平成 30 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 峻太郎 (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)、
原田 昌紀 (3)、田中 大地 (3)、小山田 隆行 (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」、「第一原理計算 Bond Dissociation Energy による化合物安定性の予測」、「第一原理計算による化合物の pKa 予測法の開発」、「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」、「MD トラジェクトリーに対する AI を利用した解析ツールの開発」、「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。GCN によって表された化学構造と予測すべき各特性値の関係性を DL 法によって学習する際に､ネットワークを何層にするか、各層のノード数をいくつにするか、学習の際に過学習を防ぐために全体の情報の何%を利用して学習するか、各ノードにおいて信号を伝達する際の活性化関数をどのような関数形にするかなど、数多くのパラメータを調整する必要があり、これらをハイパーパラメータとよんでいる。ハイパーパラメータの値として、どのような値を用いれば良いかは、学習する化学構造とその記述子および予測したい特性値の組み合わせ、それぞれに異なるので、ハイパーパラメータを系統的に変化させ、一番予測がよくなるハイパーパラメータセットを決める必要がある。ハイパーパラメータを系統的に変化させた、それぞれの条件で学習を行い、それぞれの予測モデルを構築することは計算時間を要する処理であるため、HOKUSAI の多数の CPU を並列に使用し、短時間で多数のハイパーパラメータ条件を探索することにした。予測モデルを検討するデータとして、ChEMBL の COX2 阻害を示す化合物の構造・活性データと、PubChem の化合物・溶解度特性データの 2 セットを使用した。

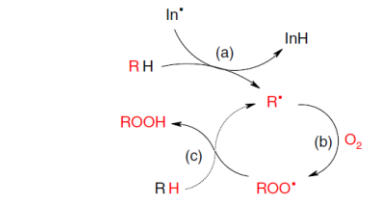
2-1-3 結果：　ハイパーパラメータの条件数が多いため、現在これらの計算は進行中である。

2-1-4 今後の計画・展望：　今後は、COX2 および溶解度の 2 つのデータセットについて、全ての条件探索を実施し、その結果を基に最適なハイパーパラメータ条件を迅速に同定するのためには、どのような要素を、どのような戦略で探索すべきかを検討する予定である。

2-2. 第一原理計算 Bond Dissociation Energy による化合物安定性の予測

2-2-1 目的：　医薬品は、製造されてから患者さんに投与・服用されるまでの全ての過程（倉庫での保管、輸送、病院や家庭での保存など）において、品質が保持されていることが重要・不可欠であり，その品質を保証するために、安定性は重要な要因である。薬剤の安定性については、その開発段階で、原薬および製剤処方条件における、温度、湿度、光等の様々な環境下で医薬品品質の経時的変化を明らかにし、その結果に基づいて貯蔵条件や有効期間を設定するために、長期保存試験や加速試験などの安定性試験が実施される。安定性について医薬品の分子設計の段階でこれを予測することが可能であれば、薬効、体内動態、安全性に加え、十分に安定な臨床候補化合物を創製できることになり、その創薬上のメリットは大きい。

2-2-2 方法：　本課題では、発生したラジカルにより化合物の水素が引き抜かれラジカル化し、ラジカル化した化合物がラジカル化していない化合物をさらにラジカル化する過程を繰り返すことによって自己分解する過程の起こりやすさを、予測する。



予測方法は、化合物から水素が引き抜かれラジカル化するエネルギー（Bond Dissociation Energy, BDE）を第一原理計算で推算することによって行う。BDE を計算するためには化合物の最安定立体構造が必要となる。このために、ソフトウェア KNIME を利用し、平面的な化学構造から立体構造の初期値を生成し、KNIME RDkit の Add conformer ノードを利用して様々なコンフォメーションを Distance Geometry 法により発生させ、発生させた立体構造を KNIME で分子力場法により構造最適化する。さらに、分子力場法により生成された各コンフォメーションを KNIME で重ね合わせることにより立体構造的によく似たコンフォメーションはこれらをまとめ、代表的なコンフォメーションを複数選択した。これらの KNIME の計算は、

KNIMEがインストールされたローカルのPC上で実施した。

複数選択した代表的コンフォメーションについて、HOKUSAI にインストールされた第一原理計算ソフトウェア Gaussian を利用して最安定構造を求めた。まず、KNIME により生成された代表的コンフォメーションそれぞれを初期構造とし、HOKUSAI の Gaussian の中に含まれる機能の一つである半経験的手法PM6により、真空中で構造最適化する。次に、その結果得られたPM6最適化構造を初期構造としてGaussianの中に含まれる機能の一つである Local Density Approximation (LDA)法により HOKUSAI で水中構造最適化し、最適化構造を用いて HOKUSAI で LDA 法により真空中エネルギーを算出する。各代表的コンフォメーションについてLDA法により算出された真空中エネルギーを比較し、最安定のコンフォメーションを選択する。

選択された最安定コンフォメーションの立体構造を基にして、各炭素原子に共有結合している水素原子を一つずつ引き抜いた構造をつくり、それを初期構造として HOKUSAI の Gaussian で Local Density Approximation (LDA)法により真空中で構造最適化し、そこで得られたエネルギー値より、各水素原子を引き抜いてラジカルになるときの Bond Dissociation Energy (BDE) を得る。得られた BDE の中でエネルギー最小のものを、その分子のラジカルへのなりやすさと考え、実験的な安定性と比較する。

2-2-3 結果：　Pharm. Res., 2015, 32, 300 に実験的安定性が報告されている 55 化合物の BDE 計算を終了した。現在、report 作成中。

2-2-4 今後の計画・展望：　上記で得られた BDE 計算値と実験的安定性の比較を行う。また、これ以外にも実験的安定性が報告されている化合物を同定し、本方法の妥当性を検証していく。

さらに、自動化された本方法を様々な化合物に適用し、どの水素が引き抜かれたときに最小の BDE が得られるかという情報と、その BDE 値の情報を蓄積する。これらの情報に対して人工知能技術を応用することにより、化学構造を入力すると、どの水素原子が最小 BDE 値を与え、その BDE 値がいくつになるかを予測するシステムを構築していく。

別の方向性として、BDE およびその部位が CYP による酸化的代謝部位と関連のあることが知られていることから、上記予測システムを CYP による酸化的代謝部位の予測に適用可能かどうかも検討していく。

2-3. 第一原理計算による化合物の pKa 予測法の開発

2-3-1 目的：　医薬品の酸解離定数 pKa は化合物の基本的な物性をあらわす特性値であり、特に製剤処方設計において重要である。また、化合物の標的タンパク質に対する相互作用も、化合物がどの pKa microstate として存在するかにより変化する。上記のように pKa は創薬上の様々な局面に関係することから、pKa について医薬品の分子設計の段階でこれを予測することが可能であれば、その創薬上のメリットは大きい。

pKa 値の予測は、実験値と構造の関係を統計・機械学習的に解析して予測する方法が主流であり、その予測値は±1 程度の誤差を有している。本課題では、第一原理計算より pKa 値を算出し、さらに pKa 計算値と実験値の関係を、人工知能技術を用いて解析することにより、精度の高い pKa 予測を行うことを目的とする。

2-3-2 方法：　第一原理計算を用いた pKa 計算方法については多くの報告がある。しかしながら、今回の最終目的は、pKa 計算値と実験値の関係について人工知能技術を用いて解析することなので、pKa 計算値を自動的に大量に計算する必要がある。大量に計算することを考えた場合は、その計算時間も重要な要素となるので、その具体的な方法については、現在検討している最中である。以下には、現在検討中の方法を記載する。

pKa 予測方法は、化合物に水素が付加する、あるいは、引き抜かれるエネルギーを第一原理計算で推算することによって行う。pKa を計算するためには化合物の最安定立体構造が必要となる。このために、現状では pKa を計算する化合物が＋あるいはーの電荷をもたない中性の状態の構造について、ソフトウェア KNIME を利用し、平面的な化学構造から立体構造の初期値を生成し、KNIME RDkit の Add conformer ノードを利用して様々なコンフォメーションを Distance Geometry 法により発生させ、発生させた立体構造を KNIME で分子力場法により構造最適化する。さらに、分子力場法により生成された各コンフォメーションを KNIME で重ね合わせることにより立体構造的によく

似たコンフォメーションはこれらをまとめ、代表的なコンフォメーションを複数選択した。これらの KNIME の計算は、KNIME がインストールされたローカルの PC 上で実施した。

複数選択した代表的コンフォメーションについて、HOKUSAI にインストールされた第一原理計算ソフトウェア Gaussian を利用して最安定構造を求めた。まず、KNIME により生成された代表的コンフォメーションそれぞれを初期構造とし、HOKUSAI の Gaussian の中に含まれる機能の一つである半経験的手法 PM6 により、真空中で構造最適化する。次に、その結果得られた PM6 最適化構造を初期構造として Gaussian の中に含まれる機能の一つである Local Density Approximation (LDA)法により HOKUSAI で水中構造最適化し、エネルギーを算出する。各代表的コンフォメーションについて LDA 法により算出された水中エネルギーを比較し、最安定のコンフォメーションを選択する。

選択された最安定コンフォメーションの立体構造を基にして、KNIME の BiosolvIT Protomer/Tautomer ノードを用いて各荷電状態の立体構造を生成する。それを初期構造として HOKUSAI の Gaussian で Local Density Approximation (LDA)法により水中で構造最適化し、そこで得られたエネルギー値より、pKa 推算値を得る。得られたそれぞれの pKa 推算値を、対応する荷電状態の実験的な pKa 値と比較する。

2-3-3 結果：　中性状態の化合物の最安定状態を PM6 の真空中コンフォメーション解析により実施する現在の方法では、pKa を計算すべき水中の最安定コンフォメーションが得られないケースがあることがわかった。

2-3-4 今後の計画・展望：　早急に pKa 自動計算のプロトコールを確立する。pKa データベース中の化合物について、pKa 値を推算し、実験値と比較する。さらに、pKa データベース中の化合物について、pKa 推算値から実験値を精度良く推算する人工知能予測システムを構築する。

2-4. 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

2-4-1 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。一方、タンパク質の立体構造を利用したドッキングの際には、X 線結晶解析などで観察された水分子を除去してドッキングを実施するのが一般的であり、その際はタンパク質周囲の水についての情報はあからさまには取り入れられていないことになる。また、X 線結晶解析の結果から重要と思われる水をタンパク質の一部として取り扱いドッキングをすることもできるが、水分子の重要性の判断に任意性があり、さらに X 線結晶解析が低解像度で結晶水を置くことができないケースや、解像度が十分であっても結晶学者が水分子を置くかどうかは任意であるケースも多く、ドッキングの際にどのように水の情報を取り込むかについては、統一的な扱いができていないのが現状である。本課題では、タンパク質周囲の水の情報をドッキングに反映させる方法を確立することを目的とする。

2-4-2 方法：　タンパク質周囲の水の情報は、X 線結晶構造からリガンドと水分子を取り除いたアポ構造に対して 3D-RISM 計算を行うことにより、タンパク質周囲に水分子がどのくらいの確率密度で存在するかによって表現することとする。

計算対象は、ドッキングの際に情報を取り込むことを考え、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 scPDB データベースのデータを用いることとした。

ドッキングに 3D-RISM の水情報を取り込む方法として AI 技術（3D Convolutional Networks, 3D-CNN）を使ったドッキングへの組み込みを考えていること、および、ドッキング予測の自動化の観点から、大量の 3D-RISM 情報を自動で生成するために、3D-RISM 計算の前処理として scPDB のデータについて、(a) リガンドと水分子を切り出し、(b) 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自動的に capping 処理をし、(c) 各原子に AMBER atom type を自動的に割り振り、(d) 自動的に水素原子を付加し、水素原子のみを構造最適化するようにした。

3D-CNN を用いた AI ドッキングシステムに 3D-RISM 水情報を付加するアプローチとは別に、既存のエネルギーベースのドッキングソフトウェア rDOCK で得られた不正解ドッキングポーズで置き換

えている水と、X 線結晶構造で得られた正解ポーズが置き換えている水に差異があるかどうかを検討した。このために scPDB のリガンドを rDOCK でドッキングし、リガンドの X 線結晶構造との RMSD が 4.5A から 5.5A の間になるドッキングポーズを不正解ポーズとして採用した。

2-4-3 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 scPDB 構造を一週間程度で計算することができた。

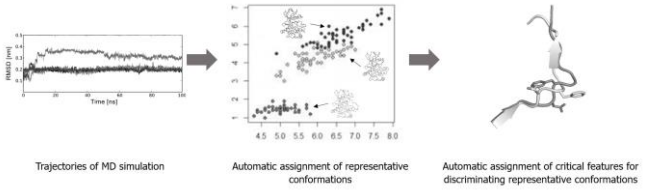
現在、3D-RISM で推算された水（の確率密度）が、結晶水、リガンド（結晶構造）、不正解ポーズのリガンドによってどのように置き換えられているかの解析を実施中。

2-4-4 今後の計画・展望：　現在進行中の解析をまとめて論文化の予定。また、別途開発中の 3D-CNN を使った AI ドッキングシステムの動作確認後、今回算出した 3D-RISM 情報を付加し、学習させ、結合親和性予測力が向上するかどうかを検討する予定。

2-5. MD トラジェクトリーに対する AI を利用した解析ツールの開発

2-5-1 目的：　タンパク質の立体構造は揺らいでおり、立体構造が変化することが、その機能と密接に結びついている場合も多い。創薬の様々な場面でも、タンパク質立体構造の揺らぎや構造変化を計算により検討する手段として、分子動力学(Molecular Dynamics, MD)シミュレーションが実施されるようになってきた。MD シミュレーションにおいては、ある時間の立体構造から、タンパク質の各原子に働く力を計算し、ほんの少し時間がたった時の立体構造を算出するというプロセスを繰り返すことにより、立体構造の時間的変化をシミュレーションする。

タンパク質各原子の各時間における立体構造座標を MD トラジェクトリーというが、そこにはタンパク質原子数×3 次元（x, y, z 座標）×時間ステップ数（例えばシミュレーション時間が 100 nano sec で 1 femto sec ステップならば 100,000）の膨大な情報が含まれている。一方、タンパク質の構造変化の理解のためには、特定アミノ酸の立体構造変化や、αヘリックスのような 2 次構造形成の有無というような、重要な構造的特徴を抽出し、理解することが求められる。これまでは膨大なトラジェクトリー情報から、解析者の経験や知識に基づいて解析し、構造的特徴を抽出してきた。そこで、膨大なトラジェクトリー情報を機械学習や AI などにより網羅的に解析し、自動的に重要な構造的特徴を抽出・同定することを目指して研究を行っている。



2-5-2&3 方法および結果：　HOKUSAI は MD トラジェクトリーの生成、および、MD トラジェクトリーの一時格納場所として利用した。

MD トラジェクトリーデータをインプットとして、アミノ酸残基間の距離を表す Distance Matrix や、タンパク質主鎖・側鎖の二面角などの記述子を算出できるスクリプトを開発した。さらに、2 つの異なった状態のトラジェクトリーにおける構造的差異を自動的に検出する手段として、構造的差異を説明する記述子として何が重要であるかを表す Gini Importance と、その記述子がどのような状態であれば構造的差異を分類できるか解析できる Decision Tree (DT)法を組み合わせて用いることが有用であることを、RORγt inverse agonist、CDK2 inhibitor、μ-opioid receptor の系などに応用することにより、見いだした。

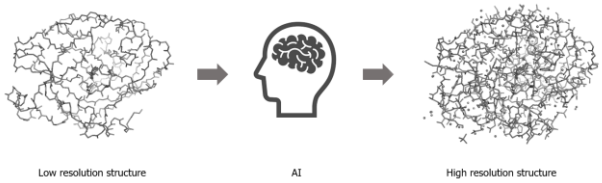
しかしながら、DT 法による解析においては、入力した記述子のほんの少しの変化により、その解析結果が変化してしまう現象が起こったことから、DT 法に代えてインプットデータの変化に強く、再現性が高い Random Forest (RF)法を採用し、解析を継続している。

2-5-4 今後の計画・展望：　今後は、RF 法による解析を続けると共に、(a) MD トラジェクトリーから、タンパク質構造だけではなく、リガンドや周囲の水分子の構造も含めて情報を抽出し記述子化して解析すること、(b) 現在は構造状態の差異を人間が指定して解析を行っているが、MD トラジェクトリーから AI 技術を用いて自動的に構造状態の差異を同定することを検討していく。

2-6. タンパク質立体構造予測

2-6-1 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。



2-6-2&3 方法および結果：　AI 用学習データを HOKUSAI で作成した。まず、正解データとして、Protein Data Bank に登録されている Trypsin の高分解能タンパク質の構造と電子密度を選択した。次に不正解例を作成するために、HOKUSAI に CCP4 ソフトウェアを用いて、高分解能タンパク質の構造と電子密度から、低分解能電子密度データを算出した。さらに、正解と不正解例を作成するため、ホモロジーモデリングソフトウェア Modeler を HOKUSAI にインストールし、この Modeler を用いて他の Serine Protease の立体構造を鋳型にして、Trypsin 構造を多数発生し、それぞれについて高分解能電子密度と低分解能電子密度を算出し、AI 用の入力データとして用いた。

2-6-4 今後の計画・展望：　今後は、Trypsin 以外の様々なタンパク質立体構造を用いて、正解・不正解電子密度データを作成し、学習データを追加・充実させ、予測モデルを構築していく予定である。

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、今まで AI 予測システムの対象となっていなかった、BDE、pKa、3D-RISM などについてのデータ作成を開始した。そのうち、BDE と 3D-RISM については、自動データ生成プロトコールが完成し、HOKUSAI の処理能力を活用し、予定していたデータセットについて作成が終了した。pKa についてはデータ作成の自動化のプロセスがまだ確立していない。

昨年度からの継続課題の MD トラジェクトリー解析については、AI による解析システムのプロトタイプが完成し、各種タンパク質・リガンド複合体の MD トラジェクトリーへの応用がなされ、その有用性が検証された。それと共に、解析方法の不備な点も明らかになり、それを受けてプロトタイプの改良も行われた。

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一種類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、別途開発中の 3D-CNN で最初の学習を行うことができた。

4.　今後の計画・展望

データセット作成が終了した BDE と 3D-RISM については、これから作成したデータセットの検証を行うと共に、これらの AI 予測システム構築に取り組んでいきたい。データ作成の自動化のプロセスがまだ確立していない pKa については、自動化プロセスを早期に確立し、HOKUSAI の処理能力を活かして早期のデータセット完成を目指したい。

MD トラジェクトリーの解析については、プロトタイプの改良により、MD トラジェクトリーから取り込んだ項目については二状態で差異がある項目を自動的に同定できることが確認できた一方、トラジェクトリーから取り込めていない情報があり、それ

らを取り込むための方法が必要であることが明確になったので、これに取り組んでいきたい。それに加え MD トラジェクトリーから自動的に代表的な状態を抽出する手法も必要と考えられたので、AI 技術を利用して、これを実現することにも着手したので、これを完成させ、さらなる自動化を目指したい。なお、MD トラジェクトリーについての、これらの取り組みにおいて HOKUSAI の利用が有用かどうかは結論がでていない。

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、別途開発中の 3D-CNN で、一種類のタンパク質のデータセットを学習したが、その結果、一種類のタンパク質のデータセットでは不十分で、多種類のタンパク質の学習用データセットが必要なのではないかと思われた。作成するデータセットはファイルサイズが膨大になることが予想されるため、作成後のデータ転送がボトルネックになる HOKUSAI でのデータ作成ではなく、AI 学習に使用するマシン上でのデータ作成についても検討する予定である。

これまで AI 予測システム構築に取り組んでこなかった対象で新規に取り組む項目については、CPU 並列による処理がデータセット作成に重要なものは HOKUSAI の利用を第一優先に考えていきたい。

他方、GPGPU による処理能力がボトルネックになる、例えば DL 法における学習などについては、その GPGPU 処理性能を考え、HOKUSAI 以外のリソースを使った処理を考えていく。

なお、AI 用データセットにおいて、そのファイルサイズが膨大になる、タンパク質立体構造予測用電子密度、MD トラジェクトリー、3D-RISM 確率密度などについては、HOKUSAI から理研ネットワーク上の AI 学習用マシンに転送するプロセスがボトルネックの一つになっていることを記しておきたい。

以上