3. 手法

3.1 構築パイプライン

ダイアグラム

自動的に生成された説明

本研究における構築パイプラインの概要を以下に示す。

1. クリプトサイトを持つ構造的に定義されたアポのデータセットを論文[1]に従って構築する。
2. タンパク質表面上のポケット検出ソフトウェアFpocket　　[4]を用いて構築したデータセットに対し、特徴量を作成する。
3. Fpocket[4]では、各タンパク質についてクリプトサイトになり得る凹みを検出するため、ホロ構造を重合せFpocket[4]の生成する点群の一部分とリガンドの距離が3Å以内にある場合、その点群をクリプトサイトになり得る凹みと定義し、ラベリングした。その他の凹みについても検出するために、Fpocket[4]の生成する点群と表面を目視で確認し、少なくとも点群の一部が凹み内部に含まれる時、その点群をその他の凹みと定義し、ラベリングした。
4. 凹み

のラベリングした。

表 ：構築パイプライン

1. 3.までで構築したデータセットを学習データとし、機械学習モデルを用いてクリプトサイトになり得る凹みか否かの分類するモデルを作成する。
2. 機械学習モデルの分類に関して、どの特徴量が分類の寄与度が高いかについて分析する。

　3.2 各工程の詳細

各工程について、詳細に説明する。

構築パイプライン1について、先行研究[1]のSupporting informationのTable S1とTable S5を参照し、それぞれ学習データとテストデータのアポ構造とホロ構造のタンパク質をProtein Data Bank[[1]](#footnote-1)(PDB)からダウンロードし、タンパク質構造データセットを構築した。

構築パイプライン2について、ポケット検出ソフトウェアFpocket[4]について以下、概要を説明する。

Fpocketは、ボロノイテッセレーションとアルファ球に基づいたオープンソースのポケット検出パッケージで、一般に利用可能なパッケージであるQhullの上に構築されている。モジュール式のソースコードは、3つの主要プログラムの基礎となる関数の中央ライブラリを中心に構成されている。

(i) Fpocket：ポケット識別を実行する。

（ii）Tpocket：既知のタンパク質-リガンド複合体のセットでポケット検出のベンチマ　ークを実施する。

(iii) Dpocket：タンパク質のセットでポケット記述子の値を収集する。

FpocketはC言語で書かれており、新しいスコアリング関数を開発したり、大規模なレベルで様々なポケット記述子を抽出したいと考えている科学コミュニティに適したプラットフォームとなっている。

Fpocket[4]のポケット検出アルゴリズム、以下の３つの主要なステップからなる。

(i) 最初のステップでは、アルファ球[[2]](#footnote-2)のアンサンブル全体がタンパク質構造から決定される。Fpocket[4]は、事前にフィルタリングされた球体のコレクションを返す。

(ii) 第2のステップでは、近接した球体のクラスターとポケットを識別して、それ以外のクラスターを削除する。

(iii) 最後のステップでは、各ポケットにスコアを付けるために、ポケットの原子から 特性を計算する。

ダイアグラム

自動的に生成された説明チョコレートチップクッキー

中程度の精度で自動的に生成された説明

図　：Fpocket のワークフロー(左図)、Fpocketによる解析例(赤:予測領域(アルファ球)、黄:実際の結合領域)(右図)

構築パイプライン3について、図　は、アポ構造のクリプトサイトになり得る凹みについて、Fpocket[4]で生成した点群と対応するホロ構造のリガンドと重合せた結果、点群の一部分とリガンドの距離が0Åのため、この点群をクリプトサイトとラベリングする実例を示している。また図　は、その他の凹みについて、Fpocket[4]の生成する点群と表面を目視で確認し、少なくとも点群の一部が凹み内部に含まれたため、その点群をその他の凹みとラベリングする実例を示している。

カラフルな絵

中程度の精度で自動的に生成された説明

図　：アポ構造:1JBUHについて、Fpocket[4]の解析例(青点群)(左)、ホロ構造: 1WUNHについて、リガンド: P5B との結合例(中) 、アポ構造とホロ構造の重合せ例(右)

カラフルな絵

中程度の精度で自動的に生成された説明

図　：アポ構造:3FDLAについて、Fpocket[4]の解析例(紫点群)(左)、ホロ構造: 2YXJAについて、リガンド: N3Cとの結合例(中) 、アポ構造とホロ構造の重合せ例(右)

構築パイプライン4について、構成したデータセットを先行研究[1]と同条件で、学習データ175個、テストデータを36個に分けた。なお、表　に、学習データとテストデータについて、クリプトサイトになり得る凹みとその他の凹みとのデータ数の内訳を示す。本研究では、テーブルデータに対する機械学習モデルとしてよく用いられるRandom Forest、XGBoost、LightGBM、Support Vector Machine(SVM)の４つのモデルを用いることにした。表　に、選定した機械学習モデルとアルゴリズム概要について示す。クリプトサイトになり得る凹みとその他の凹みの２値分類を学習し、性能を比較した。学習時の各モデルのハイパーパラメータチューニングは、Optunaを用いて行った。

表　：学習データとテストデータについて、クリプトサイトになり得る凹みとその他の凹みとのデータ数の内訳

|  |  |  |  |
| --- | --- | --- | --- |
| 学習データ | | テストデータ | |
| クリプトサイトになり得る凹み | その他の凹み | クリプトサイトになり得る凹み | その他の凹み |
| 70 | 105 | 18 | 18 |

　表　：選定した機械学習モデルとアルゴリズム概要

|  |  |
| --- | --- |
| 機械学習モデル | アルゴリズム概要 |
| Random Forest | Random Forestは、決定木に基づいた機械学習の代表的な手法である。重複を許すランダムサンプリングによって多数の決定木を作成し、各ツリーの予測結果の多数決をとることで最終予測値を決定する。 |
| XGBoost | XGBoostは、勾配ブースティングと呼ばれ、弱学習器を逐次的に構築する手法である。Random Forestは決定木を複数個構築し、その結果をアンサンブルする。 前に構築された弱学習器の結果を使用し弱学習器を構築する。 |
| LightGBM | LightGBMは、勾配ブースティングと呼ばれ、弱学習器を逐次的に構築する手法である。決定木の扱いに決定木のleaf（つまり葉）に準じて成長する「Leaf-Wise」を採用することで、他の決定木ベースの手法に比べ、学習時間が高速で、推論精度が高いことが特徴である。 |
| SVM | サポートベクターマシンは、線形入力素子を利用して2クラスのパターン識別器を構成する手法である。 訓練サンプルから、各データ点との距離が最大となるマージン最大化超平面を求めるという基準（超平面分離定理）で線形入力素子のパラメータを学習する。 |

　構築パイプライン５について、本研究では、モデルの予測を解釈するフレームワークであるSHAPを用いて因子分析を行った。SHAPについて、以下に概要を説明する。

概要

* 正式名称は「SHapley Additive exPlanation」で、解釈手法自体を指す場合と手法によって計算された値(SHAP値と呼ぶこともある)を指す場合がある。
* モデルの予測結果に対し、「なぜそのように予測したか」を解釈する局所的な説明手法。
* 予測結果に対して、どの説明変数が大きく寄与したかを明らかにする。
* Explainable AI(XAI)における局所的な説明手法としてデファクトスタンダードである。

これまでの解釈手法Additive Feature Attribution Methods(後述)に協力ゲーム理論のShapley Value(後述)を導入して改良したもので、計算効率の向上と実用上の望ましい数学的性質(後述)を満たしている。

* 協力ゲーム理論において、Shapley Valueとは各プレイヤーの貢献度合いに応じて利益を分配する指標。
* 機械学習モデルの各特徴量をプレイヤーに見立ててShapley Valueを計算することで各特徴量の貢献度合いを評価する手法。

各特徴量のSHAP値はサンプルのそれぞれのレコードに対して算出可能である。

* 特定のレコードに対して予測結果の解釈が可能(ミクロ的解釈)
* サンプル全体で各特徴量のSHAP値の分布や平均を見ることで、モデル内での各特徴量の予測値に対する影響の仕方を把握できる(マクロ的解釈)

**Additive Feature Attribution Methods**

* : 元の予測モデルでに対する局所的な代替モデルを示す。
* : 特徴量の有無によりどれだけ最終的な予測結果に影響があるかを示す。
* : モデルの平均的な予測結果E

**Shapley Value**

* が加わることによって生じる追加的な貢献度を限界貢献度と呼び、この限界貢献度を用いて計算される。
* 限界貢献度：
  + プレイヤー が加わることによって生じる追加的な貢献度のこと。
  + プレイヤーの部分集合Sに対して, v(S)をSのメンバーが協力することによって獲得できる利益とした場合に、任意の∈Nと任意の部分集合S⊆N∖{}に対して、プレイヤーの限界貢献度はv(S∪{})−v(S)で表される。
  + プレイヤー(∈N)のShapley Valueは以上の式で定義される( |S| は任意の部分集合S⊆Nに対してSの要素数)。

**機械学習モデルの解釈を目的とした場合のSHAP値**

* 機械学習モデルの解釈においては、協力ゲーム理論における各プレイヤーを機械学習モデルの各特徴量に置き換えてSHAP値を計算する。
* ある特徴量以外の全ての特徴量の組み合わせについて、予測モデルの出力が特徴量を含むことで生じる予測結果の差を評価する。

　　　　計算式は以下の通りになる。

**SHAPの性質**

　　　SHAPには以下3点の性質があり、この3点を満たす説明モデルはただ1つとなることがわかっている (SHAPの主定理)。

1: Local accuracy

* 説明対象のモデル予測結果=特徴量の貢献度の合計値(SHAP値の合計)の関係

2: Missingness

* 存在しない特徴量()は影響しない。

　　　　3: Consistency

* 任意の特徴量がモデルに与える影響が大きくなるようにモデルを変更した場合に、その特徴量の貢献度が減少しない。

1. 蛋白質、核酸、糖鎖など生体高分子の3次元構造の原子座標を蓄積している国際的な公共のデータベースである。PDBに蓄積されている構造データは、結晶解析法、核磁気共鳴法、クライオ電子顕微鏡法の3つの検証可能な手法によって実験的に決定されたデータである。 [↑](#footnote-ref-1)
2. 境界で4つの原子に接触し、内部の原子を含まない球。 [↑](#footnote-ref-2)