4. 結果

4.1 クリプトサイトになり得る凹みとその他の凹みのデータ分布比較

図　は、作成データセットに対して、Fpocket[4]より得られた全19の特徴量の分布(ヴァイオリンプロット)を示している。青色はクリプトサイトになり得る凹みのデータを表し、赤色はその他の凹みのデータを表す。

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　：作成データセットに対して、Fpocket[4]より得られた各特徴量の分布(青: クリプトサイトになり得る凹み、赤: その他の凹み)の比較。

表　は、図　の各特徴量について、クリプトサイトになり得る凹みのデータとその他の凹みのデータ分布が有意か否かを確かめるため、Kologorov-Smirnovの2標本統計に基づくP値を算出し、P値の小さい特徴量から順に示している。

表　：各特徴量とKologorov-Smirnovの2標本統計に基づくP値の関係

|  |  |
| --- | --- |
| 特徴量名 | P値 |
| Hydrophobicity score | 0.004505535750710443 |
| Mean alp. sph. solvent access | 0.012702623743886199 |
| Number of Alpha Spheres | 0.15383457971242187 |
| Score | 0.16894901636395654 |
| Mean local hydrophobic density | 0.16995554937081192 |
| Polarity score | 0.24012107232767865 |
| Alpha sphere density | 0.24012107232767865 |
| Proportion of polar atoms | 0.3787916710900978 |
| Mean alpha sphere radius | 0.39169455136082354 |
| Cent. of mass - Alpha Sphere max dist | 0.4074716655962288 |
| Polar SASA | 0.628495301602217 |
| Volume | 0.628495301602217 |
| Total SASA | 0.6297862188632384 |
| Apolar SASA | 0.7489562619763734 |
| Apolar alpha sphere proportion | 0.7859979604370235 |
| Druggability Score | 0.8179280488163592 |
| Flexibility | 0.848614167291587 |
| Volume score | 0.9179177566857942 |
| Charge score | 0.999999561098095 |

4.2 クリプトサイトになり得る凹みとその他の凹みの主成分分析(PCA)

　図　は、データセットについてPCAし、PC1とPC2に着眼し、２次元プロットを示

している。青色はクリプトサイトになり得る凹みのデータを表し、赤色はその他

の凹みのデータを表す。

グラフ, 散布図

自動的に生成された説明グラフ, 散布図

自動的に生成された説明

グラフ, 折れ線グラフ

自動的に生成された説明屋内, テキスト, 地図, テーブル が含まれている画像

自動的に生成された説明

図　：データセットについてPCAし、PC1とPC2について可視化(青：クリプトサイトになり得る凹み、赤：その他の凹み)(左上)、データセットをPCAし、PC1とPC2について可視化(青：学習データ、赤：テストデータ)(右上)、累積寄与率の可視化(左下)、PC1とPC2における各特徴量の寄与度の可視化(右下)

4.3 機械学習モデルを用いたクリプトサイト有無の学習と推論に対する性能比較

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。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とした。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を用いた。その結果図　〜　に示す。ハイパーパラメータチューニング後の最良のパラメータで学習し、テストデータに対して推論した結果を表　に示す。表　からSVMが一番性能がよく、テストデータについてF1値で約71.0%の性能を達成した。図　に4つの機械学習モデルについての混合行列の結果を示す。

グラフ, 散布図

自動的に生成された説明グラフ, 棒グラフ

自動的に生成された説明

グラフ, ダイアグラム

自動的に生成された説明　カレンダー が含まれている画像

自動的に生成された説明

図　：Random Forestについて、最適化履歴(縦軸がF1値で横軸が最適化の試行回数、赤の折れ線はそれまでの試行回数での最良のF1値)(左上)、モデルの各ハイパーパラメータの寄与度(右上)、F1値と各ハイパーパラメータの組合せ (左下)、F1値と各ハイパーパラメータ値の関係(右下)

グラフ, 散布図

自動的に生成された説明グラフ, 棒グラフ

自動的に生成された説明

カレンダー

自動的に生成された説明

ダイアグラム

自動的に生成された説明

図　：XGBoostについて、最適化履歴(縦軸がF1値で横軸が最適化の試行回数、赤の折れ線はそれまでの試行回数での最良のF1値)(左上)、モデルの各ハイパーパラメータの寄与度(右上)、F1値と各ハイパーパラメータ値の関係 (中)、F1値と各ハイパーパラメータの組合せ(下)

グラフ, 散布図

自動的に生成された説明グラフ, 棒グラフ

自動的に生成された説明

カレンダー

自動的に生成された説明

ダイアグラム

自動的に生成された説明

図　：LightGBMについて、最適化履歴(縦軸がF1値で横軸が最適化の試行回数、赤の折れ線はそれまでの試行回数での最良のF1値)(左上)、モデルの各ハイパーパラメータの寄与度(右上)、F1値と各ハイパーパラメータ値の関係 (中)、F1値と各ハイパーパラメータの組合せ(下)

グラフ, 散布図

自動的に生成された説明グラフ, 棒グラフ

自動的に生成された説明

グラフ, レーダー チャート

自動的に生成された説明グラフ, 散布図

自動的に生成された説明

図　：SVMについて、最適化履歴(縦軸がF1値で横軸が最適化の試行回数、赤の折れ線はそれまでの試行回数での最良のF1値)(左上)、モデルの各ハイパーパラメータの寄与度(右上)、F1値と各ハイパーパラメータの組合せ (左下)、F1値と各ハイパーパラメータ値の関係(右下)

表　：選定した機械学習モデルとテストデータに対する性能

|  |  |
| --- | --- |
| モデル名 | 性能(F1値) |
| Random Forest | 48.275862068965516 |
| XGBoost | 60.606060606060595 |
| LightGBM | 62.85714285714287 |
| SVM | 70.96774193548387 |

時計と文字の加工写真

中程度の精度で自動的に生成された説明グラフ, 棒グラフ

自動的に生成された説明

混合行列(Random Forest) 混合行列(XGBoost)

グラフィカル ユーザー インターフェイス, アプリケーション

自動的に生成された説明グラフ

自動的に生成された説明

混合行列(LightGBM) 混合行列(SVM)

図　：テストデータに対する4つの機械学習モデルの推論についての混合行列

4.4 機械学習モデルの予測が正答、誤答した場合のタンパク質ポケット周辺の表面構造の観察

本研究で用いた4つの機械学習モデルの内、一番良い性能を出したSVMについて、テストデータについて、予測が正答した場合と誤答した場合のタンパク質ポケット周辺の表面構造を確認した。

(図 〜 )

リンゴの絵

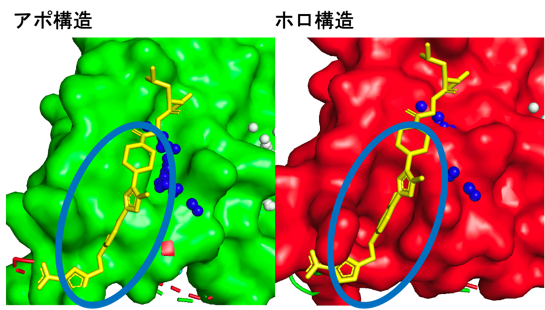
中程度の精度で自動的に生成された説明　　

図 ：モデルが予測を誤答(その他の凹みと予測)したタンパク質のポケットの例（アポ構造: 1Z92A、ホロ構造: 1PY2A、リガンド: FRH）

図 ：モデルが予測を正答したタンパク質のポケットの例（アポ構造: 1BSQA、ホロ構造: 1GX8A、リガンド: RTL）

（事例をもっと増やします）

4.5 機械学習モデルの因子分析

図　左は、SVMについてモデルが重要と判断した特徴量を重要な順に可視化したものを示している。また図　右は、各特徴量のデータが予測に与える影響度を予測全体への影響力が大きい順に示している。結果として、特徴量の内、Hydrophobicity score、Alpha sphere density、Polarity scoreが重要特徴量の上位であることがわかった。

グラフ, じょうごグラフ

自動的に生成された説明　グラフ

自動的に生成された説明

図　：SVMの重要特徴量(左図)、各特徴量のデータが予測に与える影響度(右図)

図　は、全36の各テストデータについて、どの特徴量からどの程度の影響を受けて最終的な予測値になったかの過程を示している。なお、特徴量は同じような予測パターンのサンプルをなるべく近くに表示する順で示している。

ダイアグラム

自動的に生成された説明ダイアグラム

自動的に生成された説明

ダイアグラム

自動的に生成された説明ダイアグラム

中程度の精度で自動的に生成された説明

図　：テストデータに対する予測過程(左上)、予測過程(実線：予測を正答した場合、点線：予測を誤答した場合)(右上)、予測を高確率(0.8以上)で正答した場合の予測過程(左下)、予測を誤答した場合の予測過程(実線：クリプトサイトになり得るポケットと予測した場合、点線：その他のポケットと予測した場合)(右下)