2019 年度　利用報告書

課題名（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)、髙野 浩顕 (1)、
千葉 峻太朗 (2)、池口 満徳 (2)、松本 篤幸 (2)、
吉留 崇 (3)、津田 和実 (3)、中川 寛之 (3)、小甲 裕一 (3)、宮口 郁子 (3)、中田 一人 (3)、
鹿島 亜季子 (3)、佐藤 美和 (3)、石井 裕子 (3)、小澤 基裕 (3)、小久保 裕功 (3)、藤原 崇幸 (3)、
大島 勘二 (3)、小野 聡 (3)、半田 千彰 (3)、中尾 直樹 (3)、角谷 龍展 (3)、国本 亮 (3)、古川 祐貴 (3)、
馬 彪(4)、井阪 悠太(4)

理研における所属研究室名：

(1) 医科学イノベーションハブ推進プログラム 医薬プロセス最適化プラットフォーム推進グループ
(2) 医科学イノベーションハブ推進プログラム 医薬プロセス最適化プラットフォーム推進グループ 創薬バイオメディカルインテリジェンスユニット
(3) 医科学イノベーションハブ推進プログラム 医薬プロセス最適化プラットフォーム推進グループ 分子設計インテリジェンスユニット
(4) 健康生き活き羅針盤リサーチコンプレックス推進プログラム 健康予測チーム

1. 本課題の研究の背景、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、理化学研究所健康生き活き羅針盤リサーチコンプレックス推進プログラム、医薬基盤・健康・栄養研究所、公益財団法人 都市活力研究所）に所属し、各機関と連携の上で課題を実施している。

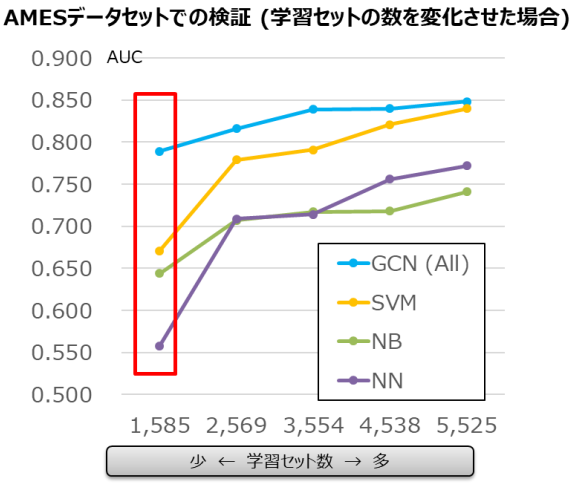
2.　具体的な利用内容、計算方法

2-1. 医薬品候補化合物の特性予測のための深層学習モデル

2-1-1 目的：創薬の際には、薬理活性、物性、体内動態および毒性など全ての特性において、薬として満足できるレベルをクリヤーする、つまり適切な投与量、投与方法、投与頻度で薬効を発揮し、安全に使用することができる化合物を早く同定する必要がある。実際の創薬プロセスは、High Throughput Screening などによる活性物質の取得後、化合物の分子設計→化合物の合成→実験による各特性値の検証を繰り返し実施し、全ての特性項目を満足するような化合物を創製するという方法で実現されている。化合物の合成と各特性の実験的検証には多くの時間と労力と費用がかかるため、これらの特性を計算機上で予測し、なるべく少ないステップ数で、全ての特性項目を満足するレベルの化合物を早期に創製することが求められている。

2-1-2 方法：そこで、本課題では、人工知能技術の近年の breakthrough である深層学習（Deep Learning, DL）法により、化学構造から各特性値を予測する予測モデルを作成している。本方法においては、化学構造を原子が結合したグラフで表し、各原子の周囲の環境を反映させる仕組みとして畳み込み(Convolution)手法を用いて化合物を記述する Graph Convolutional Networks (GCN)を用いている。Deep Learning は画像解析のように一般的に大量のデータを必要とするが、創薬初期においては何千万というような大量のデータを利用した解析が可能なわけではなく、少ないときは数百のデータで学習をしないといけない場合もある。そこで、化合物の安全性評価で重要な、化合物が DNA に変異を引き起こす性質を表す AMES 試験データ解析について、データ数を変化させ、Deep Learning 以外の複数の解析手法も含めて、その予測性に差があるかを検討した。使用した AMES データ(J. Chem. Inf. Model., 2009, 49, 2077-2081.) は、データ数 6,512 (pos: 3,503 / neg: 3,009)である。Deep Learning による学習には計算時間がかかるので、これを多数条件で行うために HOKUSAI を利用した。

2-1-3 結果：　結果を以下に示す。GCN（水色）は GCN を使った DL、SVM（黄色）は Support Vector Machine 法、NB（緑色）は Naïve Bayes 法、NN（紫色）は Neural Networks 法の結果を示している。



化合物の構造的特徴を表す Fingerprint で化学構造を記述子化し、機械学習でよく用いられる SVM、NB、NN 法で解析した場合、用いるデータ数が少ない場合は、予測力が顕著に落ちていく。一方、化学構造を GCN で記述子化し、DL で解析した場合はデータ数が少数になっても、その予測力が大きく落ちないことが明らかになった。

2-1-4 今後の計画・展望：今後は、本方法を多数のデータセットに適用し、その一般性を検討する予定である。

2-2. タンパク質リガンド結合部位の水和状態の 3D-RISM による網羅的解析

2-2-1 目的：　タンパク質とリガンド、あるいは、タンパク質同士が結合するときにはタンパク質周囲にある水和水が脱水和する過程が必要になり、脱水和エネルギーが、結合するリガンドあるいはタンパク質の結合自由エネルギー（＝結合親和性）に大きく寄与する。脱水和に要するエネルギーは、脱水和する水の水和状態に依存する。そこで、数多く報告されているタンパク質・リガンド複合体構造の水和状態を網羅的に解析することにした。水の水和状態を知るための方法として、タンパク質と水の分子動力学(MD)計算を行い水の挙動を解析する water map 法があるが、MD 計算に時間がかかるため、例えば数千のタンパク質の水和状態を解析することは難しい。そこで、計算時間の面で MD ほどの負担ではない 3D-RISM 法を用いてタンパク質の水和状態を網羅的に解析することにした。

2-4-2 方法：　タンパク質周囲の水和状態は、X 線結晶構造からリガンドと水分子を取り除いたリガンドなし構造に対して AmberTools18 suite により 3D-RISM 計算を行い、タンパク質周囲の位置 r に水分子がどのくらいの確率密度 $g_o(r)$ で存在するかで表した。最初のステップでは、水のみの water site-site correlation function の 3D-RISM 計算を実施した。水の力場とパラメータは coincident SPC/E model を用いた。水の誘電定数は 78.497、bulk density は 0.03332 $Å^{-3}$ を用いた。温度は 310 K に設定した。次にタンパク質を solute とした、タンパク質周囲の water-site correlation function $g_o(r)$の 3D-RISM 計算を実施した。タンパク質とイオンの力場とパラメータについては Amber ff99SB を用いた。タンパク質のヒスチジン残基はδ窒素に水素を付けた HID を採用した。Water box については、タンパク質から water box の端までが最低 14 Å 離れるように設定した。格子点間隔は x, y, z 方向全てで 0.5 Å とした。収束ステップ数は 20,000 とした。

計算対象は、結合親和性およびリガンドとタンパク質の結合モードが X 線結晶構造として明らかになっている PDBbind ver. 2017 の 4,154 複合体構造である。なお、PDBbind データセットにおいては結晶水以外には水素が付加されている。3D-RISM でタンパク質周囲の water-site correlation function $g_o(r)$を計算する際は、PDBBind の構造から、低分子リガンドと結晶水を除いたリガンドフリー構造を用いた。取り除いた低分子リガンド構造は SYBYL Mol2 形式と MDL SDF 形式の 2 形式で、それぞれ保存した。

3D-RISM 計算の前処理として、(a) N 端、C 端アミノ酸残基と構造が見えないアミノ酸残基の切れ目のアミノ酸残基には、N 端にアセチル、C 端にアミドを自動的に capping 処理し、(b) AMBER tools18 の tleap コマンドを用いて各原子に AMBER atom type を自動的に割り振った。tleap コマンドでエラーを発生したものは以後の計算から取り除いた。さらに AmberTools18 を用いて、重原子に 10 kcal/mol/ $Å^2$ の拘束をかけることにより水素原子のみを構造最適化した。構造最適化において溶媒効果は generalized-Born model を用いた。構造最適化計算でエラーを発生したものは以後の計算から取り除いた。最終的に 3,706 構造の 3D-RISM 計算を実施した。

結晶水の分布関数 $g_o(r)$の解析 : タンパク質表面にある結晶水の水和状態を解析するために、結晶水から 5 Å 以内にタンパク質の重原子があるもののみ解析した。その結果、2,403 タンパク質由来の約 620,000 の結晶水を解析した。結晶水の位置の $g_o(r_{cw})$を 3D-RISM 計算で計算した各格子点上の $g_o(r)$から算出するために、結晶水の位置から 0.9Å 以内にある格子点上の $g_o(r)$で最大の値をもつものを $g_o(r_{cw})$とした。$g_o(r_{cw})$の値は大きくなればなるほど、結晶水の位置にある水の存在確率が高くなる。$g_o(r)$=1 はバルクにある水の存在確率なので、もしも 3D-RISM 計算がうまくいっていれば、計算された結晶水の $g_o(r_{cw})$は 1 以上になることが期待される。$g_o(r_{cw})$の値と結晶水の環境をさらに検討するために、結晶水が接触しているタンパク質の重原子数を数えた。接触しているかどうかの閾値はプログラム HBPLUS で用いられている 3.9Å を用いた。さらに接触しているタンパク質重原子の元素の傾向を解析するために、結晶水に最も近いタンパク質重原子を同定した。

リガンド重原子の分布関数 $g_o(r_{lha})$の解析 : リガンド重原子が占有している位置の水和状態を解析した。リガンド重原子の位置の $g_o(r_{lha})$を 3D-RISM 計算で計算した各格子点上の $g_o(r)$から算出するために、結晶水の場合と同様に、リガンド重原子の位置から 0.9Å 以内に

ある格子点上の $g_o(r)$で最大の値をもつものを $g_o(r_{lha})$とした。ρを溶媒の密度としたとき、$\rho\, g_o(r_{lha})\ \Delta V$ はリガンド重原子周辺の微小体積ΔV 中の水の数となる。したがって、$g_o(r_{lha})$の値が大きくなればなるほど、リガンドが結合することによって置き換える水分子の数が多くなる。リガンド重原子の解析においては各原子を以下の SYBYL atom type に分類して実施した。

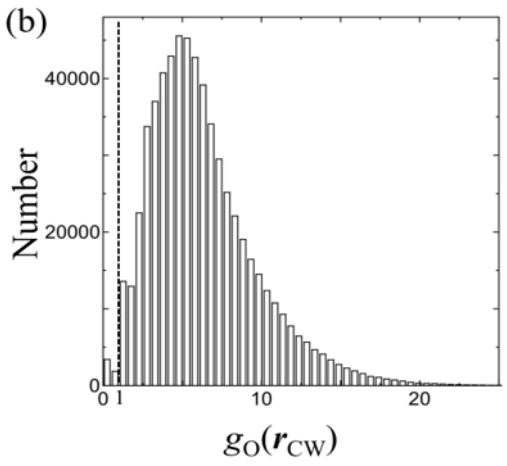
SYBYL Atom type	Notation
Hydrogen	H
Carbon sp3	**C.3**
Carbon sp2	**C.2**
Carbon sp	C.1
Carbon aromatic	**C.ar**
Carbocation	**C.cat**
Nitrogen sp3	N.3
Nitrogen sp2	**N.2**
Nitrogen sp	N.1
Nitrogen aromatic	**N.ar**
Nitrogen amid	**N.am**
Nitrogen trigonal planar	**N.pl3**
Nitrogen sp3 positively charged	**N.4**
Oxygen sp3	**O.3**
Oxygen sp2	**O.2**
Oxygen in carboxylates and phosphates	O.co2
Sulphur sp3	**S.3**
Sulphur sp2	S.2
Sulphoxide sulphur	S.o
Sulphone sulphur	**S.o2**
Phosphors sp3	**P.3**
Fluorine	**F**
Chlorine	**Cl**
Other halogens and metals	-

太字で示した atom type については、解析に十分な数のデータが得られたので、解析を行った。リガンド重原子の SYBYL atom type は SYBYL mol2 形式ファイルに記載されている atom type の情報を利用した。また、リガンド重原子が接触しているタンパク質重原子の元素の傾向を解析するために、リガンド重原子に最も近いタンパク質重原子を同定し、その元素を解析した。

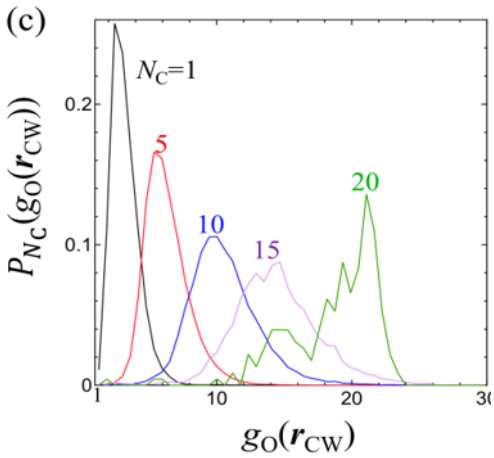
リガンドの解析は、X 線結晶構造の「正解」構造と、AutoDock vina でのドッキングによりわざと間違えた「不正解」構造の両者を比較することにより行った。不正解構造は、3,706 複合体構造に対して AutoDock vina でのドッキングを行い、リガンド X 線結晶構造との RMSD が 5 Å に一番近いドッキングポーズを、それぞれの複合体における候補ポーズとした。候補ポーズの中で RSMD が 4.5 Å~5.5 Å の範囲に入るものを最終的に「不正解」構造とした。

2-2-3 結果：自動 3D-RISM 前処理および 3D-RISM 計算は、HOKUSAI で実施することによって、約 3700 の PDBbind 構造を一週間程度で計算することができた。

結晶水の $g_o(r)$の解析：約 620,000 の結晶水の水和状態 $g_o(r)$をヒストグラムで表したものが以下の図となる。



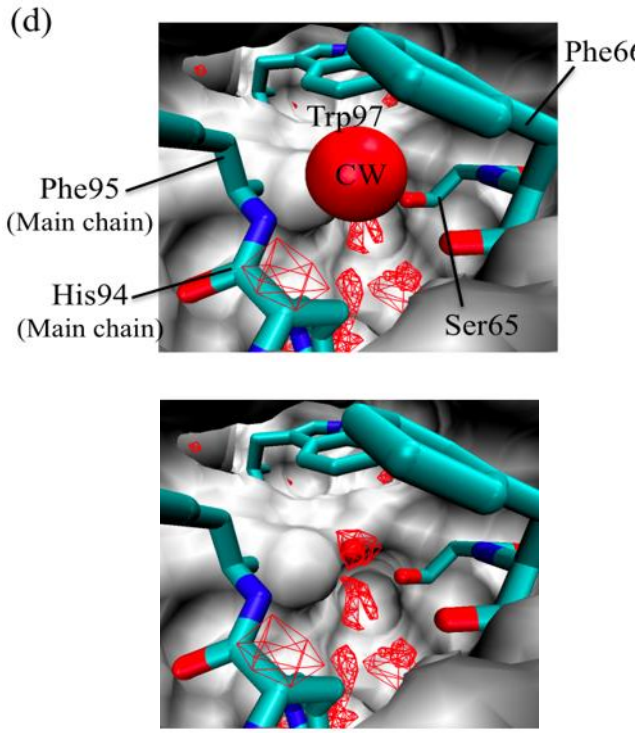
ヒストグラムは、$g_o(r)$=5 付近をピークとするポアソン分布のような形となった。点線は、水の存在確率がバルクの水と同じである $g_o(r)$=1 を表していることから、大部分の結晶水の位置の $g_o(r)$は 1 以上で、バルクの水の存在確率よりも高くなっていることから、3D-RISM 計算結果がリーズナブルであると判断した。



また、タンパク質周囲の結晶水の水和状態としては $g_o(r)$=5 付近をとるものが多く、例えば $g_o(r)$>20 となるようなケースは、かなり稀であることがわかった。また、結晶水がタンパク質重原子と接している数（N_C = 1,

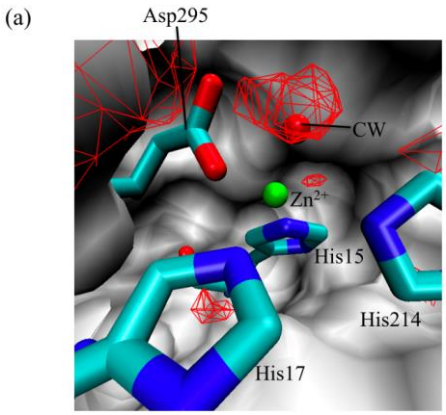
5, 10, 15, 20）と結晶水の $g_o(r)$の関係をプロットすると、以下となる。各 N_C の $g_o(r)$のピーク位置は、N_C の値が増加するにつれ、$g_o(r)$の値が増加しており、結晶水がタンパク質重原子とより多く接するにつれて、より水和が起こっている（＝水の存在確率が高くなっている）ことが明らかになった。

上記のように、結晶水がタンパク質と数多く接触しており、より水和が起こっている例、ここでは N_C=20 で $g_o(r_{cw})$=14.4、を以下に示す(PDB code: 1azm)。



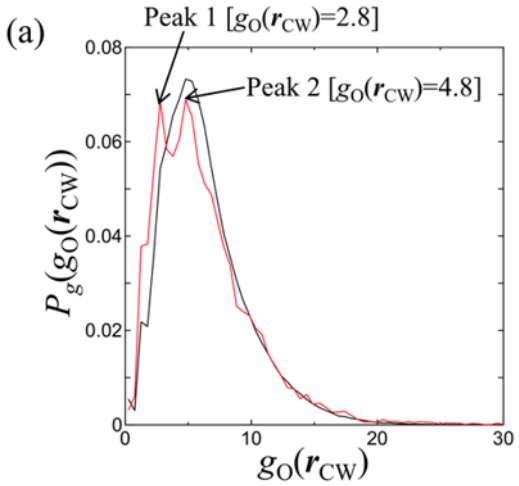
中心にある赤い球が結晶水であり、赤いメッシュは $g_o(r)\geq 6$ の領域を示している。結晶水は Ser65, Phe66, His94, Phe95, Trp97 などに囲まれており、結晶水周囲の赤いメッシュは結晶水付近に局在している。

周囲をタンパク質で囲まれることにより $g_o(r_{cw})$が高くなるケース以外に、金属の周囲にある結晶水の $g_o(r_{cw})$が高くなる例、ここでは金属は Zn で $g_o(r_{cw})\geq 30$、が見受けられたので以下に示す（PDB code: 1add）。

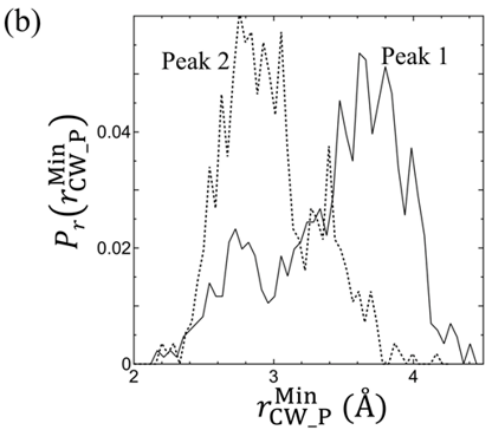


赤いメッシュは $g_o(r)>4$ の領域を示している。$g_o(r_{cw})\geq 30$ となるようなケースは、ほとんどの場合金属に配位している結晶水であり、残りの少数のケースはタンパク質内部に埋もれた結晶水であることもわかった。

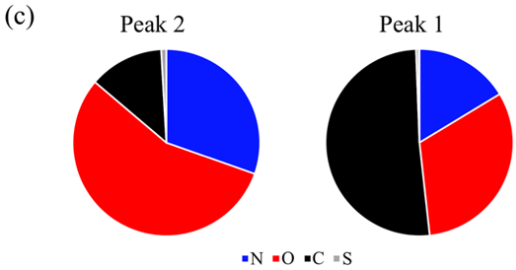
リガンド周囲の水和状態の解析：リガンドがその周囲の水和状態に与える影響を検討するために、リガンドに接する結晶水の $g_o(r_{cw})$を検討した。結晶水がリガンドと接しているかは、リガンド重原子と結晶水間の距離が 4.0 Å 以下となるかどうかで判断した。リガンドと接している結晶水の数は約 45,000 であった。リガンドと接する結晶水の $g_o(r_{cw})$の分布を赤線で、全結晶水の $g_o(r_{cw})$の分布を黒線で示した。



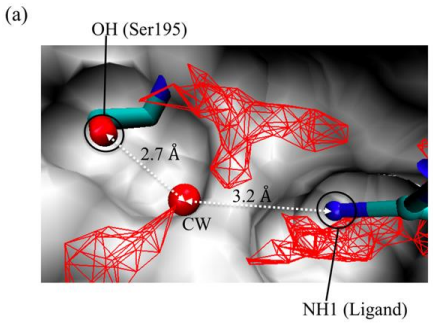
このグラフから、$g_o(r_{cw})\geq 10$ の領域では、両者にほとんど差異は見られないのに対して、$g_o(r_{cw})<10$ の領域は異なっている。リガンドと接する結晶水の $g_o(r_{cw})$の分布（赤線）においては、全結晶水の $g_o(r_{cw})$の分布（黒線）に比べて $g_o(r_{cw})$=4.8 のピーク 2 が少し低くなり、$g_o(r_{cw})$=2.8 のピーク 1 が出現している。これらのピークの由来が何であるかを検討するために、結晶水とタンパク質重原子間の距離の最小値 $r_{cw\_p}^{Min}$ と、その存在確率 $P(r_{cw_p}^{Min})$を求めてプロットしたのが以下である。



$g_o(r_{cw})$=2.8 のピーク 1（黒線）は結晶水とタンパク質重原子間の距離のピークが 3.6 Å 付近にあるが、より水和した $g_o(r_{cw})$=4.8 のピーク 2（点線）においては、結晶水とタンパク質重原子間距離 3.6 Å 付近のピークがなくなり、2.8 Å 付近のピークの割合が増加していた。さらに、より水和した $g_o(r_{cw})$=4.8 のピーク 2 の結晶水が接するタンパク質重原子の元素は、以下の左側のパイチャートのように主に窒素と酸素であるのに対し、$g_o(r_{cw})$=2.8 のピーク 1 では右図のように半分以上が炭素になっている。

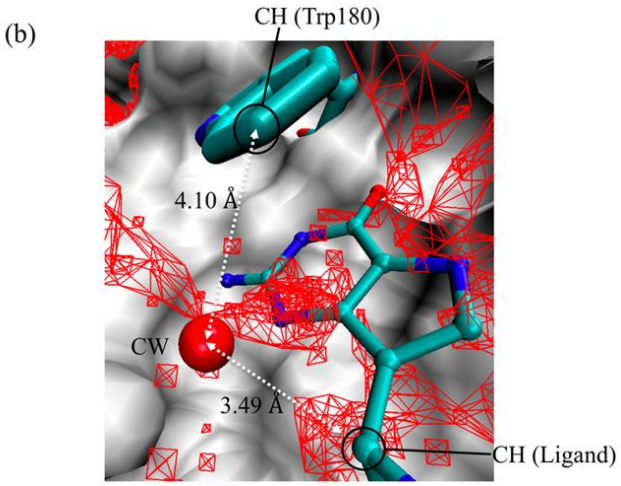


これらを考え合わせると、リガンド付近のより水和した $g_o(r_{cw})$=4.8 のピーク 2 に属する結晶水は、タンパク質の窒素あるいは酸素原子と 2.8 Å 程度の距離で水素結合しており、リガンド付近の $g_o(r_{cw})$=2.8 のピーク 1 に属する結晶水は、タンパク質の炭素原子と 3.6 Å 程度の距離で CH-O などの弱い相互作用をしていることが多いという彫像が浮かび上がってくる。$g_o(r_{cw})$=4.8 のピーク 2 に属する結晶水の具体例を以下に示す。



タンパク質は PDB code: 1a4w で、赤いメッシュは $g_o(r)\geq3.8$ で表示した。結晶水はタンパク質 Ser195 側鎖の水酸基と水素結合している。さらに、結晶水はリガンドのアミノ基とも水素結合しており、タンパク質ー結晶水ーリガンドの水素結合ネットワークを形成している。

$g_o(r_{cw})$=2.8 のピーク 1 に属する結晶水の具体例 (PDB code: 1dqn) を以下に示す。

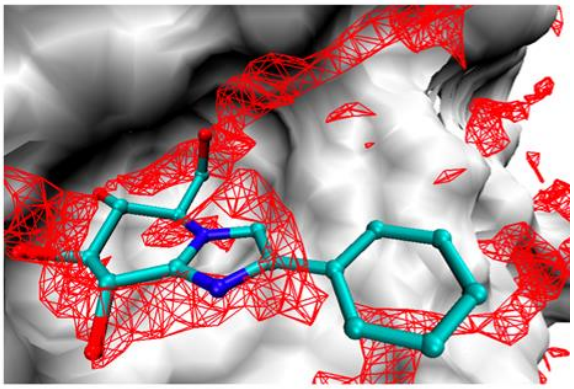


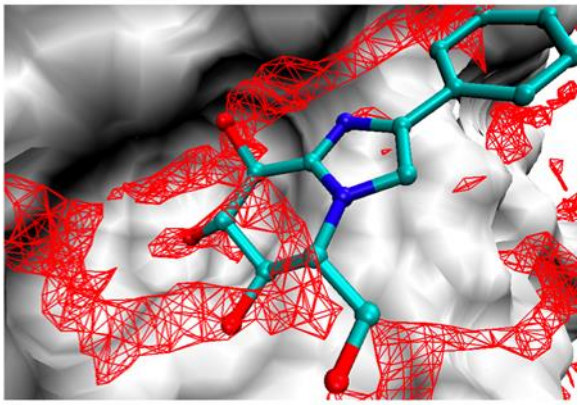
赤いメッシュは $2.3\leq g_o(r)\leq3.0$ で表示した。結晶水はタンパク質 Trp181 側鎖の CH と CH-O 相互作用している。さらに、結晶水はリガンドの CH とも CH-O 相互作用しており、タンパク質ー結晶水ーリガンドの CH-O 相互作用ネットワークを形成している。

このように 2 つの系において、タンパク質ー結晶水ーリガンドのネットワーク形成が見いだされた。これらはリガンド結合後の水の再配置によって引き起こされたものであり、相互作用ネットワーク形成により結晶水の安定化が起こっているものと考えられる。

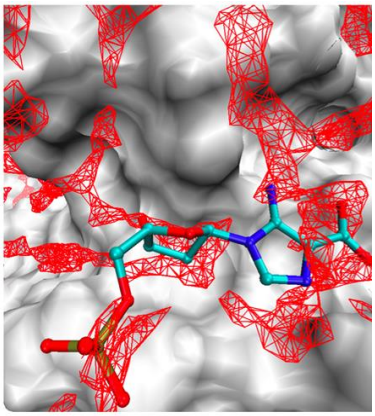
リガンド重原子の水和状態の解析：リガンドがタンパク質に結合する際の占有領域の水和状態が「正解」結合ポーズと「不正解」結合ポーズで、どのように異なるかを検討するために、正解ポーズと不正解ポーズにおけるリガンド重原子の位置の $g_o(r_{lha})$を検討した。

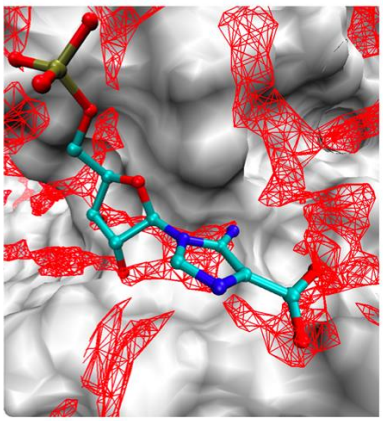
まず、一つの例としてβ-D-glucan glucohydrolase (PDB code: 1x38)を以下に示す。上段が正解ポーズ、下段が不正解ポーズであり、タンパク質周囲の $g_o(r)$ と共に表示した。赤いメッシュは $g_o(r)\geq3.7$ で、リガンドはボールアンドスティックでタンパク質表面は白色で表示した。正解ポーズでは、酸素原子や窒素原子などリガンドの極性ヘテロ原子と、より水和された赤いメッシュの領域の重なりが非常に良いのに対し、不正解ポーズではリガンド極性原子と赤いメッシュの領域の重なりがあまり良くないことが見て取れる。





もう一例として N5-Carboxyaminoimidazole ribonucleotide mutase (PDB code: 2nsl)を以下に示す。上段が正解ポーズ、下段が不正解ポーズである。

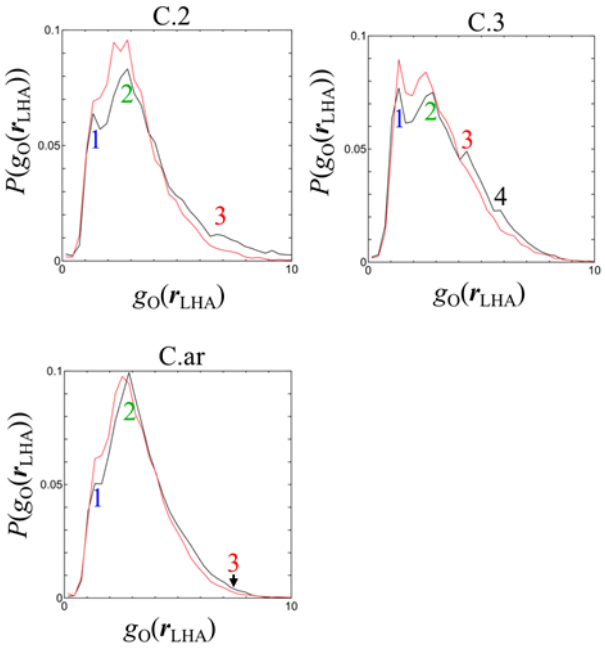




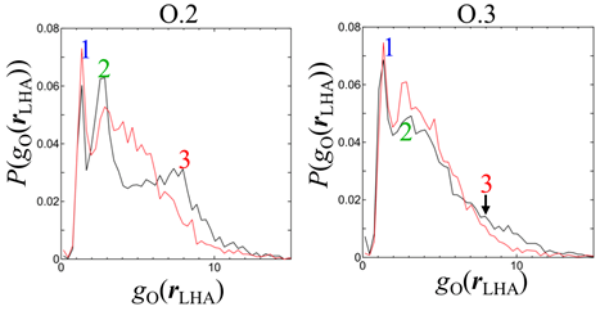
赤いメッシュは $g_o(r)\geq 3.5$ で表示した。このケースにおいても、正解ポーズではリガンド極性ヘテロ原子とより水和された赤いメッシュ領域の重なりが非常に良いのに対し、不正解ポーズではリガンド極性原子と赤いメッシュ領域の重なりがあまり良くない。

上記の例などからリガンド重原子の位置における水和状態 $g_o(r_{lha})$を網羅的に解析するため、リガンド重原子を Table 1 の SYBYL atom type に従って分類し、その水和状態 $g_o(r_{lha})$とタンパク質接触原子を検討した。

リガンド重原子が sp2 carbon (C.2), sp3 carbon (C.3), aromatic carbon (C.ar)などの炭素原子の場合、正解ポーズ（黒線）と不正解ポーズ（赤線）では大きな違いは見られない。

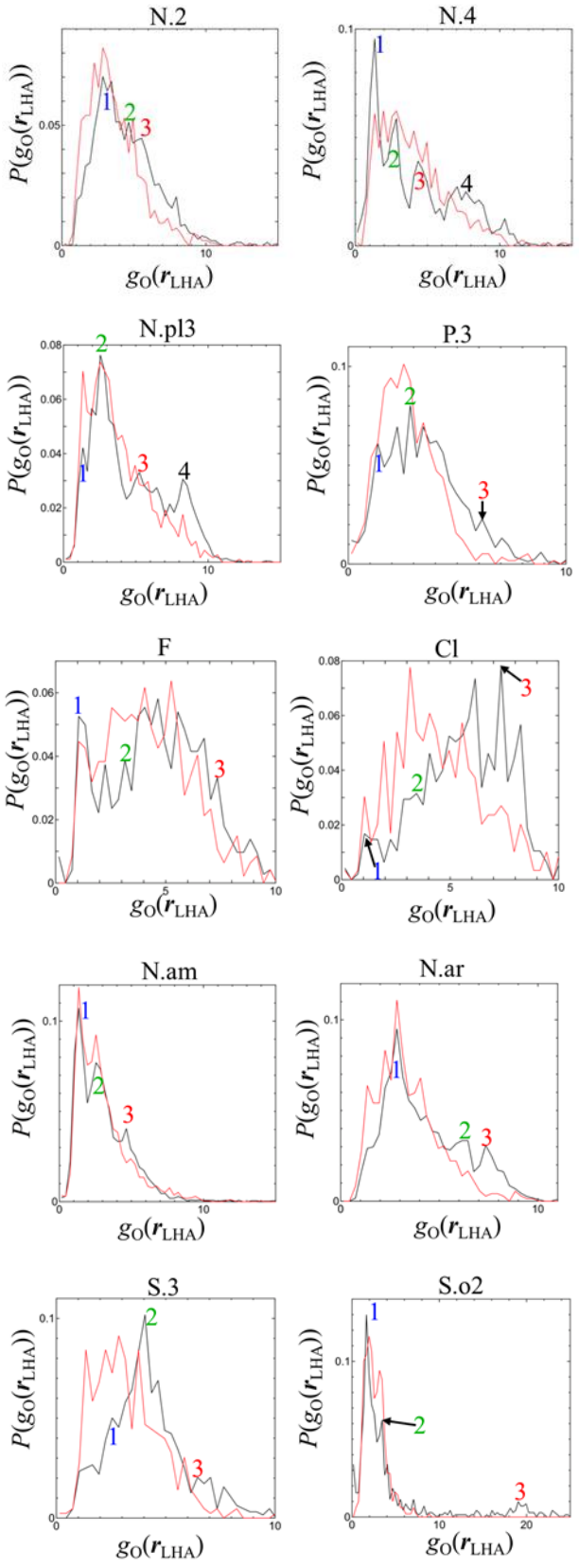


しかしながら、ヘテロ原子の場合は、特に $g_o(r_{lha})$の値が大きい、より高い水和状態にある領域で、不正解ポーズ（赤色）に比べて正解ポーズ（黒色）の確率が高くなっている。例えば、カルボニル酸素のような sp2 酸素(O.2)の場合、$g_o(r_{lha})$=1.35, 2.85, 7.95 の 3 つのピークがあるが、高い水和状態のピーク 3 ($g_o(r_{lha})$= 7.95) は正解ポーズでは高く、不正解ポーズでは低くなっている。



一方、低い水和状態のピーク 1 ($g_o(r_{lha})$= 1.35)は、不正解ポーズでは一番高く、正解ポーズでは不正解ポーズ比べて低くなっている。同様な振る舞いは sp2 窒素

(N.2, ピーク 1 と 3), カチオン性窒素(N.4, ピーク 1 と 4), 平面的 sp3 窒素(N.pl3, ピーク 1 と 4), sp3 リン(P.3, ピーク 1 と 3), フッ素(F, ピーク 1 と 3), 塩素(Cl, ピーク 1 と 3)でも見られた。



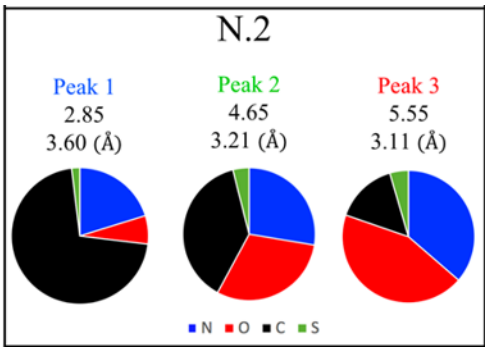
これらの結果から、リガンドヘテロ原子の位置の $g_o(r_{lha})$の情報に基づいて、正解ポーズと不正解ポーズを見分けることが可能ではないか、ヘテロ原子の $g_o(r_{lha})$の値が高いものが正解ポーズではないかということが強く示唆される。また、このような網羅的な大規模解析が可能になったのは、多くの計算コストを要する MD を利用する WaterMap 法などとは異なり、3D-RISM 法では 1 タンパク質あたり長くても 2 時間以内で終了するという比較的軽微な計算コストであることも貢献していると考えられる。

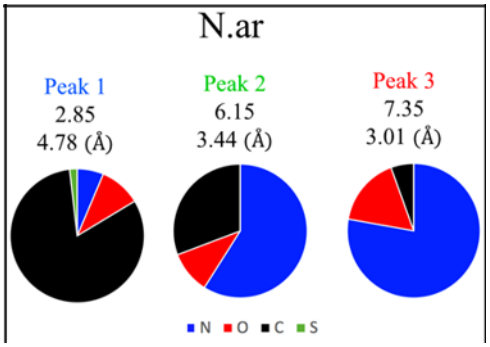
$g_o(r_{lha})$のピークの解析：各 atom type 毎の正解ポーズと不正解ポーズの $g_o(r_{lha})$のヒストグラムに、番号が大きくなると高い水和状態になるように付番したピークについて、タンパク質－リガンド間あるいは結晶水－タンパク質間の相互作用の由来を解析した。i 番目のピークにおいて、そのピークに属する、SYBYL atom type X を有するリガンド重原子とタンパク質重原子間の最近接距離 r_{lha}^{min} を求め、これらを平均した値を $\langle r_{lha}^{min}\rangle_i^X$ とした。以下に各 atom type 毎の各ピーク毎の、$g_o(r_{lha})$、$\langle r_{lha}^{min}\rangle_i^X$、リガンド重原子が接触するタンパク質重原子の元素を示す。全ての atom type において $g_o(r_{lha})$が大きくなり、より水和すると、リガンド－タンパク質間距離 $\langle r_{lha}^{min}\rangle_i^X$ が短くなっている。SO_2 イオウ原子 S.o2 を除く全ての atom type において、リガンド－タンパク質間距離 $\langle r_{lha}^{min}\rangle_i^X$ が水素結合距離(3.2 Å)以下の場合は、タンパク質のリガンド接触元素として、水素結合可能な酸素と窒素の割合が多くなっている。各ピークの $g_o(r_{lha})$の値が小さくなるにつれて、タンパク質のリガンド接触元素として、炭素の割合が多くなっている。これらの現象は、タンパク質の立体構造がその周囲の水和状態を規定し、水和状態が置き換えられるべきリガンドの元素を決めているという文脈で起きていると考えられる。

各 atom type についての解析は、まず、sp2 窒素(N.2), カチオン性窒素(N.4), アミド窒素 (N.am), 芳香族窒素(N.ar), 平面的 sp3 窒素(N.pl3)などの atom type が含まれる窒素原子から行う。

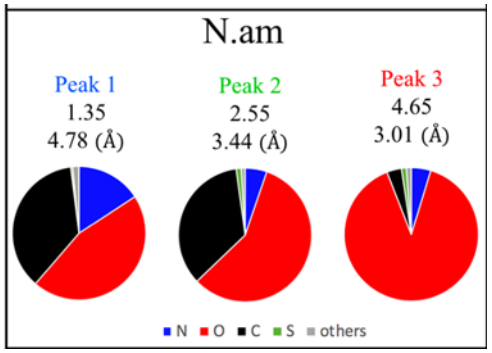
リガンド－タンパク質間距離 $\langle r_{lha}^{min}\rangle_i^X$ が水素結合距離(3.2 Å)以下のピークでは、タンパク質のリガンド接触元素はほぼ酸素と窒素である。孤立電子対を有し、水素結合アクセプターとして働く N.2 (ピーク 3 : $g_o(r_{lha})$=5.55, $\langle r_{lha}^{min}\rangle_3^{N.2}$=3.11 Å)、N.ar (ピーク 3 : $g_o(r_{lha})$=7.35, $\langle r_{lha}^{min}\rangle_3^{N.ar}$=3.01 Å)の場合、タンパク質のリガンド接触元素として、NH の形で水素結合ドナー

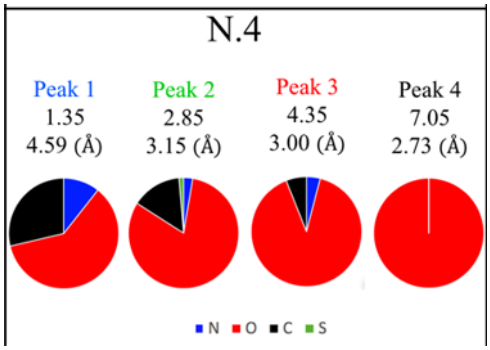
として相互作用できる窒素の割合が高くなっている。

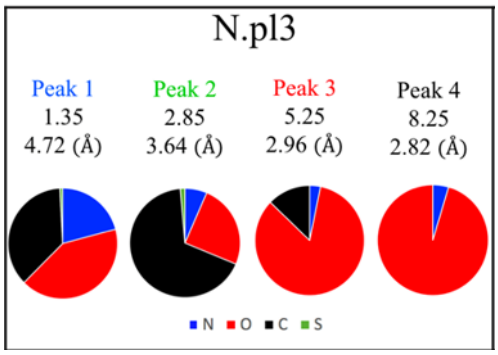




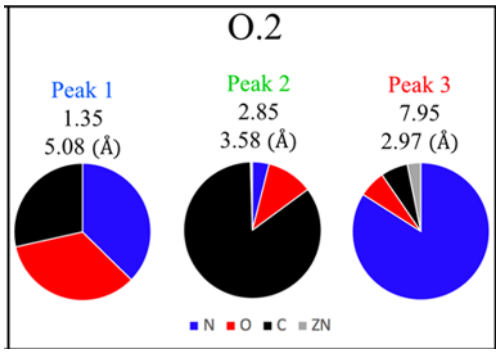
多くの場合水素が 1 つ以上結合しており、水素結合ドナーとして相互作用する N.am（ピーク 3：$g_o(r_{lha})$=4.65, $<r_{lha}^{min}>_3^{N.am}$=3.01 Å）、N.4（ピーク 2：$g_o(r_{lha})$=2.85, $<r_{lha}^{min}>_2^{N.4}$=3.15 Å、ピーク 3：$g_o(r_{lha})$=4.35, $<r_{lha}^{min}>_3^{N.4}$=3.00 Å、ピーク 4：$g_o(r_{lha})$=5.25, $<r_{lha}^{min}>_4^{N.4}$=2.73 Å）、N.pl3（ピーク 3：$g_o(r_{lha})$=5.25, $<r_{lha}^{min}>_3^{N.pl3}$=2.96 Å、ピーク 4：$g_o(r_{lha})$=8.25, $<r_{lha}^{min}>_4^{N.pl3}$=2.82 Å）の場合は、タンパク質のリガンド接触元素として、水素結合アクセプターとして相互作用できる酸素の割合が高くなっている。これらの結果から、$g_o(r_{lha})$が大きい、より水和した状態では、水素結合が主たる相互作用であり、リガンドータンパク質間相互作用距離が短くなっている様子がうかがえる。



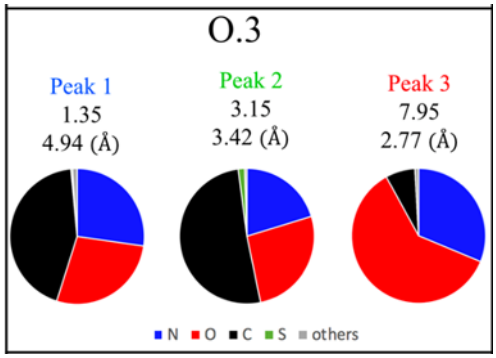




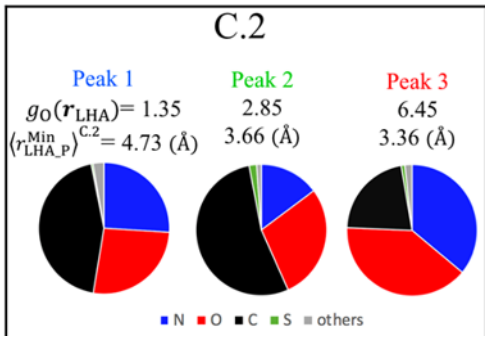
次に、O.2（sp2 酸素）と O.3（sp3 酸素）atom type を含む酸素原子について解析する。孤立電子対を 2 つ有し、水素結合アクセプターとして働く O.2（ピーク 3：$g_o(r_{lha})$=7.95, $<r_{lha}^{min}>_3^{O.2}$=2.97 Å）の場合、タンパク質のリガンド接触元素として、NH の形で水素結合ドナーとして相互作用できる窒素の割合が高くなっている。

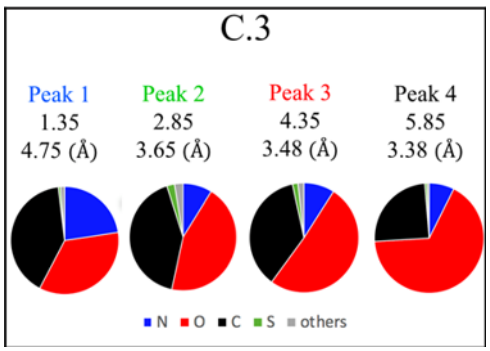


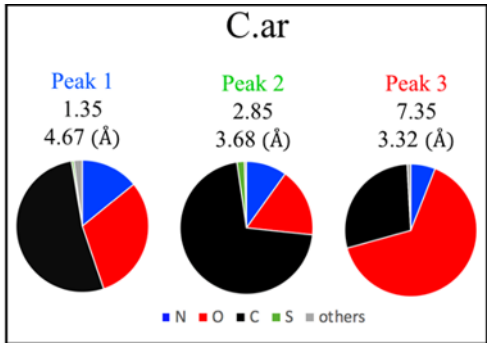
水酸基(OH)の形で、水素結合アクセプターおよびドナーの両者で相互作用可能な atom type である O.3（ピーク 3：$g_o(r_{lha})$=7.95, $<r_{lha}^{min}>_3^{O.3}$=2.77 Å）の場合、タンパク質のリガンド接触元素として、酸素および窒素の割合が高くなっている。興味深いことに、O.2 のピーク 3 の相互作用距離$<r_{lha}^{min}>_3^{O.2}$=2.97 Å は、O.3 のピーク 3 の相互作用距離 $<r_{lha}^{min}>_3^{O.3}$=2.77 Å よりも、少し長くなっている。このことは、NH-O（タンパク質 NH-リガンド O.2）の水素結合距離が、OH-O（リガンド OH (O.3) – タンパク質 O）の水素結合距離よりも、少し長いことに由来しているのかもしれない。



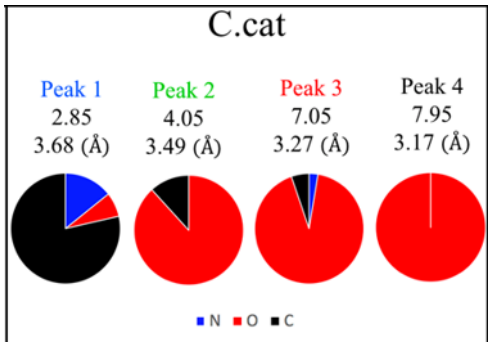
C.2 (sp2 炭素)、C.3 (sp3 炭素)、C.ar (芳香族炭素) atom type を含む炭素原子については、水素結合距離 (3.2 Å)以下の相互作用ピークは見当たらない。各 atom type の最小相互作用距離は、それぞれ C.2 のピーク 3 ($g_o(r_{lha})$=6.45, $<r_{lha}^{min}>_3^{C.2}$=3.36 Å)、C.3 のピーク 4 ($g_o(r_{lha})$=5.85, $<r_{lha}^{min}>_4^{C.3}$=3.38 Å)、C.ar のピーク 3 ($g_o(r_{lha})$=7.35, $<r_{lha}^{min}>_3^{C.ar}$=3.32 Å)である。これらの場合、主たるタンパク質のリガンド接触元素は酸素であり、CH-O 相互作用が形成されていることが強く示唆される。



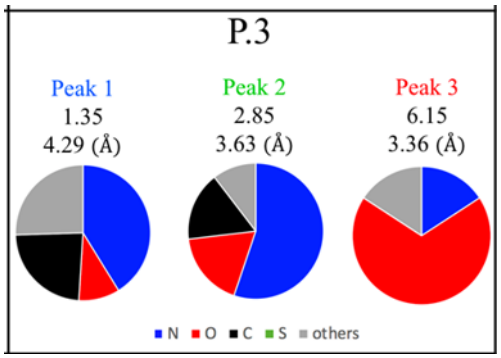




アミジノ基 ($R\text{-}C^{+}(NH_2)NH_2$) やグアニジノ基 ($R\text{-}NHC^{+}(NH_2)NH_2$) の中心にあるカチオン性炭素 (C.cat) の場合、例えば、ピーク 4 ($g_o(r_{lha})$=7.95, $<r_{lha}^{min}>_4^{C.cat}$=3.17 Å)は、$g_o(r_{lha})$=7.95 という高い水和状態に存在し、タンパク質のリガンド接触元素は酸素である。これは、リガンド C.cat のカチオンとタンパク質のアニオン性酸素間の強いイオン性相互作用に起因するものと考えられる。

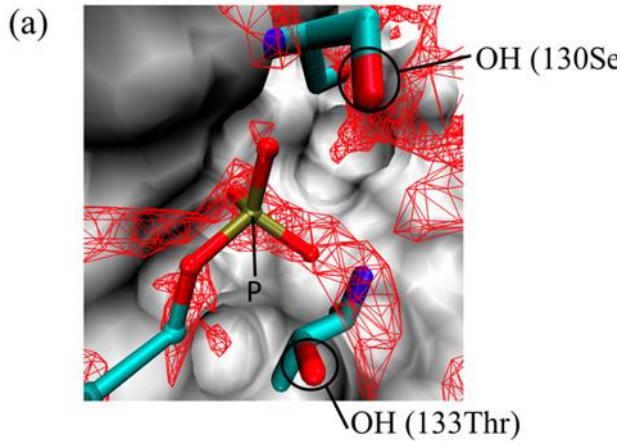


リン酸 ($R\text{-}PO_4^{3-}$)の中心によく現れる sp3 リン原子 (P.3)の場合、ピーク 3 ($g_o(r_{lha})$=6.15, $<r_{lha}^{min}>_3^{P.3}$=3.36 Å) $g_o(r_{lha})$=6.15 という高い水和状態に存在し、主たるタンパク質のリガンド接触元素は酸素である。

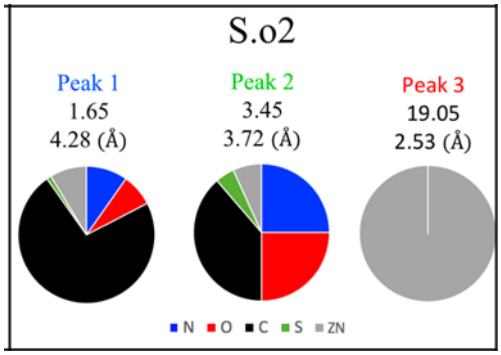


これは、タンパク質酸素原子とリガンド P.3 原子が直接相互作用している訳ではなく、タンパク質酸素原子とリガンド P.3 原子に隣接し共有結合している原子（その主たるものは酸素原子）が相互作用している結果、タンパク質酸素原子とリガンド P.3 原子が近接する現象が観測されている。その具体例を以下に示す。リガンド P.3 原子はタンパク質 Ser130 および Thr133 の側鎖の水酸基と近接しているが、それは P.3 リン原子が共有結合しているリン酸のそれぞれの酸素原子が Ser130 および Thr133 と水素結合しており、その結果、タンパク質酸素原子とリガンド P.3 原子の近接が見ら

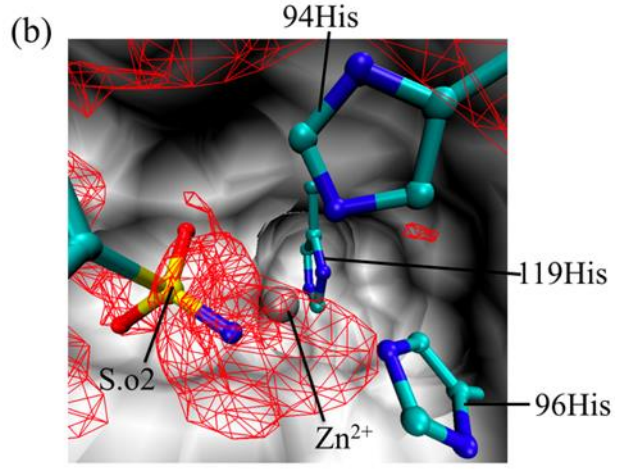
れるということになっている。この場合、Ser130 と Thr133 の間の赤色メッシュで示される高い水和領域の水が、リガンドの結合によってリガンドリン酸部位で置き換えられるということが起こっている。



イオウ原子については、スルホンイオウ原子 (S.o2) と sp3 イオウ原子 (S.3)を別々に解析する。スルホンイオウ原子 (S.o2)の場合、ピーク 3 ($g_o(r_{lha})$=19.05, $<r_{lha}^{min}>_3^{S.o2}$=2.53 Å)の最近接タンパク質重原子の元素は Zn である。

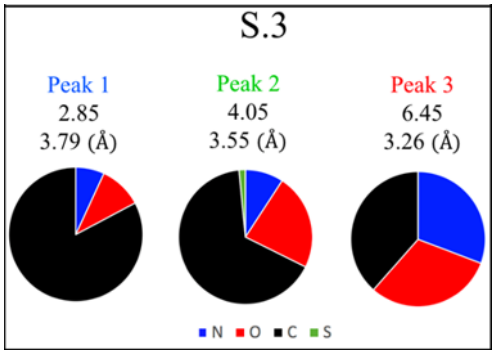


P.3 でのケースと同様に、この現象は S.o2 と Zn の直接相互作用に起因するものではない。具体例を以下に示す。

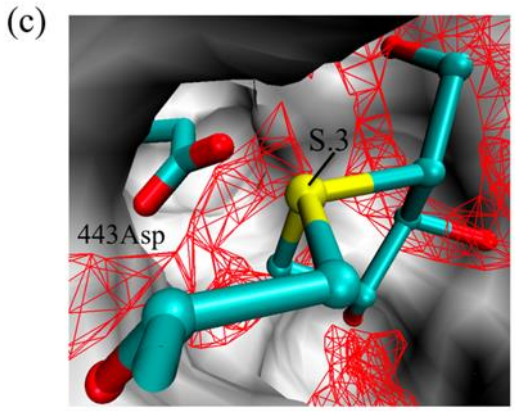


この場合スルホンアミドのイオウ原子に共有結合した酸素と窒素が Zn に配位している。この配位により、結果的に Zn と S.o2 が引き起こされている。このケースにおけるリガンド結合プロセスとしては、Zn 周囲の赤色メッシュで示される非常に高い水和領域の水が、リガンドの結合によってリガンドスルホンアミド部位で置き換えられると考えられる。

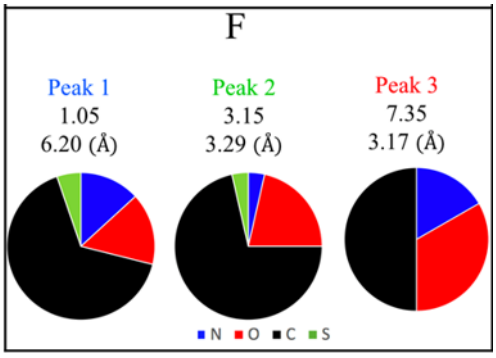
sp3 イオウ原子 (S.3)のピーク 3 ($g_o(r_{lha})$=6.45, $<r_{lha}^{min}>_3^{S.3}$=3.26 Å) の最近接タンパク質重原子の元素は、窒素、酸素、炭素がほぼ等しい割合となっている。



下図にその一例を示すが、このケースにおいては、リガンドの S.3 イオウ原子がタンパク質の Asp443 側鎖の酸素と S-O 相互作用している。

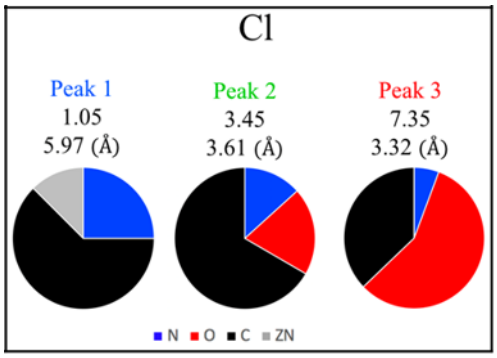


フッ素 F については、ピーク 3 ($g_o(r_{lha})$=7.35, $<r_{lha}^{min}>_3^{F}$=3.17 Å) の最近接タンパク質重原子の元素は、炭素、酸素、窒素の順となっている。



この場合のリガンドFとタンパク質との相互作用は、炭素がCH-F相互作用、酸素が直交する多極子相互作用(CF ⊥ O=C)、窒素がNH-F相互作用であると考えられる。

塩素 Cl については、ピーク 3 ($g_o(r_{lha})$=7.35, $<r_{lha}^{min}>_3^{Cl}$=3.32 Å) の最近接タンパク質重原子の元素は、酸素、炭素の順となっている。



酸素の場合は、リガンド Cl とタンパク質酸素がハロゲン結合しているものと推測される。炭素の場合は、リガンドの Cl とタンパク質の芳香環が Cl-π相互作用をしているものと考えられる。

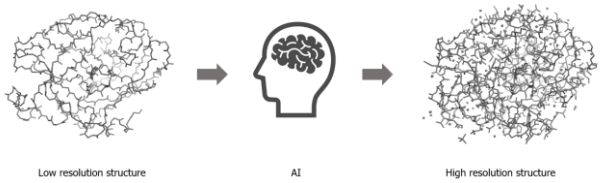
接触距離である最近接距離 $<r_{lha}^{min}>_i^X$ < 4 Å の全 atom type の全ピークについて、$g_o(r_{lha})$の値が減少するにつれて、最近接タンパク質重原子の元素が炭素となる割合が増加している。$g_o(r_{lha})$の値が減少するということは、より低い水和状態になるということであり、リガンドが置き換える水の数が減少するということである。リガンド重原子が sp2 炭素 C.2、sp3 炭素 C.3、芳香族炭素 C.ar などの中性炭素原子の場合は、この現象は、リガンド結合部位が疎水性（＝低い水和状態）になればなるほど疎水性相互作用の数が増えると考えれば、直感的に理解しやすい。

しかしながら、リガンド重原子が極性の高い窒素や酸素原子の場合は、解釈はもう少し複雑になる。エントロピーの観点からは、リガンド結合は、主に排除体積効果による水のエントロピー増加を伴う。エンタルピーの観点からは、リガンド結合により水がタンパク質から離れる時には、主に静電力とファンデルワールス項からなる脱水和エネルギーが必要となる。リガンド結合によるエネルギー損失は、リガンドータンパク質間の好ましい相互作用エネルギー、水ー水間の好ましい相互作用エネルギー、水のエントロピー増加などによって補償されなければいけない。低い水和状態、すなわち置き換えられる水の数が少ないリガンド結合部位においては、必要とされる脱水和エネルギーが小さくなる。したがって、これを補償するためには、NH-O=C のような強い水素結合は必要でなくなり、CH-O や CH-F のような弱い相互作用でも十分になると考えられる。これが、酸素や窒素においても $g_o(r_{lha})$の値が減少するにつれて、最近接タンパク質重原子の元素が炭素となる割合が増加している原因であろうと考えられる。

2-2-4 今後の計画・展望：本解析は論文化がほぼ終了しており、近いうちに投稿する予定である。また、本方法はタンパク質ータンパク質間相互作用（Protein-Protein Interaction, PPI）の系にも応用可能であり、現在計算に取りかかっている。

2-3. タンパク質立体構造予測

2-3-1 目的 : 標的タンパク質の立体構造を知ることは、創薬においてとても重要である。標的タンパク質と薬の候補化合物が結合している様子を目視し、理解することにより、その相互作用を強めるなどの様々なデザインが可能になるからである。これをタンパク質の立体構造に基づいた分子設計という意味で、Structure-Based Drug Design (SBDD)と表現している。標的タンパク質の立体構造を調べる手段としては、X 線結晶解析、NMR、低温電子顕微鏡など様々な方法があるが、創薬の場面で最も使われているのは X 線結晶解析である。X 線結晶解析ではタンパク質の結晶をつくり、X 線を照射し、その反射像から電子密度を算定し、立体構造を決定する。タンパク質の精製度が良くないなどの原因によりタンパク質結晶の質があまり良くなかったりすると、電子密度が低解像度となり、タンパク質の全原子の構造を決めることが難しくなる。このような場合、例えばアミノ酸の主鎖の構造は決まるが、側鎖構造は決まらないといったようなことも起こる。SBDD のためにはタンパク質と低分子間の、あるいは、タンパク質同士の相互作用が原子レベルで明瞭に見えている必要がある。そこで AI 技術を使って、低解像度の電子密度から、高解像度 X 線結晶構造を生成し、SBDD を可能とすることを目指して活動している。



上記目的を達成するためには、(a) 低分解能から生成されたタンパク質構造の善し悪しを部分的に評価する機能と、(b) 評価が悪かった構造を評価が良くなるように構造を発生させる機能の二つが必要となる。今回は、第一段階として、AI 技術の 3D-CNN を用いた機械学習による部分構造評価方法、つまり上記(a)の部分を検討した。

2-3-2 & 2-3-3 方法および結果：低分解能の電子密度から生成したタンパク質モデル構造の評価のために、AI 用学習データを HOKUSAI で作成した。PDB から選んだ 1.5Å分解能以上の構造を正解構造とし、そこから 2.4Åおよび 3.0Å分解能のモデル構造を作成した。これらの座標と構造因子から電子密度化したものを記述子とし、最高分解能の構造との相関を目的変数として、3D Convolutional Neural Networks (3D-CNN)を用いて、アミノ酸単位で構造評価を行った。現在、学習結果と予測性能を検討中であるが、各アミノ酸単位の予測は良好な結果を示している。

2-3-4 今後の計画・展望 : 今後は、検討結果をまとめ、論文投稿していく予定である。また、低分解能から生成したタンパク質構造で、良くないと評価された部分を修正していく方法の開発にも着手していきたい。

3.　　まとめ

本課題では、創薬プロセスの効率向上を目指して、いくつかのプロセスでの機械学習モデル作成およびそのためのデータ生成を実施している。

今年度は、3D-RISM によるリガンドフリータンパク質の水和状態の網羅的解析は計算を終了し、解析をまとめ、論文投稿の段階まで進めることができた。

低解像度電子密度から高解像度タンパク質構造の予測については、HOKUSAI を利用することにより多数のタンパク質についてデータセットを作成することができ、3D-CNN で学習させたところ、目的とした評価がほぼできそうであるところまで進展した。

4.　　今後の計画・展望

3D-RISM については、今回の方法を、タンパク質ータンパク質相互作用界面の水和状態の解析に進展させていきたい。

低解像度電子密度から高解像度タンパク質構造の予測については、低分解能データから生成したタンパク質構造の部分的評価についてまとめ、論文投稿する予定である。また、上記方法で構造が良くないと評価された部分を、評価が良くなるようにモデリングする手法にも取り組んでいきたい。

さらに、新規モダリティーとして注目されているサイクリックペプチドについて、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くことを考えている。

以上