2022 年度　利用報告書

プロジェクト名（タイトル）：

創薬プロセス効率化を目指した機械学習のための学習データの作成

利用者氏名：
○　大田 雅照 (1)、千葉　峻太朗 (2)、池口　満徳 (2)、吉留　崇 (2)、津田　和実 (1)、下道　修 (1)、小山　秀 (1)、馬 彪(3)、井阪 悠太(1)

理研における所属研究室名：
(1)　計算科学研究センター HPC/AI 駆動型医薬プラットフォーム AI 創薬連携基盤ユニット
(2)　計算科学研究センター HPC/AI 駆動型医薬プラットフォーム 分子デザイン計算知能ユニット
(3)　計算科学研究センター HPC/AI 駆動型医薬プラットフォーム　バイオメディカル計算知能ユニット

1.　本プロジェクトの研究の背景、目的、関係するプロジェクトとの関係

医薬品開発のためには多数のプロセスが存在し、承認薬はそのすべてを通過して初めて誕生する。承認までの 10～20 億 US＄とも見積もられているコストと 10 年以上という開発期間を削減するための手段として、各プロセスにおいて深層学習を含む機械学習手法およびシミュレーションを利用することは検討の価値がある。本課題では、創薬開発のプロセスでの人工知能（Artificial Intelligence, AI）技術応用の可能性を調べるため、学習に必要となるデータの作成を実施している。

特に、本課題では「医薬品候補化合物の特性予測のための深層学習モデル」、「リガンド結合によるタンパク質周囲の水の置き換えの 3D-RISM による解析」、「タンパク質構造予測」などに注目し、量子化学計算、分子動力学計算、3D-RISM 法、ホモロジーモデリング、電子密度計算などによってデータ作成および方法論開発を実施している。

本課題のメンバーは、ライフサイエンス分野で AI 技術とビッグデータ利用を推進するコンソーシアム（ライフ インテリジェンス コンソーシアム（LINC）、代表：奥野恭史、事務局：京都大学大学院医学研究科人間健康科学系専攻ビッグデータ医科学分野、理化学研究所計算科学研究センター HPC/AI 駆動型医薬プラットフォーム）に所属し、各機関と連携の上で課題を実施している。

2.　具体的な利用内容、計算方法

2-1　サイクリックペプチド AI 力場開発

サイクリックペプチドは創薬における新規モダリティー（分子形態）として注目されている。しかしながら、サイクリックペプチドは構造的柔軟度が高く、そのコンフォメーションは多様であり、N-Me アミノ酸や側鎖に天然型アミノ酸では存在しない元素を有するなどの多種多様な非天然型アミノ酸も用いられるため、従来の分子力場法では、そのエネルギーを精度良く評価することができない。そこで AI 技術を用いて、量子化学レベルの高精度のエネルギー計算を瞬時に行うシステムを開発している。

方法：サイクリックペプチドのシーケンス（非天然型アミノ酸を含む）と各アミノ酸の 2 次元構造情報から、立体構造生成システムを用いて 3 次元構造を作成する。作成された 3 次元構造を基にソフトウェア OMEGA2 (OpenEye)を用いてサイクリックペプチドの多様なコンフォメーションを発生させる。HOKUSAI を用いて、これらの多様なコンフォメーションそれぞれについて、量子化学計算法である FMO 法により、そのエネルギー値を計算する。この結果得られた FMO エネルギー値と、サイクリックペプチドの 3 次元構造座標から深層学習法を用いて AI 力場を構築する。

2-2　3D-RISM 法により計算されたタンパク質水和状態を高速に推算する AI システムの開発

タンパク質の折り畳みやリガンド結合などの生物学的プロセスに影響を与える要因の中で、タンパク質表面の水和状態を知ることは重要である。分子動力学シミュレーションや 3D-RISM 理論など、水の分布関数（＝水和状態）を計算

するための一般的な方法では、数時間から数十時間の長い計算時間が必要である。我々は、Hokusai を用いて、3706 個のタンパク質について 3D-RISM 法により、その水和状態を網羅的に計算し、その水和状態が、低分子リガンドのタンパク質との相互作用、結合ポーズなどに深く関わっていることを明らかにしてきた(J.Comput.Chem., 2020, 41, 2406-2419)。3D-RISM 法を用いた水和状態情報は有用なものであるが、その計算には HOKUSAI を利用したとしても 1 タンパク質あたり約 2 時間程度必要となる。そこで AI 技術を用いて、タンパク質立体構造情報から、その水和状態を迅速に推算するシステムを開発することを試みた。

2-2-1 計算に使用したタンパク質

PDBbind refined set (v.2017)のタンパク質-リガンド複合体 3706 個のタンパク質から、27 個のタンパク質を以下のような過程で選択した。3706 個のタンパク質から、DL モデルの原子タイプの数を限定するためにイオンを含まないタンパク質として 2718 のタンパク質を選択した。その後、27 のタンパク質をランダムに選択した。27 個のタンパク質のうち、22 個のタンパク質を学習に、残りの 5 個をテストに使用した。学習で使用した 22 タンパク質とテストで使用したタンパク質の配列類似性は、90%以下であった。各タンパク質に対して、前処理として、リガンドと結晶水を除去し、リガンドに最も近いタンパク質鎖をタンパク質構造として使用した。

PDB ID	structure title	high reso. limit (Å)	data set
1A30	HIV-1 protease complexed with a tripeptide inhibitor	2.00	train: ten proteins
1FCH	PTS1 complexed to the TPR region of human PEX5	2.20	train: 10
1PZ5	antibody in complex with an octapeptide	1.80	train: 10
2CE9	a peptide bound to the Groucho-TLE WD40 domain	2.12	train: 10
2HKF	the complex Fab M75–peptide	2.01	train: 10
2PV1	SurA complexed with peptide WEYIPNV	1.30	train: 10
2QBW	PDZ–fibronectin fusion protein	1.80	train: 10
3BZF	major histocompatibility in complex with HLA-E	2.50	train: 10
3DRF	OppA complexed with an endogenous peptide	1.30	train: 10
3DRI	OppA co-crystallized with an octamer peptide	1.80	train: 10
3ERY	H-2 class I histocompatibility antigen in complex with a peptide	1.95	train: 12
3G19	ClpS protease adaptor protein in complex with a peptide	1.85	train: 12
3IFL	amyloid β peptide/antibody complex	1.50	train: 12
3P9M	H2-Kb in complex with epitope OVA-G4	2.00	train: 12
3T6B	human DPPIII in complex with tynorphin	2.40	train: 12
3TCG	*Escherichia coli* OppA complexed with the tripeptide KGE	2.00	train: 12
3UPV	pHsp70 complex of yeast Sti1	1.60	train: 12
4EZR	*E. coli* DnaK in complex with drosocin	1.90	train: 12
4EZZ	*E. coli* DnaK in complex with peptide ELPLVKI	2.05	train: 12
4YNL	HetR in complex with the hexapeptide ERGSGR	2.10	train: 12
5E6O	*Caenorhabditis elegans* LGG-2 bound to an AIM/LIR motif	1.80	train: 12
5LSO	SPF45 UHM domain with a cyclic peptide inhibitor	2.22	train: 12
2HA2	acetylcholinesterase complexed with succinylcholine	2.05	test
2O4L	HIV-1 protease in complex with tipranavir	1.33	test
3JVR	Chk1 complexed with an allosteric inhibitor	1.76	test
3O5N	Shank PDZ domain complexed with a small molecule	1.83	test
4KAO	focal adhesion kinase in complex with an inhibitor	2.39	test

2-2-2 3D-RISM Theory

3DRISM 理論を適用し、位置 r での水・水素原子の分布関数 $g_H(r)$、または、水・酸素原子の分布関数 $g_O(r)$を計算した。本研究では、Deep Learning (DL)モデルを構築するためのターゲット変数として $g_O(r)$を使用した。タンパク質は Amber ff99SB 力場を使用し、水は coincident SPC/E model を用いて、分布関数を計算した。3D-RISM 計算では、Kovalenko-Hirata (KH) closure を使用した。誘電率、bulk density、および温度は、それぞれ 78.497、0.03332Å^{-3}、および 310K を用いた。

2-2-3 DL モデルの入力および出力形式

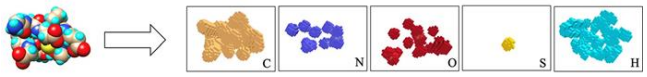
タンパク質構造を DL モデルに入力するために、タンパク質を、炭素、窒素、酸素、硫黄、および水素で構成される 5 つの原子タイプに分解し、タンパク質構造を、3D-RISM 計算で使用した water box と、同じグリッドサイズ、同じタンパク質の位置、同じサイズのボックスを用意し、ボクセル形式に変換した。次に、k 番目のボクセル n(k,i,j)に対する原子タイプ j の i 番目の原子の寄与を、式 1 に従って計算した。

$$n(k,i,j) = 1 - \exp\left[-\left(\frac{\sigma_{\mathrm{vdw},j}}{r_{ik}}\right)^{12}\right] \qquad (1)$$

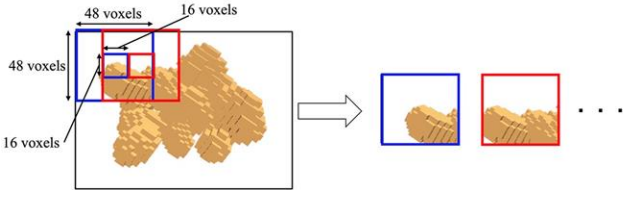
ここで、$\sigma_{\mathrm{vdw},j}$ は原子タイプ j のファンデルワールス半径、r_{ik} は i 番目の原子と k 番目のボクセルの位置の間の距離である。原子タイプ j の k 番目のボクセルへの寄与 N(k,j)は、式 2 に従って計算した。

$$N(k,j) = \sum_{i=1}^{N_j} n(k,i,j) \qquad (2)$$

ここで、Nj はタンパク質の原子タイプ j の原子数である。
この手順により、タンパク質構造を、原子タイプごとに N(k,j)の値を持つボクセルからなる 5 つのボックスに分解した。



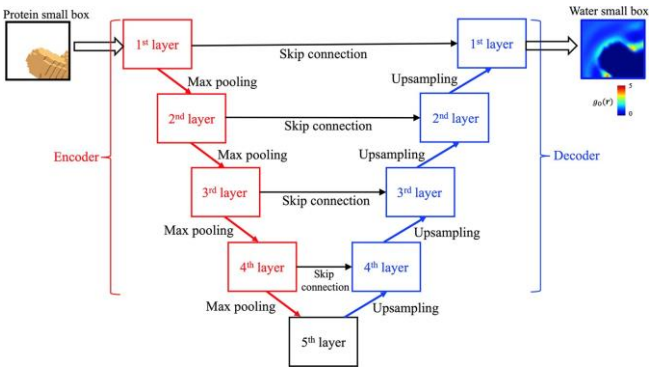
各ボックスを 483 ボクセルの小さなボックスに分解することにより、DL モデルを任意のサイズのタンパク質に適用できるようにした。以下、このボックスを「partial protein box」と呼び、タンパク質内の同じ位置にある5つの原子タイプのボックスの集合を「partial-protein-box set」と呼ぶ。



DL モデルの出力形式は、3D-RISM 理論を使用した $g_O(r)$ の water box の出力形式と同じである。DL モデルの学習では、前述のように water box も 483 ボクセルのボックスに分割した。結果として得られるボックスは、「partial water box」と呼ばれる。入力として partial-protein-box set を使用すると、DL モデルは対応する partial water box を出力する。各 partial water box の中央の 16 ボクセルを合計することによって、$g_O(r)$が得られる。

2-4-4 Deep Learning モデル

タンパク質の水和構造を予測する DL モデルのネットワークアーキテクチャには、encoder・decoder 型の U-net を採用した。DL モデルは、分布関数 $g_O(r)$(または $g_H(r)$)を予測するよう構築した。encoder と decoder は「encoder・decoder 層」と呼ばれる 4 つのレイヤーで構成され、encoder と decoder の第 4 レイヤーの間に第 5 レイヤーを用意した。各 encoder・decoder 層は 2 つの畳み込み層で構成され、それぞれに ReLU 関数を使用したアクティベーションが続く。encoder の encoder・decoder 層の 2 番目の畳み込み層の後に、2×2×2 の最大プーリング層を追加した。encoder・decoder レイヤーの最初の畳み込みレイヤーのフィルターの数は、encoder の前のレイヤーの 2 倍、decoder の前のレイヤーの半分とした。オリジナルの U-net アーキテクチャに合わせてスキップ接続を追加した。



我々の DL モデルは、オリジナルの U-net モデルとは 4 つの点で異なる。オリジナルの U-net モデルは 2 次元画像用であるが、我々のモデルは partial protein box と partial water box の 3 次元データ用に実装されている。さらに、我々のモデルでは、畳み込み層に zero padding を追加した。また、DL モデルの学習の際の過学習を減らすため、最初の畳み込み層の ReLU アクティベーション後に drop out 層を追加した。

DL モデルには、「hyperparameter set」と呼ばれる 4 つの hyperparameter があり、そのうちの 1 つは、畳み込み層のフィルターサイズ(つまり、フィルター内のボクセルの数)である。サイズは 33、43、53 の 3 種類を用意した。encoder の第 1encoder・decoder 層の原子タイプに対するフィルター数($N_{Firstfilter}$)が、もう 1 つの hyperparameter である。原子は 5 種類なので、フィルター総数は $5 \cdot N_{Firstfilter}$ である。3 番目の hyperparameter は、drop out レイヤーND で値がゼロに設定されたボクセルの数である。「drop out 比」と呼ばれる、drop out 層の ND とボクセルの総数の比率は、0.3 または 0.5 に設定した。drop out は、(i) encoder のみ、(ii) decoder のみ、または(iii) encoder と decoder の両方の各ケースについて、第 5 層、第 4-5 層、第 3-5 層、第 2-5 層、すべての層に適用した。encoder と decoder の両方に drop out を適用しない場合も考慮した。DL モデルは、TensorFlow ライブラリ(2.1.0)を使用して実装した。学習は、Adam オプティマイザーを使用してデフォルトのパラメータで実行した。GeForce RTX 2080 のグラフィックスプロセッシングユニット(GPU)を使用して計算を実行した。

2-2-5　hyperparameter 最適化

hyperparameter は、two-fold cross validation によって最適化した。学習に使用した 22 タンパク質を 2 つのセットに分割し、partial-protein-box set の総数が均一になるように、10 タンパク質と残りの 12 タンパク質に分割した。Partial protein box の数(N_{TData})は、それぞれ 6858 および 7101 だった。Partial water box 数も N_{TData} である。Cross validation では、DL モデルの学習に 10 個のタンパク質が使用された場合、残りの 12 個のタンパク質はモデルの検証に使用され、その逆も同様に実施した。以下、学習用データ、検証用データをそれぞれ「学習データ」、「検証データ」と表記する。各データセットは、Partial protein box と対応する Partial water box で構成される。

DL モデルの hyperparameter set では、epoch 数を 200 に設定し、式 3 の平均二乗誤差を損失関数 E に使用した。

$$E = \frac{1}{N_{TData} N_{Voxel}} \sum_{j=1}^{N_{TData}} \sum_{i=1}^{N_{Voxel}} \left(g_{\alpha,j}^{Model}(r_i) - g_{\alpha,j}^{3D-RISM}(r_i) \right)^2 \qquad (3)$$

ここで、N_{Voxel} はボックス内のボクセル数で、16^3 である。$g_{o,j}^{X}(r)$は、j 番目のボックスの r_i のボクセル位置での $g_O(r)$値である。上付き文字「Model」と「3D-RISM」は、それぞれ DL モデルと 3D-RISM 理論から得られた $g_O(r)$値を示す。以下、i epoch における E 値を $E_{Train}(i)$と表記する。

学習の各 epoch(i)について、検証データ[$E_{Validation}(i)$]を使用して式 3 の損失関数も計算し、学習中にオーバーフィッティングが発生しなかったかどうかを確認した。オーバーフィッティングは $E_{Validation}(200)$と $E_{Training}(200)$の値の比較により、チェックした。

4 つの hyperparameter について、162 の hyperparameter set を作成した。各 hyperparameter set に対して、次の計算を実行した。(1) 10 個のタンパク質と水・酸素原子分布関数 $g_O(r)$（または水・水素原子分布関数 $g_H(r)$）を学習データとして使用し、学習を実施、(2) $E_{Validation}(200)$値を保存、(3)学習データを 12 タンパク質に変更し、手続き(1)と(2)を繰り返し、(4)2 つの $E_{Validation}(200)$値の平均、$\bar{E}_{Validation}(200)$を計算した。162 個の hyperparameter set すべての計算の後、「最適化された hyperparameter set」と呼ばれる、最小の $\bar{E}_{Validation}(200)$値を持つ hyperparameter set を選択した。

2-2-6　テスト

「最適化された hyperparameter set」と 22 タンパク質を使用した学習を実施し、予測モデルを作成後、5 つのテストタンパク質の水・酸素原子分布関数 $g_O(r)$（または水・水素原子分布関数 $g_H(r)$）を計算した。

DL モデルの $g_O(r)$のピーク位置を 3D-RISM 理論のピーク位置と定量的に比較するために、以下の検討を行った。最初に、プログラム Placevent を使用して水・酸素原子を配置した。このプログラムでは、水・酸素原子分布関数 $g_O(r)$値に基づいて水分子を配置する。プログラムは、以下の手順で実行された。(i)水・酸素原子は、最大の $g_O(r)$値を持つ位置(r_{Max})に配置される。(ii)式 4 を満たす領域 δ を計算・同定する。

$$\int_{r_{Max}}^{r_{Max}+\delta} \rho_0 g_O(r)\, dr = 1 \qquad (4)$$

(iii)領域 δ 内の $g_O(r)$値をゼロに設定する。(iv)手順(i)、(ii)、および(iii)を、$g_O(r_{Max}) < 1.5$ になるまで繰り返す。上記を、3D-RISM 理論を使用して得られた $g_O(r)$値と、DL モデルで得られた $g_O(r)$値それぞれに対して実行し、3D-RISM 理論の水の位置、DL モデルの水の位置を得た。3D-RISM 理論

を用いて得られた i 番目の水・酸素原子の位置と、DL モデルを用いて得られた位置をそれぞれ r_{RISM} と r_{Model} で表す。配置された水・酸素原子の数は、それぞれ N_{RISM} および N_{Model} とする。

その後、3D-RISM 理論の水の位置、DL モデルの水の位置から、式 5 で定義される距離 Di を計算した。

$$D_i \equiv \min_j |\boldsymbol{r}_i^{\mathrm{RISM}} - \boldsymbol{r}_j^{\mathrm{Model}}| \qquad (5)$$

さらに Di の平均値とその標準偏差を計算し、解析した。

リガンド結合ポケットでの予測性能を Di の観点から調べるために、以下の検討をさらに行った。リガンド結合ポケットの水・酸素原子は、リガンドの重原子から 5Å 以内に配置された水・酸素原子として定義した。3D-RISM 理論と DM モデルより配置された各水・酸素原子について、Di 値を計算し、リガンド結合ポケット内の $g_O(r)$ピーク位置の比較を行った。さらに、結晶水の位置の予測性能を、タンパク質の重原子から 5Å 以内の結晶水の位置と DL モデルより水の位置の Di を比較することにより実施した。

2-2-7　計算時間の測定

$g_O(r)$の single CPU での計算時間を、5 つのテストタンパク質を用いて、3D-RISM 理論と DL モデルの間で比較した。使用した CPU は Intel Xeon Gold6230 CPU @2.10GHz である。また、GeForce RTX2080Ti の single GPU を用いて、DL モデル予測の計算時間についても検討した。

3D-RISM 理論計算は、AmberTools18 スイートを使用して実行した。3D-RISM 理論の計算は、(1)バルク水の水サイト間相関関数の計算、(2)3D-RISM 理論と KH closure を使用したタンパク質の $g_H(r)$と $g_O(r)$の値の計算という、2 つのプロセスがあるが、計算時間の比較は、後者のステップ(2)を用いて実施した。

計算時間は、DL モデルを使用した $g_O(r)$予測で Intel Xeon Gold6230 CPU @2.10GHz 2 基で構成される 40 コアを使用した場合、数十秒であった。計算時間は U-net アーキテクチャを使用した予測ステップで計測し、入力フォーマットの準備に必要な計算時間は含めなかった。

2-3. 相互作用記述子のパラメータ最適化システム構築

本グループでは、タンパク質―低分子間相互作用、タンパク質－タンパク質間相互作用、ペプチド―タンパク質間相互作用を、構造に基づき記述子化する技術を開発している。この記述子を機械学習の入力情報とすることで、結合構造構築、結合様式の妥当性判定、モデル構造の妥当性などを判別する予測機を構築できる。

特定の相互作用の有無は、事前に定義した距離や角度による定義に基づいて判定している。この定義を調整することで、予測モデルの総合的な性能を改善できる可能性がある。そこで、まず距離に関する定義に摂動を加えることで予測機の性能に変化があるか調べた。次に、記述子の定義（パラメータ）を、予測機の性能が高まるように並列ベイズ探索するシステムを構築した。

3.　結果

3-1　サイクリックペプチド AI 力場開発

当初は、1 サイクリックペプチドあたり 200 コンフォメーションの FMO 計算を実施していたが、16 cyclic peptides の AI 力場構築を試してみたところ、実測と予測の差異が大きく満足のいく AI 力場は創製できなかった。そこで、AI 用の学習データを増やすため、300K,400K,500K の 3 つの温度で molecular dynamics 計算を行い各温度で 200 コンフォメーションを抽出し、1 サイクリックペプチドあたり 800 コンフォメーションの FMO 計算を実施している。

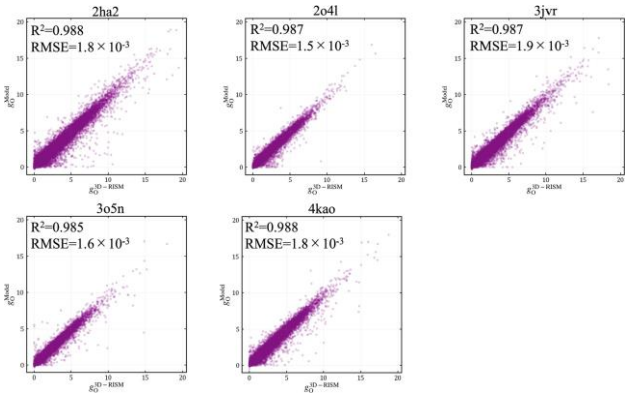
もう一つの試みとして、AI の学習においては、目的変数として FMO エネルギーを用いていたが、FMO エネルギーだけでなく、各原子にかかる force もあわせて学習するよう検討している。

3-2　3D-RISM 法により計算されたタンパク質水和状態を高速に推算する AI システムの開発

3-2-1　Prediction test

テスト用の 5 つのタンパク質における、DL モデルによる予測 $g_O(r)$値と 3D-RISM 理論で計算した $g_O(r)$値との相関関係、

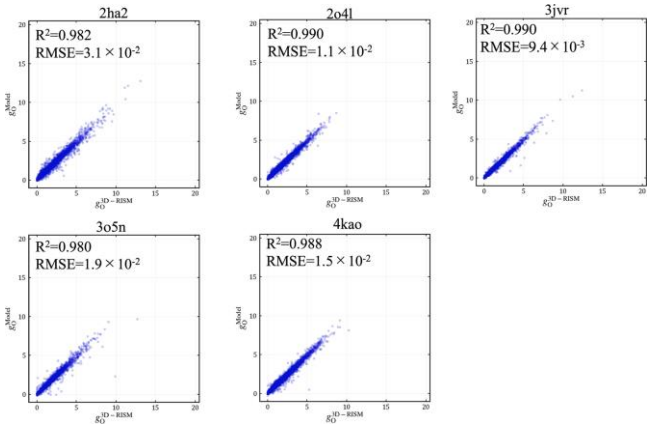
決定係数 R^2 スコア、および root mean square error (RMSE) を以下に示す。R^2 値は 0.985–0.988 と高く、ほとんどの点が y=x 付近にあった。さらに、RMSE 値は 1.5×10^{-3}～1.9×10^{-3} と DL モデルが高精度で水・酸素原子分布関数 $g_O(r)$を予測できることが示された。さらに、これらの高精度予測は、single GPU で 1 分以内と短時間で実施できた。



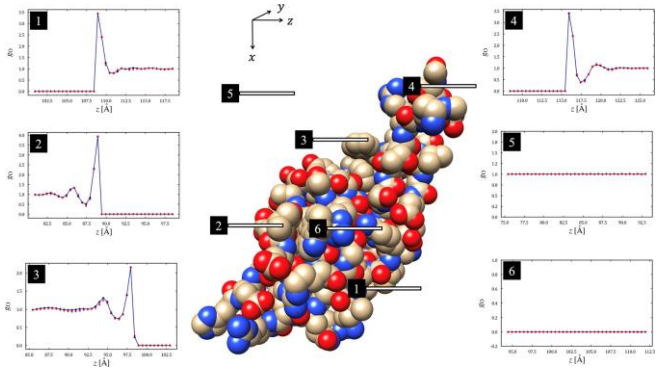
さらに、タンパク質表面での水和状態の予測性能を検討するために、タンパク質表面での R^2 スコアと RSME を計算した（以下の表）。タンパク質の重原子から 5Å 以内においても、R^2 は 0.984 以上、RMSE は 1.4×10^{-2} 以下であり、我々の DL モデルがタンパク質表面においても精度よく水・酸素原子分布関数 $g_O(r)$を予測できることを示した。

PDB ID	R^2 score	RMSE
2HA2	0.985	1.2×10^{-2}
2O4L	0.986	1.2×10^{-2}
3JVR	0.986	1.3×10^{-2}
3O5N	0.984	1.4×10^{-2}
4KAO	0.987	1.2×10^{-2}

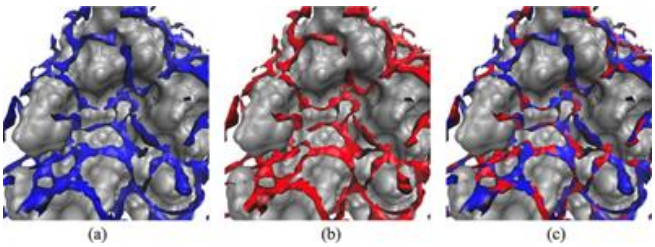
さらに、リガンド結合ポケット内（リガンドの重原子から 5Å 以内）で良好な予測ができるか検討した。以下に示すように、DL モデルの予測性能は、リガンド結合ポケットでも高く、ほとんどのポイントは y=x 付近にあり、R^2 値は高く、RMSE 値は十分に小さかった。リガンド結合部位における水・酸素原子分布関数 $g_O(r)$の高い予測精度は、結合部位において水和情報を利用して新しい化合物をデザインする際に重要な要素であると考えられる。



Shank3 PDZドメイン[PDB コード:3o5n]の結果を以下に示す。3D-RISM 理論の $g_O(r)$値（青線）と DL モデルの $g_O(r)$値（赤点）はタンパク質のどの位置において（タンパク質内部から表面：スライス1,2,3,4、タンパク質内部：スライス6、バルク水領域：スライス 5）も、よく一致していた。



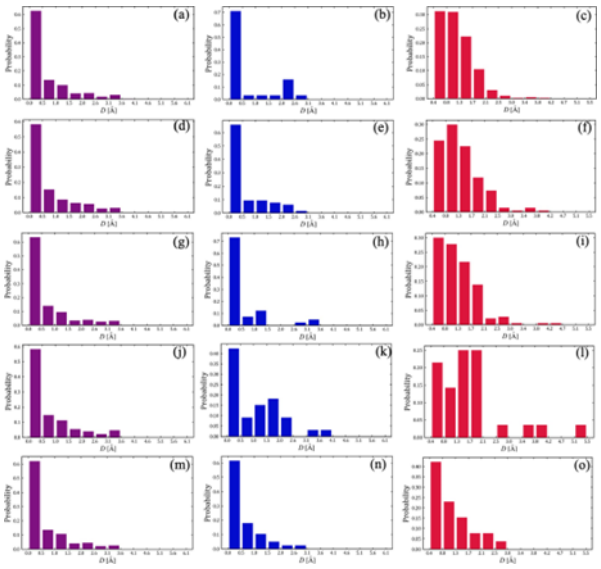
3D-RISM 理論と DL モデルの間の $g_O(r)$の良好な一致については、以下の図からも視覚的に確認できる。ここに (a) 3D-RISM、(b) DL モデル、(c)重ね合わせである。



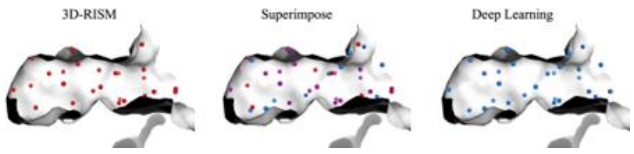
3-2-2　水分子の位置

DL モデルが水・酸素原子の分布関数 $g_O(r)$のピーク位置をうまく予測できるかを検討するため、プログラム Placevent を用いて、3D-RISM 理論または DL モデルによる $g_O(r)$に基づいて水分子を $g_O(r)$のピークに配置した。5 つのタンパク質における、3D-RISM と DL モデルの対応する水の位置間の

距離 Di のヒストグラムを以下に示す。

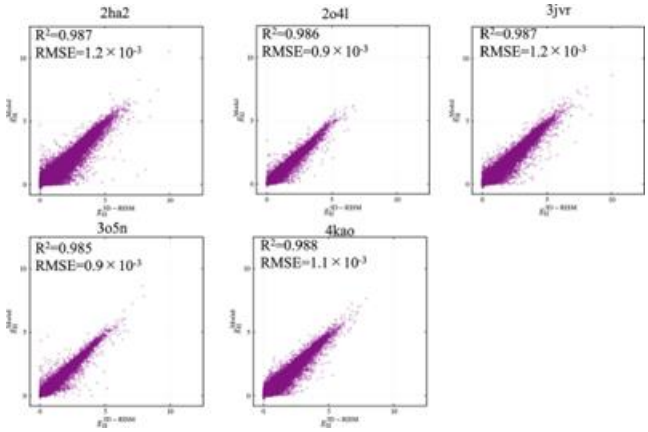


ヒストグラムより、DL モデルに基づく水の約 60%が、3D-RISM 理論に基づく水の位置から 0.5Å 以内にあることがわかった。Di の平均は、5 つのタンパク質すべてで 0.6-0.7Å であった。リガンド結合ポケットでの水の配置についても、本質的に同じ結果が得られた。以下の図からも、良好な一致が視覚的に確認された。



3-2-3　水・水素原子の分布関数の予測

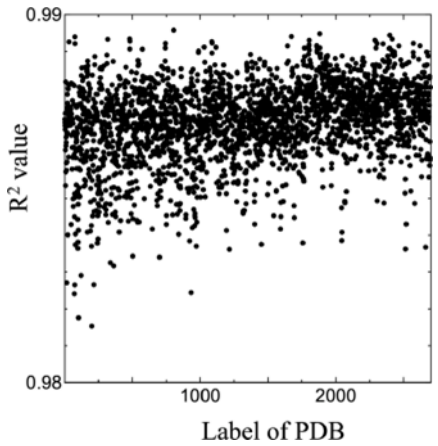
水・水素原子の分布関数 $g_H(r)$を予測するための DL モデルは、水・酸素原子の分布関数 $g_O(r)$の DL モデルで使用されたものと同じ U-net アーキテクチャを使用して構築した。22 タンパク質を使用して $g_H(r)$の学習後、5 つのタンパク質に $g_H(r)$の DL モデルを適用して $g_H(r)$を予測した。予測された $g_H(r)$値と 3D-RISM 理論による $g_H(r)$値との相関関係は、R^2 スコアと RMSE 値を以下に示す。$g_H(r)$を予測する DL モデルは、$g_O(r)$を予測モデルと同様のパフォーマンスを示した。



3-2-4　学習に含まれないタンパク質に対する予測テスト、および、異なるタンパク質を学習に用いた場合の影響

学習およびテストに選択したタンパク質が DL モデルに与える影響を検討するために、3 つの検討を実施した。

まず、これまで学習およびテストしていないタンパク質に対する DL モデルの適用性を検討するため、学習及びテストした 27 タンパク質以外の 2691 タンパク質の $g_O(r)$の予測を行った。2691 タンパク質の R^2 スコアを以下に示す。2691 タンパク質 R^2 スコアは、すべてのタンパク質で 0.98 よりも大きかった。したがって、我々の DL モデルは、学習セットに含まれていないタンパク質の大規模なセットに対しても精度よく予測できることがわかった。



次に、学習とテストに用いる 27 個のタンパク質を多数・ランダムに選択し、モデル構築、予測を行った。どのような選択を行っても、予測性能は変わらなかった。

次に、学習データ数増加の影響について検討した。学習に 44 個のタンパク質を使用したところ、R^2 スコア値がわずかに増加し、RMSE 値がわずかに減少した。ただし、変化は非常に小さく、R^2 スコアは平均 0.01、RMSE は平均 0.2×10^{-3} であった。

3-2-6 DL モデルの側鎖電荷付近の予測性能

DL モデルが電荷を有する側鎖に近い領域でも $g_O(r)$を正しく予測できるかどうかを確認するために、アルギニン側鎖 NH1 および NH2 原子、リジン NZ 原子、アスパラギン酸 OD1 および OD2 原子、グルタミン酸 OE1 および OE2 原子から 5Å 以内の $g_O(r)$値を検討した。3D-RISM と DL モデルの間の $g_O(r)$値の R^2 スコアと RMSE の値はそれぞれ 0.988 と 7.6×10^{-2} であり、電荷を有する側鎖に近い領域においても高い予測性能を示した。

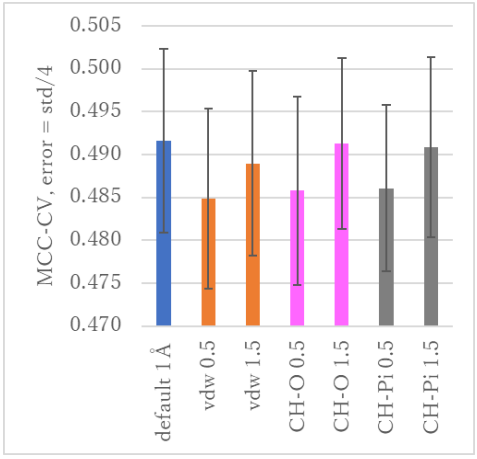
3-2-7 3D-RISM 理論と DL モデルの計算時間比較

計算時間の比較は、single CPU を使用して行った。CPU は Intel Xeon Gold6230 CPU @2.10GHz である。比較は、5 つのテストタンパク質に対して行った。比較結果は以下の表に記載した。DL モデルの計算時間は、3D-RISM 理論の計算時間の約 1/3-1/12 であった。タンパク質アミノ酸数が増えるほど、計算時間の短縮効果が顕著になる。single GPU (GeForce RTX2080Ti)を使用した場合、5 つのタンパク質の計算時間は 2～19 秒であった。

PDB code	number of residues	computation time of the 3D-RISM theory (s) [CPU]	computation time of our DL model (s) [CPU]	computation time of our DL model (s) [GPU]
2ha2	540	12,067	1163	19
2o41	100	5022	190	4
3jvr	266	4567	586	8
3o5n	94	435	190	2
4kao	258	5064	418	6

3-3. 相互作用記述子のパラメータ最適化システム構築

別のプロジェクトで利用している多数のタンパク質の側鎖の構造モデル（正解構造と不正解構造）に対して、正解構造の記述子と不正解構造の記述子を利用して、機械学習による判別モデルを構築した。この時のクロスバリデーションの性能は、Matthew's correlation coefficient (MCC)で評価した。



相互作用定義と機械学習の性能にどの程度の関連があるか調べるため、vdW 相互作用または CH-Pi 相互作用の距離に関する定義をわずかに変化させ、性能に変化があるか調べた。図に示される通り、距離に関する相互作用を変化させるだけで性能に変化が見られた。この場合は、初期定義（default）を利用した場合の性能が最も高かったが、そのほかの相互作用の定義を変化させた場合や、複数の定義を同時に変化させた場合に、機械学習の性能がさらに高まる可能性があることが示唆された。

次に、記述子の定義（パラメータ）を、機械学習の性能が高まるように並列ベイズ探索するシステムを構築した。このシステムのプロダクションランによって、機械学習の性能を高める定義の探索を実施予定である。

4.　まとめ

4-2　3D-RISM 法により計算されたタンパク質水和状態を高速に推算する AI システムの開発

本研究では、U-net アーキテクチャに基づいてタンパク質周囲の水和構造を予測するための DL モデルを提案した。タンパク質の 3D 構造のみを入力とし、水和状態を表す、水・酸素原子分布関数 $g_O(r)$、あるいは、水・水素原子分布関数 $g_H(r)$を出力する。

我々の DL モデルは、学習セットに含まれていない 5 つのタンパク質と、2691 のタンパク質について、3D-RISM 理論による $g_O(r)$と $g_H(r)$を、決定係数 R^2 スコアの値で、それぞれ 5 タンパク質で約 0.98、2691 タンパク質で 0.98 と精度よく予測できた。さらに、水・酸素原子の位置を DL モデルと 3D-RISM 理論の間で比較したところ、3D-RISM 理論の $g_O(r)$のピーク位置を正確に予測した。3D-RISM 理論の水

の位置と DL モデルの水の位置の間の距離 Di の平均(0.6-0.7Å)は、水・酸素原子のサイズ 3Å と比較して小さかった。水・水素原子の分布関数 $g_H(r)$についても、精度よく予測できた。タンパク質全体について、DL モデルは平均して single GPU を使用して 1 分以内に $g_O(r)$を予測した。さらに、リガンド結合ドメインなどの関心のある領域のみの $g_O(r)$を予測することも可能である。

我々の DL モデルは、single GPU を使用して短時間で計算可能なため、リガンド結合プロセス、タンパク質構造変化、アミノ酸変異に伴うタンパク質周辺の水和状態変化など、多数の系の水和状態を検討することが可能となった。

以下、我々の DL モデルと関連する 3 つの方法論の比較について述べる。

Sosnin らによって提案された DL と 3D-RISM 理論のハイブリッド法は、3D-RISM 理論を使用して得られた $g_O(r)$と $g_H(r)$を入力として、有機分子の生物濃縮係数値を予測する。したがって、予測する対象が全く異なる方法である。さらに、Sosnin らが使用した DL モデルも、我々のものとは異なっていた。

Ghanbarpour らは、タンパク質周辺の水和構造を予測するための DL モデルを提案した。彼らは、水和状態を水の占有率、つまり水分子が特定の位置で見つかる確率によって特徴付けた。水占有率は、その定義から、分布関数 $g_O(r)$に関連しているものではあるが、同一ではない。さらに、我々の研究は、次の点で Ghanbarpour らによる研究とは異なる。

1. 我々の DL モデルは $g_O(r)$を予測するが、Ghanbarpour らの DL モデルは水占有率のクラスを予測する。
2. 我々の研究では、Placevent を使用し $g_O(r)$に基づいて水・酸素原子を配置し、DL モデルが $g_O(r)$のピーク位置をうまく予測できているかを議論したが、Ghanbarpour らの DL モデルでは不可能である。
3. 我々の DL モデルはタンパク質全体の $g_O(r)$を予測できるが、Ghanbarpour らの DL モデルは、リガンド結合ポケットでのみ予測できる。したがって、我々の DL モデルは、水和自由エネルギーや Partial Molar Volume などタンパク質全体の水和熱力学的量を計算できる。

1 から、我々の DL モデルは定量的な議論に適しており、Ghanbarpour らの DL モデルは定性的な議論に適していると言える。例えば、水・酸素原子を使った議論が必要な場合、我々の DL モデルが適している。一方、水占有率に応じた水分子ランクの議論が必要な場合は、Ghanbarpour らの DL モデルが適している。

丸山らは、GPU を使用する 3D-RISM 計算高速化アルゴリズムを提案し、タンパク質の 3D-RISM 計算を、Tesla-K40 GPU を使用して数分以内に終了することができた。我々の DL モデルは、丸山らと比較して、2 つの利点がある。まず、single CPU を使用しても、計算は迅速に完了する。さらに、我々の DL モデルは、タンパク質全体が 483 ボクセルの小さなボックスに分割されているため、リガンド結合ポケットなど、タンパク質の特定の領域で $g_O(r)$の計算が可能であるが、丸山らの方法ではこのような計算は実行不可能である。

5.　今後の計画・展望

5-1　サイクリックペプチド AI 力場開発

現在のところ、cyclic peptide 各コンフォメーションのエネルギーのみを目的変数とするより、各原子の force も目的変数に加えて実施したほうがよさそうであるので、この方向で進めていきたい。

また、作成した AI 力場を用いた構造最適化のプロトタイプが完成したので、ここで最適化した構造が、初期構造に比べて実際にエネルギー的に低くなっているかを FMO 計算し確認していきたい。

5-2　3D-RISM 法により計算されたタンパク質水和状態を高速に推算する AI システムの開発

今回開発した DL モデルの制限の 1 つは、使用できる原子タイプが、炭素、窒素、酸素、硫黄、水素に限定されていることです。したがって、他の原子(金属、リン酸化アミノ酸のリン、セレノメチオンのセレン、イオン、リガンドのハロゲンなど)を含むタンパク質に対しては、現在の DL モデルは適用不可能である。DL モデルの適用範囲を広げるためには、原子タイプの数を増やし、学習を行う必要がある。これらについては将来検討していく。

また、熱力学量の計算についても、DL モデルの応用として重要である。例えば、水和自由エネルギーは、$g_O(r)$、$g_H(r)$、および Singer と Chandler によって導出された水和自由エネルギーの解析式による直接相関関数を使用して計算できる。水和エントロピーと水和エネルギーは、水和自由エネルギーを使用して計算できる。また、Partial Molar Volume

(PMV)は、$g_O(r)$、$g_H(r)$、および Kirkwood−Buff 溶液理論で計算できる。DL モデルを使用した、これら熱力学的量の高速計算についても、将来的に実施していきたいと考えている。

5-3　相互作用記述子のパラメータ最適化システム構築

記述子の定義（パラメータ）を、機械学習の性能が高まるように並列ベイズ探索するシステムを構築した。このシステムのプロダクションランによって、機械学習の性能を高める定義の探索を実施予定である。

6.　利用がなかった場合の理由