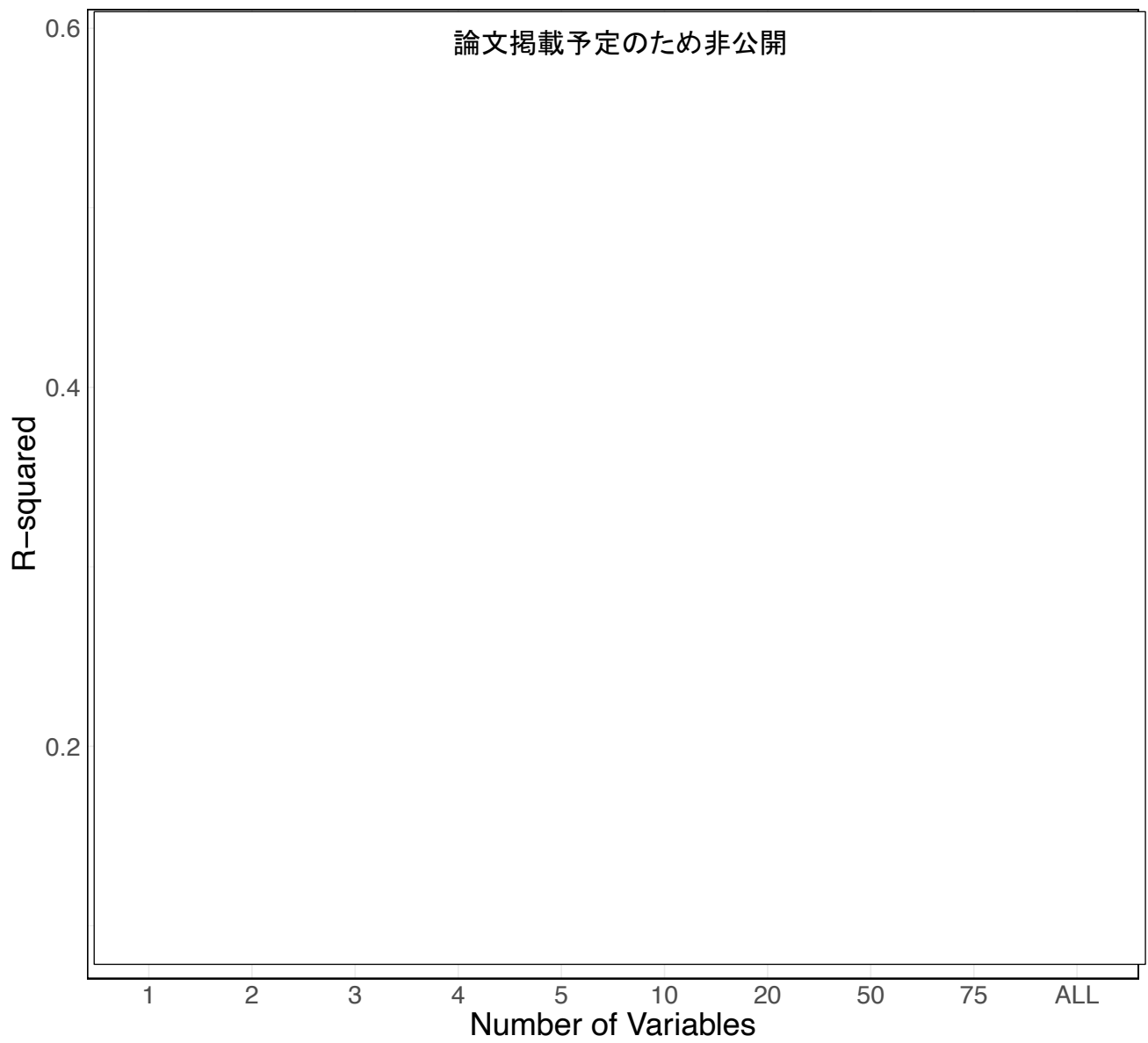


図 4 BCF 予測モデルの精度評価

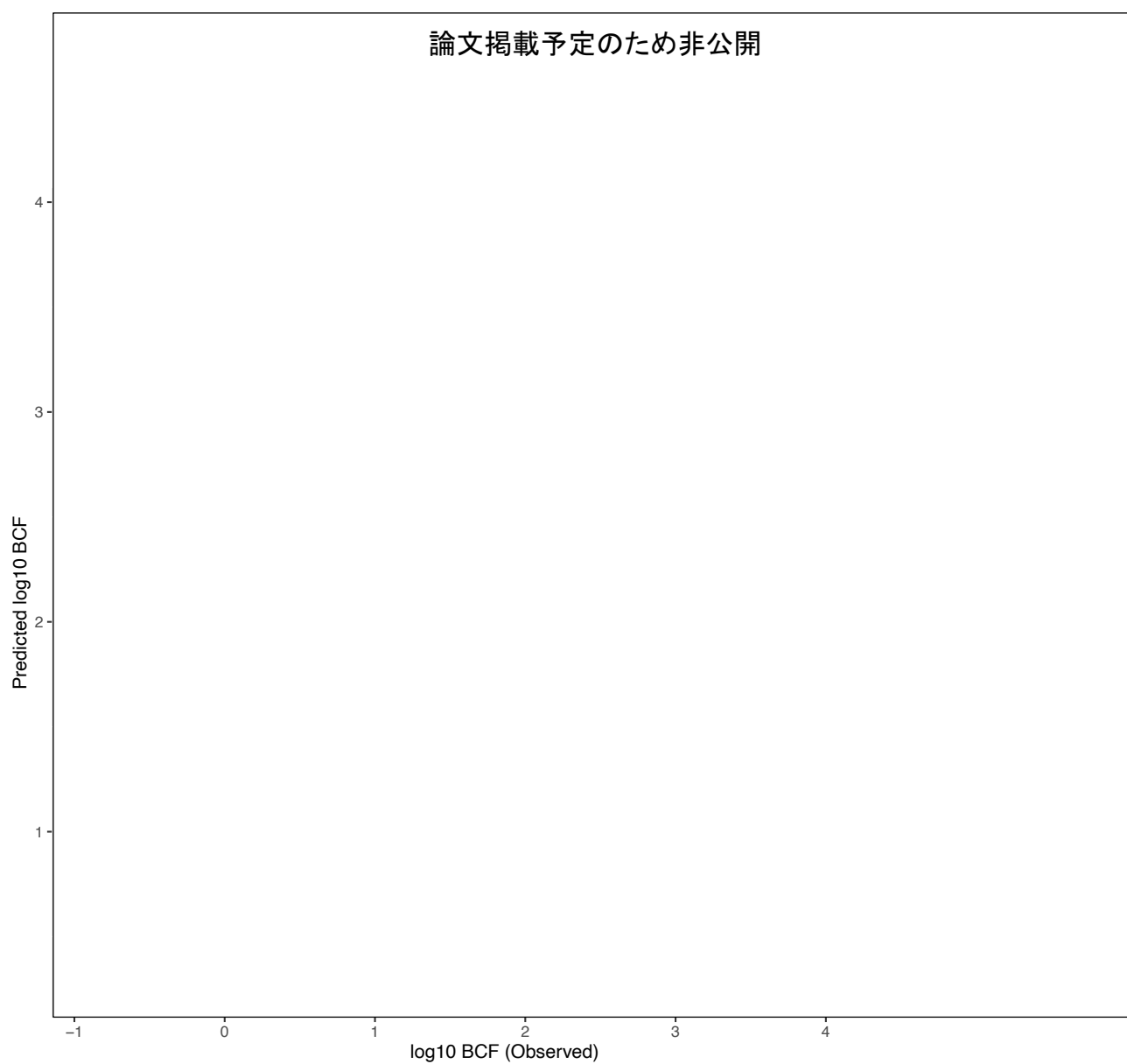


図5 予測された BCF と実際の BCF の散布図

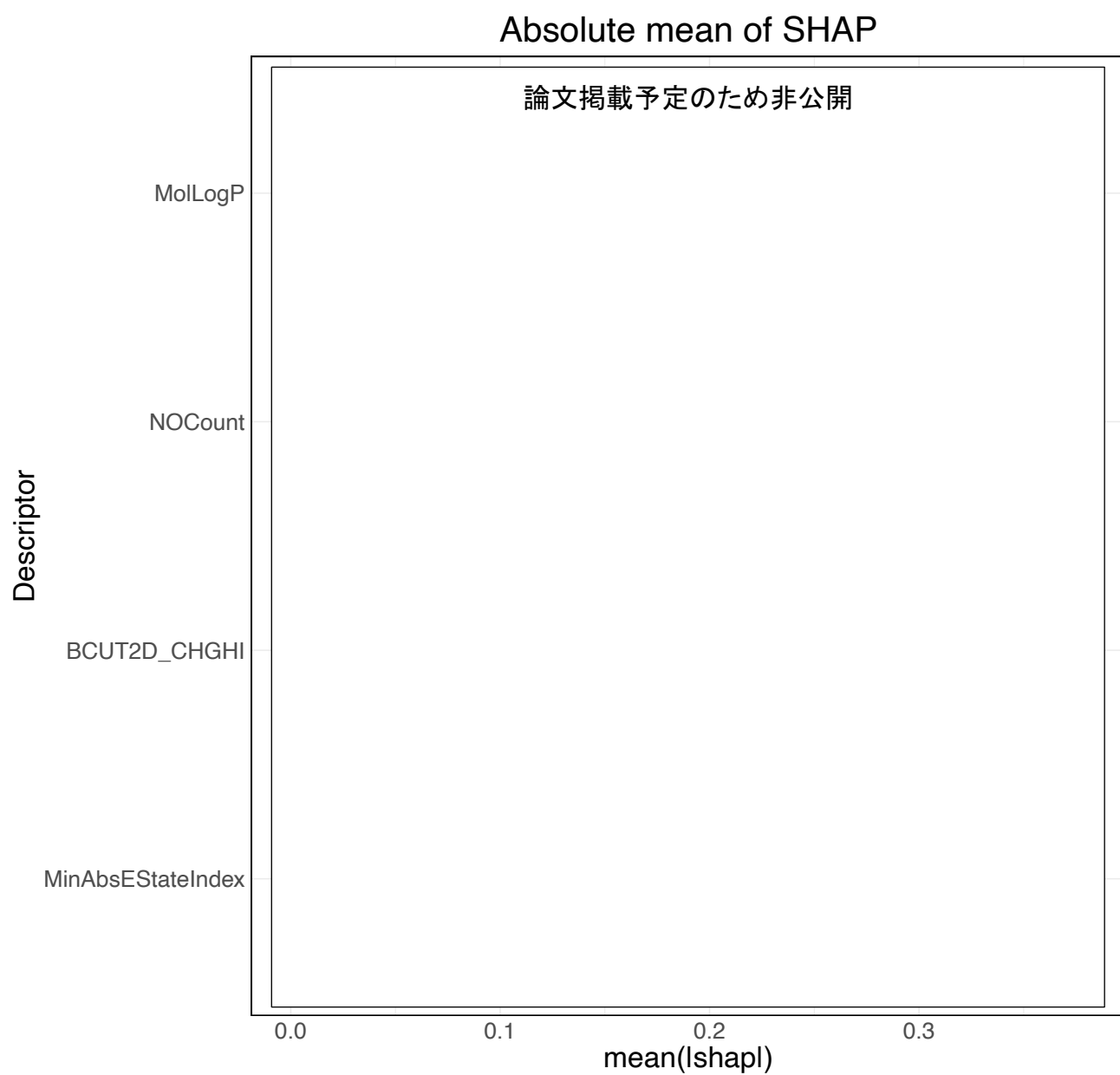


図 6 BCF 予測に重要な記述子

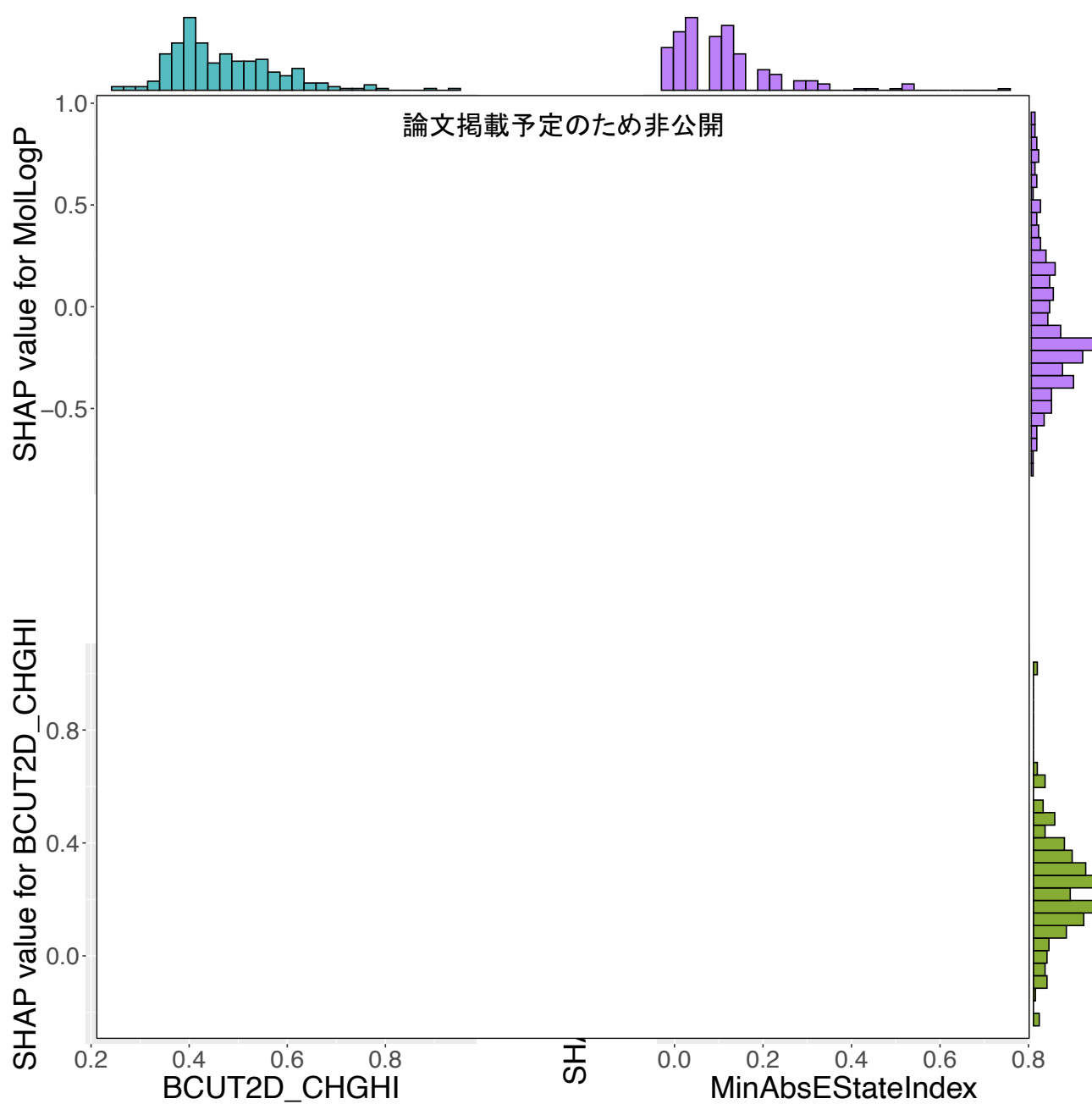


図 7 SHAP 値と分子記述子の値の散布図

SHAP 値が正の場合 → 該当する分子記述子が BCF を増加させる

SHAP 値が負の場合 → 該当する分子記述子が BCF を減少させる

論文掲載予定のため非公開

図 8 予測誤差が大きかった化学物質の予測 BCF と実測 BCF 値の散布図

表 1 カテゴリー分類（CERI 分類法）令和 6 年度九州工業大学卒業論文 大槻竜裕(資料 1 を元に記載)

経済産業省以外の第三者の知的財産権が関与するため非公開

参考文献

- Arnot, J.A., Gobas, F.A.P.C., 2006. A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. *Environ. Rev.* <https://doi.org/10.1139/A06-005>
- Comesana, A.E., Huntington, T.T., Scown, C.D., Niemeyer, K.E., Rapp, V.H., 2022. A systematic method for selecting molecular descriptors as features when training models for predicting physiochemical properties. *Fuel* 321, 123836. <https://doi.org/https://doi.org/10.1016/j.fuel.2022.123836>
- Kobayashi, Y., Yoshida, K., 2021. Development of QSAR models for prediction of fish bioconcentration factors using physicochemical properties and molecular descriptors with machine learning algorithms. *Ecol. Inform.* 63. <https://doi.org/10.1016/j.ecoinf.2021.101285>
- Lunghini, F., Marcou, G., Azam, P., Patoux, R., Enrici, M.H., Bonachera, F., Horvath, D., Varnek, A., 2019. QSPR models for bioconcentration factor (BCF): are they able to predict data of industrial interest? *SAR QSAR Environ. Res.* 30. <https://doi.org/10.1080/1062936X.2019.1626278>
- Maeda, K., Hirano, M., Hayashi, T., Iida, M., Kurata, H., Ishibashi, H., 2024. Elucidating Key Characteristics of PFAS Binding to Human Peroxisome Proliferator-Activated Receptor Alpha: An Explainable Machine Learning Approach. *Environ. Sci. Technol.* 58. <https://doi.org/10.1021/acs.est.3c06561>
- Pore, S., Pelloux, A., Chatterjee, M., Banerjee, A., Roy, K., 2024. Machine learning-based q-RASAR predictions of the bioconcentration factor of organic molecules estimated following the organisation for economic co-operation and development guideline 305. *J. Hazard. Mater.* 479, 135725. <https://doi.org/https://doi.org/10.1016/j.jhazmat.2024.135725>
- 製品評価技術基盤機構, n.d. カテゴリーアプローチによる 生物濃縮性予測に関する報告書 [WWW Document]. URL https://www.mhlw.go.jp/shingi/2009/12/dl/s1218-16m_0001.pdf

調査結果の発表状況

国内学会

1. 大槻 竜裕，仲山 慶，田上 瑠美，飯田 緑，魚の蓄積性に関する研究，金沢.
(2024 年 12 月)