機械学習を用いたタンパク質クリプトサイトの予測法の開発

大規模知識発見分野
熊田　匡仁

背景

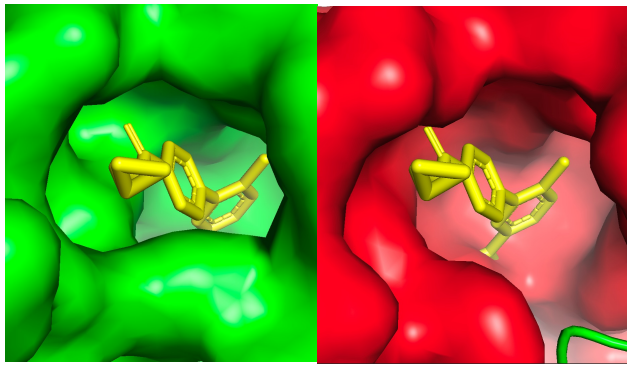
- 医薬品設計の基本原理である「鍵と鍵穴理論」は、タンパク質のポケット(鍵穴)の構造情報に基づいて医薬品分子(鍵)を設計する創薬研究で合理的な戦略の一つである。

- 鍵穴を同定することは創薬標的タンパク質の探索の最初の重要課題とされている。

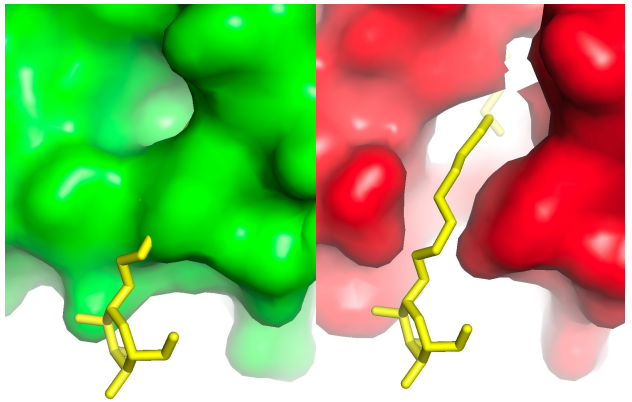
- 近年、一般的な鍵穴構造と異なり、通常(アポ構造)は閉じているが薬剤が結合したとき(ホロ構造)に形成される隠れたリガンド(薬剤)結合部位である**クリプトサイト**が存在することが知られ、新たな創薬標的としての応用が期待されている。

一般的な鍵穴構造例

アポ構造　　**ホロ構造**



クリプトサイト構造例



PDB: 2ZB1　　**PDB: 2NPQ**

背景

- これまで発見されているクリプトサイトの多くは、構造生物学解析によって決定されたリガンドと標的タンパク質のホロ構造とアポ構造の比較によって、偶然確認されているため、アポ構造情報のみからクリプトサイトタンパク質の予測が望まれている。

- 現在、クリプトサイトを誘導する特徴的なフラグメント分子を共溶媒した実験や分子動力学シミュレーション等により、クリプトサイトを予測する手法の開発への取り組みがなされているが、フラグメント分子の汎用性や大規模なシミュレーション時間を要するなど課題が多い。

※Kimura SR *et al.*, *J Chem Inf Model.* **57**, 1388-1401 (2017).

アポ構造および対応する混合溶媒で生成された代表的な構造に対して行った SiteMap 分析の結果※

system	apo (rank, Dscore)[d]	mixed-solvent (rank, Dscore)
exonuclease I[b]	NF	9, 0.81
Niemann-Pick C2 protein	NF	1, 0.93
Staphylococcal nuclease	1, 0.81	1, 0.80
toluene-4-monooxygenase	NF	2, 0.74
TETR-like transcriptional regulator LFRR	NF	1, 1.02
kinesin Eg5	NF	1, 0.75
Cdc4[c]	NF	13, 0.32
P38α	5, 0.78	1, 1.78

[a]Unless otherwise noted, default SiteMap settings were used to run the calculations. [b]Search extended to the top 20 sites. [c]Search extended to the top 20 sites and at least 10 points required to define a site. [d]NF = not found.

- ２つの系のみ、クリプトサイトを特定することに成功。残り６つの系については、クリプトサイトを特定できなかった。

研究目的

- アポ構造のタンパク質構造を入力として、クリプトサイトの有無を分類する機械学習モデルを作成する。

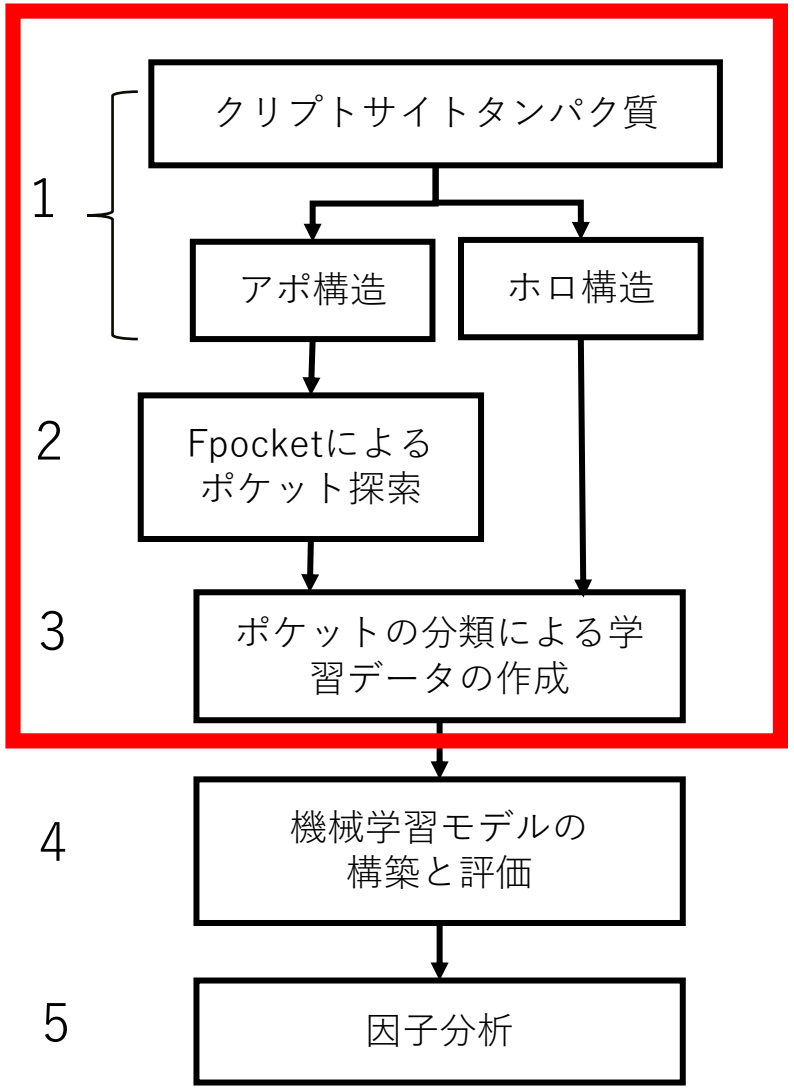
  ⇒　新たな創薬標的タンパク質同定システムとして創薬支援に活用

- 作成した機械学習モデルからクリプトサイト の因子評価を試みる。

  ⇒　クリプトサイト形成メカニズムと新しい鍵と鍵穴理論への理解

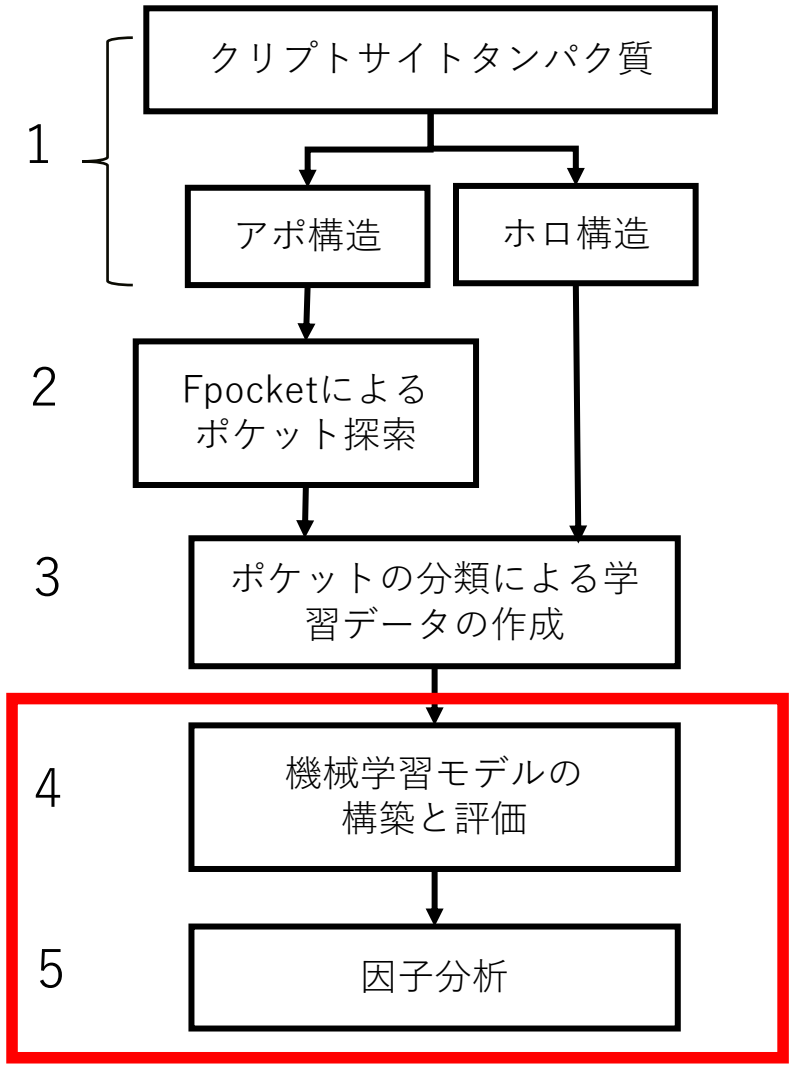
構築パイプライン

1. クリプトサイトを持つタンパク質のアポ構造のデータセットに先行研究論文に従って構築する。
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の解析結果について、
   - クリプトサイトになり得る凹みを検出するため、ホロ構造を重ね合せ Fpocket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、ラベリングする。
   - その他の凹みについても検出するために、Fpocketの生成する点群に対し、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、ラベリングする。



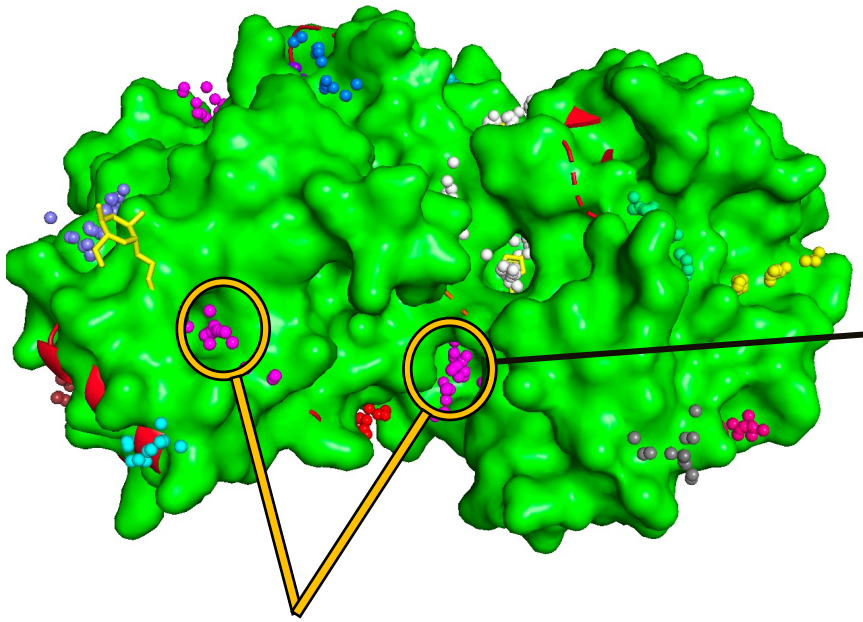
構築パイプライン

4. 3.までで構築したデータセットを学習データとし、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。
5. 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 。



2. Fpocketについて

Fpocket：タンパク質表面の幾何学的特徴からポケットを検出するオープンソースソフトウェア。



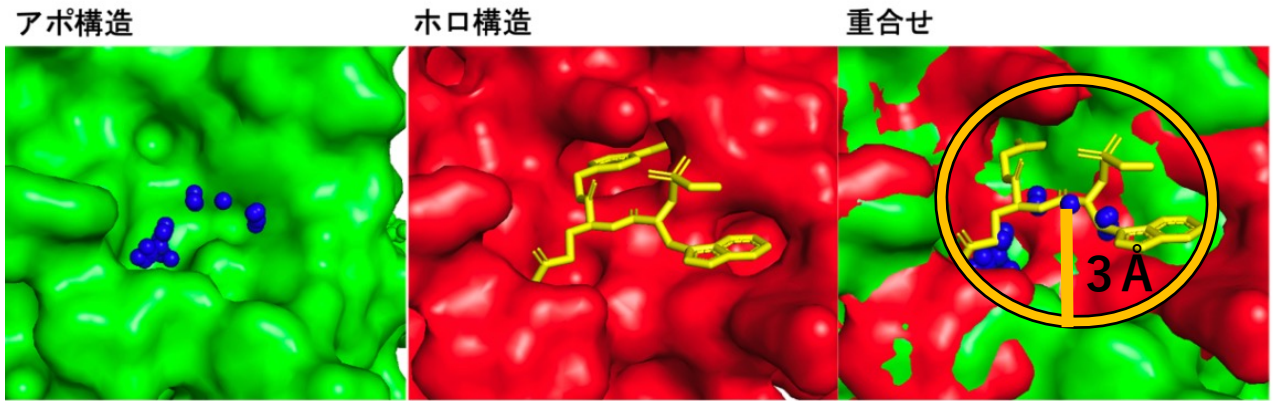
アポ構造：2ZB1Aについて、
Fpocketで解析した結果

算出される物理化学的特徴量(19種類)

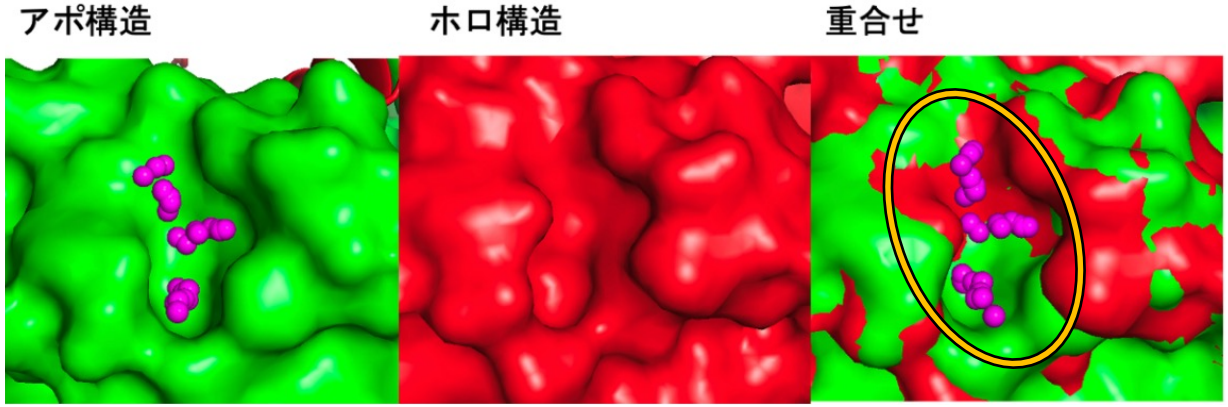
1. Score,
2. Druggability Score,
3. Number of Alpha Spheres,
4. Total SASA,
5. Polar SASA,
6. Apolar SASA,
7. Volume,
8. Mean local hydrophobic density,
9. Mean alpha sphere radius,
10. Mean alp. sph. solvent access,
11. Apolar alpha sphere proportion,
12. Hydrophobicity score,
13. Volume score,
14. Polarity score,
15. Charge score,
16. Proportion of polar atoms,
17. Alpha sphere density,
18. Cent. of mass - Alpha Sphere max dist,
19. Flexibility

3. データセット構築について

クリプトサイトになり得る凹みのアノテーション例



その他の凹みのアノテーション例



データセットの内訳

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

データセット集計

- 学習データ　：175
- テストデータ：36

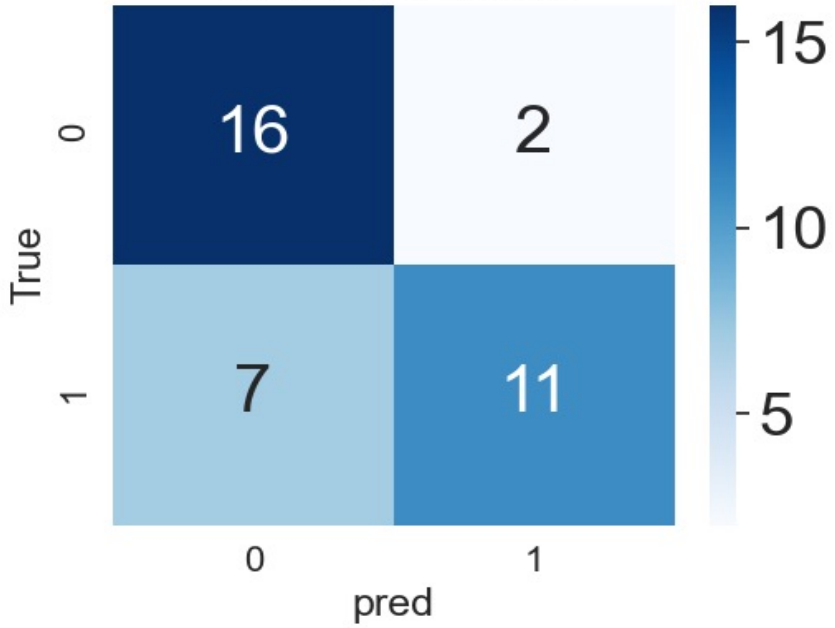
4. 機械学習モデルについて

選定モデル名とテストデータに対する性能

選定モデル名	性能(F1値)%
Random Forest	48.3
XGBoost	60.6
LightGBM	62.9
Support Vector Machine(SVM)	71.0

結果：
SVMがテストデータに対して最も精度が高く、
F1値: **71.0%** の精度を達成した。

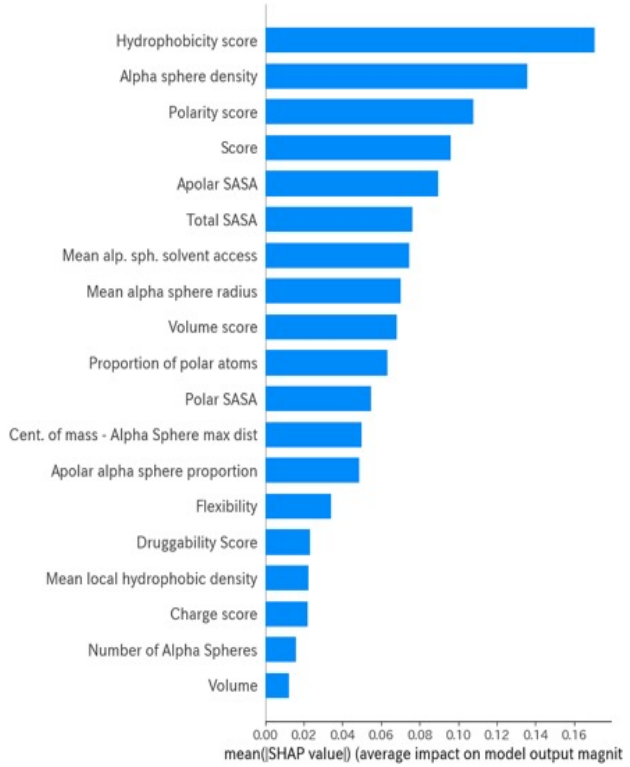
SVMの混合行列

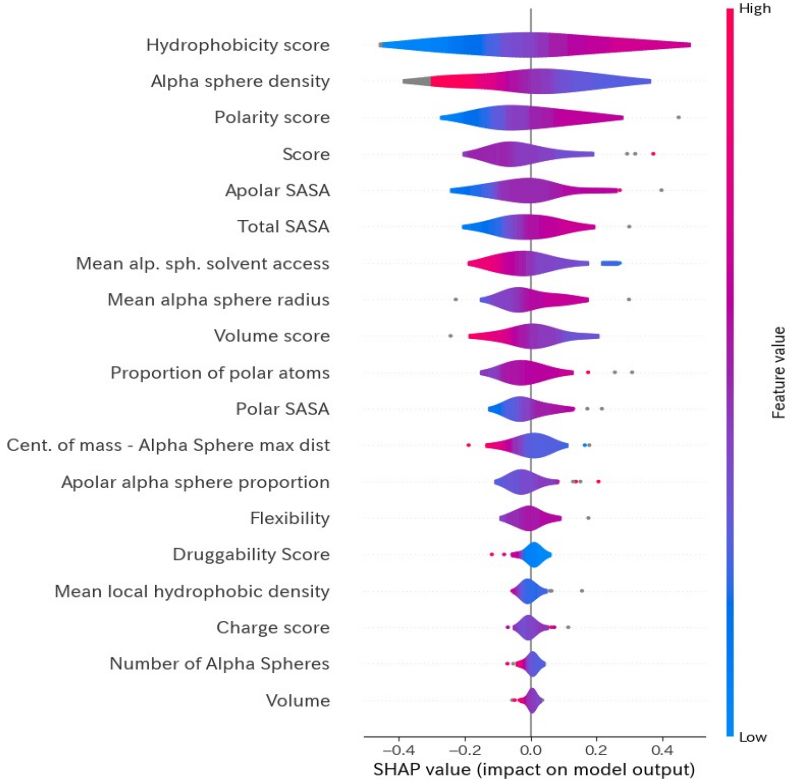


- 1: クリプトサイトになり得る凹み
- 0:その他の凹み

5. SVMの因子分析

- 機械学習モデルの予測を解釈するフレームワークである SHAP※を活用。
- テストデータの各特徴量が予測に与える影響度が平均的に高い順に可視化。





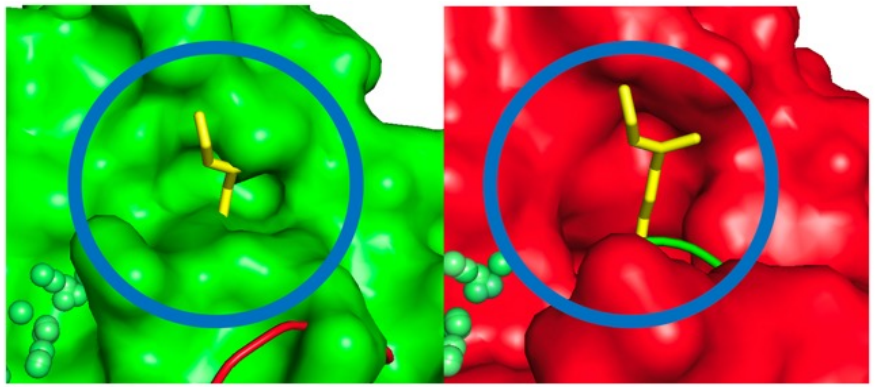
- 特徴量の内、Hydrophobicity score、Alpha sphere density、Polarity scoreが重要特徴量の上位である。
- モデルの出力に対して、Hydrophobicity scoreとPolarity scoreは正の相関があり、Alpha sphere densityは負の相関がある。

※ SHapley Additive explanation(SHAP) ; Lundberg, *et al.*, *NeurIPS*. **30,** 4768-4777 (2017).

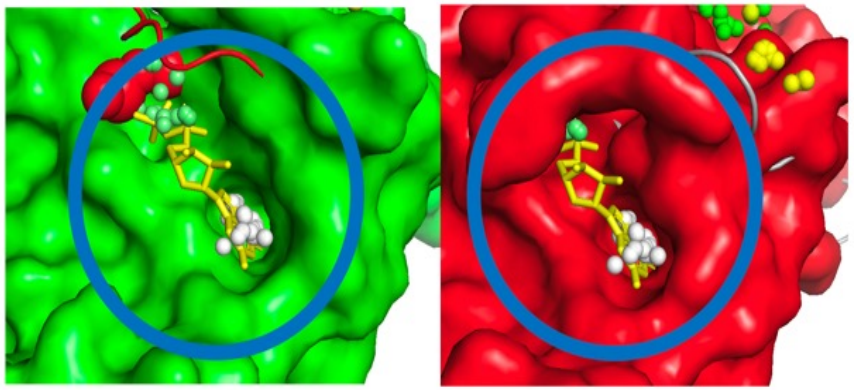
5. SVMの因子分析

予測を高確率(0.8以上)で正答したポケット例

アポ構造: 1BSQA、ホロ構造: 1GX8A



アポ構造: 1KZ7D、ホロ構造: 1GRNA



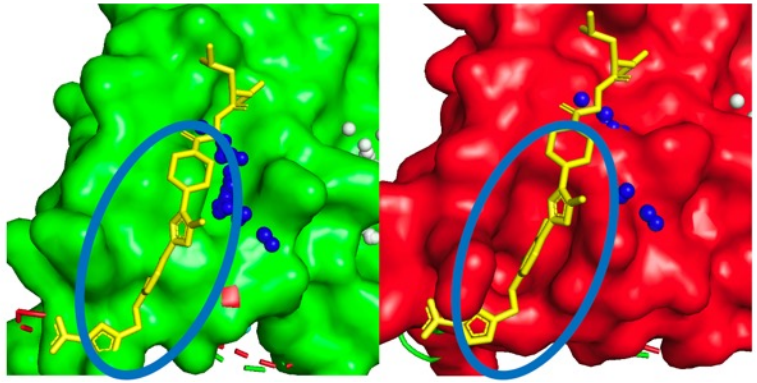
- 予測を正答した場合のアポ構造とホロ構造を観察すると、アポ構造のクリプトサイトにリガンドが誘導し、リガンドが結合したホロ構造では、凹みの表面積と深さがアポ構造に比べ、大きくなっている。

⇒ リガンドがドリルとして働き、クリプトサイトを掘削するという本研究の前提に適った結果であると考えられる。

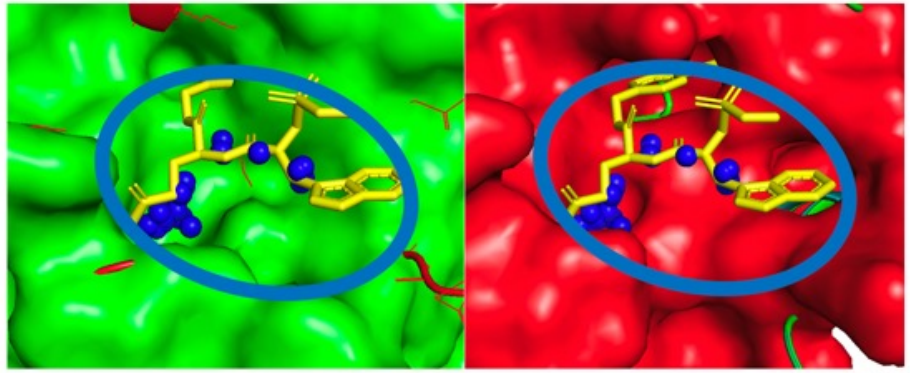
5. SVMの因子分析

予測を誤答(その他の凹みと予測)したポケットの例

アポ構造: 1Z92A、ホロ構造: 1PY2A



アポ構造: 1JBUH、ホロ構造: 1WUNH

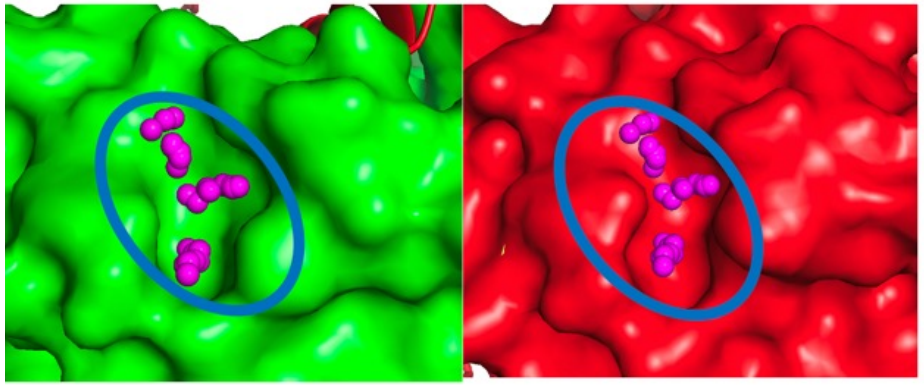


- 予測を誤答した(その他の凹みと予測)場合は、
  アポ構造においてクリプトサイトになり得る凹みがモデルの予測が正答した場合に比べて浅かったため、その他の凹みと判定を誤ったと考えられる。

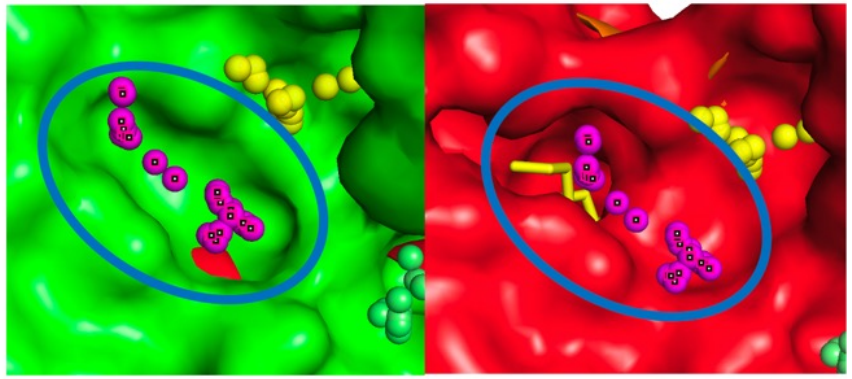
5. SVMの因子分析

予測を誤答(クリプトサイトになり得る凹みと予測)したポケットの例

アポ構造: 3FDLA、ホロ構造: 2YXJA



アポ構造: 1E2XA、ホロ構造: 1H9GA



- 予測を誤答した(クリプトサイトになり得る凹みと予測)場合は、
  - アポ構造の凹みがホロ構造の凹みと殆ど変わらないほど凹みの表面積と深さが大きいこと(左)
  - アポ構造よりもホロ構造の方がむしろ凹みの表面積が大きくなってしまっていること(右)

から、アポ構造の凹みをクリプトサイトになり得る凹みと判定を誤ったと考えられる。

まとめ

- アポ構造のタンパク質構造を入力として、クリプトサイトの有無を分類する機械学習モデルを作成。予測精度(F1値)は、SVMが最も高く、71.0% であった。
- 作成した機械学習モデルはFpocketの特徴量の内、以下が重要特徴量の上位であった。
  - Hydrophobicity score、Alpha sphere density、Polarity score
- Hydrophobicity scoreとPolarity scoreの値が正の値であればあるほど、Alpha sphere densityの値が負であればあるほど、モデルは１(クリプトサイトになり得る凹み)を出力しやすくなるということが考えられる。
- テストデータの一部について、実際のタンパク質のアポ構造とホロ構造を確認し、モデルの予測が正答、誤答した場合それぞれについて考察した。
- 本研究で用いた学習データ数が少ないことを考慮すると、今回のようなアポ構造の場合にモデルが一定の予測を誤答してしまうのは、避けられないと考えられる。

展望

- 今後は、得られた知見をもとに、まずはFpocketをカスタマイズし、クリプトサイト検出精度の向上に取り組みたい。
- 上記で得られた知見をもとに、最終的には独自のクリプトサイト検出ソフトウェアの開発に試みたい。
- 本研究で作成したデータや機械学習モデルの実装、出力結果について、以下のGitHub上に公開している。
    - https://github.com/MasahitoKumada/research/tree/main/2021

補足

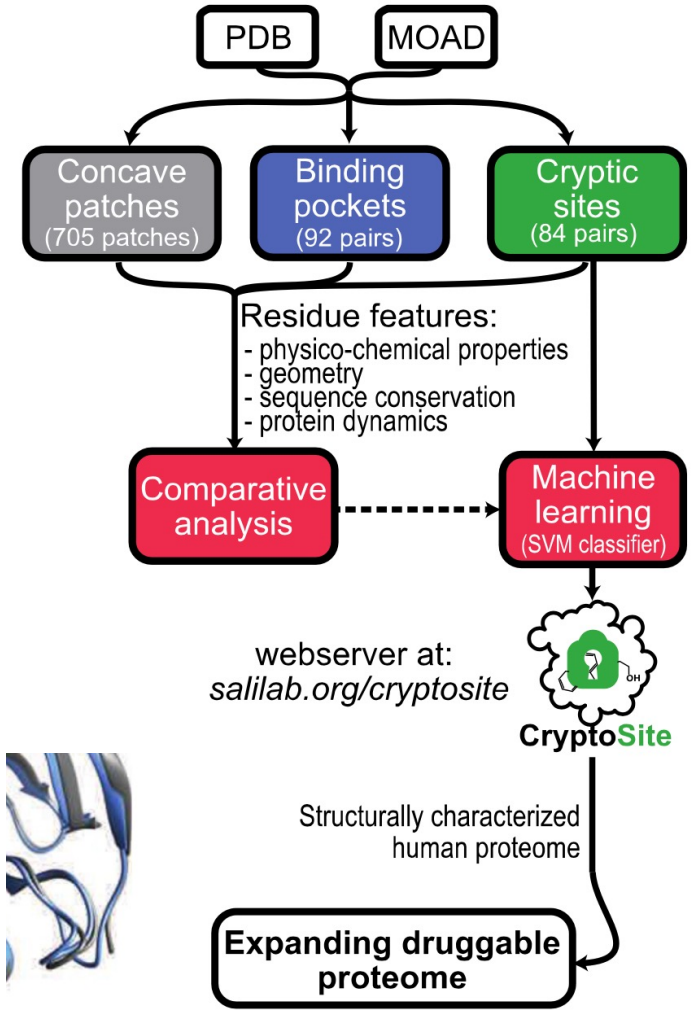
Appendix

先行研究※

- Protein Data Bank およびMOADデータベースから、84個の暗号的結合部位、92個の結合ポケット、および705個の凹面パッチの既知の例の代表的なデータセットを作成することから始めた。その中から、リガンドが生物学的に関連性のある暗号部位と結合ポケットを選択した。
- 個々の残基とその近傍の配列、構造、ダイナミクスを記述する58の特徴量のセットを設計した。
- 機械学習アルゴリズムを用いて、残基を暗号部位に属するか否かを分類した。
- 構造的に特徴づけられたヒトプロテオーム全体の暗号部位を予測した。

※ Peter Cimermancic *et al.*, *JMolBiol*.**428**, 709-719(2016).

先行研究のパイプライン



タンパク質構造からリガンド結合部位を予測する既存ソフトウェアの開発の歴史(2009年以降)※

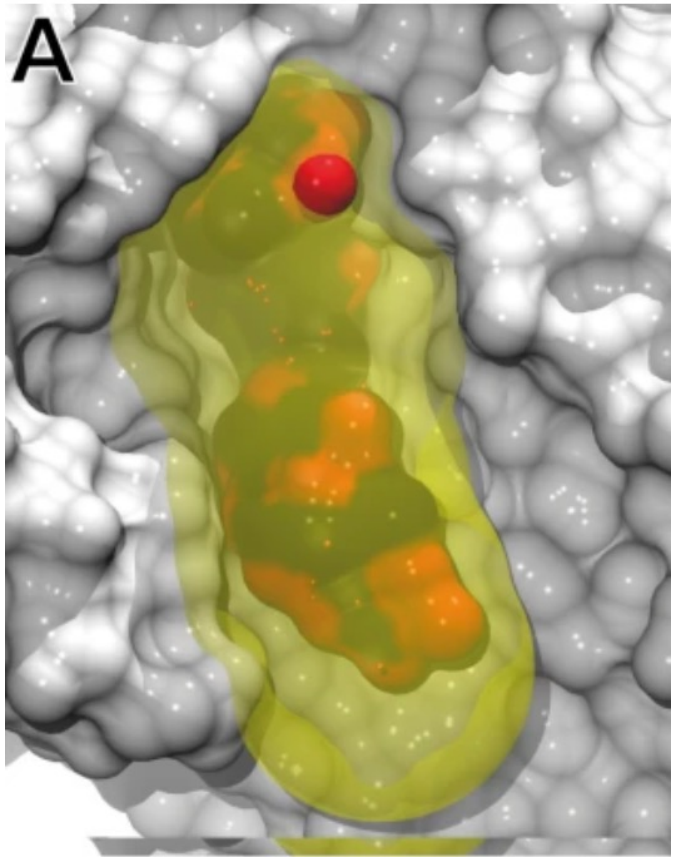
Name	Year	Type	Web server	Stand-alone	Fully automated†	Source Code
SiteMap [35]	2009	Geometric	–	Yes	Yes	–
Fpocket [18]	2009	Geometric	Yes	Yes	Yes	Yes
SiteHound [28]	2009	Energetic	Yes	Yes	Yes	Yes
ConCavity [36]	2009	Conservation	Yes	Yes	–	Yes
3DLigandSite [37]	2010	Template	Yes	–	–	–
POCASA [38]	2010	Geometric	Yes	–	–	–
DoGSite [39]	2010	Geometric	Yes	–	–	–
MetaPocket 2.0 [27]	2011	consensus	Yes	–	–	–
MSPocket [81]	2011	Geometric	–	Yes	Yes	Yes
FTSite [40]	2012	Energetic	Yes	–	–	–
LISE [41]	2012	Knowledge/conservation	Yes	Yes	–	–
COFACTOR [42]	2012	Template	Yes	Yes	Yes	–
COACH [43]	2013	Template† †	Yes	Yes	Yes	–
G-LoSA [44]	2013	Template	–	Yes	–	Yes
eFindSite [45]	2013	Template	Yes	Yes	–	Yes
GalaxySite [46]	2014	Template/docking	Yes	–	–	–
LIBRA [47]	2015	Template	Yes	Yes	–	–
P2Rank (this work)	2015*	Machine learning	–**	Yes	Yes	Yes
bSiteFinder [48]	2016	Template	Yes	–	–	–
ISMBLab-LIG [32]	2016	Machine learning	Yes	–	–	–
DeepSite [33]	2017	Machine learning	Yes	–	–	–

※ Krivák, R., Hoksza, D. *J Cheminform* **10,** 1-12 (2018).

Fpocket※1のアルゴリズム概要

Fpocketは、以下の３つの主要なステップからなる。

- 最初のステップでは、アルファ球※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。
- 第2のステップでは、近接した球体のクラスターを識別し、ポケットを識別し、それ以外のクラスターを削除する。
- 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。



Fpocketによる解析例
- 赤: 予測領域、
- 黄: 実際のリガンド領域

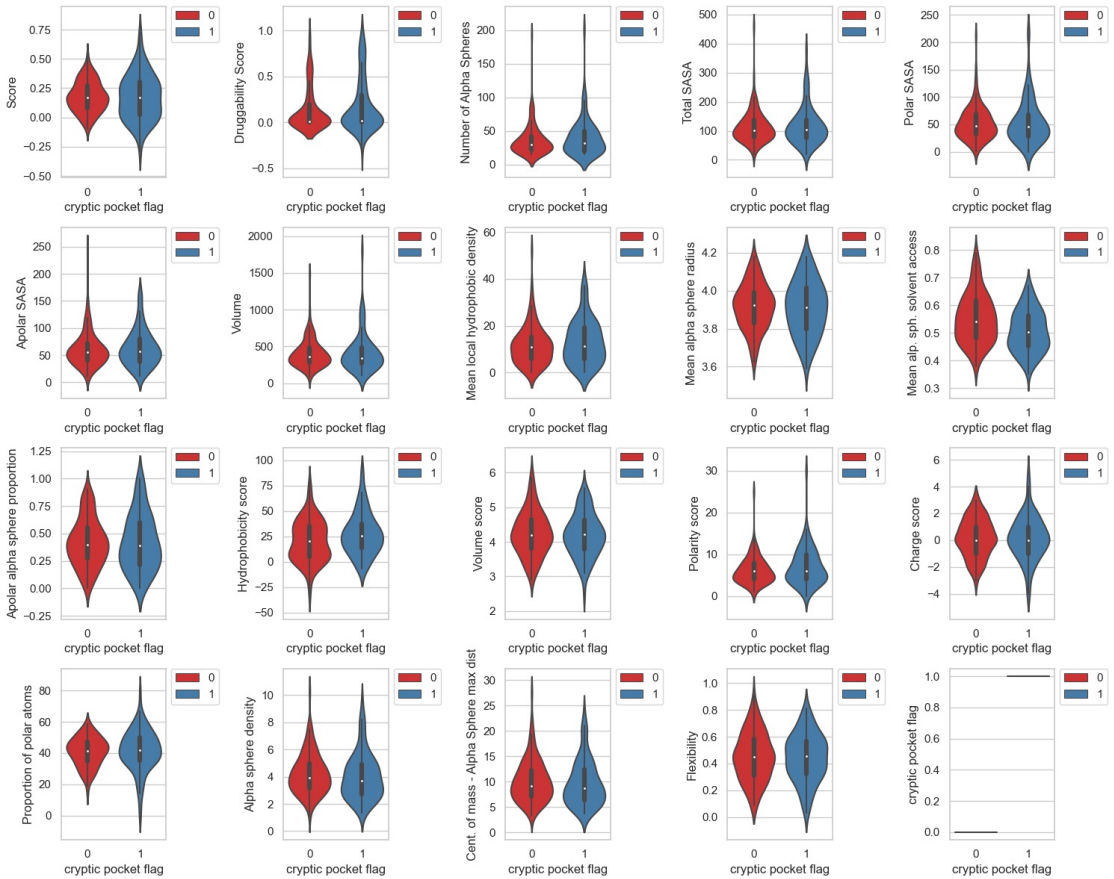
※1 Le Guilloux *et al.*, *BMC Bioinformatics*. **10**, 168 (2009).

※2アルファ球：その境界で4つの原子に接触し、内部の原子を含まない球。

5. 因子分析

作成データセットに対して、特徴量の分布

- 青色:クリプトサイトになり得る凹みのデータ
- 赤色:はその他の凹みのデータ

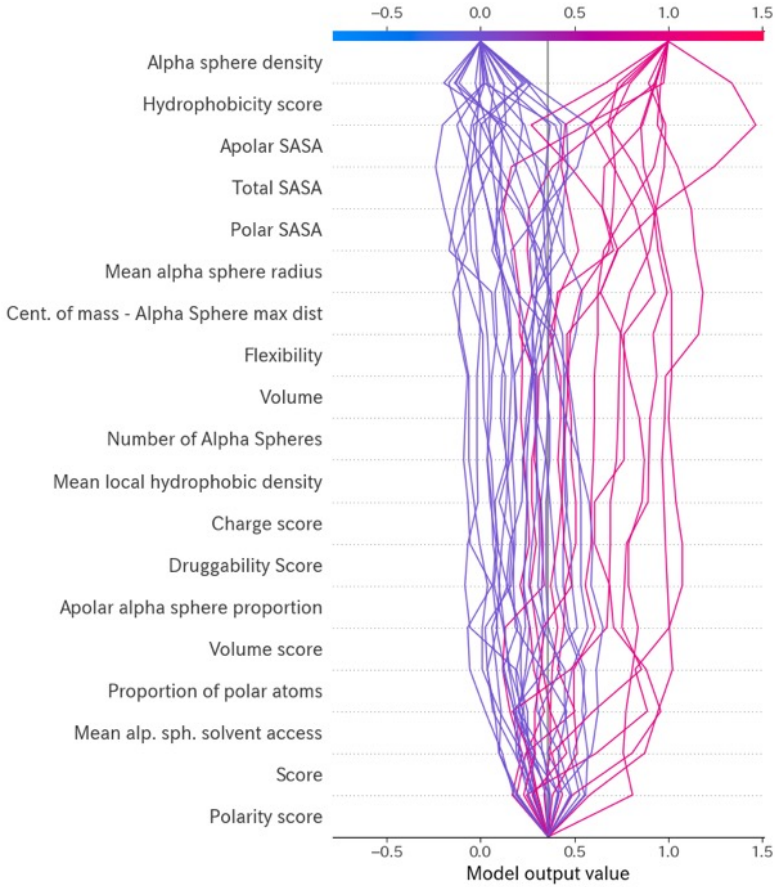


コロガロフ・シルノフの2標本統計に基づくP値の関係

特徴量名	P値
Hydrophobicity score	0.00451
Mean alp. sph. solvent access	0.0127
Number of Alpha Spheres	0.154
Score	0.169
Mean local hydrophobic density	0.170
Polarity score	0.240
Alpha sphere density	0.240
Proportion of polar atoms	0.379
Mean alpha sphere radius	0.392
Cent. of mass - Alpha Sphere max dist	0.407
Polar SASA	0.630
Volume	0.628
Total SASA	0.630
Apolar SASA	0.749
Apolar alpha sphere proportion	0.786
Druggability Score	0.818
Flexibility	0.849
Volume score	0.918
Charge score	1.00

5. SVMの因子分析

各テストデータについて特徴量の寄与を分析

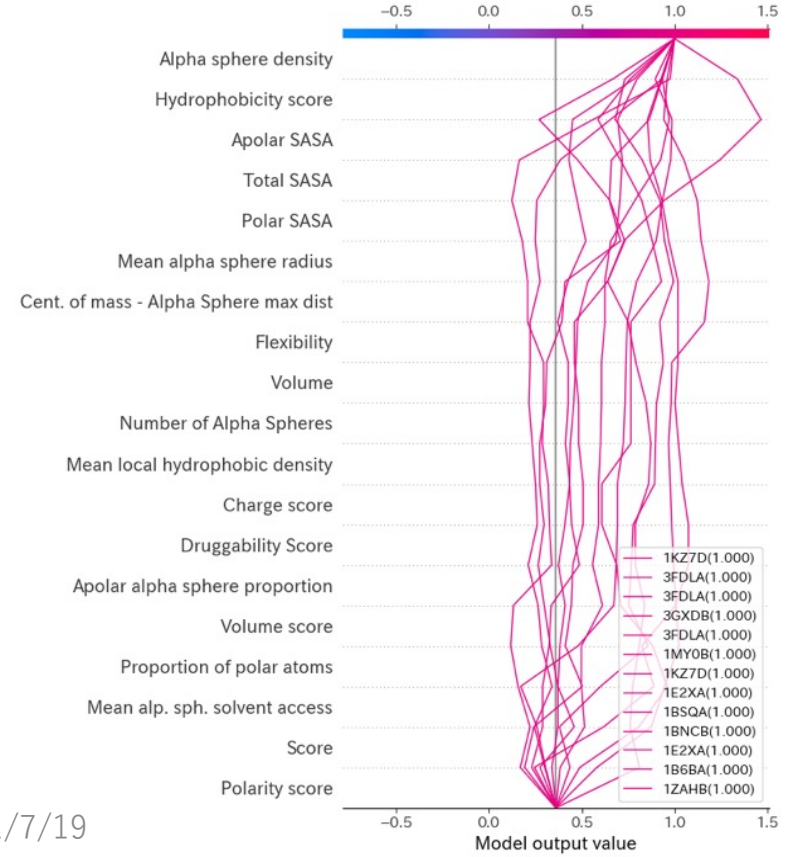


- 全36の各テストデータについて、どの特徴量からどの程度の影響を受けて最終的な予測値になったかの過程を示している。
- ある特徴量に対し、モデルの出力値が大きく変化したとき、その特徴量が予測に寄与したといえる。

5. SVMの因子分析

各テストデータについて特徴量の寄与を分析

予測を高確率(0.8以上)で正答した場合の予測過程



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