論文 / 著書情報
Article / Book Information

題目(和文)	機械学習とシミュレーションを組み合わせた薬物代謝酵素(CYP3A4)の結合様式予測法の研究
Title(English)	
著者(和文)	佐藤敦子
Author(English)	Atsuko Sato
出典(和文)	学位:博士(学術), 学位授与機関:東京工業大学, 報告番号:甲第11541号, 授与年月日:2020年3月26日, 学位の種別:課程博士, 審査員:山村 雅幸,小野 功,青西 亨,瀧ノ上 正浩,関嶋 政和,小長谷 明彦,本間 光貴
Citation(English)	Degree:Doctor (Academic), Conferring organization: Tokyo Institute of Technology, Report number:甲第11541号, Conferred date:2020/3/26, Degree Type:Course doctor, Examiner:,,,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

2019 年度博士論文

機械学習とシミュレーションを組み合わせた薬物代謝酵素(CYP3A4)の結合様式予測法の研究

東京工業大学　情報理工学院　情報工学系　知能情報コース

佐藤　敦子

指導教員

山村雅幸　教授

本間光貴　特定准教授

2020 年 3 月

論文概要

本論文は、「機械学習とシミュレーションを組み合わせた薬物代謝酵素(CYP3A4)の結合様式予測法の研究」と題し、6 章から構成される。

第 1 章「序論」では、研究の社会的、技術的な背景、研究の目的と全体像について述べている。本研究ではタンパク質の薬物結合ポケットの動きを考慮して薬物の結合様式を予測する方法の開発を目的とし、題材としてポケットが大きく、主鎖レベルで構造の揺らぐ CYP3A4 を用いて以下 2 つのアプローチを検討した。1 つ目のアプローチはタンパク質と化合物複合体の 3 次元構造を用いた機械学習による結合様式予測である。2 つ目のアプローチは複数の短時間 MD シミュレーションを並列させた結合様式予測である。また本研究の背景として、創薬研究における結合様式予測効率化の重要性と、予測の難しいターゲットタンパク質に対する技術的な課題について述べている。

第 2 章「関連研究」では、結合様式予測の難しいターゲットに対する技術的な課題を、「機械学習による結合様式予測」と「シミュレーションによる結合様式予測」に分け、関連研究を整理し、本研究の位置づけを示した。機械学習による結合様式予測では、3 次元構造情報を保持したタンパク質構造および化合物構造の入力手法と、訓練データとなる 3 次元構造情報が不十分な場合の学習方法ならびに高品質データセットの作成方法論の必要性について述べた。シミュレーションによる結合様式予測では、結合する薬物の形状に応じて結合ポケットが柔軟に変化する予測難度の高いターゲットタンパク質に対するアプローチを整理しつつ、十分な構造をサンプリングする方法論の必要性を示した。

第 3 章「機械学習による結合様式予測」では、3 次元構造情報を用いた機械学習スコアリングとして 3D-CNN を利用した CYP3A4 の結合様式予測に取り組んだ。従来研究では、ターゲットに依らないデータセットでの予測モデル構築が検討されたが、本研究では、訓練データが不十分である場合のデータセットとしてさまざまな特徴とデータサイズを有する 4 つの訓練セットを選択し、CNN モデルの事前学習とファインチューニングを検討した。その結果、従来の結合様式評価法（エネルギーベースのドッキングポーズ評価、ターゲットに依らないデータセットでの予測モデルによる評価 ）と比較して大幅に予測精度が向上し、その要因の考察を通じてデータセット選択の方法論を構築した。

第 4 章「MD シミュレーションによるタンパク質の動きを考慮した結合様式予測」では複数の MD 初期構造を用いた短時間シミュレーションによる CYP3A4 の結合様式予測に取り組んだ。従来研究では、化合物の位置に着目して MD 初期構造が選択されたが、本研究ではより柔軟な構造を持ち、結合様式予測の難度が高いケースにおいて、タンパク質と化合物間の相互作用に着目して初期構造の選択を行った。その結果、従来法とは異なり、MD 初期構造の選択には化合物側よりも相互作用に着目した選択が有効であり、また、短時間の MD シミュレーションを複数並行して実施する手法が結合様式予測の難度が高い化合物においても有用であることを示した。

第 5 章「総合討論」では第 3 章、第 4 章で構築した手法の応用および機械学習で用いたデータセットの考察について述べた。まず、機械学習とシミュレーションを組み合わせて結合様式予測を効率化させる方法の提案に向け、第 3 章で得られた機械学習モデルを MD 初期構造の選択に適用した結果を確かめた。その結果、従来法で選択した方法と比較して結合様式を正しく予測できることが示された。第 3 章の機械学習アプローチで作成したデータセットに関する考察では、オキシドリダクターゼの比率、ポケットサイズに注目した比較の他に、複数の観点でデータの分布を調査した。そのうちタンパク質側に注目した切り口では、タンパク質ファミリーセットが標的に依らないデータセットをよく補完していることを再確認し、一方で化合物側に注目した切り口では、データセット間の分布に偏りがないことを示した。

第 6 章「結論」では、第 5 章までの主要な結果をまとめ、今後の展望や課題について述べた。本研究では、結合サイトが柔軟で大きな高難度ターゲットの結合様式予測を目的に、CYP3A4 を題材として機械学習による結合様式予測に向けたデータセット作成のための方法論と、複数初期構造を用いた MD シミュレーションによる結合様式予測法を示した。また、機械学習とシミュレーションそれぞれについて、今後の課題と展望を示した。将来、シミュレーションと機械学習を真に融合させた高難度ターゲットの結合様式予測として、1）シミュレーションの強みである実験では得られないような仮想大量データを生成、2）その大量データによる結合様式予測機械学習モデルの訓練、3）その結合様式予測機械学習モデルによる MD 初期構造の選択、というサイクルを繰り返すことにより、予測手法の高速化、高精度化が期待され、汎用的な手法として創薬研究に貢献すると考えている。

謝辞

本研究の一部は、国立研究開発法人日本医療研究開発機構（AMED）創薬等ライフサイエンス研究支援基盤事業 創薬等先端技術支援基盤プラットフォーム（BINDS）の課題番号 0210 の支援を受けました。また、本研究で使用されている 3D-CNN モデルは、財団法人先端医療振興財団からみずほ情報総研株式会社への委託成果物を使用させていただきました。

この論文は、筆者の東京工業大学博士後期課程での研究成果をまとめたものです。協和キリン株式会社に在籍し企業研究を遂行しながら、学生としてアカデミア研究に取り組み論文を執筆することは、多くの方々の支援がなければ成し得ないことでした。お世話になりましたすべての方々に深く感謝の意を表します。

本論文を作成するにあたり、終始ご指導を頂きました本学小長谷明彦名誉教授に心より感謝致します。特に、機械学習の研究において、その概念から細部に至るまで丁寧にご指導いただき、また研究を進めるに当たっての環境整備にも心を配っていただきました。学位取得に向けた諸々のご助言をいただき、また研究室移籍後の様々な手続きを進めて下さいました本学山村雅幸教授に深謝いたします。スパコン TSUBAME の使用に便宜を図っていただき、論文審査にあたっても親身に前向きに議論、ご指導くださいました本学関嶋政和准教授に深く感謝いたします。本学瀧ノ上正浩准教授、小野功 准教授、青西亨准教授に感謝いたします。博士論文の審査を通じた様々なアドバイス、議論を通じ、論文の質を大幅に向上させることができました。TSUBAME の使用に関しご指導くださいました本学我妻竜三特任助教に感謝いたします。機械学習、特にファインチューニングの実施にあたり、手順を丁寧に教えてくださいました小長谷研究室 Zhang Yuhui 氏に深く感謝いたします。

理化学研究所生命機能科学研究センター本間光貴チーム長、本学特定准教授に心より感謝致します。学位取得のきっかけを与えてくださり、研究者としての心構え、計算科学者として習得すべきスキルなど様々なアドバイスをいただきました。また研究が思うように進まなかった期間もポジティブな気持ちを維持できるよう、支えてくださいました。本間研究室、幸瞳博士に深謝いたします。研究の初期からドッキング、MD シミュレーションの基本的な考え方、研究の進め方に関して多くの議論、サジェスチョンをいただきました。本間研究室、渡邉千鶴博士に深く感謝いたします。特に、MD シミュレーション、量子化学計算の実施にあたり、とても丁寧にご指導くださいました。本間研究室、保田真由子さんに深謝いたします。理研内外の事務手続き全般を円滑に進めてくださいました。本間研究室、高谷大輔博士には MD シミュレーションの指導、ハードウェア、ソフトウェア両面で環境整備をしてくださいました。ここに深く感謝いたします。本間研究室、森脇寛智博士に感謝いたします。機械学習、特にディープラーニングモデルの改良、結果の解釈など多くの議論、アドバイスをいただきました。TSUBAME の使用方法、長時間 MD シミュレーションの実行方法について丁寧にご指導くださいました理化学研究所の千葉峻太朗博士に深く感謝いたします。

みずほ情報総研株式会社谷村直樹博士に深く感謝いたします。3D-CNN に関する研究初期

のプログラムの迅速な実装を進めてくださり、結果に関する多くの議論に貴重なお時間をいただきました。

3D-CNN による結合様式予測に関し様々な議論にお付き合いくださいましたライフインテリジェンスコンソーシアム（LINC）のプロジェクトメンバーに心より感謝いたします。

協和キリン株式会社、齋藤純一博士に心より感謝いたします。入学の許可ならびに会社への推薦、学業と業務の両立に向けたアドバイスをいただきました。また業務上の配慮、大学院講義への出席を許可してくださいました協和キリン株式会社、網城宣善博士、阿部真之博士に感謝いたします。

4 年間もの間、家事、育児、学校・地域役員等の多くの仕事を引き受け、精神的にも支えてくれた夫に深く感謝いたします。業務と学業の両立に苦労していた期間も温かく見守り、支え続けてくれた両親に感謝いたします。休日に一緒に過ごす時間が少ない中、応援し続けてくれた長男と次男に心より感謝します。

研究業績

査読付き英語論文

Sato, A.; Tanimura, N.; Honma, T. and Konagaya, A. Significance of data selection in deep learning for reliable binding mode prediction of ligands in the active site of CYP3A4. *Chemical and Pharmaceutical Bulletin*, **2019**, *67*, 1183-1190. (第 3 章)

Sato, A.; Yuki, H.; Watanabe,C.; Saito, J.; Konagaya, A. and Honma, T. Prediction of the site of CYP3A4 metabolism of tolterodine by molecular dynamics simulation from multiple initial structures of the CYP3A4-tolterodine complex. *Chem-Bio Informatics Journal*, **2017**, *17*, 38-52. (第 4 章)

学会発表

Sato, A.; Tanimura, N.; Honma, T. and Konagaya, A role of data selection in deep learning based CYP3A4 binding mode prediction. CBI2019, Tokyo, 2019.　(第 3 章)

Sato, A.; Yuki, H.; Watanabe,C.; Saito, J. and Honma, T. The site of CYP3A4 metabolism prediction of tolterodine using MD and its application to compound design. CBI2015, Tokyo, 2015.　(第 4 章)

Sato, A.; Yuki, H.; Watanabe,C.; Saito, J.; Konagaya, A. and Honma, T. Reactivity evaluation of donepezil in the oxidation by CYP3A4 based on QM calculation. CBI2016, Tokyo, 2016.

招待講演

佐藤敦子、幸瞳、渡邉千鶴、齋藤純一、本間光貴、分子動力学計算を用いた化合物の CYP3A4 代謝部位予測とデザイン、MOE フォーラム、東京、2016 年 7 月　(第 4 章)

目次

第1章　序論

1.1　緒言

本研究全体を通しての課題はタンパク質と化合物の結合様式予測である。従来、タンパク質と化合物の結合様式を予測するためにドッキングシミュレーションが用いられてきた。しかしながら、結合ポケットが大きく、また結合する化合物によりポケット形状が大きく変化するような標的タンパク質（Figure 1.1）では結合形式の予測精度は不十分となっている。このような予測の難しい標的タンパク質に対して、本研究では機械学習による MD 初期ポーズ選択と MD シミュレーションを併用した結合様式予測に取り組んだ。

本章では、はじめに、本論文の主要テーマである「CYP3A4 の結合様式予測」の理解に必要な背景知識として、*in silico* 技術から AI 創薬への創薬研究の変遷、ドッキングシミュレーションから結合様式予測への技術課題の変遷、シトクロム P450 3A4（CYP3A4）における結合予測問題の重要性について述べる。次に、機械学習と MD シミュレーションを用いたシトクロム P450 3A4（CYP3A4）の結合予測における本研究の独自性について概述し、最後に、本研究に関する背景知識のまとめと本博士論文の章構成について述べる。

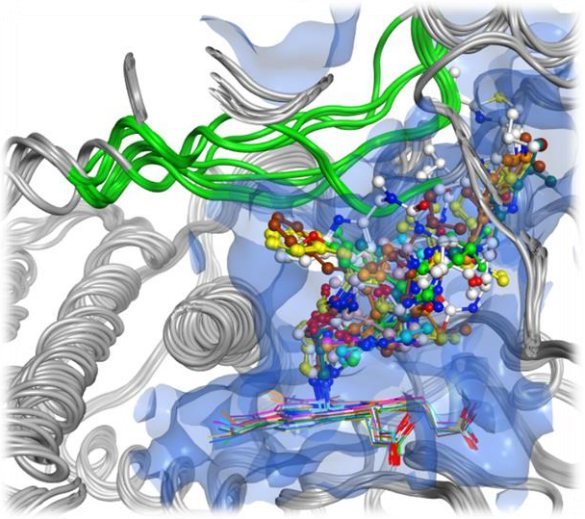


Figure 1.1　結合ポケットが大きく、化合物によりポケット形状が大きく変化するタンパク質 (例:CYP3A4)

結合サイトを青色、化合物を ball and stick で、構造変化するループを緑色で示す。

1.2　*In silico* 技術から AI 創薬への変遷

医薬品の開発に必要とされる期間は 10〜15 年、数千億規模の開発費用がかかる一方で、リード化合物（医薬品のタネ） から臨床試験を経て承認に至る成功確率は約 25,000 分の 1 と非常に低く（Figure 1.2)、その難易度も年々上昇している[1]。そのため、医薬品の研究開発における成功確率を向上させて開発期間とコストを下げ、早期に患者に薬を届けるべく様々な創薬イノベーションが求められている。創薬プロセスには、ターゲット探索から薬のタネとなる分子の創製、試験管内（*in vitro*）または動物体内（*in vivo*）での効果の評価からヒトでの臨床試験と多くのステップが含まれる。これらのステップを効率化、高精度化させると期待されているのが計算機を活用した種々の *in silico* 技術である。

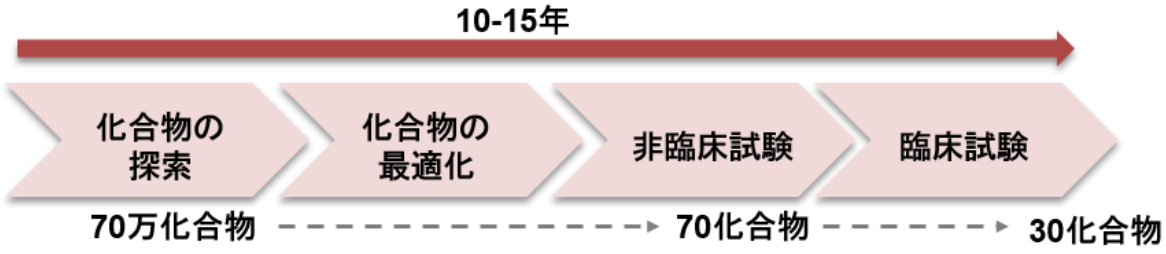


Figure 1.2 医薬品開発のプロセス、期間と開発段階別の化合物数[1]

2019 年現在、創薬プロセスにおいて *in silico* 技術が期待される範囲は創薬テーマの創出から制御分子創製、臨床試験、市販後の調査と非常に幅広い。計算機を活用した創薬研究の支援自体は新しいものではなく、その歴史は 1960 年代の定量的構造活性相関[2], [3]に遡る。1990～2000 年代ごろまでは、ケモインフォマティクス、分子モデリングなど薬物候補の効果や毒性などを評価する個々の「*in silico* 技術」としての活用に焦点が当てられた。さらに 2000 年～2010 年代にかけての計算機性能の飛躍的向上と創薬研究におけるデータの増大に伴い、これまでの「*in silico* 技術」から計算機の中で薬を産み出し評価する「*in silico* 創薬」[4]として活用され、創薬プロセス全体が底上げされつつある。

これまで、*in silico* 創薬技術の中核として注目されていたのが数々のシミュレーション技術である。シミュレーション技術の強みは実験や観察が難しい課題をコンピュータ上で再現することにより、最適な解決策を導ける点である。創薬分野においては、薬理活性化合物を探索する計算技術 （virtual screening）に加えて、活性化合物を最適化する計算技術、吸収、分布、代謝、排泄、毒性（ADMET） を予測する計算技術などが活発に研究されている。特に、生体内でタンパク質や薬物分子がどのような挙動をするか、どのような薬物分子を設計すべきかの問いに関するシミュレーション技術は創薬研究において関心が高く、様々な技術が開発されてきた。これらのシミュレーション技術は一般に多くの計算機資源を必要とし、計算速度の向上が課題となっていた。昨今では「TSUBAME」（東工大） [5]、「京」（理化学研究所）[6]などのスーパーコンピュータの開発によって大規模なシミュレーションも可能となっている。

一方、創薬分野における各種データの増大に伴い、創薬プロセスにおける機械学習の適用が急速に拡大している。これまでにも、疾患ターゲットの探索、タンパク質構造または化合物ベースからの薬物設計等に機械学習手法が応用されてきた。これら従来の適用法に加えて、データの標準化、情報の加工と管理における科学技術と倫理面の両方の対策が進んだことにより、電子カルテ情報からの各種所見と疾患の関連分析、細胞形態変化の画像解析による薬効評価、既存薬物の適用疾患の拡大 （ドラッグリポジショニング）など、いわゆる「AI 創薬」の取り組みが加速化している。すでに海外の大手製薬企業においては、大量に蓄積された社内データを利用し、IT 企業との協働による新薬の探索や標的タンパク質の同定などで成果を挙げている[7] 。 国内でも「AI 創薬」を目指したベンチャー企業、アカデミアと製薬企業の共同研究、コンソーシアム活動などを通して機械学習の創薬応用が展開されている。特に、機械学習とシミュレーション技術を組み合わせたドッキング法の効率化は、従来型 *in silico* 創薬手法だけでは解決できない高難度ターゲットに対する AI 創薬のイノベーションとして期待されている。

1.3 結合様式予測手法の変遷

In silico 創薬において、ドッキングシミュレーションは、薬物、特に低分子医薬品の設計における基盤的な計算科学手法として長年研究されてきた。標的とするタンパク質と医薬品候補化合物（リガンド） はしばしば「鍵穴」と「鍵」に例えられ、ドッキング法では両者の形を補償する構造を計算機内で探索するアルゴリズムが用いられる[8]。この「鍵穴」の情報であるタンパク質の立体構造から医薬品候補化合物との結合様式を正しく特定することができれば、標的タンパク質に対するリガンドの作用機序を原子レベルで理解することができる。さらに、これらの立体構造情報を活用することにより、医薬品候補化合物の結合親和性や他タンパク質種との選択性向上のための合理的な誘導体設計が可能となる。このようにドッキングシミュレーションは医薬品設計に非常に有用な手段である一方で、その精度には未だ多くの課題を残していることも現状であり、現在でもドッキングシミュレーションの精度向上を目指した研究が活発に行われている[9]。

しかしながら、「鍵穴」であるタンパク質の結合ポケットが広く、かつポケットを形成する表面構造が変化するような高難度標的タンパク質では、「鍵」となる化合物の設計が非常に難しい[10], [11]。このため、高難度ターゲットに対するタンパク質と化合物の正しい結合様式の予測法の開発が大きな課題となっている。なお、本論文では「結合様式」を結晶構造中のタンパク質と化合物の位置関係およびドッキングシミュレーションにより得られたタンパク質と化合物の最終的な予測結果の位置関係の意味で用いる。そして、結合様式の候補となる、ドッキングシミュレーションにより得られる一時的なタンパク質と化合物の位置関係を「ドッキングポーズ」と呼ぶ。

結合様式予測では、一つのタンパク質についても様々なポケット形状を考慮し、最良の活性を示す化合物の複合体構造を求める必要がある。一般に、一つのタンパク質構造のみを用いてドッキングを実施しただけでは、たまたまポケットの形状がフィットしなかったために悪いスコア値を示す可能性がある。この問題を解決するために、タンパク質の動きを考慮した結合様式予測法として、これまでに、「アンサンブルドッキング」、「partially flexible ドッキング」、MD シミュレーションおよび機械学習法が提案されている。

「アンサンブルドッキング」は、あらかじめ準備しておいた複数のタンパク質結晶構造に対し、並行してドッキングを行う手法である[12]。この方法では疑似的にタンパク質の柔軟性を取り入れることができる一方で、実験で多数のタンパク質構造が解かれているターゲットタンパク質に限定される問題がある。

「partially flexible ドッキング」は、結合ポケット周辺のアミノ酸側鎖の自由度を考慮してドッキングシミュレーションを実施する方法である[13]。具体的には、結合ポケット内の自由度の高いアミノ酸残基を一時的に小さなアラニンに置換して初期ドッキングポーズを得、置換したアラニンを元に戻して再ドッキングを実施する [14]。この方法ではアミノ酸側鎖の動きは考慮できるものの、アミノ酸主鎖も含めた動きは考慮できない問題がある。

分子動力学（MD）シミュレーションは、アミノ酸主鎖を含めてタンパク質の柔軟性を扱う最もシンプルで強力な手段である。しかしながら、古典的な MD シミュレーションによって薬物候補化合物とタンパク質の結合を再現するには通常 μs～ms オーダーのシミュレーションが必要であり、結合様式予測に必要な計算コストが高い[15], [16]。また、近年の計算機性能の高度化で長時間の MD シミュレーションが実現可能となっているものの、化合物が含まれない構造（アポ体） での長時間シミュレーションではポケット構造が崩れた事例が報告されている[17]。さらに、タンパク質と化合物の複合体を使用した長時間 MD では化合物がポケット内で安定に結合したのち他の結合様式に遷移せず網羅的な結合様式が抽出できない例も確認されている（付録.A） 。このため、MD シミュレーションに必要な適切なタンパク質-化合物の初期構造をいかにして効率良く選択するかが大きな課題となっている。

機械学習による結合様式推定では、加速度的に増加した、タンパク質構造データバンク（PDB）に収載される結晶構造（Figure 1.3）を活用して、結合様式を推定する。2000 年代より、結合様式を予測するための機械学習法では、タンパク質構造および化合物構造は 1 次元の記述子およびフィンガープリントとして扱われてきた[18]。これらの方法では、3 次元空間におけるタンパク質-化合物の空間配置情報が失われ、タンパク質-化合物間の相互作用が十分に表現できないことが懸念される[19]。そのため、最近では 3 次元構造情報を用いた機械学習法も試みられている[20]。しかしながら、結合ポケットが大きく、化合物によりポケット形状が大きく変化するタンパク質の場合には、これまでの機械学習を用いた結合様式推定法では、何故そのような結合様式が推定されたのかを説明することは難しく、結果の解釈ならびに予測精度を高めるための方法論の開発が強く求められている。次に、そのようなタンパク質の代表例であるシトクロム P450 3A4 について詳述する。

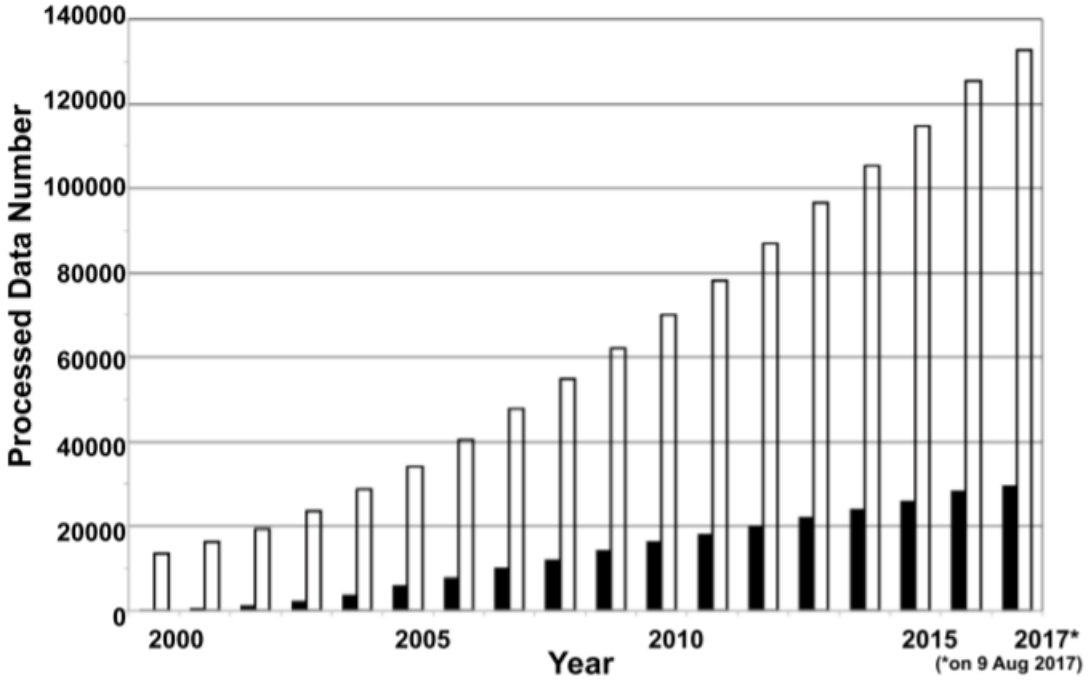


Figure 1.3 タンパク質構造データバンク（PDB）に収載される結晶構造数の推移[21]
白棒が PDB 登録数（黒棒は、PDBj（https://pdbj.org/） によって登録編集が行われたデータ数）

1.4 シトクロム P450 3A4（CYP3A4）における結合予測問題

本研究では、機械学習による結合様式予測方法開発の課題として、タンパク質の結合ポケットが広く、かつポケットを形成する表面構造が変化するシトクロム P450 3A4（CYP3A4）を題材として取り組んだ。シトクロム P450 は、本来生体内に侵入した異物を無毒化（代謝）する酵素であるが、投与された薬物が小腸や肝臓を通過する際に薬が異物として代謝・排泄されるために、薬物の効果が減弱したり失活したりする問題が起こる。CYP3A4 の結合ポケットは化合物形状に合わせて変化し（Figure 1.4）、かつ結合ポケットは最大級である（Figure 1.5）。このため CYP3A4 は多様な薬物を代謝する。実際、CYP3A4 は医薬品候補化合物の代謝への寄与が小腸で 80%、肝臓で 40%と CYP ファミリー内で最も大きく[22]、CYP3A4 に代謝されにくい化合物の設計は重要な創薬課題となっている（Figure 1.6）。このことから、CYP3A4 は薬物候補化合物の結合様式予測が高難度であり、結合様式予測法の開発が喫緊の課題となっている。

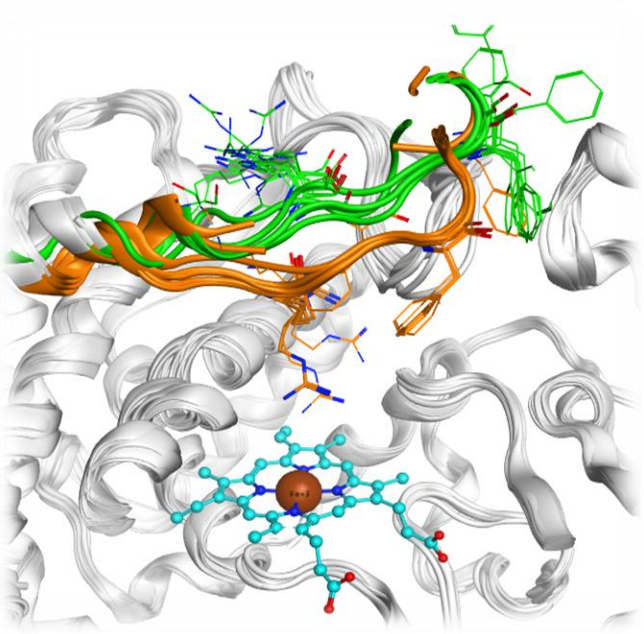


Figure 1.4 CYP3A4 の結晶構造重ね合わせ

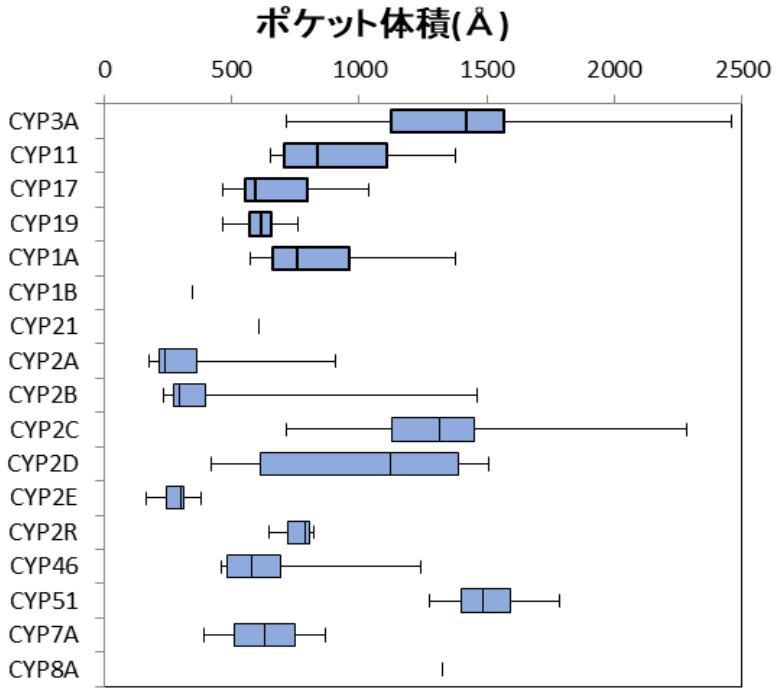


Figure 1.5 CYP ファミリーごとのポケット体積分布

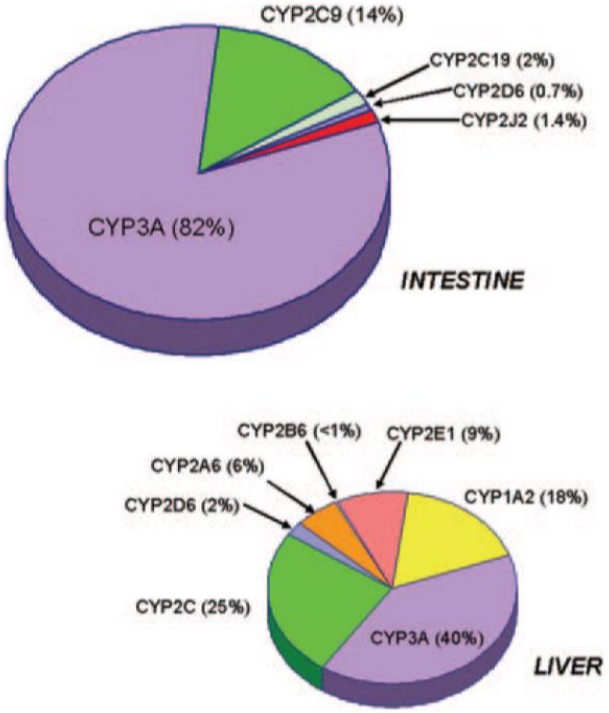


Figure 1.6 個々の CYP 酵素の寄与率[22]

1.5 結言

以上で述べたように、従来の結合様式予測ではドッキングシミュレーションが主に使用されているが、結合ポケットのサイズが大きく、動きのある標的タンパク質での結合様式予測が難しい。一方で長時間の MD シミュレーションのみでは結合様式の網羅性に問題がある。さらに、これまでの機械学習による結合様式予測では、結合ポケットのサイズが大きく、動きのある標的タンパク質に対しては結果の解釈が困難という問題がある。そこで、本研究では結合ポケットの動きを考慮して化合物の結合様式を予測する方法の開発を目的とし、題材としてポケットが大きく、主鎖レベルで構造の揺らぐ CYP3A4 を用いて以下 2 つのアプローチを検討した。1 つ目のアプローチはタンパク質と化合物複合体の 3 次元構造を用いた機械学習による結合様式予測である。2 つ目のアプローチは複数の短時間 MD シミュレーションを並列させた結合様式予測である。最後に機械学習により得られた結合様式予測モデルを MD 初期ポーズの選択に適用し、機械学習とシミュレーションを組み合わせた結合様式予測を試みた。

機械学習アプローチでは、CYP3A4 を題材とし、ターゲットタンパク質のデータが少ない場合のデータセット作成の方法論を検討した。シミュレーションアプローチでは、27 個の MD 初期ポーズから短時間の MD シミュレーションを並列して実行した結合様式予測を検討した。化合物が実際に代謝を受ける部位と予測結果を照合し、矛盾がないことを確認している。さ

らに、機械学習で得られた CYP3A4 の結合様式予測モデルで MD 初期ポーズを半数に絞り込んだ結果、多数の初期ポーズからのシミュレーション結果と同等の結合様式予測結果が得られた。これにより、シミュレーションと機械学習の 2 つのアプローチを組み合わせることで、ポケットの揺らぎを考慮した高精度な結合予測プロセスを効率よく実施できることを示した。

本論文は全 6 章から構成される（Figure 1.7）。第 2 章ではシミュレーションを利用した結合様式予測と 機械学習による結合様式予測の各課題に対する関連研究と本論文の立ち位置を整理する。第 3 章では、機械学習アプローチによる結合様式予測において、CYP3A4 を題材とし、ターゲットタンパク質のデータが少ない場合のデータセット作成の方法論について述べる。次いで第 4 章ではシミュレーションアプローチによるポケット揺らぎを考慮した結合様式予測の中で、CYP3A4 を題材として短時間 MD を多数実施した試みを述べる。さらに第 5 章では結合ポーズ予測を効率化させる方法論の提案に向け、第 3 章で得られた機械学習モデルを適用して MD 初期ポーズを選択した結合様式予測、第 3 章で選択したデータセットに関する考察について述べる。最後に第 6 章では本研究で得られた結果を総括すると共に今後の展望について述べ、結論とする。

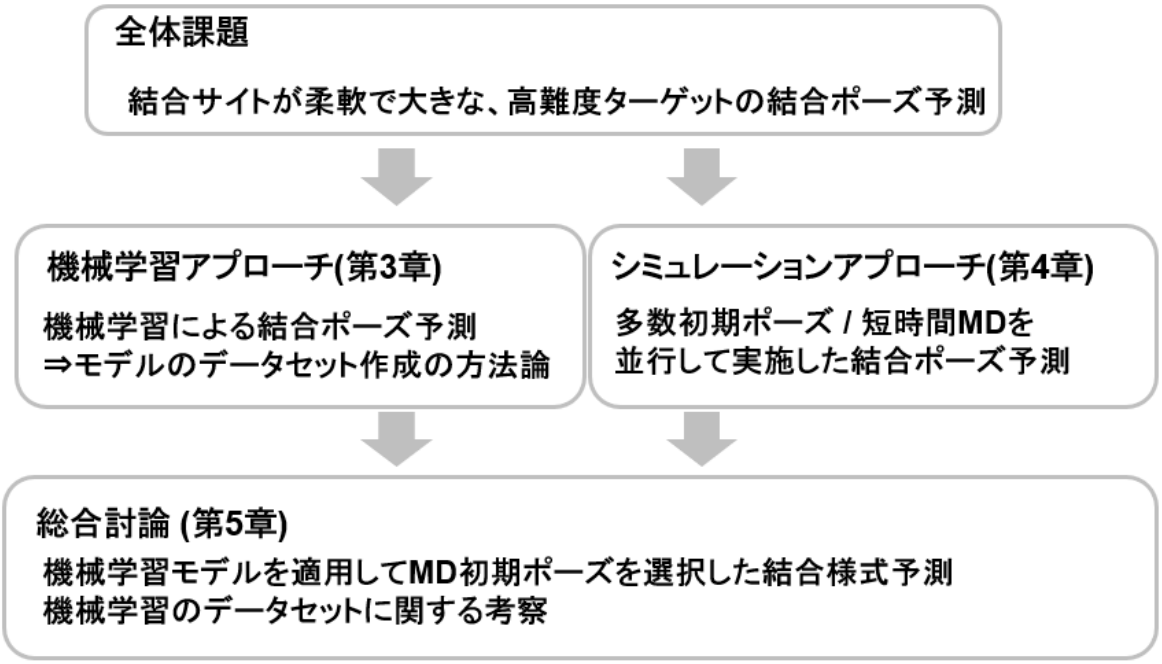


Figure 1.7　論文の構成

第2章　関連研究

2.1　緒言

第 1 章では、ポケットサイズが大きく、タンパク質主鎖を含めた構造揺らぎを考慮した化合物の結合様式予測が必要であると述べた。第 2 章では、この問題を解決するためのアプローチとして立てた、機械学習による結合様式予測とドッキングシミュレーションによる結合様式予測とシミュレーションを活用した結合様式予測に関連する研究について述べる。

2.2　機械学習によるポーズ正誤予測に関する既往の研究

ここ数年で、機械学習の分野は理論的な研究から実社会の応用へと移行しつつある[23]。自動車や金融など多くの消費者サービス業界は早期より機械学習を採用してきたが、製薬業界での活用は最近まで遅れていた。その要因として、倫理的な理由でヒトに関するデータへのアクセスが難しかったこと、製薬企業内で取得された実験データは機密事項とされ大規模なデータが集めにくいこと、さらに製薬企業側が機械学習で何ができるのかを理解していなかったことなどが挙げられる。近年の Merck Kaggle challenge[24]、IPAB[25], [26]など公共のコンテストおいて、ディープラーニングなどの新しい機械学習アルゴリズムを利用した薬物探索の著しい成果が示された[27]。これらの成果をきっかけとして創薬開発の短期化と成功率向上を期待とした機械学習の利活用が急速に進み、現在では臨床試験を含む薬物の研究開発の全てのプロセスにおいて、機械学習の応用が検討されつつある（Figure 2.1）。

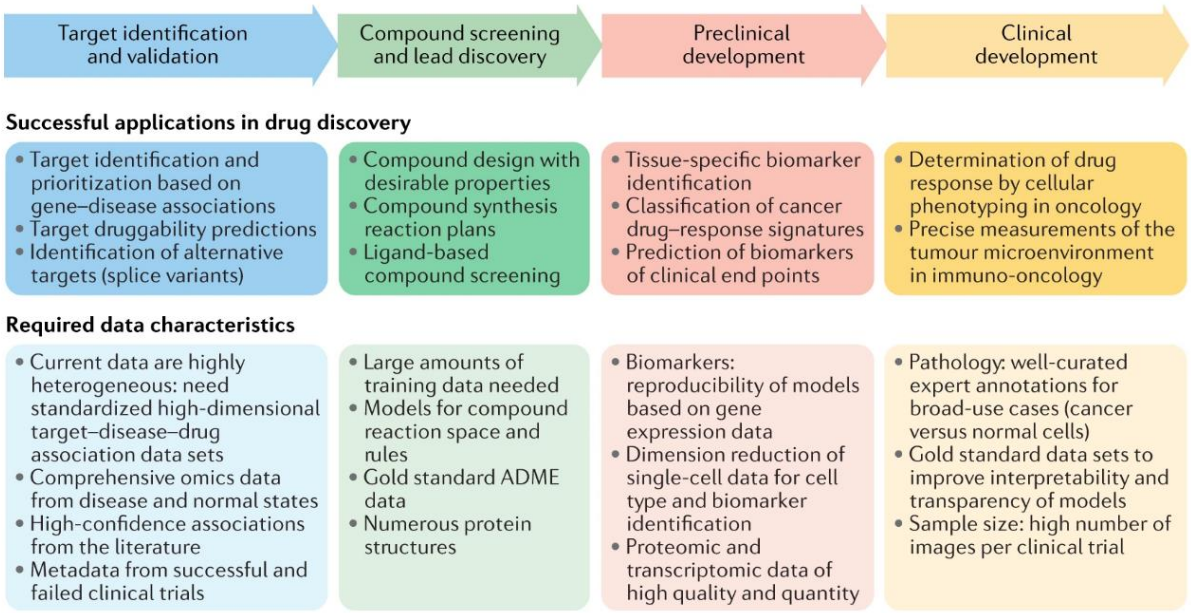


Figure 2.1 創薬パイプラインにおける機械学習の応用とそれらに必要なデータ特性[23]

昨今の「AI 創薬」への期待が高まる以前から、実験データ取得が比較的容易な創薬の初期段階では機械学習の適用が進められてきた。特に、医薬品ターゲットと良く結合する化合物を計算機中で評価するバーチャルスクリーニング技術は様々な手法が開発され、薬の「タネ」となるヒット化合物の取得率はランダム選択と比較し数十～数百倍向上するなど初期の創薬研究に欠かせない技術となっている[28]。タンパク質構造情報を利用したバーチャルスクリーニング（Figure 2.2）では、2.3 で述べたドッキング・スコアリング方法の改良により、スクリーニングプロセスの大部分は化合物が機械的にフィルタリングされている[29]。しかしながら、ドッキングされた化合物のランク付け性能が未だ不十分であることから、500～1000 化合物程度に絞り込んだ後の目視による最終チェックが依然として必要となっている[30]。この目視検査は、目視検査者の専門分野により主観的に判定される（同じ人間が同じドッキング結果を検査した場合でも別の日に実施すると判定結果が変わる）とともに非常に労力のかかる作業となる[31]。そこで従来のスコア関数に加えて機械学習によって「妥当な」結合様式および結合活性を予測する方法の開発が種々検討されている。

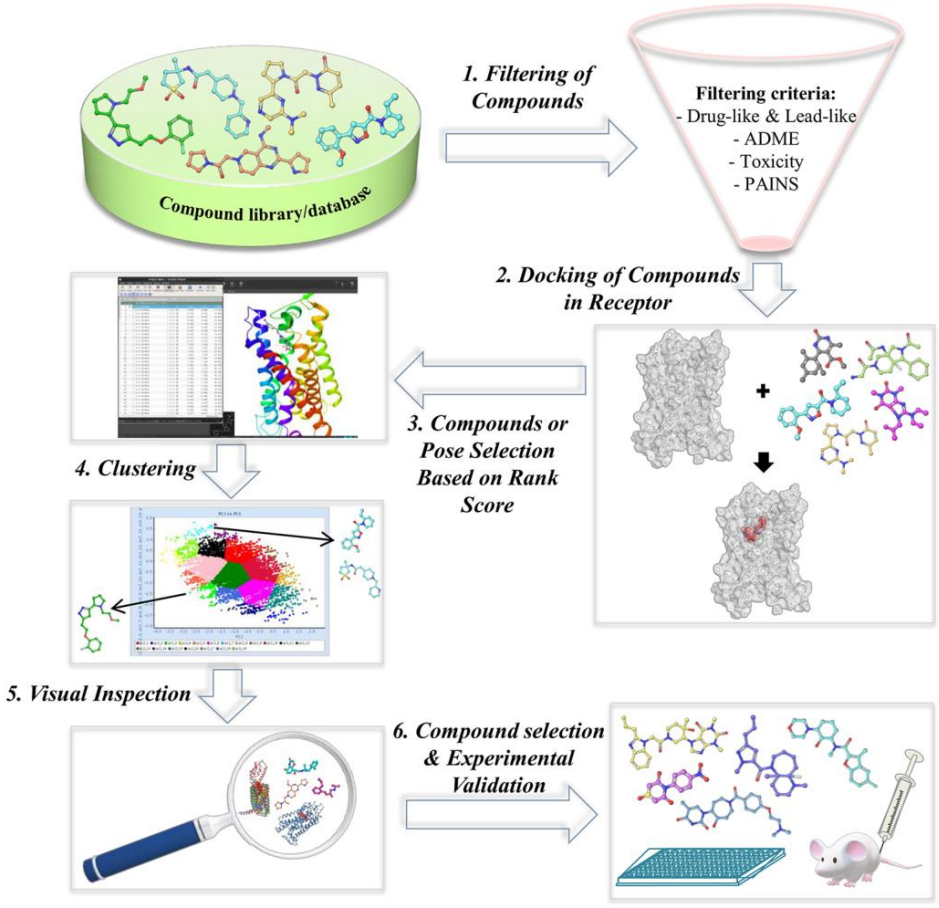


Figure 2.2　タンパク質構造ベースのバーチャルスクリーニングの典型的なワークフロー[29]

機械学習ベースのスコア関数は上で述べた古典的なスコア関数と比較してよい性能を示すことが報告されている[30], [32], [33]。機械学習ベースのスコア関数の多くはバーチャルスクリーニングにおいてタンパク質と化合物が良く結合しうるかどうか（結合活性）を評価する[18]。一方、薬物設計を目的とした結合様式予測のためのスコア関数の開発は結合活性のスコア関数ほど多くはないが、2015 年ごろまでタンパク質構造および化合物構造は 1 次元の記述子およびフィンガープリントとして扱われてきた[34]–[36]。これらの方法では、3 次元空間におけるタンパク質-化合物の空間配置情報が失われ、タンパク質-化合物間の相互作用が十分に表現できないことが懸念される[19]。そのため、最近では 3 次元構造情報を用いた機械学習法の開発が試みられている。

驚くべきことに、結晶構造学者など多数の「正しい」結晶構造に触れる経験を重ねたり適切な訓練を受けたりした人間は、ドッキング後の目視検査において構造を物理的または化学的方程式を用いることなく視覚的に分析し、化合物結合を特徴付けて結合様式を正しく評価することができる[34]。計算モデルは脳の複雑さには及ばないが、タンパク質-化合物複合体構造の正しさを識別する脳の力を、3 次元構造情報を入力としたニューラルネットワークモデルで模倣できると期待されている。Wallach らは、化合物の結合サイトをグリッドで表現し

て構造的特徴を表す値を保持させ、deep convolutional neural network で活性予測を検討した最初の機械学習モデル「AtomNet」を開発した[37]。さらに Ragoza らは、AtomNet と類似した手法でタンパク質-化合物相互作用の 3 次元（3D）表現を入力として使用する畳み込みニューラルネットワーク（CNN）で結合様式を識別するスコアリング関数を開発した[20]。その他、バーチャルスクリーニングを志向したスコアリング関数の開発も含めてこの数年で非常に盛んな研究分野となっている[38]–[42]。

タンパク質 - 化合物複合体の訓練データは、目的のドッキング用途およびモデリング戦略に応じて、その品質、タンパク質ファミリーメンバ、構造および結合データの種類によって選択することができる[35]。古典的なスコア関数は通常、あらゆるファミリーのタンパク質との複合体をスコア付けするために、結合定数とともに最高品質の数百の X 線結晶構造を採用している。これとは対照的に、機械学習スコア関数のデータ選択は、はるかに多様で（Figure 2.3）、興味深いことに従来スコア関数の開発に有害と考えられていた低品質の構造および相互作用データの活用は機械学習のスコア関数の予測性能にとっては有益であることが確かめられている[43]。しかしながら、創薬現場で機械学習を利用する際に、（特に新規のターゲットである場合など）学習データとなる 3 次元構造情報（結晶構造）が不十分なケースが多く、そのような場合にどのようなデータを作成すべきかについてはほとんど報告例がない。

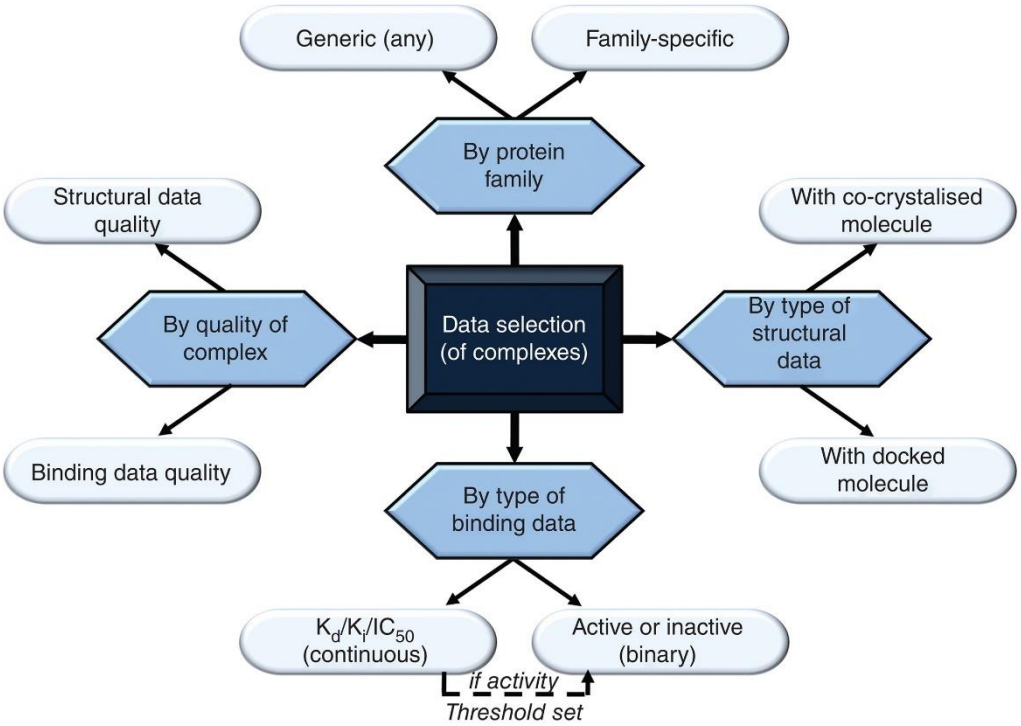


Figure 2.3　スコアリング関数を構築および検証するためのデータ選択[35]

2.3　ドッキングシミュレーションによる結合様式予測に関する既往の研究

ドッキングシミュレーションは、ドッキングポーズの生成とドッキングポーズの評価 （スコアリング） の 2 つのプロセスから成る（Figure 2.4）[9]。ドッキングポーズの生成では、化合物のとり得る配座を生成した後、タンパク質の結合ポケットに様々な向きでフィットさせて仮の結合様式（ドッキングポーズ） を得る。その後のドッキングポーズ評価では、ドッキングポーズの生成で得られた仮の結合様式を、タンパク質との相互作用や化合物の配座の安定さなどで評価する。本節ではこの 2 つのプロセスそれぞれについて述べる。

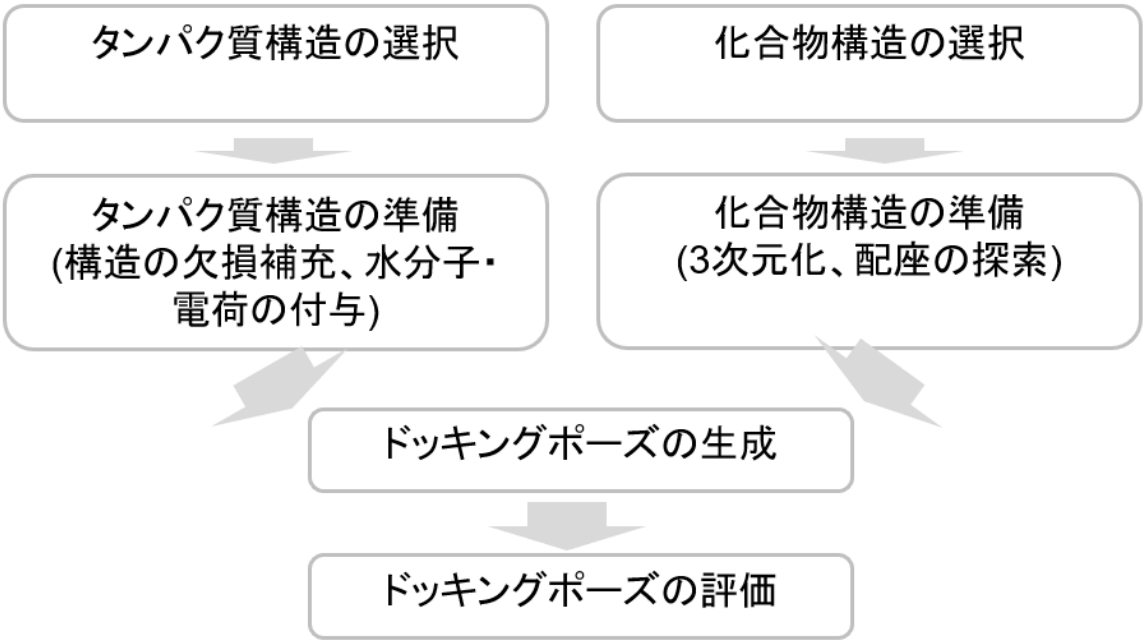


Figure 2.4　一般的なドッキングシミュレーションのワークフロー

ドッキングシミュレーションとして、Figure 2.5 に示す 4 つの方法がある。ドッキング法が発明された初期のモデルでは、計算機性能の限界によりタンパク質と化合物を剛体として仮定し扱う「鍵と鍵穴モデル（lock-and-key model）」が用いられた[44]–[46]。その後、Koshland によって提案された「induced fit」理論は、化合物と受容体がドッキング中に柔軟であるとして扱われるべきであることを示唆した[47], [48] 。しかしながら、受容体/化合物両方を完全に柔軟なものとして扱う induced-fit model は、lock-and-key model よりも自由度が格段に高く、莫大な計算コストを必要とする。Induced-fit model と lock-and-key model の中間を取り、側鎖の柔軟性のみを考慮した partially flexible protein model[13], [14]も考案されているが、現在ではタンパク質側は固定し、結合ポケット内の化合物の柔軟性を考慮する flexible ligand model が

標準的に用いられている。

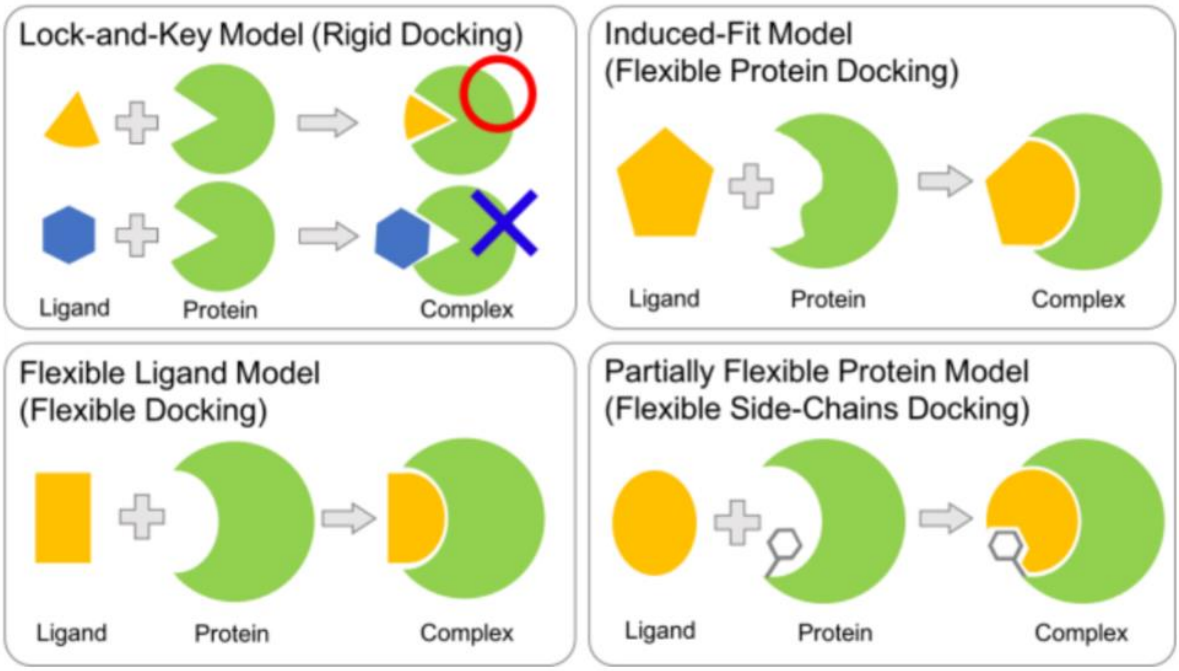


Figure 2.5　ドッキングシミュレーションの 4 つの方法[9]

過去 20 年にわたり、*FlexX*[49]、*GOLD*[50]、*DOCK*[51]、*eHiTS*[52]、*LigandFit*[53]、*AutoDock*[54]、*Surflex*[55]、*Glide*[56]、*FRED*[57]、*MOE-Dock*[58]、*LeDock*[59]、*AutoDock Vina*[60]、*rDock*[61]、*UCSF Dock*[62]など、学術用途と商業用途の両方で 60 を超えるさまざまなドッキングプログラムが開発されてきた。ドッキングポーズの生成の観点でこれらのプログラムを大まかに分類すると、形状に基づいたアルゴリズム、遺伝的アルゴリズム、系統的またはモンテカルロ法に基づく探索アルゴリズム、フラグメント探索法に分けられる[63]。*DOCK* に代表される形状に基づいたポーズ生成は、局所的形状特徴を用いて化合物を受容体に配置する方法である[64]。 遺伝的アルゴリズムは、生物の適応進化に関するダーウィンの自然選択説を計算機上で模擬する最適化手法であり[65]、*GOLD*、*AutoDock* ではこれをドッキングポーズの配置探索に利用している。*Glide*、*LigandFit*、*AutoDock Vina* は系統的または乱数により系の状態を生成するモンテカルロ法[66]に基づく探索アルゴリズムによるポーズ生成法を採用している。フラグメント探索法は、化合物をフラグメントに分割してタンパク質表面との相補性を探査し推定ポーズを生成する方法で、*Surflex*、*FlexX* などがこの方法を採用している。これらのプログラムによるドッキング構造の探査は非常に高速であるが、成功率は探査アルゴリズムの性能に大きく依存する。特に回転可能結合数が大きい化合物に対するドッキング構造の探査は困難であり、ドッキングシミュレーションの成功率は低下する傾向にある[9]。またこれらドッキングプログラムは基本的にタンパク質側を剛体として扱っており、タンパク質の柔軟性の扱いは依然として大きな課題となっている[63]。

先に述べたように、化合物構造の柔軟性が大きくなるほど化合物配座の探索空間が広がり、正しいドッキングポーズを生成することが難しくなる。ドッキングプログラムによる結合構造の予測性能は、これまでに様々な研究によって評価されている[10], [11], [67]–[72]。このうち、Plewczynski らは 1300 種類のタンパク質-化合物複合体結晶構造を利用し、7 つのドッキングプログラム（*GOLD*、*AutoDock*、*Surflex*、*LigandFit*、*Glide*、*eHiTS*、*FlexX*）について化合物の柔軟性（回転可能結合数）とドッキング正解率を比較している[11]。なおドッキングプログラムによる結合構造の予測性能評価は、Kramer らの分類[73]により RMSD （Root Mean Square Deviation） 最大 2.0 Å までを「正解」としている例が多く、Plewczynski らの研究もこの基準を踏襲している。回転可能結合数 5 以下の比較的「固い」化合物と回転可能結合数 6 以上の柔軟な化合物のドッキング成功率を比較すると、全てのプログラムにおいて回転可能結合数 6 以上のドッキングポーズ正解率は回転可能結合数 5 以下の化合物と比較して大きく減少している（Table 2.1）。この問題に対しては、ドッキングプログラムの探査アルゴリズムを改良することによって性能が向上する結果も報告されている[74]–[76]。

Table 2.1　化合物の回転可能結合数で分類したドッキングポーズ正解率の比較[11]

ドッキングプログラム	トップスコアのポーズの正解率（RMSD < 2 Å） [%]	
	回転可能結合数 5 以下	回転可能結合数 6 以上
eHiTS	64.8	37.7
FlexX	49.0	26.0
Glide	49.6	34.5
GOLD	67.1	48.6
LigandFit	51.3	28.4
Surflex	53.9	41.3
AutoDock	61.1	27.5

ここまではタンパク質を剛体として扱う方法（rigid docking） を中心に関連研究を紹介したが、実際には化合物だけでなくタンパク質構造も動いているため、より精緻なドッキングシミュレーションを行うためにはタンパク質の立体構造の変化を扱う必要がある。タンパク質の柔軟性を扱うためのドッキング方法として、タンパク質側鎖の柔軟性を扱う partially flexible ドッキング[13], [14]、複数の結晶構造もしくは MD シミュレーションによりあらかじめ複数の構造のタンパク質構造を準備しておき、並列にドッキングを行うアンサンブルドッキング[12], [77]がある。各手法をタンパク質の動きの大きい順に整理した（Figure 2.6）。

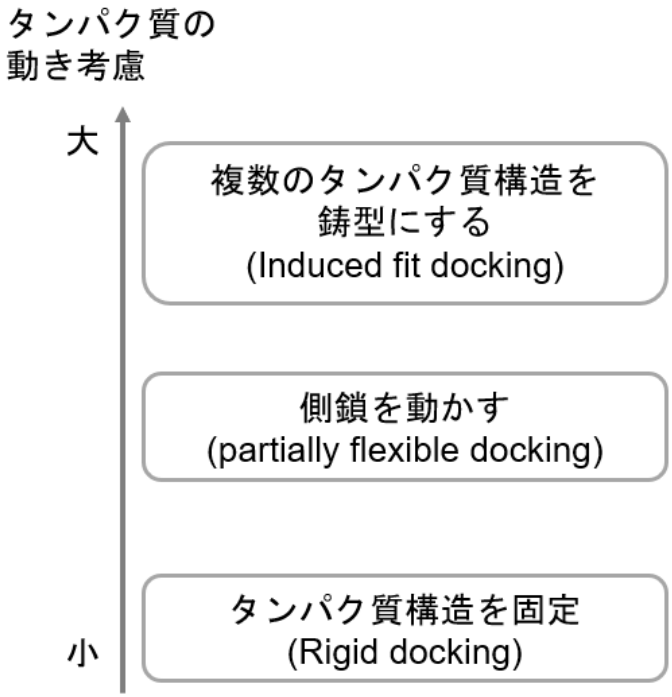


Figure 2.6 タンパク質の柔軟性を扱うためのドッキング方法と考慮する動きの大きさ

Partially flexible ドッキングは、結合ポケット周辺のアミノ酸側鎖の自由度を考慮してドッキングシミュレーションを実施する方法である（Figure 2.7）[13], [14], [78]。Sander らは、柔軟なアミノ酸側鎖の回転異性体をサンプリングしたのち、多数のポケット構造を作成してドッキングし、化合物とポケット間の衝突が最小となるように構造最適化することでタンパク質構造の柔軟性を考慮した手法を開発した[78]。「Fleksy」と呼ばれるこの方法により、35 種類のタンパク質-化合物複合体について rigid docking と「Fleksy」での結晶構造の再現率を比較した結果、rigid docking では 44%であったのに対して「Fleksy」では 78%と大幅に向上していた。この他、ドッキングプログラム（*Glide*）と蛋白質立体構造予測プログラム（*Prime*）を相互に組み合せた解析プロトコルも考案されている[14]。この方法の特徴は、活性サイト内にて induced-fit の原因となる自由度の高い残基を一時的にアラニン残基に置換して初期配座探索を行い、側鎖構造をモデリングして元に戻し構造最適化する点である。これらの方法は、いずれも結合ポケット内の主鎖構造は硬く、限られた数の可動側鎖を有すると仮定している。4000 ほどの結晶構造を用いてタンパク質主鎖と側鎖の動きを解析した最近の報告では、タンパク質主鎖の運動は微小であり、側鎖は柔軟に動くことが報告されており[79]、ほとんどのケースでは Partially flexible ドッキングが合理的と言える。しかしながら、主鎖構造の動きも観られるタンパク質に対してはその柔軟性の考慮は十分でない。

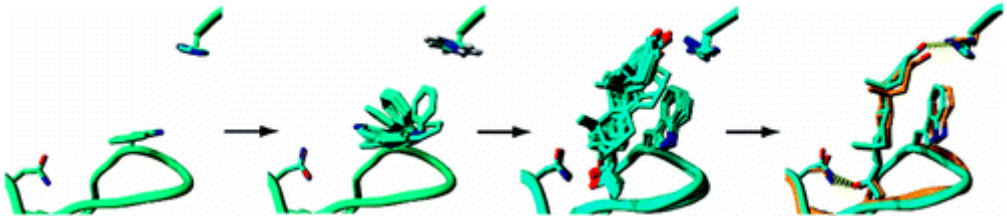


Figure 2.7　Partially flexible ドッキングのイメージ[78]

可動性の側鎖をサンプリングしたのち、複数ポケット内でのドッキングポーズがリスコアリングされ、最終ポーズが得られる。

主鎖の動きも考慮したアンサンブルドッキング（Figure 2.8） の手法開発は 1990 年代に始まり[80]、酵素阻害剤[81]やタンパク質表面の揺らぎが大きいタンパク質-タンパク質相互作用阻害分子の探索[82], [83]において成功例が報告されている。1999 年の論文[80]は、複数の MD 構造または複数の結晶構造に基づくコンセンサスファーマコフォアモデルが、単一の立体配座に基づくモデルよりも結合予測に成功していることを示した。2002 年にはタンパク質のアポ体 MD から得た多数のスナップショットをドッキング鋳型に使用した「“relaxed-complex” scheme （RCS）」が検討された。多くの MD スナップショットの使用により受容体の柔軟性を説明することができたが、同時にドッキングのための計算コストも増加させた。この問題に関しては、タンパク質の原子位置に基づくクラスタリングアルゴリズムを使用することにより、アンサンブルドッキングの効率が大幅に向上している[84]。この方法の問題点として、シミュレーションによって到達可能なタイムスケール（通常は数マイクロ秒） とターゲット構造の変化の遅さの大きな差によって、ターゲットタンパク質の配座空間のサンプリングが不十分であることが挙げられる[77]。ANTON[85]のような使用目的を MD シミュレーションに絞ったコンピュータの構築は、タンパク質の MD シミュレーションをミリ秒のタイムスケールに拡張することを可能にした。しかしながら、最近の報告では単一タンパク質の内部ダイナミクスは、非平衡的かつ非周期的であり、1 ミリ秒のシミュレーションでも十分ではないことが示唆され[86]、十分なサンプリングは依然として課題となっている。

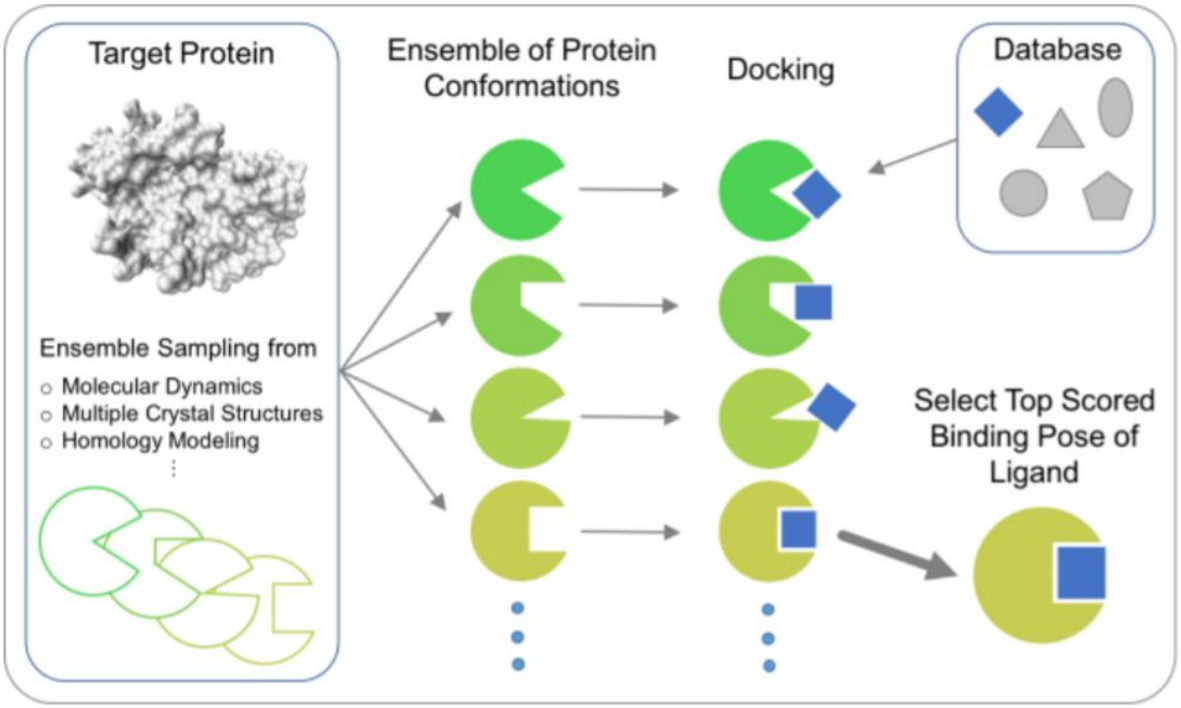


Figure 2.8　アンサンブルドッキングの概念図[9]

ドッキングポーズの評価に利用されるスコア関数はこれまでに 100 以上報告されており、これらのスコア関数の種類や特徴については複数の総説で詳細に解説されている[87]–[89]。大まかに分類すると、物理（力場）ベース、経験ベース、知識ベース、機械学習ベースの 4 つに分けられる。物理（力場）ベースのスコア関数はタンパク質と化合物の原子間に働くファンデルワールス相互作用や静電相互作用を直接計算する[46], [73]。これらの古典的な方法に加え、近年溶媒和/脱溶媒和効果を考慮した計算法[90]–[95], [12]や量子力学に基づいたスコア関数[96]–[98]が開発されている。経験ベースのスコア関数では、水素結合、疎水効果、立体衝突などの分子間の寄与を様々な経験的ポテンシャルとして記述する[99]–[102]。比較的短時間でスコアが求められるため、ChemScore （*GOLD*）、 X-Score, LigScore （*LigandFit*）, GlideScore （*Glide*）、FlexX （*FlexX*）などタンパク質-化合物のドッキングプログラムを用いたバーチャルスクリーニングで利用されている[9]。知識ベースのスコア関数では、PDB から得られる大量の複合体結晶構造をデータセットとして利用し、複合体の 2 原子間のポテ ンシャルを統計的に定める[103], [104]。これらのスコアリングのためのトレーニングセットは構造情報のみからなり、実験的な結合親和性データとは無関係である。そのため、知識ベースのスコア関数は実験条件によって生じる可能性のある結合親和性曖昧性を回避できることから、結合親和性予測よりも結合ポーズ予測に適している[89]。機械学習ベースのスコア関数では、サポートベクターマシン、ランダムフォレスト、ニューラルネットワーク、ディープラーニングなどの手法によりドッキングポーズをスコアリングする（Figure 2.9）。機械学習ベースのスコア関数はトレーニングデータセットに依存しているため、直接ドッキングソフトウェアに組

み込まれておらず、通常リスコアリングのために使用されている[105]。機械学習ベースのスコア関数は上で述べた従来法と比較してよい性能を示すことが報告されており[30], [32], [33]、現在様々な手法の開発が進められている。これらについての詳細は、2.2 で述べた。

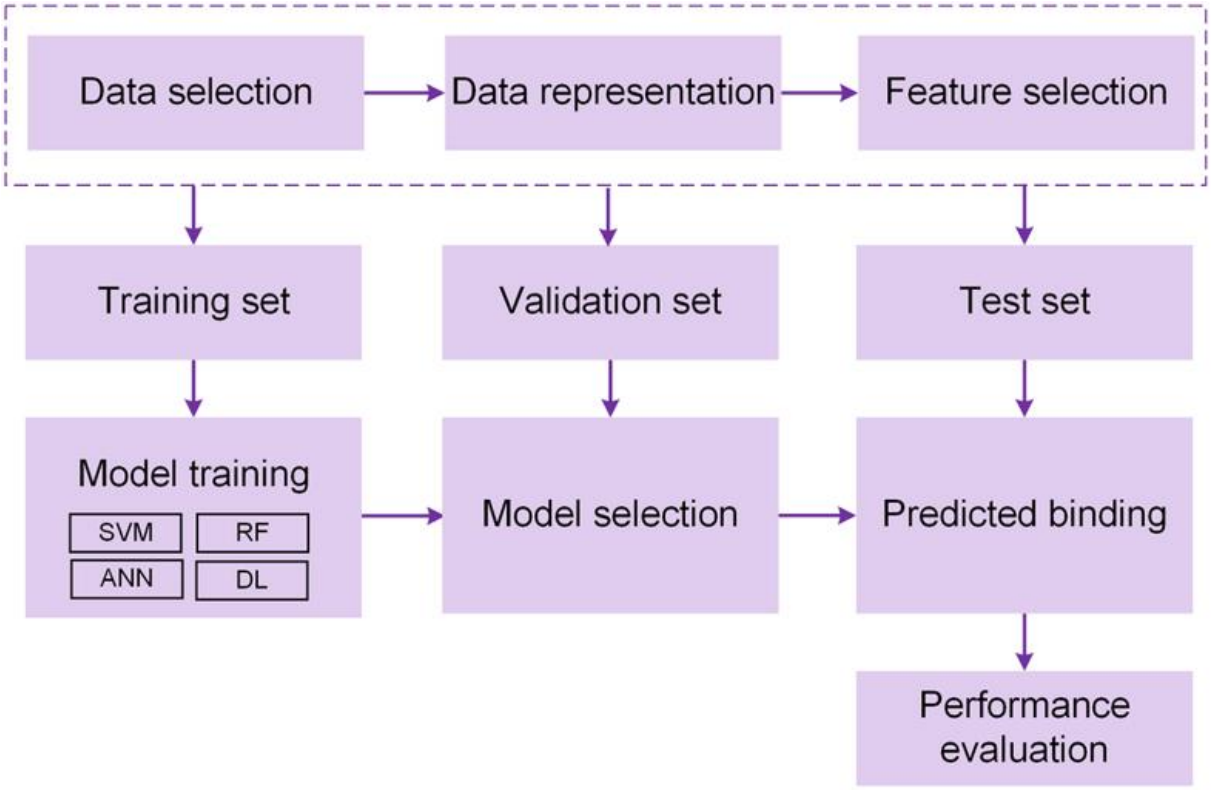


Figure 2.9　機械学習ベースのスコア関数をトレーニングするワークフロー[89]

2.4　　シミュレーションを活用した結合様式予測に関する既往の研究

MD シミュレーションは、原子間相互作用を支配する物理学の一般的モデルに基づいて、タンパク質または他の分子系内のあらゆる原子が経時的にどのように移動するかを予測する計算科学手法である[106]。本手法は新しいものではなく、タンパク質の最初の MD シミュレーションは 1970 年代後半に行われ[107]、これらのシミュレーションを可能にした基礎的な研究[108], [109]は 2013 年のノーベル化学賞を受賞している。近年 MD シミュレーションは、特に実験的な構造生物論文に頻繁に現れ始めており（Figure 2.10）、実験結果の解釈と実験作業の指針の両方に用いられている。MD シミュレーションにおけるタイムステップは数値的な安定性を確保するために典型的にはそれぞれわずか数フェムト秒（10^{-15} 秒） である。一方で、生化学的に興味深い事象、例えばタンパク質の機能的に重要な構造変化などは、ナノ秒、マイクロ秒、またはそれ以上のタイムスケールで起こる。したがって、典型的なシミュレーションは、数百万から数十億のタイムステップを含む。1 つのタイムステップで評価される何

百万もの原子間相互作用と組み合わさり、MD シミュレーションは非常に計算上の要求が厳しい。

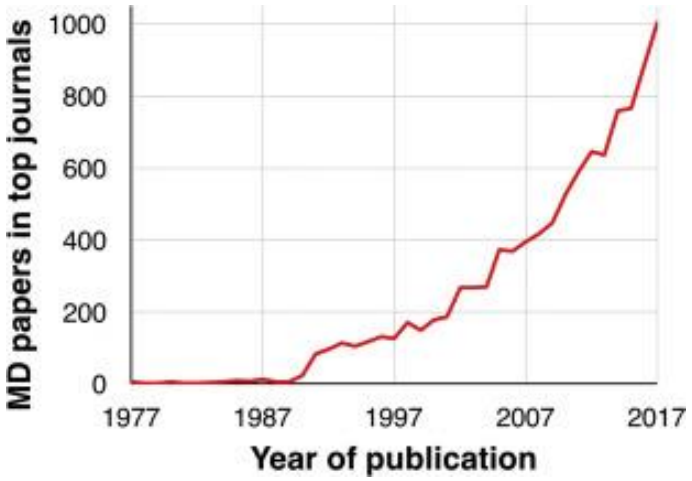


Figure 2.10　MD シミュレーションが利用されている構造生物学論文[110]

過去数十年にわたる計算ハードウェアおよび MD に使用されるアルゴリズムおよびソフトウェアの改良により、より長くより安価なシミュレーションが可能になっている。高度に特殊化されたハードウェアは、達成可能な最大速度を大幅に向上させた[85], [111]。また GPU の登場により生物学的に意味のあるタイムスケールのシミュレーションをこれまで以上に多くの研究者が利用できるようになった[112]。このような状況から、今後必要となるのはシミュレーションによってどの質問に対処できるかを考え出し、これらの質問に対処するためのシミュレーションを設計し、シミュレーション結果を解釈することである。シミュレーション結果の解釈、すなわち揺れ動く原子の質量を記述する大量の軌跡データ（トラジェクトリ）から生物学的洞察を得ことは困難な場合があり[110]、また単純な「ブルートフォース」シミュレーションでは解決できない問題に対処するために、さまざまな高度なシミュレーション技術が開発されている。

多くの MD シミュレーション研究は、作用中の生体分子プロセス、特に化合物結合、化合物または電位によるコンフォメーション変化、タンパク質フォールディング、または膜輸送など、実験で直接対処することが難しい重要なタンパク質機能プロセスの観察を目的としている[110]。最も基本的で直感的なシミュレーションの応用は、生体分子のさまざまな領域の移動性や柔軟性を評価することであり、これによりタンパク質機能および化合物結合にとって重要なタンパク質の構造変化や水分子、イオンの動的挙動を明らかにすることができる[113]–[115]。シミュレーションで安定している化合物の結合ポーズは不安定であるポーズよりも正確である可能性が高く[116]、cryo-EM における曖昧な化合物ポーズを決定するのに有効であったとの報告もある[117]。またシミュレーションは新しい実験的研究につながる仮説を生み出すことができ、これらの仮説は EPR （電子スピン共鳴）分光法、NMR（核磁気共鳴）

分光法、蛍光クエンチング法、水素-重水素交換など、生体分子の構造の集合体またはダイナミクスに関する構造特性を直接調べる生物物理学的手法により検証される[110]。

創薬分野においては、シミュレーションが実験を推進する興味深い例が多く報告されている[118], [119]。特に、薬物の効果を強くしたり副作用を軽減したりする化合物最適化においてその有効性または他の特性を改善するための様々な情報を得ることができる。定性的には化合物によって誘導される結合ポケットの揺らぎを予測し、鍵となる相互作用を予測して化合物の結合ポーズを精緻化する[120], [121]。いくつかのケースでは、完全な薬物結合プロセスのシミュレーションにより、結合部位および化合物のポーズを明らかにした例もある[15], [122], [123]。定量的には、シミュレーションベースの方法は、ドッキングなどの他の計算手法よりも実質的に正確な化合物結合親和性推定値を提供する。自由エネルギー摂動法（Free energy perturbation）では、一連のシミュレーションを通じて 1 つの化合物が徐々に別の化合物に「変換」され、結合エネルギーが最も正確に推定される[124]。しかしこの方法は同様の骨格を共有する配位子間の相対的結合エネルギーを計算する場合にのみ一般に信頼性が高い[125], [126]。化合物（候補薬物）が標的タンパク質と結合する過程での自由エネルギー変化を求め、結合の強度を定量的に推計する MP-CAFEE 法も提案されているが[127] 、計算コストが高い。水分子を明示的に表現せず、連続体溶媒モデルを利用した方法（MM- PBSA 法、MM- GBSA 法）は、精度が落ちるもののかなり高速に結合自由エネルギーを求めることができる[128]。

MD シミュレーションを実際に実行することは比較的簡単になりつつあるが、実行する前に拡張サンプリング技術を含めてどのシミュレーションを実行するか、そしてどのように結果を分析するかの選択が課題となる。シミュレーションでより長い時間スケールのイベントをとらえるために、生体分子を所望の初期立体配座から所望の最終立体配座へ引っ張るターゲット MD[129]、すでに訪れた立体配座空間の領域からシミュレーションを遠ざけるメタダイナミクス、ある程度の自由度が伴う効率的な温度上昇、一旦系の温度を高温にして古典 MD よりも幅広く構造をサンプリングするレプリカ交換 MD[130], [131]や加温（temperature associated）MD[132]、また近年、局所探索による短時間化を目指した supervised MD （suMD）[133]、力場を変えてエネルギー障壁の高さを減らす加速 MD[134]などが報告されている。これらは、関心のある特定の反応座標を前もって特定することができる場合には、非常に有用であることが証明されている[110]。これらの拡張サンプリング技法は、任意の長いタイムスケールに到達するよう調整される一方で、精度の低下を招く懸念も指摘されており[135]、精度を低下させず、かつ十分な構造のサンプリングが可能なシミュレーション方法の工夫の必要性が依然として残されている。

2.5 結言

本章では、機械学習による結合様式予測、ドッキングシミュレーションによる結合様式予測ならびに MD シミュレーションによる結合様式予測に関連する研究について述べた。タンパク質を剛体として扱う方法（rigid docking）やタンパク質側鎖の動きを扱う partially flexible ドッキングでは、主鎖構造の動きも観られるタンパク質に対し柔軟性の考慮は十分でない。一方、機械学習によるドッキングポーズ正誤識別においては、創薬現場で機械学習を利用する際に、学習データとなる 3 次元構造情報（結晶構造）が不十分なケースが多い問題がある。本論文では、これらの解決策として、創薬現場でのニーズが高く、ポケットサイズが大きくかつ主鎖構造を含めて構造の揺らぎが見られ、従来法では結合様式の予測が難しい CYP3A4 を題材として、機械学習法と MD シミュレーションの両方を用いる新規結合様式予測法を提案する。第 3 章において、ターゲットタンパク質の結晶構造の数が限られている場合に機械学習に基づく結合様式予測のための適切な訓練データセットを選択する方法論を提示する。さらに、第 4 章において、分子動力学（MD）計算を用い、タンパク質原子の動きやすさも考慮したドッキングポーズ予測方法について提示する。

第3章　機械学習による結合様式予測

3.1　緒言

第 2 章（2.2）で述べたように、機械学習ベースのドッキングスコア関数は上で述べた古典的なスコア関数と比較してよい性能を示すことが報告されている[30], [32], [33]。そこで第 3 章では CYP3A4 を題材として、3 次元構造情報を用いた機械学習スコアリングによる、ドッキングポーズの正誤識別に取り組んだ。

2.2 で、1 次元の記述子およびフィンガープリントでタンパク質構造や化合物構造を扱うことでタンパク質-化合物間の相互作用が十分に表現できないことが懸念され、3 次元構造情報を用いた機械学習法の開発が盛んに検討されていることを述べた[20], [37]–[40] 。3 次元構造ベースの機械学習法の主な利点は、イオン結合、水素結合、ファンデルワールス相互作用など、ターゲット原子とその周囲の原子との間の相互作用を用いる点である。これにより、ドッキングの本質であるタンパク質および化合物の空間配置情報を失うことなく、ドッキングポーズの正確さを評価することを可能にする。しかしながら、創薬において 3 次元構造を用いた機械学習を行う場合、特に新たな創薬ターゲットについては、訓練データのサイズが限られている場合が多い。このような場面においては、どのようにトレーニングデータセットを作成するかが問題となっている。

Wallach らはドッキングポーズ予測のための最初の機械学習モデル AtomNet を開発した[37]。この手法は、ボクセルグリッドとしてタンパク質構造を扱い、コンボリューションニューラルネットワーク（CNN）を使用している。化合物の結合サイトは 20 Å の立方体として切り取られ、1 Å 間隔のグリッドが配置され、各グリッドには格子は構造的特徴を表す値が保持される。深い CNN アーキテクチャーを階層的に構築することにより、AtomNet は、タンパク質 - 化合物結合の複雑な非線形現象をモデル化し、さまざまなベンチマークデータセットに対する従来のドッキングアプローチを大幅に上回った。

Ragoza らはタンパク質-化合物相互作用を 3D で表現し AtomNet 類似の CNN スコア関数を開発した[20]。このスコア関数では、24 Å の辺を持つ立方体中に 0.5 Å の間隔でグリッドが配置され、ドッキングプログラム smina [136]のアトムタイプ密度が個々のグリッドに割り当てられる。この検討では、AutoDock Vina[60]でよりも優れたドッキングポーズ予測とバーチャルスクリーニング結果が得られた。また、Ragoza らのアーキテクチャーを使用して CNN ベースのスコアを個々の原子に対する寄与度に分解し可視化する方法も開発されている[137]。

この可視化方法は 3.5 の考察において使用している。

ディープニューラルネットワークを使用するもう 1 つの利点は、訓練データセットが少数の場合に、事前に多数のデータセットを使用して訓練されたモデルを基礎として再訓練できる点である。再訓練に関する最近の集中的な研究により、転移学習、帰納的学習、マルチタスク学習などの概念が生み出された。例えば、事前学習は、関連データで既にトレーニングされたモデルを利用することにより、新しいタスクの予測パフォーマンスを向上させる転移学習技術である[138]。しかしながら、CNN を含むディープニューラルネットワークの訓練パフォーマンスは、訓練セットの偏りの影響を強く受け、特に訓練データセットのサイズが小さい場合はしばしば過学習が起こる。この問題に対処するため、さまざまな化合物を含む大規模なデータセットを使用した事前学習により、さまざまな標的タンパク質に適用可能な一般的なドッキングポーズ予測モデルを構築し、過学習させずに標的タンパク質の特定の特徴を捉えたモデルを構築できるようファインチューニングを検討した。事前学習とファインチューニングの組み合わせで構築されたモデルは、ファインチューニングされていないモデルと比較して優れたパフォーマンスを示すことが報告されている[139]–[141]。

2017 年時点で、CYP3A4 について報告されている結晶構造は約 30 のみであった。よって、CNN モデルの事前学習とその後のファインチューニングのために、CYP3A4 の特徴を有する多数の他のタンパク質の結晶構造を別途準備する必要がある。この問題を解決するために、本研究では、さまざまな特徴とデータサイズの 4 つの訓練セットを使用した。ファインチューニングの有無に関わらず、これら 4 つのデータセットをさまざまな組み合わせで使用することで、データセットの構築、事前学習とファインチューニングが、CYP3A4 と結合したドッキングポーズを予測するためのニューラルネットワークモデルの精度にどのように影響するかを調べた。さらに、モデルで学習した結合様式の特徴を視覚化した。データセットの準備からモデル評価までをまとめたワークフローを Figure 3.1 に示す。

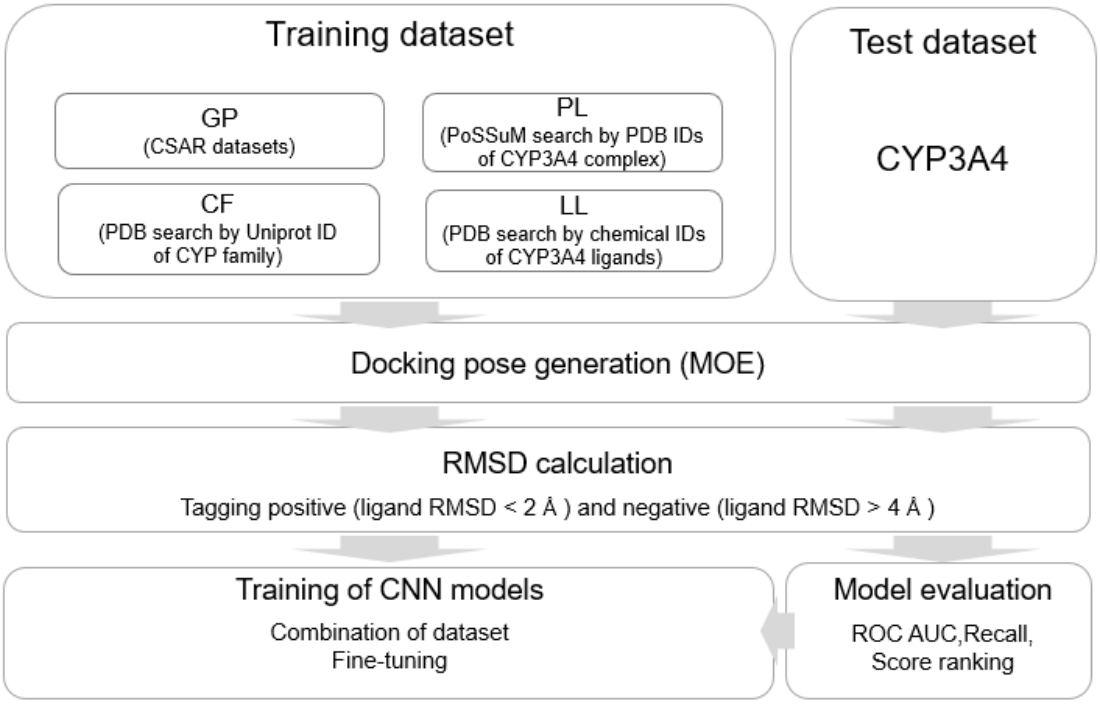


Figure 3.1　データセット作成からモデル評価までのワークフロー

3.2　データセット

本研究では、4 つのトレーニングデータセットと 1 つのテストデータセットを使用した（Table 3.1）。最大のデータセットである標的に依らないタンパク質データセット（GP データセット、Table 付録 1） は、モデルがタンパク質-化合物複合体の一般的な特性を学習するために構築した。データセットは、Ragoza らの方法[20]に従って選択され、CSAR-NRC HiQ および CSAR HiQ データセット[142]から 10 μM より強い結合親和性を持つ 337 のさまざまなタンパク質複合体が抽出された。

2 番目のデータセット（PL セット、Table 付録 2）は CYP3A4 様のポケットの特徴を学習するために構築され、CYP3A4 に類似したポケットを含むさまざまなタンパク質種の 64 の複合体が含まれる。このデータセットは、PoSSuM データベース[143]を使用して抽出された。PoSSuM データベースでは、ポケットの形状を様々な形の三角形として表し、その三角形の頂点と辺に疎水性、荷電性などの性質を割り当ててポケット形状を表現している。2014 年 9 月時点で、PDB[144]から取得した 5,513,691 の既知および推定の結合部位が登録されていた。PoSSuM データベースにて CYP3A4 の PDB ID を使用し「Search K」モードで既知の化合物に類似した結合部位を検索し、共有結合する化合物複合体を削除した。できるだけ多くの複合体を取得するために、「cosine similarity cut-off」値を最低値の 0.77 に設定した。ポケット体積

は、*MOE*[145]の Site Finder プログラムによって計算された。

3 番目のデータセット（LL セット、Table 付録 3） は CYP3A4 とその化合物間の結合様式の特徴を学習するために構築され、既知の CYP3A4 化合物を含む 28 の複合体が含まれる。このデータセットは、17 個の既知 CYP3A4 化合物をクエリーとして使用し、PDB データベースの「chemical ID」検索を介して構築された。

4 番目のデータセット（CF セット、Table 付録 4）は GP セットのサイズの約 30%で、Pfam[146]データベースにて Cytochrome P450 と判定されるタンパク質すべての「Uniprot ID」をクエリーとした PDB 検索によって取得された CYP を含む 116 の複合体が含まれていた。このデータセットは、CYP と化合物の相互作用に共通する特徴を学習するためにモデル用に構築した。

テストデータセット（CYP3A4 セット、Table 付録 5）には、ヘムグループの近くに化合物が結合した 22 個（2017 年当時）の CYP3A4 複合体が含まれる。複合体が 2 つの化合物を含み、両方がヘムの近くに結合した場合、一つ一つ再ドッキングして 2 つの異なる複合体として扱った（4K9T、4K9U）。

Table 3.1　トレーニングおよびテストデータセットの概要（Positive、negative の基準は後述）

		Shortened names	Number of complex structures	Number of positive poses	Number of negative poses
Training sets	General protein complexes	GP	337	2338	2807
	Complexes with CYP3A4-like pockets	PL	64	708	2842
	Complexes with CYP3A4-binding ligands	LL	28	157	1174
	CYP family complexes	CF	116	1086	4945
Test set	CYP3A4	-	22	22	693

すべてのドッキングポーズは *MOE*[145]のドッキングプログラムを使用して作成した。ドッキングサイトとして「ligand（化合物）」を指定し、化合物原子を除く他のすべての原子を「receptor」として指定した。Alpha Triangle アルゴリズム[58]を使用して化合物を結合部位に配置し、LondonΔG スコア関数に従って結合エネルギーをスコアリングした。ドッキングポーズを最小化するため、Amber10：EHT 力場を使用した。GP セットの各結晶構造に対しては、Ragoza らの先行研究[20]と同様に最大 20 のポーズを出力した。他の 4 つのデータセット（PL、LL、CF、CYP3A4） のデータ数は GP セットのデータ数よりも少なかったため、最大 70 のドッキングポーズを出力した。ドッキング後、不適切なプロトン化状態や、結合次数の正し

くない不適切な化合物構造を手動で修正した。なお、すべてのデータセットに「正解ポーズ（positive)」が含まれるよう、結晶構造の化合物立体配座をデータセットに追加した。その理由は、非常に柔軟な化合物が大きな結合ポケットに結合する場合、ドッキングによって正しいポーズを取得することが難しかったためである[147]。訓練セットは、結晶構造のポーズから 2Å 未満の RMSD を持つ姿勢をポジティブな （正解の）ポーズとしてタグ付けされ、4 Å より大きい RMSD を持つ姿勢はネガティブな （不正解の）ポーズとしてタグ付けされた（Figure 3.2.)。テストセットの CYP3A4 セットでは、結晶構造の化合物立体構造のみをポジティブとして使用した。各データセットのポジティブポーズとネガティブポーズの数を Table 3.1 に示す。

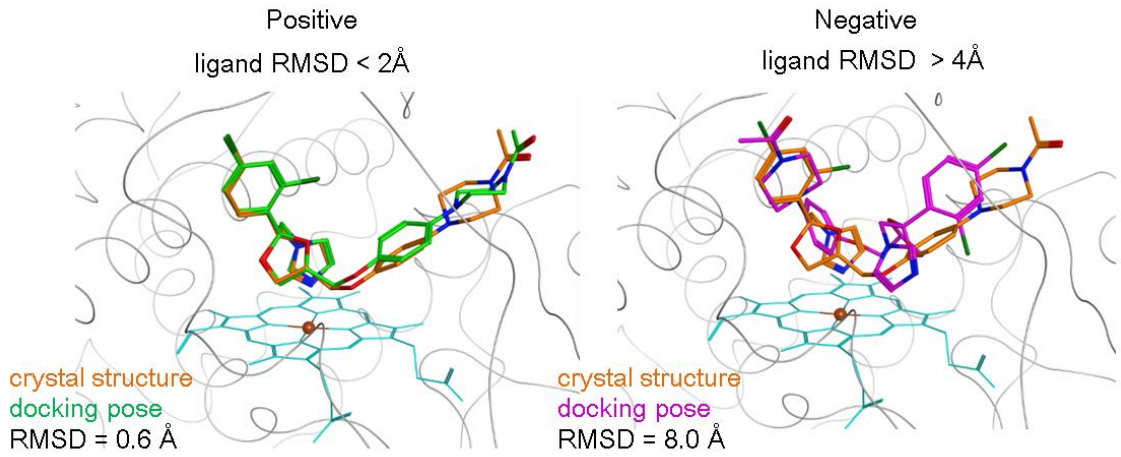


Figure 3.2.　ポジティブな （正解の）ポーズとネガティブな （不正解の）ポーズの例

Ragoza らの方法[20]に従って、タンパク質-化合物構造の原子タイプを各グリッドポイントに割り当て、CNN の入力として使用した。使用したグリッドは、結合部位を中心とした解像度 0.5 Å、24 $Å^3$ の立方体である。3 次元グリッドの各ポイントには、タンパク質原子（16 原子タイプ）および化合物原子（18 原子タイプ）の 34 の原子タイプ密度があてはめられた。原子タイプの定義（Table 3.2） と原子タイプ密度の計算式（Figure 3.2） は、Ragoza らが使用したものと同じものを使用している。

Table 3.2 原子タイプ一覧

Receptor	Ligand
AliphaticCarbonXSHydrophobe	AliphaticCarbonXSHydrophobe
AliphaticCarbonXSNonHydrophobe	AliphaticCarbonXSNonHydrophobe
AromaticCarbonXSHydrophobe	AromaticCarbonXSHydrophobe
AromaticCarbonXSNonHydrophobe	AromaticCarbonXSNonHydrophobe
	Bromine
Calcium	
	Chlorine
	Fluorine
	Iodine
Iron	
Magnesium	
Nitrogen	Nitrogen
NitrogenXSAcceptor	NitrogenXSAcceptor
NitrogenXSDonor	NitrogenXSDonor
NitrogenXSDonorAcceptor	NitrogenXSDonorAcceptor
	Oxygen
OxygenXSAcceptor	OxygenXSAcceptor
OxygenXSDonorAcceptor	OxygenXSDonorAcceptor
Phosphorus	Phosphorus
Sulfur	Sulfur
	SulfurAcceptor
Zinc	

$$A(d,r) = \begin{cases} e^{-2d^2/r^2} & 0 \le d < r \\ \dfrac{4}{e^2 r^2} d^2 - \dfrac{12}{e^2 r} + \dfrac{9}{e^2} & r \le d \le 1.5r \\ 0 & d \ge 1.5r \end{cases}$$

Figure 3.2 原子タイプ密度の計算式

各原子は、関数 A（d、r）として記述される。d は原子の中心からの距離、r はファンデルワールス半径である。A は、ガウス（原子の中心からファンデルワールス半径まで）と二次関数（半径の 1.5 倍でゼロ）の連続的な区分の組み合わせで表される。

3.3 モデル

CNN モデルは、GitHub（https://github.com/gnina/models）で公開されている最適化されたモデルのアーキテクチャー[20]を使用した。このモデルは ReLU を活性化関数とする 3 つの 3×3×3 の畳み込み層を持つ（Figure 3.3）。最初の畳み込み層では、34 個の特徴マップが生成される。各畳み込み層の後には、プーリング層（カーネルサイズ 2、最大プーリング）が続き、畳み込み層とプーリング層のそれぞれを通過した後、完全接続層に出力され、予測スコア値（CNN スコア）が 0 から 1 の合計 1 になるようスケーリングされる。

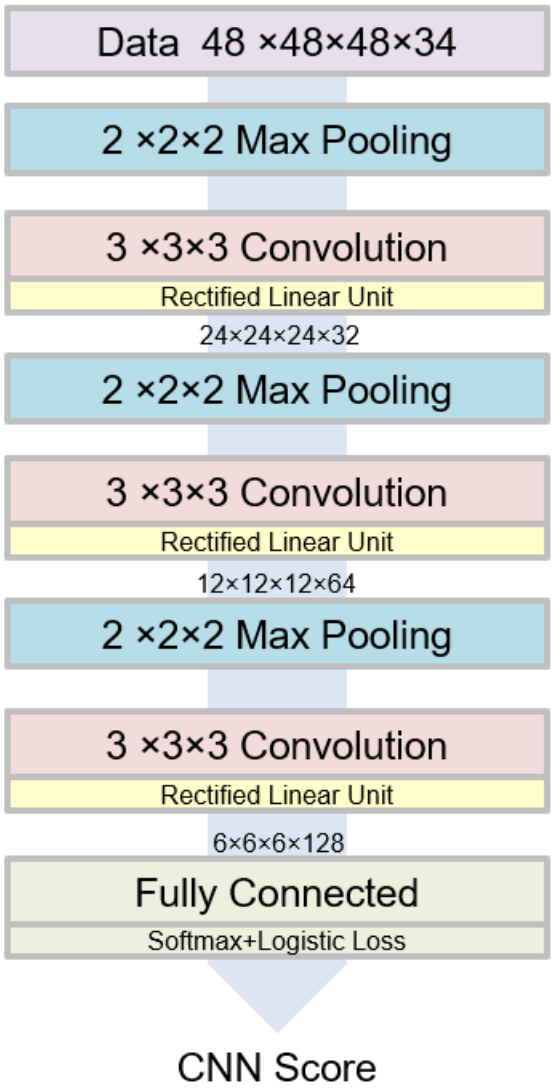


Figure 3.3　モデルの構造

3.4　　アルゴリズム

CNN モデルの構築には、Caffe ディープラーニングフレームワーク[148]と MolGridData レイヤー入力フォーマットが使用され[20]、 CNN モデルのすべてのネットワークは、多クラス

のロジスティック回帰を最小化するために確率的勾配降下法を使用してトレーニングされました。モデルのバッチサイズは 10 で、繰り返し回数は 10,000 回であった。Ragoza ら[20]の研究に従い、すべてのモデルに使用されたパラメーターを次に示す。学習率= 0.01、モーメンタム=0.9、学習係数の低減（inverse learning rate decay）: power = 1、gamma = 0.001、weight_decay = 0.001、dropout_ratio = 0.5。入力グリッドは、入力構造を 24 回ランダムに回転および並進させることにより拡張（オーギュメント） された。訓練されたモデルは、3 分割交差検証により評価された。

ファインチューニングとは、特異性の低い多くのデータセットで事前に訓練された後に、2 より特異性の高いデータセットを使用してモデルをさらに訓練する手法である。本研究において、GP セットまたは GP + CF セットの組み合わせによる事前学習により、タンパク質-化合物複合体の標的に依らない特徴を学習した。次に、CYP 固有のデータセット（PL、LL、または CF セット） のいずれか、またはそれらの組み合わせを使用して、ファインチューニングを実行した。事前学習の有効性を調べるために、 (1) 識別層（最終層） のみをファインチューニングし、他のすべての層のパラメーターを固定した場合、および（2） 全ての層をファインチューニングした場合、の 2 種類のファインチューニング方法を使用した。（**Figure 3.4**）

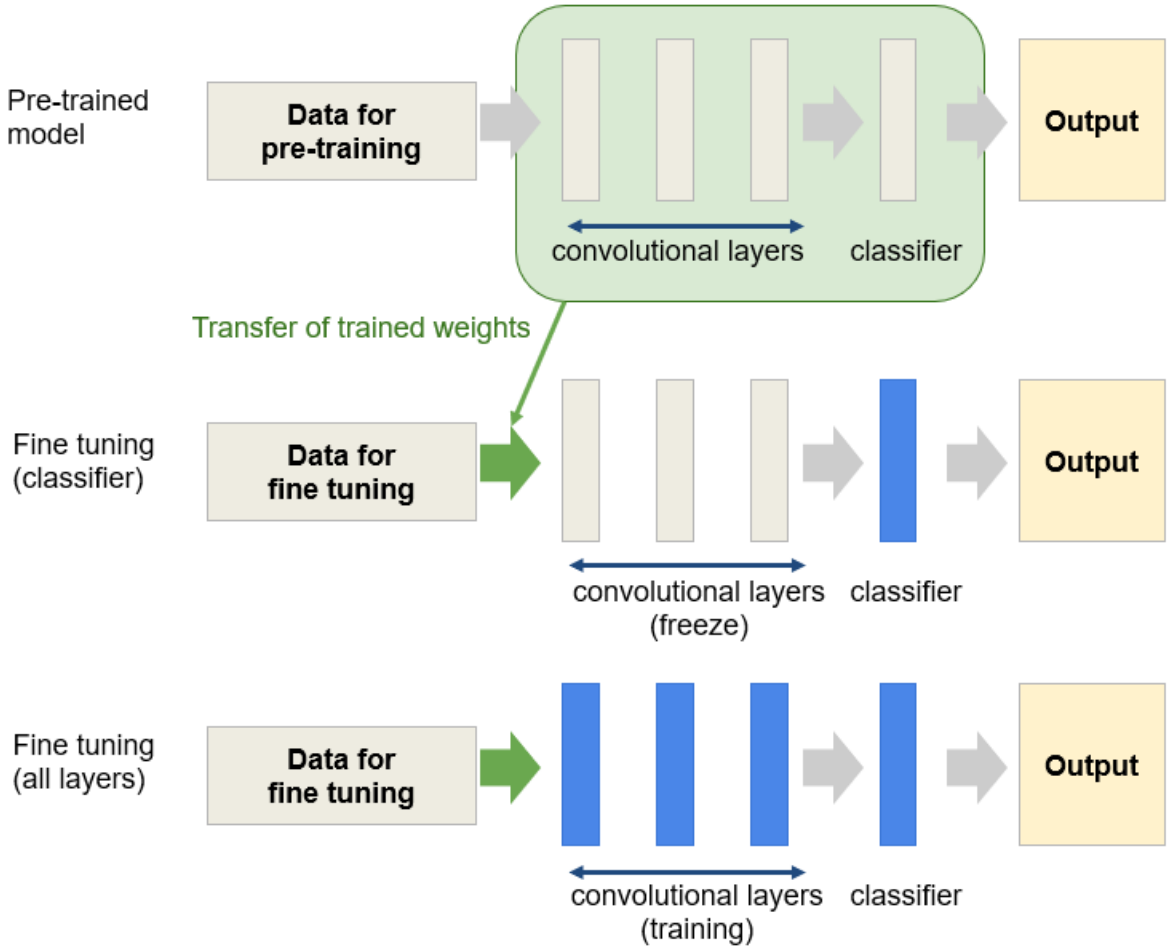


Figure 3.4 モデルの構築に使用される訓練スキームのイメージ

CYP3A4 セットは、訓練された CNN モデルの結合様式予測を評価するために設計された。CYP3A4 セットでの訓練、評価は、3 分割交差検証により行った。予測精度を向上させるために、CYP3A4 セット内の複合体を，同様の結合様式の Ritonavir アナログ複合体、ヘムの近くに化合物が結合した複合体、およびその他すべての複合体の 3 つのグループに分割した。CYP3A4 セットはすべて同じアミノ酸配列を持っているため、配列類似性によるデータセット分割は行わなかった（Figure 3.5）。

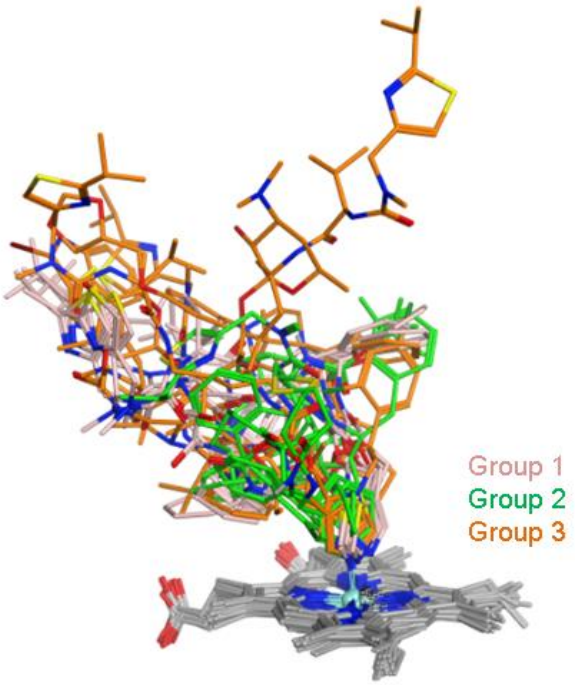


Figure 3.5　CYP3A4 セットの 3 つのグループへの分割

Group1（ピンク）：同様の結合様式の 7 つの Ritonavir アナログ複合体（PDB ID：3NXU、4I4G、4I4H、4K9V、4K9W、4K9X、5VC0）。Group 2（緑）：ヘムの近くに化合物が結合した 7 つの複合体（PDB ID：1W0G、3TJS、4D6Z、4D75、4D78、4D7D、5TE8）。Group 3（オレンジ）：Group 1 および 2 には適用されない 8 つの複合体（PDB ID：2J0D、2V0M、3UA1、4K9T（2 化合物）、4K9U（2 化合物）、5VCE）。

複数の観点で予測性能を評価するために、ROC 曲線下面積（ROC AUC）、recall、およびスコアランキングの指標を使用した。 ROC 曲線[149]は、偽陽性率に対して真陽性率をプロットすることにより、全体的な分類性能を示す。ROC 曲線の下の領域は性能測定基準であり、曲線下の領域（AUC）= 1 は完全な分類を表し、AUC <0.5 はランダム選択と同程度であることを示す。recall は、ポジティブ等と予測されたポーズ（本研究では CNN スコア> 0.5 と定義）のうち、実験的に正しいポーズの割合として定義された。数学的な定義は recall = TP /（TP + FP） であり、TP と FP は、それぞれ真陽性と偽陽性の数を示す。スコアランキングは、CNN スコアにより正解ポーズが濃縮されるか確認するために使用した。濃縮係数は、CNN スコアで順序付けられた場合、topX%に含まれる正しいポーズの CNN スコアの合計と平均によって示した。

3.5 結果と考察

この研究では、データセットの組み合わせ、ファインチューニングがモデルの予測性能に与える効果を AUC と recall を使用して比較し、正解ポーズの濃縮効果をスコアランキング（topX%および CNN スコア）で分析した。

Table 3.3 に、GP、PL、LL、CF セットまたはそれらの組み合わせで訓練したモデルの予測性能を示した。真値を取得するため、CYP3A4 セットを使用した。まず、*MOE*[145]のドッキングプログラムによるスコアリングの AUC は 0.579 であった。結晶構造をそのまま入力とした場合、分子力場に関して不安定な部分があり、これらのドッキングスコアが過度に低下した（AUC 0.127)。そのためドッキング時と同じ分子力場を使用し、構造最適化した結晶構造でスコアリングした。CYP3A4 セットの 3 分割交差検証の AUC は、わずか 0.540 であった。これは、データセットが小さすぎて適切なニューラルネットワークの機械学習ができなかったためと考えられる。GP、PL、LL、CF セット単独で訓練したモデルの AUC は、それぞれ 0.610、0.470、0.745、0.647 であった。LL セットでトレーニングされたモデルは最も高い AUC 値を示したが、その recall 値（0.273）は低かった。これらの結果は、各データセットで個別に訓練した場合、CYP3A4 と化合物間の結合様式の特性を十分に補足していないことを示唆している。データセットをさまざまに組み合わせた場合の予測性能への影響を調べた結果、組み合わせたデータセットは、個々のデータセットを使用した場合に得られた値と比較して改善された AUC 値を示す傾向が見られた。最も高い AUC 値（0.831） は、GP セットと CF セットの組み合わせの場合に得られたが、再現率（0.273） は低かった。これは、補完的なデータセット（すなわち、標的に依らないタンパク質データセット（GP）および CYP ファミリーデータセット（CF)）を組み合わせた場合に、高い AUC が期待できることを示唆した。このことから、その後の分析で GP + CF セットを組み合わせて事前学習に使用した。

Table 3.3　AUC と Recall によるデータセットの組み合わせのモデル性能の比較（Recall の評価では、0.5 以上の CNN スコアをポジティブとして設定した）

Dataset	AUC[a)]	Recall
CYP3A4(three-fold cross validation)	0.540	-
GP	0.610	0.136
PL	0.470	0.364
LL	0.745	0.273
CF	0.647	0.182
GP + PL	0.635	0.045
GP + LL	0.617	0.364
GP + CF	0.831	0.273
PL + LL + CF	0.713	0.045
GP + PL + LL + CF	0.781	0.227
docking score (MOE dock)	0.127[b)]	
	0.579[c)]	-

a) AUC, area under the curve.
b) Scored the crystal structure (positive) as it is.
c) Crystal structures (positive) were optimized in the same force field as other docking poses and scored.

これまでの結果から、CNN スコアと従来のエネルギーベースのスコアの予測パフォーマンスを比較する。データセットが小さい場合、CNN スコアはエネルギーベースのスコアよりも劣る（例: CYP3A4 の 3 分割交差検証、AUC 値 0.540）。一方、本検討により CYP3A4 のような予測の難しいターゲットでも、適切なデータセットを組み合わせることにより予測精度が向上することが確認された。この点は、エネルギーベースの均一なスコアリングとは異なる。また、エネルギーベースのスコアリングで性能を評価する際に結晶構造をそのまま使用すると、力場の違いの問題がスコア値に顕著に現れる。このことから、エネルギーベースのスコアリングと比較し、CNN スコアは力場の違いによる軽度の不安定性に対して堅牢であることが分かる。

ファインチューニングは、事前学習用のサイズの大きなデータセットと、小さなバイアスのかかったデータセットを使用して訓練する強力な手段である[138]–[141]。Table 3.4 は、組み合わせにより最も高い AUC 値を示した GP+CF セットで事前学習し、2 つの CYP 固有のデー

タセット（PL および LL データセット）を使用しファインチューニングして得られた予測性能を示す。2 つのファインチューニングアプローチのうち、識別層（classifier）のみをファインチューニングしたモデルは、AUC と recall の両方に関して、検討したすべてのデータセット（PL、LL、または PL＋LL セット）において、全層でファインチューニングしたモデルよりも優れた予測性能を示した（例えば、LL セットの AUC 0.769 対 0.829 および recall 0.270 対 0.500）。最高の予測性能が見られたモデルは、GP +CF セットを使用した事前学習に続いて、LL セットを使用して識別層のみファインチューニングしたモデル（pre(GP+CF)/ft-cl(LL)）であった。

Table 3.4　データセット組み合わせと AUC に基づく微調整のテストセットパフォーマンスの比較

Pre-training	Dataset	Fine tuning			
		All layers		Classifier	
		AUC	Recall	AUC	Recall
GP+CF	PL	0.677	0.000	0.765	0.182
Area under ROC curve: 0.831	LL	0.769	0.270	0.829	0.500
Recall: 0.273	PL + LL	0.792	0.091	0.808	0.273

AUC, area under the curve; ROC, receiver operating characteric.

GP＋CF セットのみでトレーニングされたモデルと pre(GP+CF)/ft-cl(LL)でトレーニングされたモデルの違いを調査するために、正解ポーズの濃縮と、スコア値分布の 2 種の指標を追加して多面的に評価した。正解ポーズの濃縮は、スコアの良い順に topX%内の正しいポーズの割合で示す。スコア値の分布は、各モデルが正解ポーズをどの程度ポジティブと判定しているかを示す。正解ポーズの濃縮においては、GP＋CF モデルと pre（GP＋CF）/ ft-cl（LL）モデルは同程度の性能を示した（Figure 3.6、Y 軸）。一方で、CNN スコアの比較では、ファインチューニングしたモデル（pre(GP + CF)/ ft-cl(LL)モデル） はファインチューニングしなかったモデル（GP + CF） のスコアの約 2 倍のスコアを示した。この結果から、事前学習に続いてファインチューニングを行うと、正解ポーズの CNN スコアが大幅に向上し、実験的に正しいポーズを高いスコアで検出できることが示された。これらの分析から、pre（GP + CF）/ ft-cl（LL）モデルが最も実用的なモデルであると考えられる。

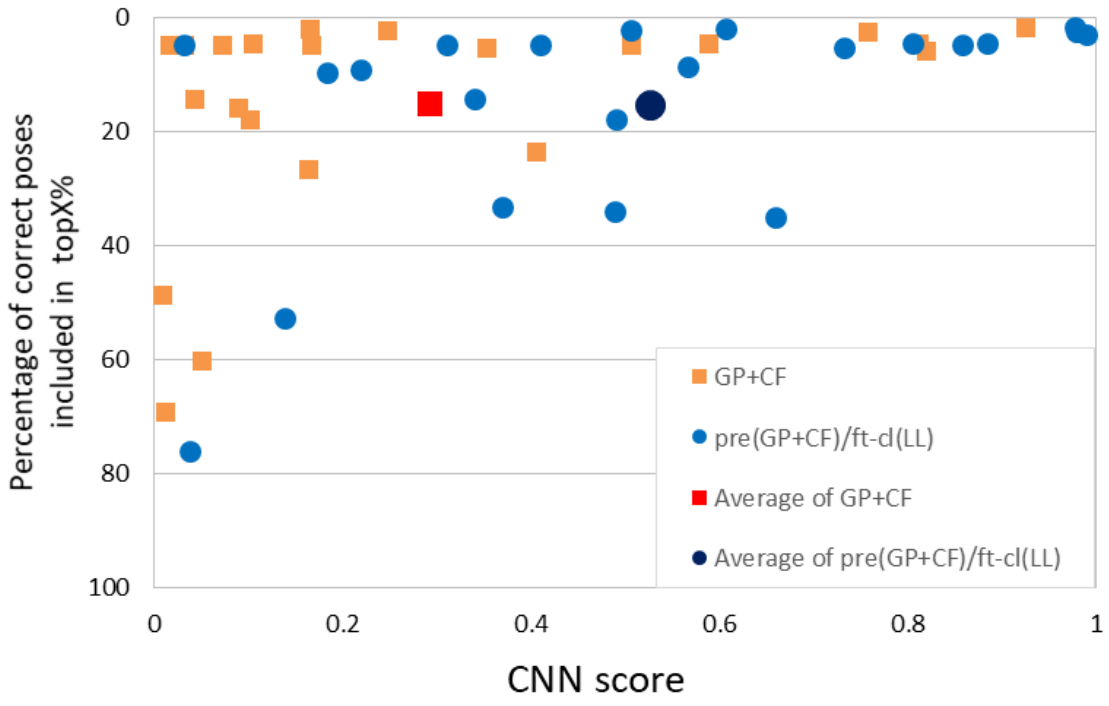


Figure 3.6　スコアの良い順に topX%以内に含まれる正しいポーズの割合と CNN スコアの散布図

TopX%内の正しいポーズの平均割合：GP＋CF モデル、15.1%。 pre（GP＋CF）/ ft-cl（LL）モデル、15.2%。 平均 CNN スコア：GP＋CF モデル、0.293; pre（GP＋CF）/ ft-cl（LL）モデル、0.526。

全体として、適切な一般データセット(GP+CF セット)が事前学習に使用され、ターゲット固有のデータセットがファインチューニングに使用された場合、ファインチューニングにより結合様式の予測性能が向上することが本検討の結果から示された。GP+CF セットは事前学習において AUC と recall を大幅に改善し、LL セットは正解ポーズの予測スコア値を改善した。これらの検討結果を考察するため、タンパク質の分類とポケットサイズに関するデータセットの詳細な分析を実施した。

GP+CF セットの組み合わせがトレーニング前の GP セットよりも優れている理由を考察するため、各データセットの組成を調査した。Protein Data Bank に登録されているタンパク質構造は機能分類と紐づけられている[150]。CYP3A4 はオキシドリダクターゼに分類されていることから、各データセットにオキシドリダクターゼが含まれる割合を Figure 3.7 に示した。CF セットには主に CYP が含まれていたため、そのデータセットのタンパク質の 89%はオキシドリダクターゼであった。対照的に、GP セットではたった 4%のみがオキシドリダクターゼであった。オキシドリダクターゼの割合は、GP セットと CF セットを組み合わせると 25%に増加し、この割合は GP + LL セット（4%）および GP + PL セット（7%）よりもはるかに高かった。このことから、タンパク質の機能とデータサイズのバランスが組み合わせ上重要

であったと考えられる。

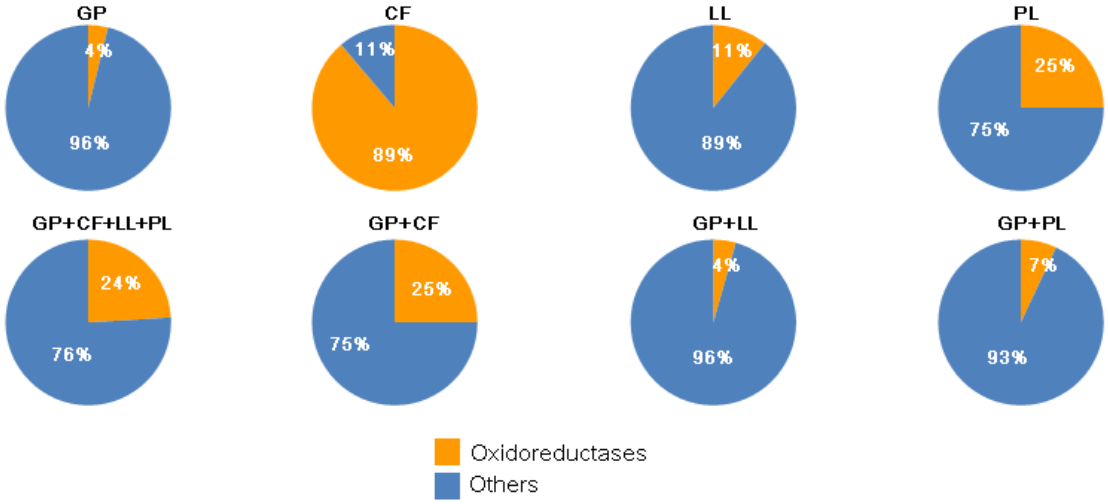


Figure 3.7　各データセットにオキシドリダクターゼが含まれる割合

Figure 3.8 に、CF セットを GP セットに追加した場合の、結晶構造の 3 つの化合物ポーズ（5VCE、4D75、4K9T）の CYP3A4 ポケットの CNN スコアの変化を視覚化した。本モデルの可視化方法として、Hochuli らにより masking、gradient、conserved layer-wise relevance propagation (CLRP)の 3 つの方法が開発されている[137]。Masking は、元の入力と元の入力から一部の原子やフラグメントがマスクされた状態間の予測出力スコアの差を計算することにより、その原子またはフラグメントがスコアに与える寄与度を評価する方法である。Gradient は、ネットワーク入力上の勾配を計算し、勾配を各原子上の 3 次元ベクトルとして視覚化する方法であり、ネットワークが特定の入力をより良くするために何を必要としているかを判断するのに役立つ。また CLRP は分類確率などのニューラルネットワークの出力を元の入力に戻すことにより、入力内の空きスペースの影響を分析する方法である。本研究では、このうちもっとも直感的で、解釈の比較的容易な masking を使用して結果の可視化を実施した。Masking は、元の入力と元の入力から一部の原子やフラグメントがマスクされた状態間の予測出力スコアの差を計算することにより、その原子またはフラグメントがスコアに与える寄与度を評価する方法である。使用したプログラムは、gnina プロジェクト（http://github.com/gnina）の中で、オープンソースライセンス下で「gninavis」として利用可能となっている。この図から、CF セットが GP セットに追加されると、ヘムグループの近くの原子が CNN スコアにより多く寄与することが示された。

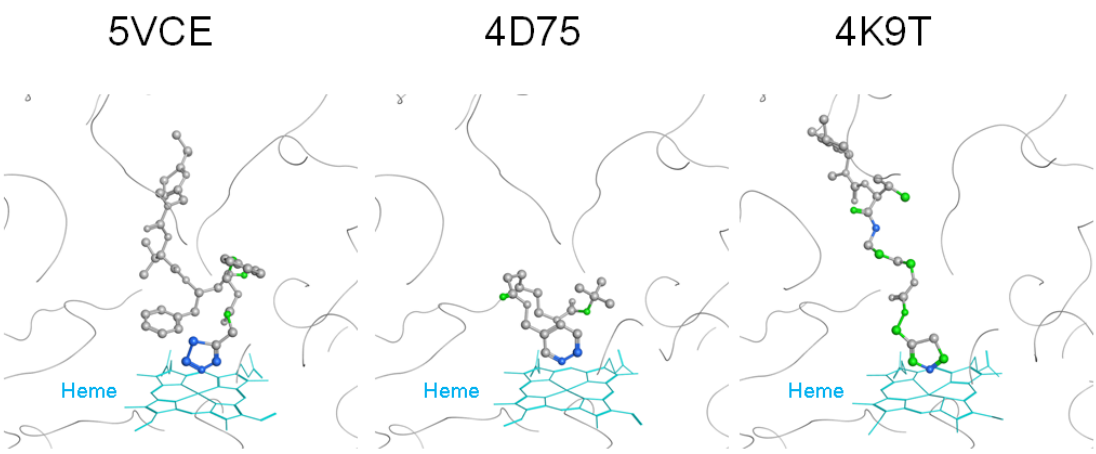


Figure 3.8 CF セットが GP セットに追加されたときの CNN スコアの変化の視覚化
両モデルでスコアの差が大きかった 3 つの結晶構造における化合物ポーズを示す。青色 ： 大幅な増加、 緑色 ： わずかな増加、灰色:変更なし。

次に、ファインチューニングのために LL セットが PL セットよりも優れている理由を調べた。ここで、CYP3A4 類似ポケットとの複合体を含むデータセット（PL セット、25%）のオキシドリダクターゼの割合は、CYP3A4 結合化合物との複合体を含むデータセット（LL セット、11%）よりも高いことに注意する必要がある。これは、PL セットでファインチューニングした後、結合様式予測の改善が見られなかったことについて、タンパク質の機能分類以外に要因があることを示唆する。CYP3A4 セットの重要な特徴は、そのポケット体積（約 1500Å^3）が GP、PL、および CF セットのほぼ 2 倍である点である（Figure 3.9）。PL セットは CYP3A4 に似たポケット形状の複合体を含むように設計されたが、ポケット体積の範囲は GP および CF タセットの範囲と非常に似ていた。対照的に、LL セットのポケット体積ははるかに大きく、CYP3A4 のポケット体積の範囲を完全にカバーしていた。また、LL セットによるファインチューニングの効果を確認するため、GP + CF モデルと pre（GP + CF）/ ft-cl（LL）モデル間の CNN スコアの変化を可視化した（Figure 3.10）。その結果、CF セットとは対照的に、ヘムから離れた箇所の結合様式を学習していることが示された。さらに、LL セットでの学習で最もスコアの向上が観られたエリスロマイシンを取り上げ、訓練セットの結合サイトを観察した（Figure 3.11）。その結果、芳香族アミノ酸、脂肪族アミノ酸など化合物周辺の環境が CYP3A4 と LL セット間で類似していることが分かった。このことから、LL セットでは、狙い通りリガンドとタンパク質間の位置関係を学習できたのではないかと考えている。以上より、高度に相補的な GP セットと CF セットでの事前学習と、それに続く LL セットでのファインチューニングにより、CYP3A4 のタンパク質と化合物の一般的な相互作用および CYP3A4 特有の相互作用の両方の特徴が捕捉されたことが示唆された。

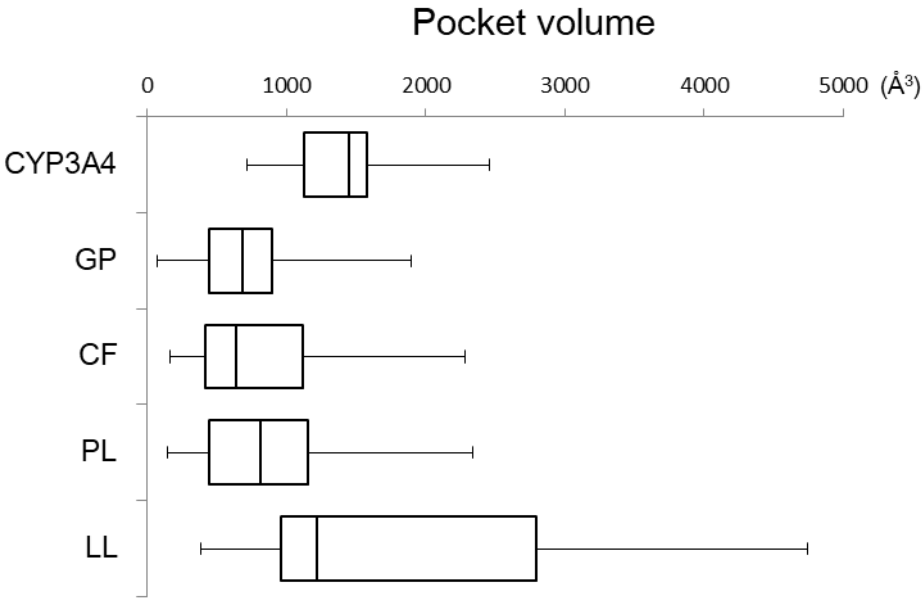


Figure 3.9 トレーニングデータセットとテストデータセットのタンパク質のポケットサイズの比較

ボックスプロットの線は、左から最小、下四分位、中央値、上四分位、および最大を表す。

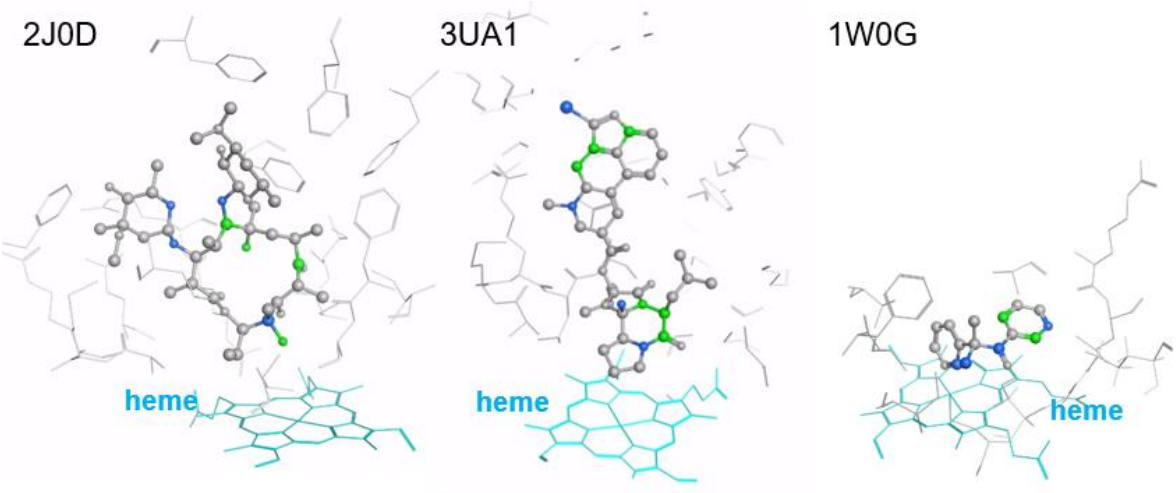


Figure 3.10 GP+CF セットで事前学習し、LL セットでファインチューニングしたモデルと GP+CF セットで訓練したモデルとの CNN スコアの変化の視覚化

Figure 3.8 と同様、両モデルでスコアの差が大きかった 3 つの結晶構造における化合物ポーズを示す。青色 ： 大幅な増加、 緑色 ： わずかな増加、灰色:変更なし。

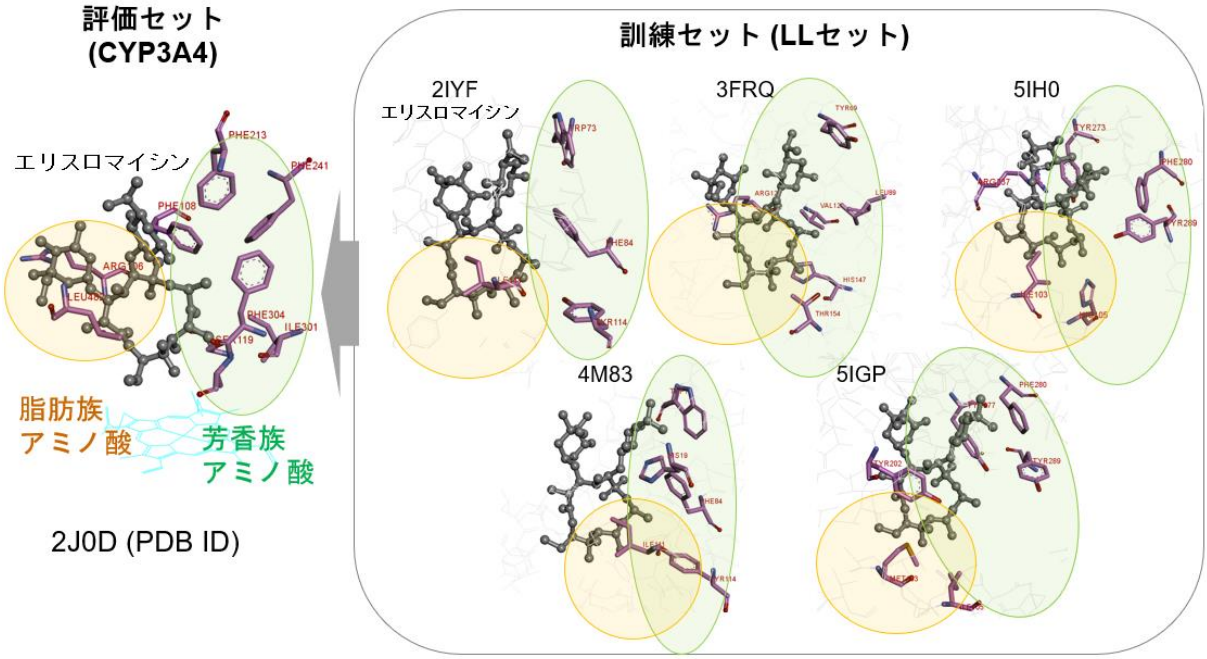


Figure 3.11　エリスロマイシン複合体の評価セットと訓練セットの化合物周辺環境の比較

3.6　結言

第 3 章では、CYP3A4 を題材として、標的タンパク質に関連するタンパク質-化合物複合体の結晶構造の数が限られている場合に、サイズと特異性の異なるデータセットを組み合わせてファインチューニングを行うと、CNN モデルの結合様式予測にどのように影響するかを調べた。データセットの組み合わせの効果の調査では、さまざまなタンパク質複合体を含む一般的なデータセット（GP セット）と CYP タンパク質ファミリーとの複合体を含むターゲット固有のデータセット（CF セット）の組み合わせでトレーニングされたモデルは最も高い ROC AUC 値を示した。このことは、予測機能を最適化するためにタンパク質の機能とデータサイズに関してバランスの取れたデータセットが重要であることを示唆している。GP＋CF セットでの事前トレーニングと、それに続く CYP3A4 結合化合物（LL セット）を含むデータセットでのファインチューニングにより、CNN スコアのランキングと ROC AUC の両方で大幅に予測精度が向上した。これは、CYP3A4 と同程度に結合ポケットサイズの大きな結晶構造データセットが CYP3A4 の結合様式の予測に重要であったことを示唆している。

今回の結果が他の標的タンパク質に適用できることを確認するには、対象化合物に類似した LL セットのみを選び直して CNN モデルを作成するなどさらなる検討が必要である。また、同じタンパク質ファミリーで構成されたデータセットはタンパク質側の情報にあまり変化が

ないため化合物側の性質に大きく影響を受ける可能性もある。これに対し、化合物とタンパク質の空間的配置に本質的に影響し、両者を紐づけする相互作用に関する記述子の開発などの改善が必要である。効果的な薬物設計のためには、正確なドッキングポーズ予測モデルが必要である。適切な訓練セットの設計戦略の開発により、ニューラルネットワークはこのニーズに対処する有望なアプローチとなり得る。

第4章　MD シミュレーションによるタンパク質の動きを考慮した結合様式予測

4.1　緒言

CYP3A4 のようにポケットサイズが大きい標的タンパク質については、タンパク質の構造揺らぎを考慮した化合物の結合様式予測が必要である。従来の Partially flexible ドッキングでは側鎖構造の動きの考慮は検討されているが、タンパク質主鎖構造の動きを含めたドッキングポーズの生成はできない。また、長時間のシミュレーションのみでは結合様式サンプリングの網羅性に問題がある。そこで本研究では多数のドッキングポーズから比較的短時間の MD シミュレーションを利用した化合物の結合様式予測に取り組んだ。

シトクロム P450 （CYP）は薬物および内因性化合物の代謝に関与する一群の酵素である。CYP は、小腸粘膜、肺、腎臓、脳、嗅覚粘膜、および皮膚において高発現しているが、主に薬物代謝の主な器官である肝臓に集中している[151]。 多くの CYP のうち、CYP3A4 は薬物代謝にとって最も重要なものの 1 つであり、臨床的に使用される薬物の 30%を超える代謝に寄与している[152]。 経口薬の初回通過代謝において重要な役割を果たすことに加えて、CYP3A4 を含む CYP は薬物間相互作用（例えば CYP 阻害）にも関与しており、これらの相互作用の悪影響は薬物開発にとって重要な障害となる。

多くの薬物候補の開発は、この薬物動態学的問題のために初期段階で中止または遅延されており、CYP による代謝に関して薬物候補の安定性を改善することは、創薬における最優先課題の 1 つである。創薬化学者は、CYP による薬物の代謝、排泄を減らすことを目的として、リード化合物の誘導体を合成する。 例えば環サイズの縮小、水酸基の付加、代謝部位へのかさ高い置換基やハロゲン原子の導入、代謝部位の環化、代謝不安定な部位の除去など様々なアプローチが試みられる[153]。置換基の導入を含むこれらのアプローチの多く[154], [155]は、代謝部位、結合様式の正確な予測が必要となる。

CYP によって基質の代謝部位を予測する際には、2 つの要素を考慮する必要がある[156]。1 つ目は、化合物の各原子のヘム鉄への接近可能性、2 つ目は基質内の各原子の酸化反応性である。 後者の予測方法では、化合物構造のみが考慮されるため、薬物設計に必要なタンパク質と薬物との複合体構造が得られない[157]–[160]。前者の各原子のヘム鉄への接近可能性を考慮しながら代謝部位を予測する方法は、代謝的部位を予測するのみならず、CYP-薬物複合体の結合様式についての情報も提供することができる。タンパク質を剛体として扱うドッキ

ングシミュレーションの他[161]、側鎖と化合物の柔軟性を考慮した代謝予測法も提案されているが[162], [163]、これらの代謝部位・結合様式予測方法ではタンパク質主鎖の柔軟性の影響は考慮されていない。2017 年当時、25 の CYP3A4 の X 線構造が Protein Data Bank に登録されていた。 これらの構造を重ね合わせると（Figure 4.1）、オレンジ、緑で示した F-G ループは結合する化合物に合わせ側鎖だけでなく主鎖構造を含めて動いている。このことから、CYP3A4 に対する代謝部位、結合様式予測はタンパク質主鎖も含めた柔軟性の影響を考慮する必要がある。

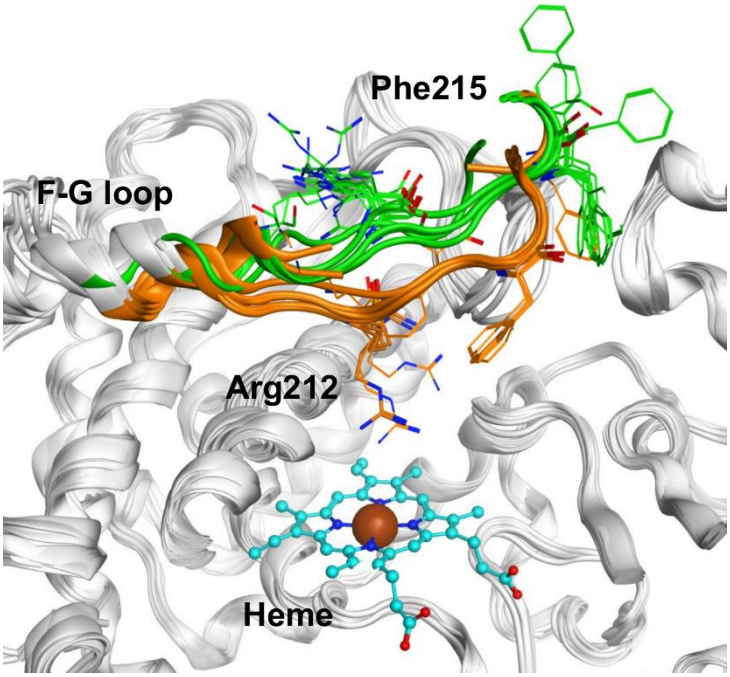


Figure 4.1　CYP3A4 の結晶構造の重ね合わせ

本研究に最も近い先行事例として、カルバマゼピン（CBZ）の CYP3A4 による代謝部位をドッキングと MD シミュレーションを用いて予測する最初の手法が報告されている[164]。彼らはドッキングシミュレーションだけでは CYP3A4 の分子の柔軟性を正確に説明するのに十分ではなく、MD シミュレーションが必要であると結論づけている。さらに、複合体の複数の初期ポーズから始まる MD シミュレーションが、可能な結合様式を効率的に探索するために有効であることを報告した。彼らは「固い」カルバマゼピンの代謝部位を首尾よく予測したが、一般に結合様式の予測は、固い化合物よりも柔軟な化合物のほうが困難であり、本手法の堅牢性を検証するために柔軟な化合物への適用が必要である。カルバマゼピンは回転可能な結合を 1 つだけ含む固い構造をしており、立体配座の違いは各原子のヘムへの近接性にほとんど影響を与えない。本研究は、手法の堅牢性検証と改良を目的として、より柔軟な分子トルテロジンを選択した[165]。 CYP3A4 は、窒素上のアルキル基の 1 つを酸化することによってトルテロジンを *N*-脱アルキル化トルテロジンに代謝することが実験で明らかとなっている[166]（Figure 4.2）。

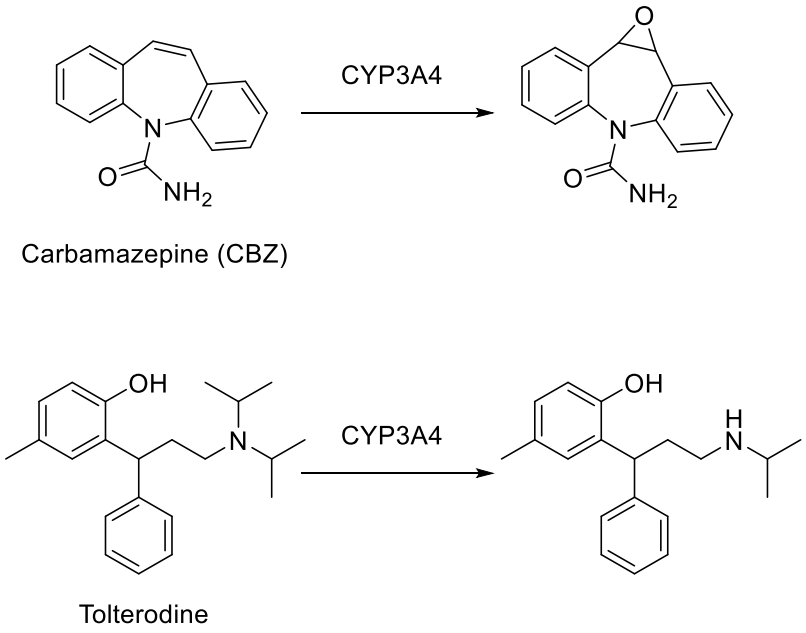


Figure 4.2　ヒト肝ミクロソームにおけるカルバマゼピン、トルテロジンの代謝生成物

最近の計算能力の向上により、大規模な MD シミュレーションが可能となっているものの、MD 初期ポーズの効果的な選択は、薬物設計のような実際的な用途に向け計算時間を最小にするため依然として重要となっている。また、先行事例[17]、予備検討（付録.A）において長時間 MD では配座サンプリングが十分でないことも確認されている。そこで本研究では MD の初期ポーズを抽出して選択するために、ドッキングシミュレーションによって生成された複数の CYP3A4 構造とトルテロジンのドッキングポーズのクラスター化方法を検討した。従来、RMSD に基づくクラスタリング方法は、類似の立体配座特徴を共有する構造のグループに分割するために使用されてきた。近年、相互作用フィンガープリント（SIFt）[167]またはタンパク質-化合物相互作用フィンガープリント（PLIF）がドッキングポーズや MD のトラジェクトリのクラスター化に使用されつつある。この方法は、複合体の構造におけるタンパク質-化合物相互作用のパターン間の類似性を明らかにすることができる。 カルバマゼピンの代謝部位予測の場合は、RMSD クラスタリングに基づいて選択された初期ポーズを含む 5 つの MD シミュレーションを実施し、ヘム鉄へのカルバマゼピン原子の接近可能性を評価した。本研究では、RMSD に加え PLIF を用いたクラスタリングによる初期ポーズ選択を比較した。また、カルバマゼピンよりも柔軟な化合物であるトルテロジンの代謝部位、結合様式を正確に予測するために必要十分な初期ポーズの数も調べた。ワークフローを Figure 4.3 に示す。

Figure 4.3　多数の MD 初期ポーズを用いた結合様式評価フロー

4.2　研究方法

CYP3A4 の分子動力学シミュレーションをするにあたって、CYP3A4 タンパク質構造に関して、下記の準備を行った。CYP3A4 のアポ体（PDB ID: 1TQN） の結晶構造において、Arg212 の側鎖はトルテロジンのポケットへの結合を阻害すると考えられた。そこで、1TQN（PDB ID）と化合物複合体 1W0G（PDB ID） を重ね合わせ、*MOE* [145]を使用して、1TQN 構造の Arg212 座標を 1W0G 構造のものに置き換えた（Figure 4.4）。 ドッキングシミュレーションの前に、結晶構造中の水分子は除去した。

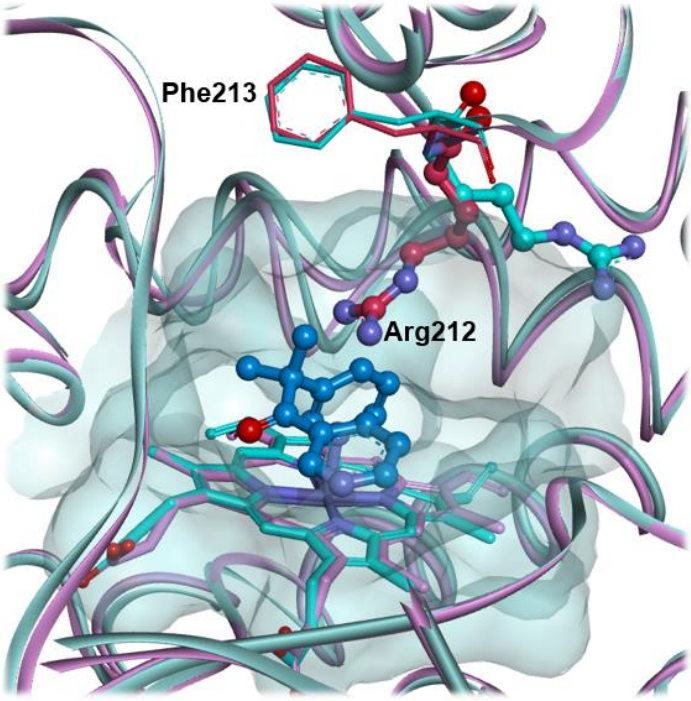


Figure 4.4　1TQN（マゼンタ） と 1W0G（シアン） の間の Arg212 の位置の比較
(1TQN の Arg212 はそのヘムポケットの中心部を占める)

次に、MD シミュレーション用の CYP3A4-トルテロジン複合体の初期ポーズ候補を生成するため *MOE* のドッキングプログラムを利用した[145]。ドッキングサイトとして、(a) SiteFinder[168] によって検出されたダミー原子、(b) Phe213、および (c) ヘム中の鉄原子、の 3 つを指定した。他のすべての設定にはデフォルト値を使用した。ドッキング部位を特定するために Phe213 を選択した主な理由は、以前の研究で行われた MD シミュレーションの間、CBZ がこの残基と安定的に相互作用したためである[164]。Alpha Triangle アルゴリズムを用いてトルテロジンを基質結合ポケットに置き、結合エネルギーを London のΔG 関数に従って最大 100 個の異なるドッキングポーズを生成させ[58]、構造を最小にするために Amber 12 : EHT 力場を用いた。ドッキングポーズの足切りのため、スコア（GBVI / WSA ΔG）が -5 kcal / mol 以下で、ヘム鉄といずれかのトルテロジン炭素原子との最短距離が 7 Å 以下である CYP3A4 - トルテロジンドッキングポーズを選択した。

ドッキングポーズは 2 つの方法でクラスタリングした。まず、*MOE* を用いて PLIF 分析を行い、Arg105、Ala370、Arg372、Glu374 とトルテロジン間の相互作用から 4 つの相互作用を抽出した。これらの相互作用を用いて、k-means アルゴリズム[169]を用いてクラスター分析を行った。次に *MOE* を使用し、化合物の原子座標の RMSD に基づいてトルテロジンのドッキングポーズのクラスタリングを行い、各クラスターにおけるドッキングスコア（GBVI / WSA ΔG）の最も良い（値の小さい）ポーズを代表的構造として選択した。このプロセスによって選択したドッキングポーズを、その後の MD シミュレーションの初期ポーズとして使用した。

MD シミュレーションのプロトコルに関しては、先行事例[170]に倣い、ヘム基およびトルテロジン分子の電荷は *Gaussian03* ソフトウェア[171]を用いて B3LYP / 6-31G ** レベルで実施した量子力学計算によって導き、RESP 法[172]によって各原子上に当てはめた。CYP3A4-トルテロジン複合体構造の各モデルは、タンパク質からの最小距離が 18 Å となるよう直方体の TIP3P water box に挿入した[173]。カウンターイオン（Na^+、Cl^-）を加えて系を中和した。MD シミュレーションは ff03 力場[174]を用い、Amber 12 ソフトウェアパッケージの PMEMD モジュールを使用して実行した[175]。この系の水素原子は、最急降下法および共役勾配法により、それぞれ 5000 ステップにてエネルギー最小し構造最適化した。側鎖および溶媒の最適化について上記と同じ方法でそれぞれ 5,000 ステップ、タンパク質全体と溶媒の最適化について同じ方法でそれぞれ 30,000 ステップ、化合物を含む系全体について同じ方法でそれぞれ 30,000 ステップにて構造最適化した。この系を NVT アンサンブル条件下で 130 ps、300 K に加温し、NPT 条件下でトルテロジンに 10 kcal mol^{-1} $Å^{-2}$ の位置拘束をかけながら 1.1 ns 平衡化した。平衡化後、拘束を徐々に解放するため 100 ps で 5 kcal mol^{-1} $Å^{-2}$、100 ps で 2 kcal mol^{-1} $Å^{-2}$ の化合物位置拘束下で追加のシミュレーションを実行した。次に、プロダクトランとして、拘束条件のないシミュレーションを 10 ns 実行した。 10 ns のシミュレーション中、スナップショットを 1 ps ごとに保存し、分析に使用した。シミュレーションを通し、水素原子を含む結合を拘束するために SHAKE アルゴリズム[176]を採用した。タイムステップは 1 fs に設定した。非結合項のカットオフ距離は 10Å に設定した。

MD シミュレーションの解析においては、Amber ソフトウェアパッケージの ptraj モジュールは、RMSD の計算や距離の測定などの MD トラジェクトリの解析に使用された。 溶媒効果を考慮しながら各結合様式の安定性を比較するために、Amber の MM/PBSA 法を使用して各モードの結合自由エネルギー（$\Delta G_{binding}$） を計算した。 一般に、CYP3A4 とトルテロジンの間の結合に対する$\Delta G_{binding}$ は、次の式で与えられる。

$$\Delta G_{binding} = \Delta G_{MM} + \Delta G_{sol} - T\Delta S$$

ここで、ΔG_{MM} は分子力学項、ΔG_{sol} は溶媒和の項、そして $T\Delta S$ はエントロピーの項である。ΔG_{MM} およびΔG_{sol} は、Amber で実施された MM/PBSA 法によって計算した。トルテロジン-CYP3A4 複合体の MD シミュレーションでは、化合物分子が同一であるため、$T\Delta S$ の計算を考慮していない。 結合自由エネルギーの計算は、10 ns のプロダクトランから 1 ps 間隔で取得した 10,000 のスナップショットについて実施した。

4.3 MD 初期ポーズの作成

MD シミュレーションの結果は初期ポーズに大きく影響される。本研究では、以下の観点から CYP3A4 の初期ポーズを選択した。

シトクロム P450 による酸化のメカニズムおよび触媒サイクルについては多くの研究がな

されている[177]。 基質がヘムに接近するときのヘム鉄の酸化状態は重要である。 基質結合後の酸化状態の実験的決定は困難であるため、鉄が基質結合後に酸化されるか、または基質が存在しない間に酸化されるかは、決定的されていない。 この研究におけるシミュレーションを単純化するために、本研究では基質が接近している間にヘム鉄が酸化されていないと仮定した。2017 年の時点で 25 の CYP3A4 の X 線構造が報告されているが、本研究では 2013 年時点で入手可能であった結晶構造（PDB ID : 1TQN[178]、1W0E、1W0F [179]、2J0D、2V0M [168]、3NXU[181]、3UA1[182]、3TJS [33]、4I4H、4I4G、4I3Q [34]、4K9X、4K9W、4K9V、4K9U、4K9T [35]）から鋳型構造を選択することとした。1TQN（PDB ID）は、これらの中で最も良い分解能（2.05 Å）の結晶構造である。構造の欠損はなかったが、アポタンパク質の構造であり、Arg212 の側鎖はヘムポケットの内側に突き出ている。 この立体配座では、Arg212 がトルテロジンのポケットへの結合を阻害すると考えられたため、当時利用可能であった構造の中で唯一化合物との複合体を形成していた 1W0G の構造に 1TQN の構造を重ね合わせ、Arg212 側鎖を置き換えた（Figure 4.4）。

化合物ドッキングの前に、先行研究[164]と同様に結晶構造中の水分子を除去した。*MOE* のドッキングプログラムを利用し[145]、236 個の MD シミュレーション用の CYP3A4-トルテロジン複合体の初期ポーズ候補を生成した。さらにドッキングスコア（GBVI/WSA ΔG）が-5 kcal/mol 未満で、ヘム鉄と少なくとも 1 つのトルテロジン炭素原子の間の最短距離が 7 Å より小さい 154 個のドッキングポーズを選択した。154 個のトルテロジンドッキングポーズについて、化合物の原子座標の RMSD と PLIF に基づいてクラスタリングを行い、カットオフ値を調整して 23 個のクラスターを生成した。いくつかのクラスターは複数のドッキングポーズドッキングパターンを含んでいたので、我々はさらに目視検査によってクラスターを分割した。この手順により、各クラスターで最も低いドッキングスコアを持つ 27 個の代表的な構造を選択した（Figure 4.5）。これらの 27 の構造は、MD シミュレーションの初期ポーズとして使用された（Figure 4.6）。

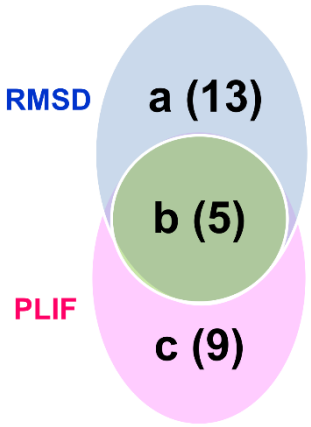


Figure 4.5　RMSD および PLIF クラスタリングによって生成された 27 の初期ポーズのベン図
（括弧内の数字は複合体構造数）

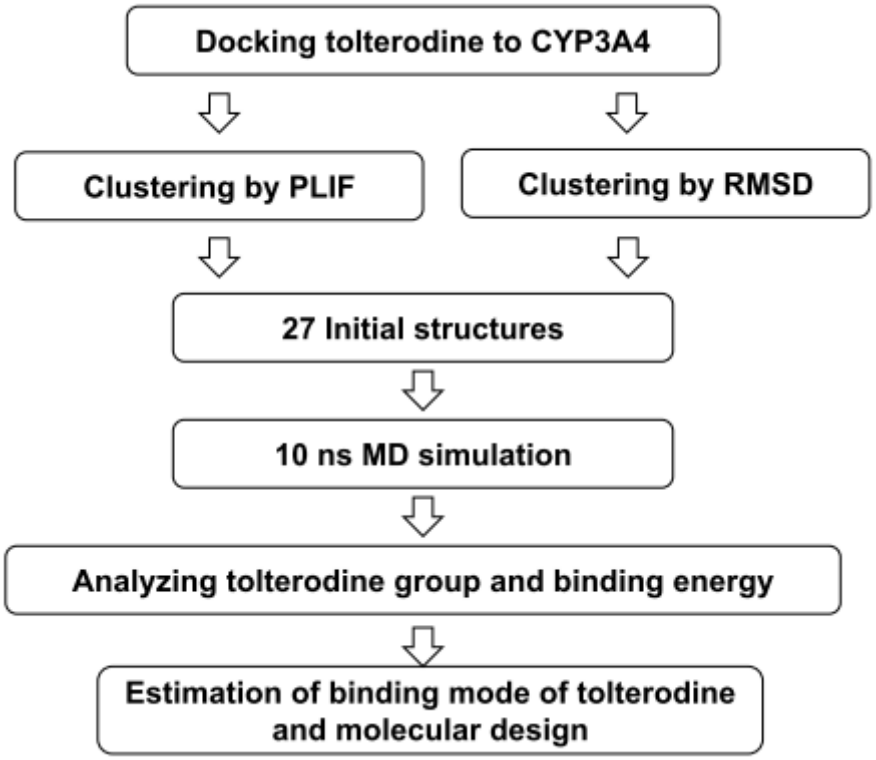


Figure 4.6 複数の代表構造選抜方法を用いて MD 初期ポーズを作成した結合様式評価フロー

4.4 MD シミュレーションの解析

MD シミュレーションは、Amber 12 ソフトウェアパッケージの PMEMD モジュールを使用し、先行研究[170]と同様の方法で実行した。カルバマゼピンのシミュレーションでは、平衡化後の追加シミュレーションで、カルバマゼピンと CYP3A4 の間に水素結合が存在したと考えられ、距離拘束を適用した[164]。対照的に、本研究ではこの化合物と酵素との間の重要な水素結合が予測されなかったので、トルテロジンの初期座標のみを拘束した。この方法は広範囲の化合物に適用することができるため、将来の研究に適用させることができる。平衡化後のプロダクトランでは拘束を外して 10 ns のシミュレーションを行った。

10 ns のシミュレーション中、スナップショットを 1 ps ごとに保存し、Amber の ptraj モジュールを用いて MD トラジェクトリの分析に使用した。溶媒効果を考慮しながら各ドッキングポーズの安定性を比較するために、MM/PBSA を使用して$\Delta G_{binding}$値を計算した。トルテロジン-CYP3A4 複合体の MD シミュレーションで$\Delta G_{binding}$値を計算する目的は、同じ化合物の異なるドッキングポーズの安定性の指標を示し、化合物のヘム鉄への近接可能性を予測することである。 同じタンパク質と同じ化合物の場合、エントロピーはΔG 結合にほとんど影響

を与えないため本研究では考慮していない。

10 ns の MD プロダクトラン実行中、構造の崩れがないかを確認するため主鎖 RMSD 値、$\Delta G_{binding}$ 値を評価した。主鎖 RMSD 値の評価からは、急激な構造的変化は見られないことを確認している（Figure 4.7）。またシミュレーション中にトルテロジンが CYP3A4 から遊離しなかったことを確認するために、各シミュレーションについて$\Delta G_{binding}$値の時間経過を分析した（Figure 4.8）。全てのシミュレーションにおいて、距離拘束が解かれた瞬間の著しいエネルギー変化はなかった。この結果は、シミュレーション中に構造の急激な変化が発生しなかったことを示している。これらの経時的分析の結果は、MD シミュレーションが適切に行われ、CYP3A4-トルテロジンスナップショットが選択された初期ポーズから CYP3A4 へのトルテロジン結合の再現が期待できることを示唆している。シミュレーション中の MD トラジェクトリは、トルテロジン炭素原子のヘムへの接近可能性の分析に使用された。

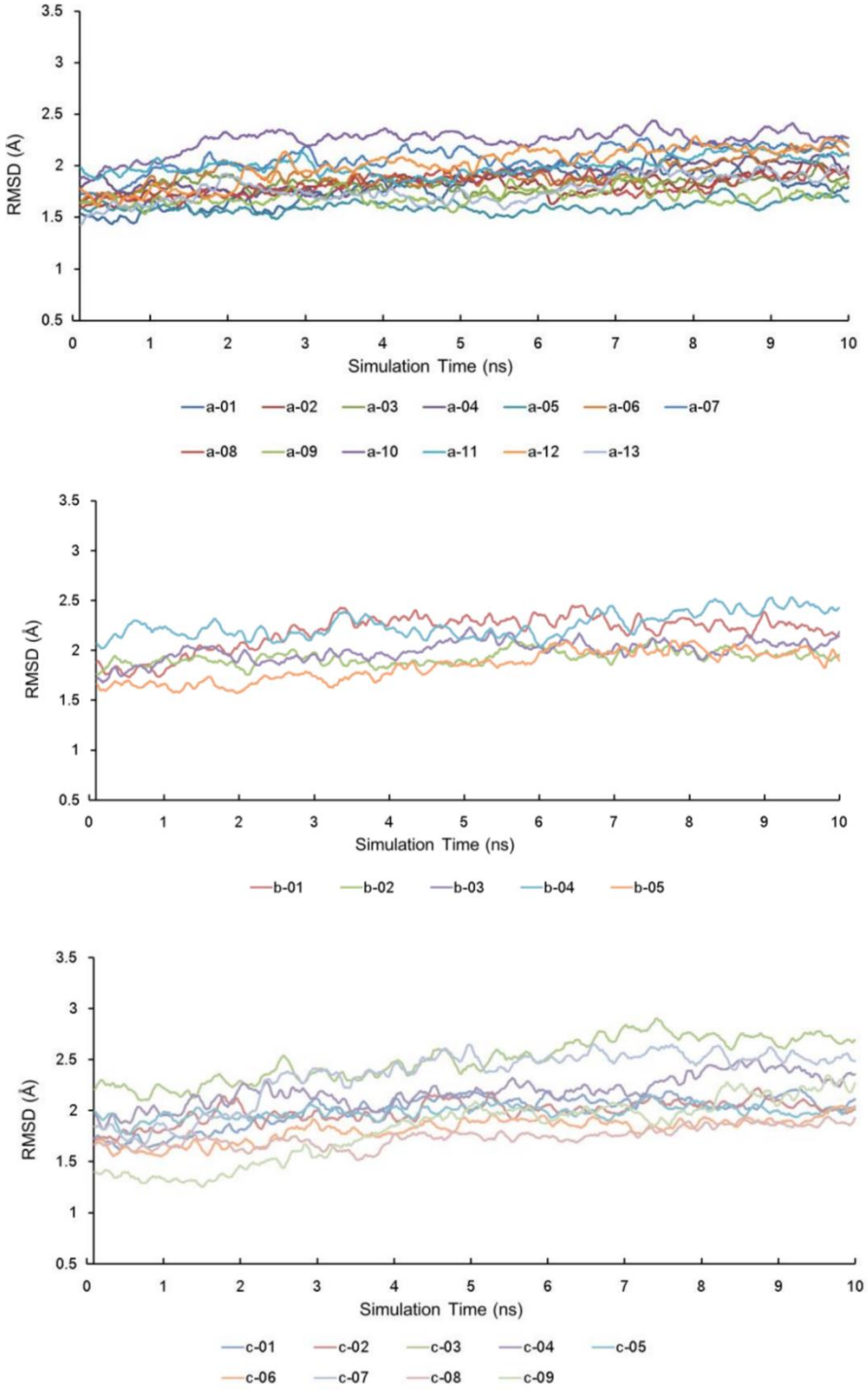


Figure 4.7 10ns MD シミュレーション中の主鎖原子の RMSD の時間経過

「a」、「b」、「c」は、Figure 4.5 の MD 初期ポーズのカテゴリに対応する。a-01～a-13：RMSD で初期ポーズを選択した 13 シミュレーション（上のパネル）、b-01～b-05： PLIF＋RMSD（中央のパネル）で初期ポーズを選択した 5 シミュレーション、c-01～b-05 c-09： PLIF で初期ポーズを選択した 9 シミュレーション（下のパネル）。

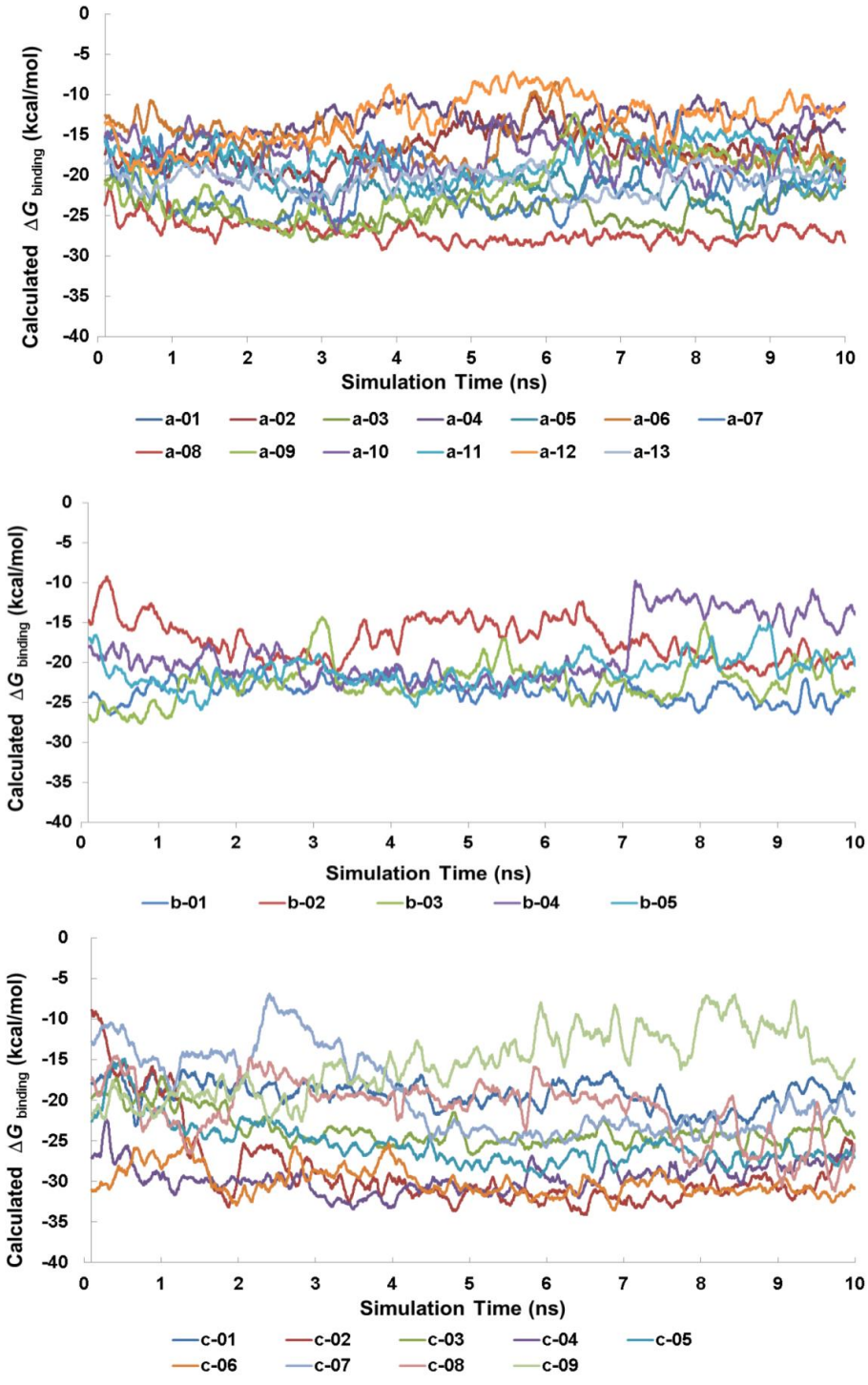


Figure 4.8　10 ns のシミュレーション中の$\Delta G_{binding}$ の時間経過

「a」、「b」、「c」は、Figure 4.5 の MD 初期ポーズのカテゴリに対応する。a-01～a-13：RMSD で初期ポーズを選択した 13 シミュレーション（上のパネル）、b-01～b-05： PLIF＋RMSD（中央のパネル）で初期ポーズを選択した 5 シミュレーション、c-01～b-05 c-09： PLIF で初期ポーズを選択した 9 シミュレーション（下のパネル）。

4.5　CYP3A4 ヘム鉄への近接性評価

トルテロジンの炭素原子に番号を付与し、4 つのグループに分類した（グループ A-D、Figure 4.9）。基質のヘム鉄への接近可能性の指標として、4 グループの炭素原子それぞれにおいて少なくとも 1 つの原子がヘム鉄の 6 Å 以内にあるスナップショット（各シミュレーションにつき 10,000 ずつ）数をカウントし、対応する$\Delta G_{binding}$を計算した。なお、酸化反応が起こり得るカットオフ距離として 6 Å が適切であることは先行研究[164]において確認されている。少なくとも 1 つの炭素原子がヘム鉄から 6 Å 以内にある 179,019 個のスナップショットを選択し、さらにヘム鉄から最も近い炭素原子のグループに基づいて分類した。

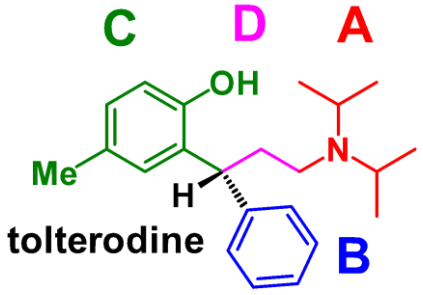


Figure 4.9　トルテロジンの炭素原子に割り当てられた位置ラベル

SOM を予測するために、グループ A～D のカットオフ距離内の出現頻度と CYP3A4-トルテロジン錯体の$\Delta G_{binding}$を計算した。27 の MD シミュレーションすべてについてスナップショットを分析したところ、グループ A がヘムに最も近いスナップショットの割合が最も高く（40.8%）、グループ B、C、および D の対応する割合は 20.9%、37.9%、および 0.4%であった。Table 4.1 と Figure 4.10 に各群の MM / PBSA 法による平均$\Delta G_{binding}$値を示す。これらの図表から、トルテロジン分子中のグループ A の原子がヘム鉄に最も近いポーズが、他の 3 つのグループの原子がヘム鉄に近いポーズよりも平均$\Delta G_{binding}$値が低いことが読み取れる。4 つのグループの平均$\Delta G_{binding}$値は、A <D <C <B の順序で増加した。したがって、グループ A のトルテロジン炭素原子がヘムに最も近づきやすいと結論づけた。

Table 4.1　ヘムへの近接性

各グループにおける炭素原子位置がヘム鉄から≦6Å に位置する頻度。10 ns の MD シミュレーションを通して、距離を 1 ps ごとに監視した。

$\Delta G_{binding}$ は、MM/PBSA を用いて計算した。

	Distance $_{Fe\text{-}C}$ ≤ 6 Å			
	Frequecy		$\Delta G_{binding}$ (kcal/mol)	
Group	No. of snapshots	%	Mean	Standard deviation
A	72,980	40.8	–24.10	5.89
B	37,455	20.9	–16.84	5.68
C	67,887	37.9	–20.98	4.99
D	697	0.4	–22.01	2.92

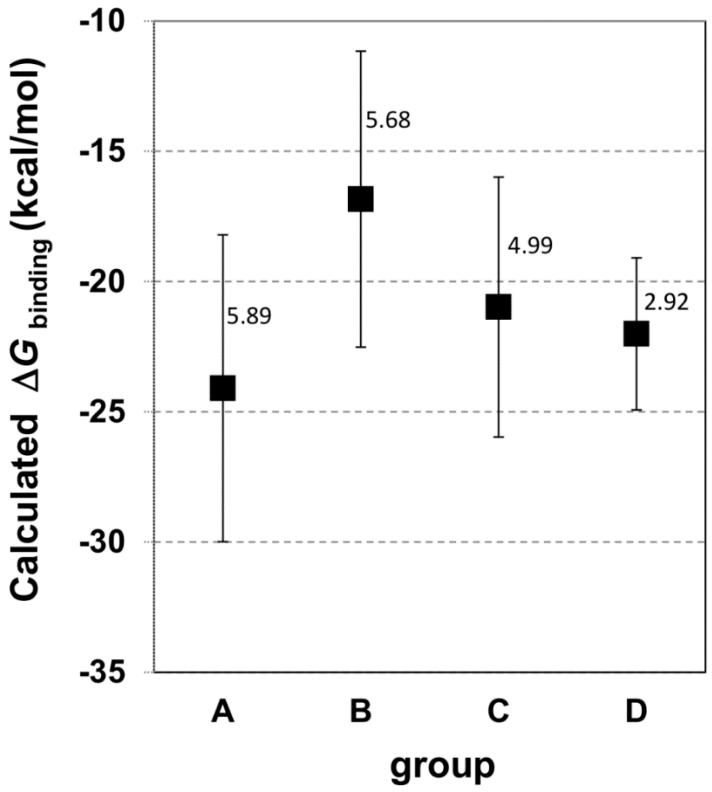


Figure 4.10 炭素原子がヘム鉄から≦6Å に位置したときの、平均$\Delta G_{binding}$ 値

エラーバーは標準偏差を示す。

4.6　結合様式の予測

前節において、グループ A のトルテロジン炭素原子がヘムに最も近づきやすいと推察された。これに加え、代謝的に安定な薬物の設計に役立てるため酸化反応時の結合様式の推定を試みた。トルテロジンの合理的な結合様式を決定するために、2,700 個の MD スナップショット（100 ps ごと）をサンプリングし、グループ A の炭素原子がヘムに近い 734 個のスナップショットを抽出した。これらスナップショットから PLIF [167]のビットを生成し（Figure 4.11）、*PipelinePilot*[183]で類似した結合様式を持つスナップショットが同じクラスに集まるよう閾値レベルを目視で調整しつつクラスタリングを行った結果、16 個のクラスターが生成された。Table 4.2 に、各クラスターに含まれるスナップショット数と、平均$\Delta G_{binding}$値および平均最小 Fe- C 間距離を示す。このうちクラスター1 が最も多くのクラスタメンバーを含み、2 番目に低い平均$\Delta G_{binding}$値、および最小の平均 Fe- C 間距離を示した。この結果から、クラスター 1 にトルテロジンの結合様式を表す複合体が含まれると推測した。また、クラスター7 の複合体は、スナップショットの 1%超を占めるクラスターの中で最も小さい平均$\Delta G_{binding}$値を有していた（Table 4.2）。

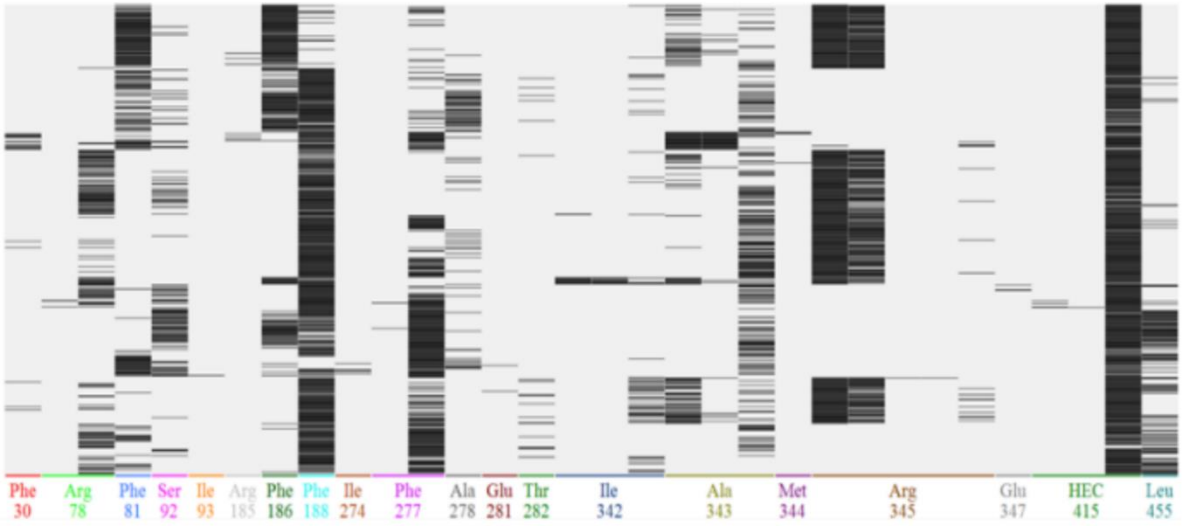


Figure 4.11　グループＡがヘムに近い（≤6 Å）　複合体についての PLIF のヒートマップ
横座標は相互作用するアミノ酸残基を示し、縦座標は MD スナップショットを示す。

Table 4.2　グループ A がヘムに近づく MD スナップショットの PLIF クラスタリング結果

Cluster	No. of snapshots	%	Average $\Delta G_{binding}$ (kcal/mol)	Average minimum Fe-C distance (Å)
1	**483**	**65.80**	**-25.57**	**4.18**
13	107	14.58	-21.61	4.40
11	61	8.31	-22.22	4.69
16	26	3.54	-25.26	5.36
9	14	1.91	-18.19	5.71
15	14	1.91	-13.11	5.10
7	**9**	**1.23**	**-29.46**	**5.82**
12	6	0.82	-13.40	4.08
14	4	0.54	-27.65	4.57
8	3	0.41	-12.16	3.75
6	2	0.27	-13.27	4.28
2	1	0.14	-16.58	3.82
3	1	0.14	-23.98	5.99
4	1	0.14	-22.00	3.76
5	1	0.14	-17.75	3.77
10	1	0.14	-14.97	3.46

Figure 4.12 は、クラスター 1（青） およびクラスター 7（灰色） における最も安定な複合体を示す。クラスター1 から抽出された複合体は、すべての複合体の中で最も低い$\Delta G_{binding}$値を示した（-38.56 kcal/ mol）。一方、クラスター7 の最も安定な複合体の$\Delta G_{binding}$値は-33.05 kcal / mol であった。スナップショット数と結合エネルギー値より、我々はクラスター1 の複合体が酸化反応時のトルテロジンの結合様式である可能性が高いと推定した。この構造から、化学修飾が標的タンパク質に対する化合物の活性に影響を及ぼさない限り、CYP3A4 とトルテロジンの間の結合を妨げる可能性のあるいくつかの薬物設計が考えられる。例えば、Arg345 とトルテロジンのヒドロキシル基との間の水素結合を切断すると（Figure 4.12、①）、CYP3A4 への結合親和性が低下する可能性がある。さらに、結合ポケットとの間の空間が狭いベンゼン環上の水素原子（Figure 4.12、②）またはイソプロピル基（Figure 4.12、③）をかさ高い置換基に変換すると、結合親和性を効果的に低下させる可能性がある。

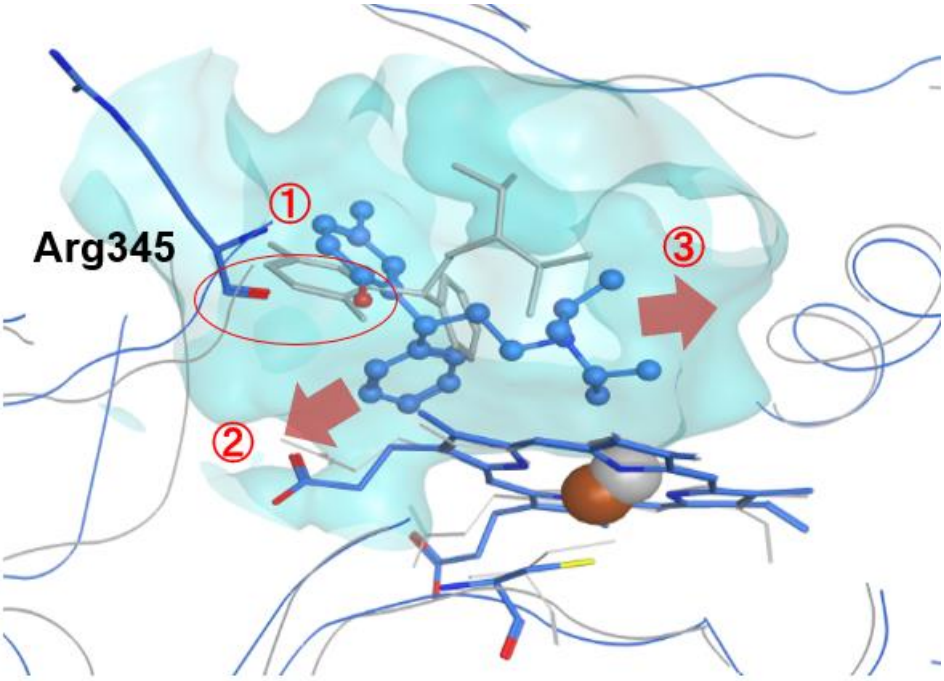


Figure 4.12　クラスター1（青）、およびクラスター7（灰色）における最も安定な複合体
クラスター1 の複合体は、すべての複合体の中で最も結合自由エネルギーが低い。

4.7　MD 初期ポーズの効果的な選択

本研究では、トルテロジンのドッキングポーズから RMSD と PLIF の両方のクラスタリング手法により、27 個のトルテロジンドッキングポーズを選択し（Figure 4.5）、MD シミュレーションの初期ポーズとして使用した。なお RMSD クラスターと PLIF クラスターの両方で 5 つの初期ポーズが選択されているが、これは各クラスターの中でドッキングスコアの最も良いものを代表として選択したためである。

Figure 4.13 は、27 個の MD シミュレーションについて、シミュレーション開始からの時間とヘム鉄から 6Å 以下の距離（酸化に必要な距離内）にあったグループ A〜D の炭素原子のスナップショットを対応させたヒートマップを示す。全体として、グループ A の炭素は、最も長い時間ヘムに近づいていた。興味深いことに、長い出現時間は、PLIF クラスターでのみ見られた初期ポーズで特に顕著であった（Figure 4.13 の上部パネルの「c」を参照）。グループ B の炭素原子がヘムに近づく時間はグループ A の約半分であった。次にヘムに近づく時間が長かったのは、グループ C であった。グループ C の場合、PLIF クラスターから選択された構造と RMSD クラスターから選択された構造との間に大きな違いはなかった（それぞれ、Figure 4.13 の c および a）。グループ D の炭素原子はトルテロジン分子の中心に位置し、置換基で囲まれているためヘムに近づくスナップショットはほとんどみられなかった。

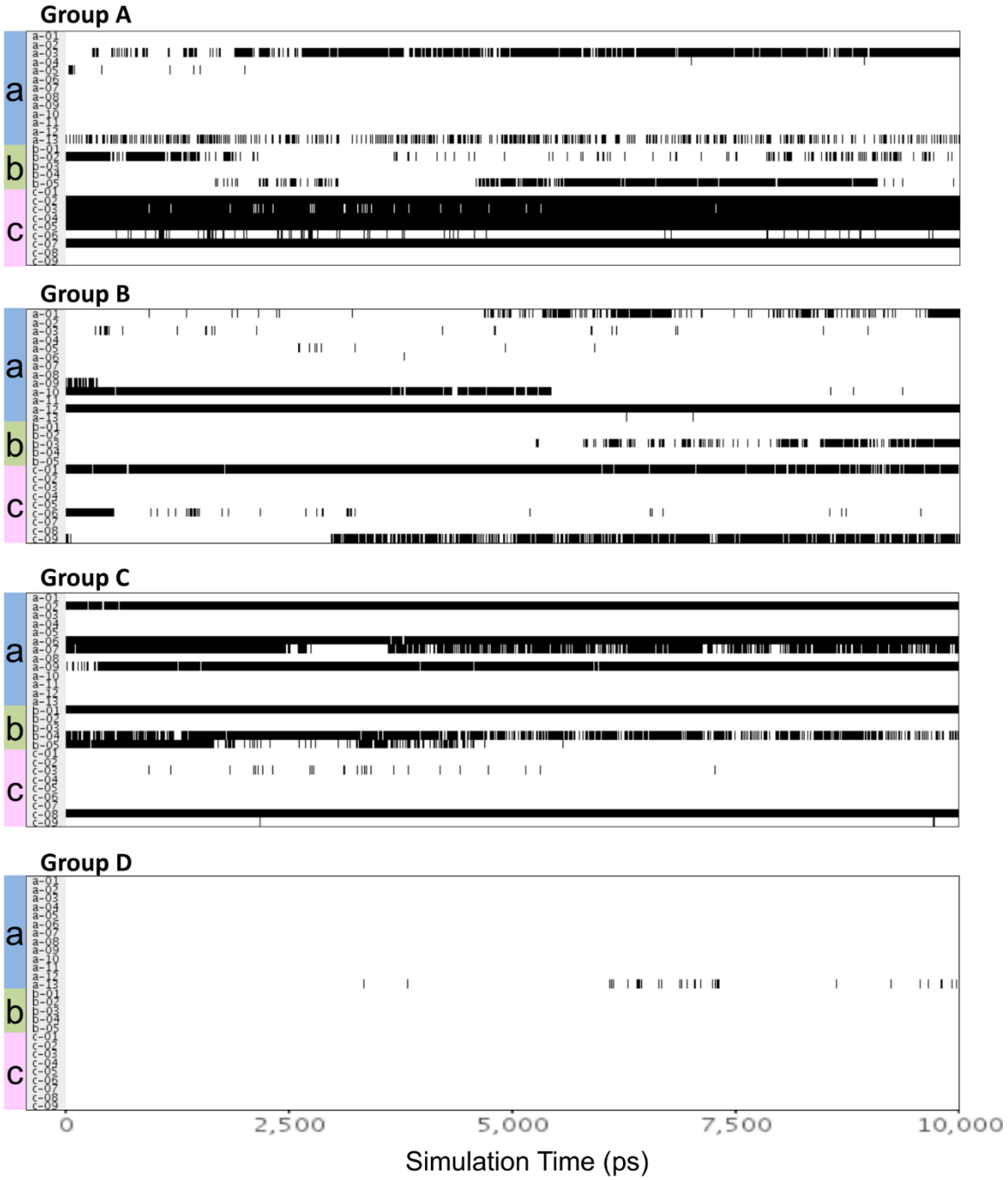


Figure 4.13　炭素原子位置がヘムの近くに位置したスナップショットのヒートマップ

それぞれの黒い線は、トルテロジンの少なくとも 1 つの炭素原子がヘムから 6 Å 以下で、もっともヘムに近い原子が属するグループ（A- D）のスナップショットを示す。10 ns の MD シミュレーションを通して 1 ps ごとに原子間距離をモニターした。 27 個の初期ポーズを選択するために使用されたクラスタリング方法は、各ヒートマップの左側に示されている（a、b、c の文字と色は Figure 4.5 と対応している）

これらの結果に基づいて、MD 初期ポーズの数を最小にするための RMSD および PLIF ク

ラスタリング法の有効性を比較した（Table 4.3）。PLIF クラスターから選択された 14 個の初期ポーズからの MD シミュレーションで得られたヘムへの近接時間および$\Delta G_{binding}$ の順序の順位は、27 個すべての初期ポーズからの MD シミュレーションで得られた順位と同じであったが、RMSD でのクラスタリングから得た 18 個構造で得られた順位は同じでなかった（正しくなかった）。

Table 4.3 MD の初期ポーズ数を最小にするためのクラスタリング方法の有効性の比較

a この列の小文字は Figure 4.5 を参照

クラスタリング方法	Group	ヘムへの近接時間			$\Delta G_{binding}$ (kcal/mol)		初期構造数
		スナップショット数	%	Rank	平均	Rank	
PLIF + RMSD (all)[a]	A	72,980	40.8	1	–24.10	1	27
	B	37,455	20.9	3	–16.84	4	
	C	67,887	37.9	2	–20.98	3	
	D	697	0.4	4	–22.01	2	
	Total	179,019					
RMSD (a + b)[a]	A	22,493	22.1	2	–21.28	2	18
	B	20,855	20.5	3	–16.09	4	
	C	57,598	56.7	1	–20.94	3	
	D	697	0.7	4	–22.01	1	
	Total	101,643					
PLIF (b + c)[a]	A	57,796	53.8	1	–24.54	1	14
	B	18,929	17.6	3	–18.32	3	
	C	30,716	28.6	2	–21.50	2	
	D	0	—	—	—	—	
	Total	107,441					

比較的柔軟な化合物であるトルテロジンについての比較分析から、PLIF クラスターからの初期ポーズの選択が、従来の RMSD クラスタリングを使用する選択よりもより効果的かつ堅牢であることが示された。すでに述べたように、PLIF はタンパク質-化合物相互作用を形成する、またはしないパターンを表しており、PLIF を使用して生成されたクラスターはドッキングポーズを効果的に識別することができる。これとは対照的に、RMSD クラスタリングは単純な位置の変化を検出するが、その変化は必ずしも結合様式にとって意味がないと言える。トルテロジンに関する我々の比較分析から、代謝部位の予測および結合様式の推定のための MD シミュレーションには相互作用パターンに焦点を合わせ初期ポーズが選択されるべきで

あることが示された。

4.8 結言

本章で述べた検討から、2 つの結論が得られた。第 1 に、複数の初期ポーズで得られた MD シミュレーションから CYP3A4 中のヘム鉄への近接性を分析し、カルバマゼピンよりも柔軟な化合物であるトルテロジンの代謝部位ならびに結合様式を首尾よく予測した。実験的に代謝される炭素が最も長い時間ヘム鉄に接近し、かつ最も安定した結合を示した。これにより、複数の MD 初期ポーズから短時間のシミュレーションを行い代謝部位、結合様式を予測する手法が柔軟な化合物にも適用できることを示した。

第 2 に、本検討から MD シミュレーションのための初期ポーズの選択に有効な手段を見出した。PLIF クラスタリングによって選択された初期ポーズを使用した MD シミュレーションは、RMSD クラスタリングから選択された構造を使用したものよりもはるかに正確なヘムへの近接性予測を達成した。初期ポーズの最小の数を適切に選択するためのより堅牢なプロトコルを確立するためには、他の事例での検証が必要であるが、PLIF クラスタリングは有望であると考えられた。

本研究では、短時間の古典 MD でも実験結果に合致した結果が得られた。今後古典 MD では網羅的なサンプリングに限界がある場合、次のステップとして複数の結晶構造を用いる、または 2.4 で述べたような広い範囲の構造探索が可能な拡張アンサンブル法が選択肢として考えられる。また、本研究では CYP3A4 により代謝を受ける化合物の結合様式予測を実施したが、CYP3A4 の結合サイトの中で、ヘムから遠い場所に結合し、他の化合物の代謝を阻害するタイプの化合物もある。このような場合、高難度ではあるがポケットへの近づき方を長時間 MD でとらえ、その結合経路を抑えて CYP3A4 に結合しにくくするアプローチも有効と考えられる。

次に、第 5 章では結合様式予測を効率化させる方法論の提案に向け、第 3 章で得られた CYP3A4 の機械学習モデルを適用して MD 初期ポーズを選択した結合様式予測について述べる。

第5章　総合討論

5.1　緒言

第 3 章の機械学習による結合様式予測において、モデル構築のためのデータセット作成の方法論を提案し、第 4 章の MD シミュレーションによるタンパク質の動きを考慮した結合様式予測では短時間 MD を多数実施する手法により、結合様式を予測することができた。本章では、第 3 章、第 4 章で構築した手法の応用、機械学習のデータセット考察について述べる。5.2 では、機械学習とシミュレーションを組み合わせてポーズ予測を効率化させる方法の提案に向け、第 3 章で得られた機械学習モデルを MD 初期ポーズの選択に適用した結果を確かめた。5.3 では、第 3 章の機械学習アプローチで作成したデータセットの分布に関し示した。

5.2　機械学習モデルとシミュレーションを組み合わせた結合様式予測

本節では本論文の主課題である、機械学習による MD 初期ポーズ選択と MD シミュレーションを組み合わせた結合様式予測の例として、トルテロジン-CYP3A4 複合体を題材としたときの結果について述べる。具体的には従来法（RMSD）で MD 初期ポーズを選択した場合と機械学習で MD 初期ポーズを選択した場合の予測結果を比較する。

MD 初期ポーズ選択に用いた機械学習モデルには、第 3 章の GP +CF セットを使用した事前学習に続いて、LL セットを使用して識別層のみファインチューニングしたモデル（pre(GP+CF)/ft-cl(LL)）を選択した。第 4 章において用いた 27 個の MD 初期ポーズについてこの機械学習モデルを適用し、CNN スコアの上位半数（13 個）を選択して MD シミュレーションの結果を解析した。結果の解析方法は、CYP3A4 ヘム鉄への近接性評価（本論文 4.5）に準じた。

MD シミュレーションの解析結果を Figure 5.1、Figure 5.2 に示す。MD シミュレーション中、代謝を受けるグループ A がヘムに近づく時間を比較すると、機械学習で MD 初期ポーズを選択した場合は、代謝部位 A が他の部位と比較してヘムに近づく時間が長く、かつその時の結合自由エネルギー値は低く安定し、実験事実と合致した。一方、従来法（RMSD）で選択した場合は、代謝部位ではないグループ C がヘムに近づく時間が長く、結合自由エネルギー値はグループ A と同程度に安定であり、実験事実と合致しなかった。以上の結果から、機械学習モデルによる MD 初期ポーズの選択により、従来法よりも結合様式を正しく予測できること

が示された。

今後の改善点として、今回の検討ではトルテロジンが含まれる結晶構造が含まれなかったため第 3 章で得られた機械学習モデルをそのまま MD 初期ポーズ選択に使用したが、トルテロジンまたはその類縁体が含まれる結晶構造でファインチューニングしたモデルが作成できれば、さらなる効率化、高精度化が期待できると考える。

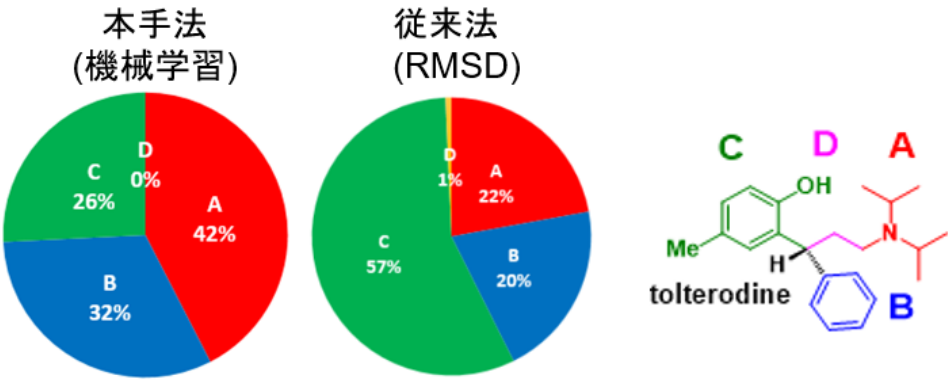


Figure 5.1　MD シミュレーション中、代謝部位 A がヘム鉄に近づく時間の比較

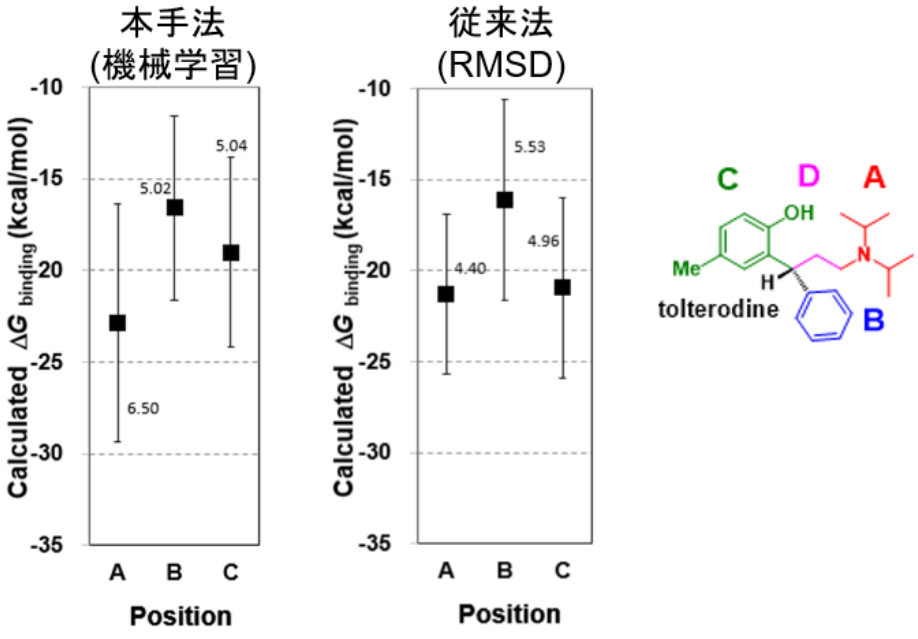


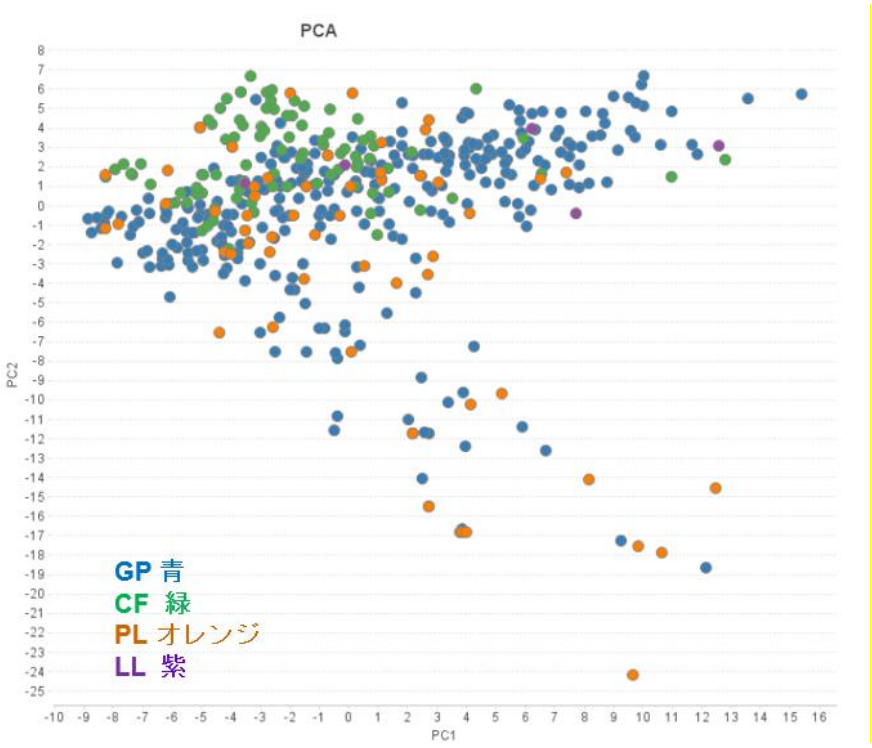
Figure 5.2　MD シミュレーション中、各グループがヘム鉄に近づくときの結合自由エネルギー比較

5.3　機械学習のデータセットに関する考察

本節では、第 3 章 機械学習による結合様式予測で選択したデータセットについて、さらに以下の追加調査、考察を加えた。まず、第 3 章で述べた訓練データセットに関し、タンパク質、化合物の両面からデータの分布を調査した。さらにデータセットの positive、negative のバランス調整について考察する。

第 3 章の機械学習による結合様式予測で、訓練データセットの効果についてオキシドリダクターゼの比率、ポケットサイズに注目し比較した。本小節では、さらに訓練データセットの分布に関し、化合物空間（ケミカルスペース）とタンパク質機能の両面で調査した。

ケミカルスペースは、分子を分子量子数と呼ばれる 42 個の記述子によって定義される 42 次元のプロパティ空間から原子、結合、極性基、トポロジフィーチャーのさまざまなカテゴリをカウントし、サイズ、剛性、極性によって分子を分類し 2 次元に投影したマップである[184], [185]。データセットに含まれる化合物構造から重複を除き、次元削減の方法としてよく利用されている主成分分析と、近くのクラスター間の分布間の差異を表す手法として比較的最近考案された t-SNE[187]を用いて GP セット（309 化合物）、CF セット（93 化合物）、PL セット（56 化合物）、LL セット（5 化合物）のリガンドのケミカルスペースを比較した（Figure 5.3）。その結果、いずれの方法でも PL セット（オレンジ）は GP セット（青）と分布が類似していたが、CF セット（緑）、LL セット（紫）は特段の偏りは見られなかった。



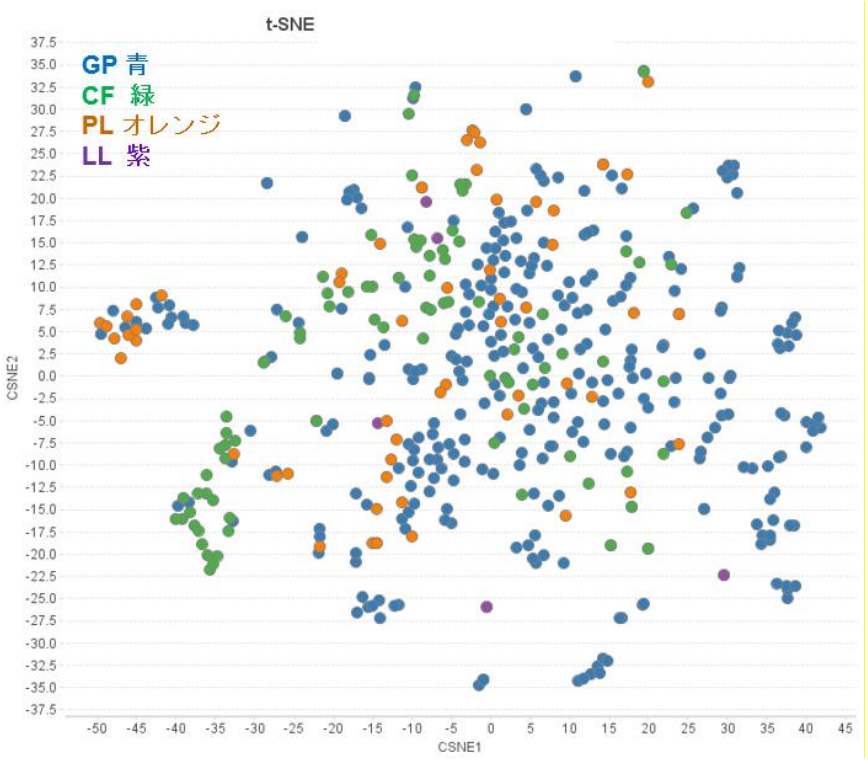
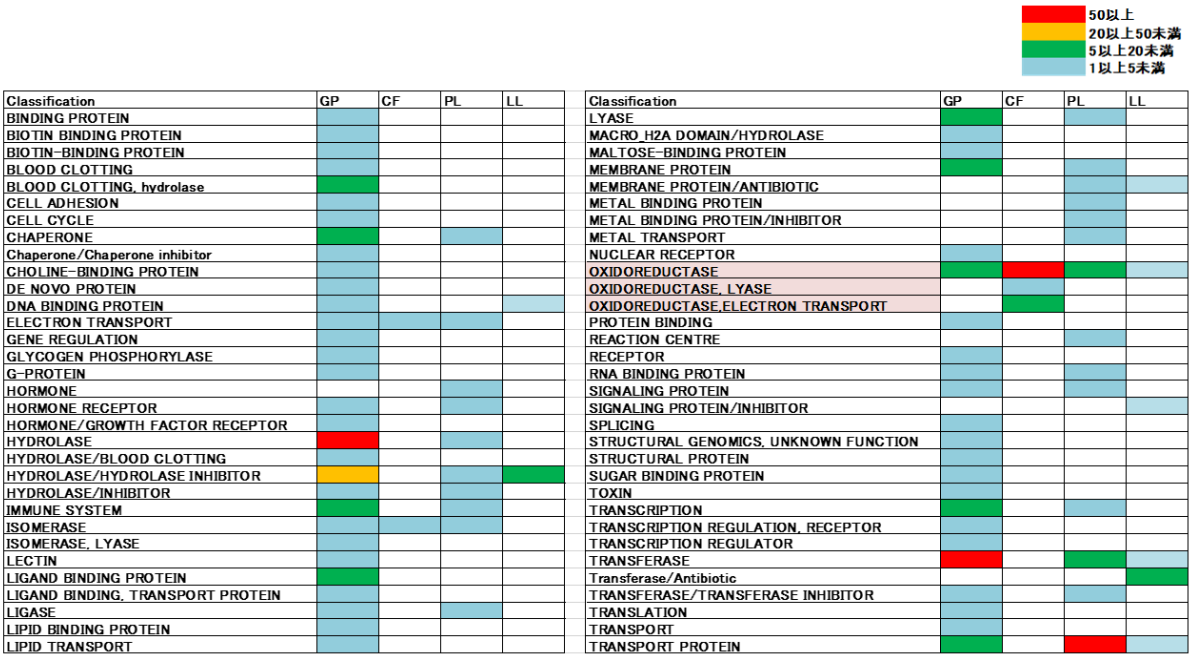


Figure 5.3　4 種のデータセットについて、化合物側の性質に関する主成分分析、t-SNE によるマッピング

タンパク質機能の分布に関し、第 3 章では各データセットに CYP3A4 が属するオキシドリダクターゼが含まれる比率を比較した。本小節では、さらに各データセットについて PDB の機能分類とその分類に含まれるデータ数を調査した。その結果、Figure 5.4 に示すように、CF セットに「OXIDOREDUCTASE」が集中し、GP セットとの組み合わせで補完していることを再確認できたが、PL、LL セットに特に偏った分布は見られなかった。

これまでの調査の結果から、タンパク質側に関してはデータセットの組み合わせで GP セットと CF セットが補完の相性が良いことが再確認された。一方、化合物側のプロパティについては分布の差は見られなかった。



Classification	GP	CF	PL	LL
BINDING PROTEIN				
BIOTIN BINDING PROTEIN				
BIOTIN-BINDING PROTEIN				
BLOOD CLOTTING				
BLOOD CLOTTING, hydrolase				
CELL ADHESION				
CELL CYCLE				
CHAPERONE				
Chaperone/Chaperone inhibitor				
CHOLINE-BINDING PROTEIN				
DE NOVO PROTEIN				
DNA BINDING PROTEIN				
ELECTRON TRANSPORT				
GENE REGULATION				
GLYCOGEN PHOSPHORYLASE				
G-PROTEIN				
HORMONE				
HORMONE RECEPTOR				
HORMONE/GROWTH FACTOR RECEPTOR				
HYDROLASE				
HYDROLASE/BLOOD CLOTTING				
HYDROLASE/HYDROLASE INHIBITOR				
HYDROLASE/INHIBITOR				
IMMUNE SYSTEM				
ISOMERASE				
ISOMERASE, LYASE				
LECTIN				
LIGAND BINDING PROTEIN				
LIGAND BINDING, TRANSPORT PROTEIN				
LIGASE				
LIPID BINDING PROTEIN				
LIPID TRANSPORT				

Classification	GP	CF	PL	LL
LYASE				
MACRO_H2A DOMAIN/HYDROLASE				
MALTOSE-BINDING PROTEIN				
MEMBRANE PROTEIN				
MEMBRANE PROTEIN/ANTIBIOTIC				
METAL BINDING PROTEIN				
METAL BINDING PROTEIN/INHIBITOR				
METAL TRANSPORT				
NUCLEAR RECEPTOR				
OXIDOREDUCTASE				
OXIDOREDUCTASE, LYASE				
OXIDOREDUCTASE,ELECTRON TRANSPORT				
PROTEIN BINDING				
REACTION CENTRE				
RECEPTOR				
RNA BINDING PROTEIN				
SIGNALING PROTEIN				
SIGNALING PROTEIN/INHIBITOR				
SPLICING				
STRUCTURAL GENOMICS, UNKNOWN FUNCTION				
STRUCTURAL PROTEIN				
SUGAR BINDING PROTEIN				
TOXIN				
TRANSCRIPTION				
TRANSCRIPTION REGULATION, RECEPTOR				
TRANSCRIPTION REGULATOR				
TRANSFERASE				
Transferase/Antibiotic				
TRANSFERASE/TRANSFERASE INHIBITOR				
TRANSLATION				
TRANSPORT				
TRANSPORT PROTEIN				

Figure 5.4　各データセットに含まれるタンパク質の機能分布。(機能はアルファベット順)

一般に機械学習でモデルの訓練を行うためには訓練データの positive、negative の数が均衡していることが望ましい。第 3 章において、訓練データセットの positive、negative の比は平均して 1:3、最大（LL セット）で 1:7.5 となっている（Table 3.1）。訓練セットの分類の基準は一律に結晶構造のポーズから RMSD 値 2Å 未満のポーズを positive なポーズとしてタグ付けされ、RMSD 値 が 4 Å より大きいポーズは negative なポーズとしてタグ付けされている。しかしながら、Figure 5.5 に示すように同じ RMSD 値でも重原子数（水素以外の原子数）の大きな分子ほどわずかな位置のずれで negative ポーズと分類される。LL セットは CYP3A4 と結合する大きな化合物との複合体構造で構成されることから、他の類似セットと比較して相対的に negative の数が多くなる。

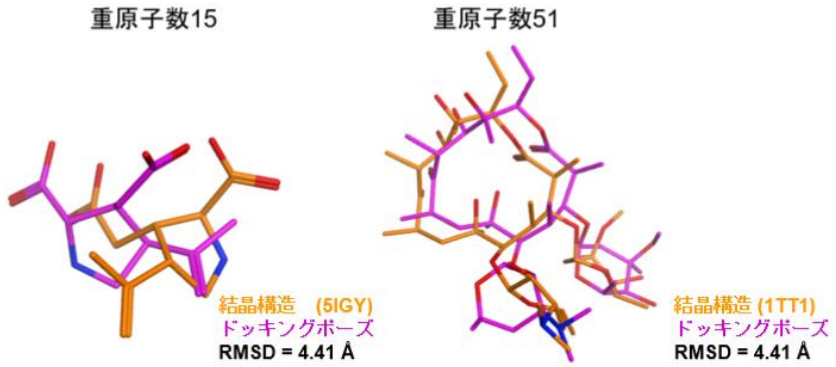


Figure 5.5　分子サイズの大小と同じ RMSD 値を持つ negative ポーズの比較。左 ： 小さな分子では結晶構造と分子の向きが反転、右 ： 大きな分子では、分子の向きは同じ

データ作成当初、positive、negative の比が最大 1:60 程度であったことから、本研究では negative をランダム選択して positive の 6-7 倍に収まるように調整した。今後の改善点として、結晶構造以外の NMR 構造、MD 計算で出力した構造を positive に追加するなどの工夫によりデータセットのバランスを調整するなどが考えられる。

上記の数に関する問題の他、positive、negative の定義についても改善の余地がある。分子量の大きな化合物と小さな化合物では同じ RMSD 値でもずれ幅が異なり、単に RMSD 値が小さくても結合に重要な相互作用が失われている可能性もある。一方で、相互作用のみでの定義では位置関係の要素がなくなり、エラーも増える懸念もある。今後の課題として、RMSD については例えば重原子数で割ってスケーリングし、かつ PLIF（protein-Ligand interaction fingerprint）など相互作用の要素も考慮することで、正例と負例を公平に定義することができると考えられる。

5.4 結言

本章では、第 3 章、第 4 章で構築した手法の応用、機械学習のデータセット考察について述べた。5.2 では、機械学習シミュレーションを組み合わせてポーズ予測を効率化させる方法の提案に向け、第 3 章で得られた機械学習モデルを MD 初期ポーズの選択に適用した結果を確かめ、従来法よりも予測精度が向上していることが確認された。5.3 では、第 3 章の機械学習アプローチで作成したデータセットに関してさらに深堀して解析した。一方で、データセットの作成にはさらなる工夫の余地がある点も示した。

第6章　結論

6.1　結論

本研究では、結合サイトが柔軟で大きな高難度ターゲットの結合様式予測を目的に、CYP3A4 を題材として機械学習による結合様式予測と MD シミュレーションによる構造のサンプリングが有効であることを示した。さらに、機械学習による結合様式予測モデルを適用して MD 初期ポーズを選抜する高難度ターゲットの結合様式予測を試みた。これにより、機械学習とシミュレーションを組み合わせて効率化させる一つの方法論を示した。

第 3 章の機械学習による結合様式予測では、標的タンパク質に関連するタンパク質-化合物複合体の結晶構造の数が限られている場合に、サイズと特性の異なるデータセットを組み合わせてファインチューニングを行うと、CNN モデルの結合様式予測にどのように影響するかを調べた。データセットの組み合わせの効果の調査では、予測機能を最適化するためにタンパク質の機能とデータサイズに関してバランスの取れたデータセットが重要であることが示唆された。さらに、ターゲットタンパク質に結合する化合物を含む他のターゲットタンパク質複合体構造で構成されたデータセットでファインチューニングすることにより、ROC AUC と CNN スコアの両方で大幅に予測精度が向上した。CYP3A4 を題材とした場合には、ターゲットタンパク質と同程度に結合サイト体積の大きな結晶構造データセットが CYP3A4 の結合様式の予測に重要であったことを示唆している。

第 4 章の MD シミュレーションによるタンパク質の動きを考慮した結合様式予測では、CYP3A4 を題材として短時間 MD を多数実施した手法について検討した。本検討から、複数の MD 初期ポーズから短時間のシミュレーションを行い代謝部位、結合様式を予測する手法が先行研究よりも柔軟で高難度な化合物にも適用できることを示した。さらに、初期ポーズの選択には従来法に適用された RMSD によるクラスタリングよりも、PLIF によるクラスタリングの方が高精度に結合様式を予測できることを見出した。

第 5 章では、第 3 章、第 4 章で構築した手法の応用、機械学習のデータセット考察について述べた。機械学習による結合様式予測モデルを適用して MD 初期ポーズを選抜する高難度ターゲットの結合様式予測を試みた結果、従来法と比較し予測精度の向上が確かめられた。さらに、機械学習モデルにに寄与したデータセットの分布も複数の観点で示した。

6.2 今後の展望

機械学習による結合様式予測については、記述子、モデル構造に関して改良の余地を残している。本研究で扱った記述子は各グリッドにおける原子タイプのみとなっており、タンパク質-化合物間の相互作用に関する情報などは含まれない。今後はグリッドに載せる記述子情報として、相互作用情報や水分子影響を追加することにより、さらなる予測精度の向上が期待される。こういった記述子の情報量増加に対応し、例えば DenseNet[188]のように複雑なフィーチャーマップを形成して効率的に訓練するようなモデル構造の改良も必要と考えられる。また、現在のモデルでは、別のプログラムで発生させたドッキングポーズを評価しているのみであるが、グリッド上に好ましい原子タイプ、または部分構造を *de novo* で生成させるようなモデルが作成できれば、効率的な分子設計につながる。

シミュレーションによる結合様式予測について、本研究では短時間 MD を多数実施することにより網羅的な結合様式を抽出し、結合ポケットが大きく、揺らぎのあるターゲットに対する結合様式推定を実施した。本研究では MD 初期ポーズの作成のために 1 つの結晶構造を鋳型としたドッキングポーズを用いたが、タンパク質構造のより大きな動きを考慮するためには、複数のタンパク質構造に対するドッキングポーズを用いることでさらに網羅性が高まると考えられる。さらに、この手法で得られた推定結合様式を活用した薬物設計も次の課題となる。

今後、シミュレーションと機械学習を真に融合させた高難度ターゲットの結合様式予測として、シミュレーションの強みである実験では得られないような仮想大量データの生成と、その大量データにより訓練した結合様式予測の機械学習モデルを適用して MD 初期ポーズ選択サイクルを回すことにより、予測手法の高速化、高精度化が期待される。さらに、CYP3A4 以外の通常のドッキングで結合様式予測ができなかったタンパク種、化合物へ適用できれば、汎用的な手法として創薬研究を後押しすると考えている。

付録

付録.A 長時間 MD

付録 A では全体課題 「結合サイトが柔軟で大きな、高難度ターゲットの結合様式予測」の中で、短時間 MD を並列した効果を確認するための長時間 MD との比較について述べる。CYP3A4 に関する長時間 MD の事例として、apo 体（化合物が結合していない状態）について 10 μs の MD シミュレーションを実施し、スナップショットに対してドッキング計算を行い、代謝部位予測を試みた事例が報告されている[17]。しかしながら、5μs 以降で急激にポケット体積、ポケット内疎水性が大きくなり、それに伴って予測精度が低下し、ポケット構造が崩れていることが示唆された。そこで、本検討ではトルテロジンと CYP3A4 の複合体に対し、仮説として「短時間 MD と比較して、 結合様式の網羅性が優れているのではないか?」を設定し、短時間 MD （10 ns）の 100 倍となる 1000 n 秒 （1 μs）の MD を実施し検証した。

MD 初期ポーズは、4.3 で作成した 27 の初期構造のうち、最もドッキングスコアが低く、かつ実験的に代謝をうける「グループ A（Figure 付録 1）」がヘムに近づいていないドッキングポーズを選択した（Figure 付録 2）。

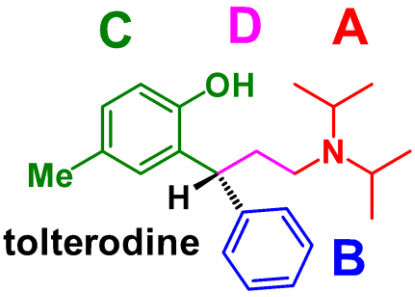


Figure 付録 1 トルテロジンの炭素原子に割り当てられた位置ラベル（再掲）

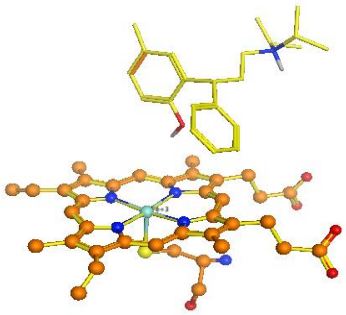


Figure 付録 2　選択した MD 初期ポーズ（グループ B が最もヘム鉄に近づいている）

MD シミュレーションのプロトコルは 4.2 とほぼ同様であるが、細かな点が異なるため下記に計算条件を記す。ヘム基とトルテロジン分子の電荷は、B3LYP / 6-31G **レベルで Gaussian03 ソフトウェア[57]を使用して実行された量子力学的計算によって導き出され、RESP 法によってそれぞれの原子に適合した[172]。CYP3A4-トルテロジン複合体のドッキングモデルは、長方形の TIP3P ウォーターボックス[173]に、境界から最小 18Å の距離で挿入された。系を中和するために、対イオン（Na^+、Cl^-）を追加した。力場は ff03 力場[174]を使用した。MD シミュレーションは、AMBER 12 ソフトウェアパッケージ[175]の GPU で加速された PMEMD モジュールで実行した。水素最適化のために、最急降下法と共役勾配法それぞれ 5,000 ステップにて系のエネルギー最小化計算を行った。さらに、側鎖と溶媒の最適化のために、同じ方法で各 5,000 ステップ、タンパク質と溶媒の最適化のために同じ方法でそれぞれ 30,000 ステップ、最後に、リガンドを含むシステム全体の最適化のために、それぞれ同じ方法で 30,000 ステップ、系のエネルギー最小化計算を行った。系を積分時間ステップ 1 fs の NVT アンサンブル条件下で 130 ps かけて 300 K に加温し、10 kcal mol^{-1} $Å^{-2}$ の位置拘束をトルテロジンに適用して、NPT 条件で 100 ps で平衡化した。平衡化後、100 ps の場合は 5 kcal mol^{-1} $Å^{-2}$、100 ps の場合は 2 kcal mol^{-1} $Å^{-2}$ のリガンド位置拘束下で追加のシミュレーションを実行した。この追加のシミュレーションの目的は、拘束を徐々に解除することである。次に、プロダクトランとして 1 µs の拘束なしのシミュレーションを実行した。1 µs のシミュレーション中、スナップショットを 10 ps ごとに保存し、分析に使用した。シミュレーション全体を通じて、SHAKE アルゴリズム [176]を採用し、水素原子を含む結合を拘束した。積分時間ステップは 2 fs に設定した。 非結合項のカットオフ距離は 10Å に設定した。

MD シミュレーションの解析は、4.4 と同じ方法で構造の崩れがないかどうかの確認を行った。さらに、プロダクトラン開始後 100 ps、1 ns、10 ns、100 ns、500 ns、1 µs のスナップショットについて結合様式を確認した。

まず、MD シミュレーション自体に問題がないかを確認するため、MD プロダクトラン実行中の主鎖 RMSD 値、E_{total}（系の全体エネルギー値）を評価した。主鎖 RMSD 値の評価からは、時間経過とともにわずかに大きくなる傾向はあるが（Figure 付録 3Figure 1.1）、E_{total} は、経時的に小さく（安定化）しており（Figure 付録 4）、シミュレーション自体に問題はないと判断

した。

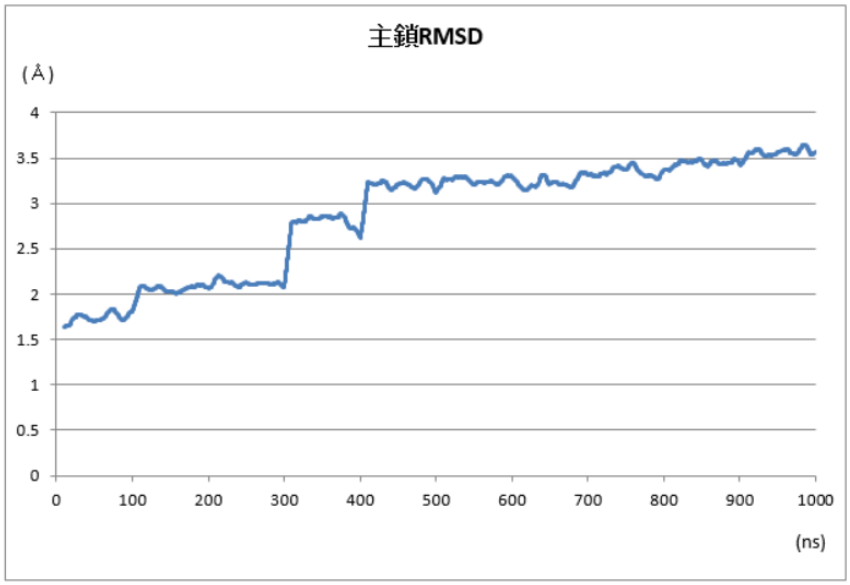


Figure 付録 3　MD プロダクトラン実行中の主鎖 RMSD 値変化

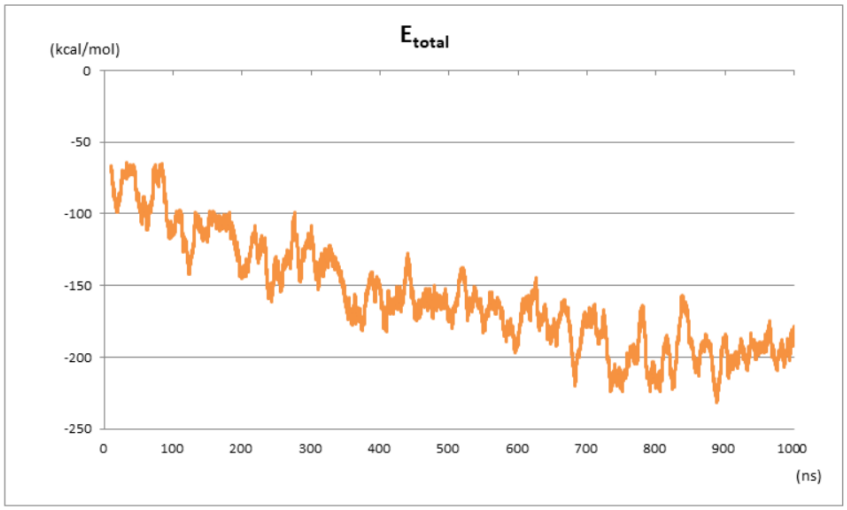


Figure 付録 4　MD プロダクトラン実行中の E_{total}（系の全体エネルギー値）変化

シミュレーション実施の目的である、長時間 MD での結合様式の網羅性確認のため、ヘム鉄に近づく炭素原子のバリエーションを確認した。プロダクトラン開始後 100 ps、1 ns、10 ns、100 ns、500 ns、1 µs のスナップショットについて結合様式を確認した。その結果、初期ポーズから 1000 ns までグループ C の炭素原子がヘムの方を向き、結合様式は大きく変化しないことが分かった（Figure 付録 5）。

仮説として「短時間 MD と比較して、 長時間 MD では結合様式の網羅性が優れているの

か?」を設定し、短時間 MD （10 ns）の 100 倍の長さ 1000 ns（1 μs）の MD を実施し検証した。その結果、短時間 MD の 100 倍シミュレーションをしても、結合様式がほとんど変わらず、結合様式の網羅性は得られなかった。1 例のみの実施ではあるが、結合様式を網羅するためには、複数の初期構造から短時間 MD を並列して実施する方法によって、効率よく結合様式を網羅できることが示された。

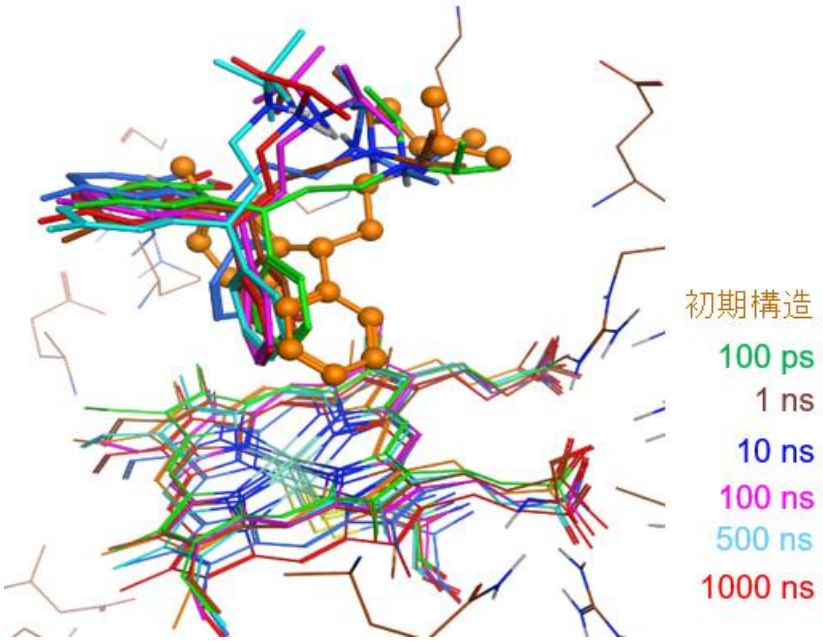


Figure 付録 5　プロダクトラン開始後 100 ps、1 ns、10 ns、100 ns、500 ns、1000 ns（1 μs）の結合様式

付録.B 機械学習モデルのデータセット

第 3 章で用いたデータセットを下記に記載する。

Table 付録 1　GP セットの PDB ID

PDB ID							
10GS	1NW5	1XW6	2FDP	2JJ3	2QHZ	2X8Z	3DX3
1A8I	1O5A	1Y0L	2FF1	2NMX	2QI1	2XA4	3E92
1A99	1O5R	1Y20	2FLR	2NN1	2QI3	2XC4	3EHY
1B6J	1OIM	1Y6B	2FQT	2NNQ	2QI4	2XD9	3EJP
1B6L	1OW4	1YC1	2FQW	2NQ7	2QI5	2XM1	3EJQ
1B6M	1P1N	1YDK	2FQX	2O4J	2QI6	2Y7X	3EJR
1CW2	1P1O	1Z95	2FQY	2OAG	2QMG	2YEK	3EJT

PDB ID							
1D2E	1PXP	1ZHX	2FV5	2OJG	2QNQ	2Z4B	3EKO
1D4I	1Q0Y	1ZHY	2FVD	2OVV	2QRY	2ZMM	3EKR
1D4J	1Q4W	2AFX	2FXD	2OYM	2QTA	2ZN7	3ELM
1DUV	1Q72	2AYR	2FXU	2P09	2QTG	3A6Q	3ENE
1EBY	1QKT	2B07	2GKL	2P16	2QU6	3A6R	3EQR
1EBZ	1QXL	2B1Z	2GM1	2P3T	2QVU	3AR4	3F3D
1EC0	1R5Y	2B3F	2GZ2	2P4Y	2R4F	3B27	3F48
1EC1	1S38	2BBF	2H15	2P7G	2R5P	3B50	3F8C
1EC2	1S39	2BFR	2H21	2P95	2R6W	3B7R	3FAS
1ENU	1S50	2BOH	2H6B	2P98	2R75	3B92	3FAT
1FCX	1S7Y	2BOJ	2H6T	2PGZ	2RCA	3BE2	3FH5
1FH8	1S9T	2BRC	2HD6	2POG	2RDE	3BEX	3FZS
1FH9	1SR7	2BVD	2HJ4	2POU	2REG	3BGQ	3FZT
1FHD	1SW1	2BYH	2HZL	2POV	2UWL	3BGZ	3GI5
1G2K	1SWK	2BYI	2HZY	2POW	2UWP	3BHX	3GI6
1GI9	1SYH	2BYR	2I0A	2PQC	2UY5	3BL0	3H0W
1GJ8	1SYI	2BYS	2I0D	2PSV	2V59	3BRN	3HV6
1GJA	1TR7	2C1Q	2ICA	2PVU	2V77	3BU1	3I25
1GJD	1TT1	2C92	2IDZ	2Q2A	2V7T	3BXE	3IK3
1GPK	1TXF	2C94	2IL2	2Q3C	2V7U	3C7I	3IOK
1H22	1U1W	2CEM	2ILZ	2Q54	2V7V	3CCN	3K5C
1H23	1UGX	2CEN	2ISW	2Q6B	2V8Y	3CCW	3KDM
1HA2	1UI0	2CF8	2IWX	2Q6M	2VHW	3CCZ	3KEK
1HNN	1ULD	2CF9	2J2U	2Q88	2VKM	3CD0	3KF7
1HWK	1URG	2CGF	2J34	2Q89	2VPN	3CD7	3L0E
1I7Z	1USI	2CJI	2J4I	2Q8Z	2VTI	3CDA	3OT8
1II5	1USK	2CJW	2J78	2Q93	2VTQ	3CIC	3P8O
1IY7	1UWU	2CN0	2J7G	2Q94	2VU3	3CZV	3PD2
1JGL	1UZ1	2D3U	2JBJ	2Q96	2VW5	3D2R	3PJC
1KZK	1UZV	2D3Z	2JDM	2QBQ	2W26	3D83	3PXY
1LHW	1V0L	2E2R	2JDU	2QBR	2W3K	3DDG	3QBH
1LNM	1VOT	2EPN	2JFZ	2QBS	2W67	3DP4	3QOX
1LRH	1X70	2F3F	2JG0	2QDT	2W6N	3DUY	3R7X
1NC1	1X8R	2F5T	2JGB	2QEH	2W97	3DX0	3VGW
1NC3	1XL5	2FAI	2JH6	2QHY	2X2R	3DX2	3VHI
							4UBP

Table 付録 2　PL セットの PDB ID

PDB ID			
1EZV	2HPY	3H0J	4GX2
1GUY	2J7X	3H52	4MNI
1HQY	2PCK	3HB4	4MS2
1JBV	2PJ9	3HB5	4PBG
1JTE	2Q7K	3IBH	4PDZ
1JTV	2W12	3ISH	4PJ3
1OGV	2WHW	3KPY	4PJA
1OHV	2X58	3P3L	4Q40
1R6T	2XFH	3R9B	4QBI
1VZ4	2YEV	3V94	4R0C
1W29	2YQS	3WCI	4U4W
1YLJ	3AOD	4BO0	4XH4
2C97	3CE0	4CUM	4XIG
2E9D	3EKD	4D1A	4YTZ
2FFU	3FQQ	4D9T	5DKN
2GMV	3GT8	4DSG	5EK0

Table 付録 3　LL セットの PDB ID

PDB ID		
1HXW	3NDW	4ZJL
1N49	3NDX	4ZJO
1PHG	3PRS	4ZJQ
1RL8	3Q70	5IGP
1SH9	3TNE	5IGT
2B60	3U5K	5IGY
2IYF	4EYR	5IH0

2JJP	4M83	5IWU
3AOC	4ZF8	5YK2
3FRQ		

Table 付録 4　CF セットの PDB ID

PDB ID				
1OG5	3JUS	3T3S	4NKX	5JKW
1R9O	3KOH	3T3Z	4NKY	5JL6
1Z10	3LC4	3TBG	4NKZ	5JL7
1Z11	3LD6	3TDA	4NZ2	5JL9
2FDU	3MDM	3UA5	4RQL	5K7K
2FDV	3MDR	3V8D	4RRT	5TFT
2FDW	3MDT	4DVQ	4RUI	5TFU
2FDY	3MDV	4EJG	4UHI	5UAP
2HI4	3N9Y	4EJH	4UHL	5UDA
2NNH	3N9Z	4EJI	4WNT	5UEC
2NNI	3NA0	4EJJ	4WNU	5UFG
2NNJ	3NA1	4ENH	4WNV	5UYS
2P85	3PM0	4FDH	4WNW	5VEU
2Q9F	3QM4	4FIA	4XRY	5W0C
2VN0	3QOA	4GL5	4XRZ	5WBG
3B6H	3QU8	4GL7	4Y8W	5X23
3C6G	3RUK	4GQS	4ZGX	5X24
3CZH	3S79	4I8V	4ZV8	5XXI
3DL9	3S7S	4I91	5A5I	6CHI
3E6I	3SN5	4J14	5A5J	6CIR
3EBS	3SWZ	4KQ8	5IRQ	6CIZ
3EQM	3T3Q	4NKV	5IRV	6DWM
3GPH	3T3R	4NKW	5JKV	6DWN
3IBD				

Table 付録 5　テストセットの PDB ID

PDB ID		
1W0G	4D75	4K9V
2J0D	4D78	4K9W
2V0M	4D7D	4K9X
3NXU	4I4G	5TE8
3TJS	4I4H	5VC0
3UA1	4K9T	5VCE
4D6Z	4K9U	

参考文献

[1] “日本製薬工業協会「てきすとぶっく製薬産業 2016-2017」”.

[2] Toshio. Fujita, Junkichi. Iwasa, and Corwin. Hansch, “A New Substituent Constant, π, Derived from Partition Coefficients,” *J. Am. Chem. Soc.*, vol. 86, no. 23, pp. 5175–5180, Dec. 1964.

[3] Corwin. Hansch and Toshio. Fujita, “p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure,” *J. Am. Chem. Soc.*, vol. 86, no. 8, pp. 1616–1626, Apr. 1964.

[4] 山崎一人, “創薬における数学・AI 活用の現状と将来ビジョン,” 九大-理研-福岡市・ISIT 三者連携シンポジウム「数理・AI が解く未来!~計算科学の展開と期待~」, 2018.

[5] https://www.gsic.titech.ac.jp/tsubame. [Accessed: 29-Jun-2019].

[6] https://www.r-ccs.riken.jp/jp/k/. [Accessed: 29-Jun-2019].

[7] 澤田拓子, “製薬企業から見た AI の可能性,” 第 391 回 CBI 学会講演会　創薬分野における AI 活用の可能性と実際, 2018.

[8] 齋藤大明, “分子ドッキング法を用いたリガンド結合構造予測と分子認識,” 分子シミュレーション研究会会誌“アンサンブル,” vol. 17, no. 2, p. 77, 2015.

[9] 上原彰太 and 田中成典, “タンパク質ーリガンドドッキングの現状と課題,” 日本化学会情報化学部会誌, vol. 34, no. 1, p. 10, 2016.

[10] E. Kellenberger, J. Rodrigo, P. Muller, and D. Rognan, “Comparative evaluation of eight docking tools for docking and virtual screening accuracy,” *Proteins Struct. Funct. Bioinforma.*, vol. 57, no. 2, pp. 225–242, 2004.

[11] D. Plewczynski, M. Łaźniewski, R. Augustyniak, and K. Ginalski, “Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database,” *J. Comput. Chem.*, vol. 32, no. 4, pp. 742–755, 2011.

[12] S. Tian *et al.*, “Assessing an Ensemble Docking-Based Virtual Screening Strategy for Kinase Targets by Considering Protein Flexibility,” *J. Chem. Inf. Model.*, vol. 54, no. 10, pp. 2664–2679, Oct. 2014.

[13] W.-H. Shin and C. Seok, “GalaxyDock: Protein–Ligand Docking with Flexible Protein Side-chains,” *J. Chem. Inf. Model.*, vol. 52, no. 12, pp. 3225–3232, Dec. 2012.

[14] 島田裕三, “Induced-Fit を考慮した Glide/Prime によるタンパク質－リガンド相互作用解析,” *Mol. Sci.*, vol. 6, no. 1, p. NP0022, 2012.

[15] Y. Shan, E. Kim, M. P. Eastwood, R. O. Dror, M. A. Seeliger, and D. E. Shaw, “How Does a Drug Molecule Find its Target Binding Site?,” *J. Am. Chem. Soc.*, vol. 133, no. 24, pp. 9181–9183, Jun. 2011.

[16] R. O. Dror *et al.*, “Pathway and mechanism of drug binding to G-protein-coupled receptors,” *Proc. Natl. Acad. Sci.*, vol. 108, no. 32, pp. 13118–13123, Aug. 2011.

[17] H. Saito, “Development of a pharmacokinetics prediction system using multiscale integrated modeling:5. Prediction of sites of drug metabolism by cytochrome P450 by molecular simulation,” presented at the Chem-Bio Informatics Society(CBI) Annual Meeting 2017, Tokyo, 2017.

[18] Q. U. Ain, A. Aleksandrova, F. D. Roessler, and P. J. Ballester, “Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening,” *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 5, no. 6, pp. 405–424, 2015.

[19] J. Gabel, J. Desaphy, and D. Rognan, “Beware of Machine Learning-Based Scoring Functions—On the Danger of Developing Black Boxes,” *J. Chem. Inf. Model.*, vol. 54, no. 10, pp. 2807–2815, Oct. 2014.

[20] M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri, and D. R. Koes, “Protein–Ligand Scoring with Convolutional Neural Networks,” *J. Chem. Inf. Model.*, vol. 57, no. 4, pp. 942–957, Apr. 2017.

[21] Nakamura H. and Kurisu G., “Data Science and PDBj Activities,” *Seibutsu Butsuri*, vol. 58, no. 2, pp. 071–077, 2018.

[22] M. F. Paine, H. L. Hart, S. S. Ludington, R. L. Haining, A. E. Rettie, and D. C. Zeldin, “The human intestinal cytochrome P450 ‘pie,’” *Drug Metab. Dispos. Biol. Fate Chem.*, vol. 34, no. 5, pp. 880–886, May 2006.

[23] J. Vamathevan *et al.*, “Applications of machine learning in drug discovery and development,” *Nat. Rev. Drug Discov.*, vol. 18, no. 6, p. 463, Jun. 2019.

[24] “Kaggle Competitions.” [Online]. Available: https://www.kaggle.com/competitions. [Accessed: 20-Jul-2019].

[25] 関嶋政和, “オープンイノベーションによる IT 創薬：コンテスト形式による薬剤候補化合物の探索,” 情報管理, vol. 58, no. 12, pp. 900–907, 2016.

[26] “並列生物情報処理イニシアティブ (IPAB).” [Online]. Available: http://www.ipab.org/. [Accessed: 20-Jul-2019].

[27] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, “Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships,” *J. Chem. Inf. Model.*, vol. 55, no. 2, pp. 263–274, Feb. 2015.

[28] T. Zhu *et al.*, “Hit Identification and Optimization in Virtual Screening: Practical

Recommendations Based Upon a Critical Literature Analysis," *J. Med. Chem.*, vol. 56, no. 17, pp. 6560–6572, Sep. 2013.

[29] Basith S. *et al.*, "Exploring G Protein-Coupled Receptors (GPCRs) Ligand Space via Cheminformatics Approaches: Impact on Rational Drug Design," *undefined*, 2018. [Online]. Available: /paper/Exploring-G-Protein-Coupled-Receptors-(GPCRs)-Space-Basith-Cui/d735116885bc97f55fc0df884dd6353341c0d718. [Accessed: 20-Jul-2019].

[30] T. Cheng, Q. Li, Z. Zhou, Y. Wang, and S. H. Bryant, "Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review," *AAPS J.*, vol. 14, no. 1, pp. 133–141, Jan. 2012.

[31] G. Bouvier, N. Evrard-Todeschi, J.-P. Girault, and G. Bertho, "Automatic clustering of docking poses in virtual screening process using self-organizing map," *Bioinformatics*, vol. 26, no. 1, pp. 53–60, Jan. 2010.

[32] M. A. Khamis, W. Gomaa, and W. F. Ahmed, "Machine learning in computational docking," *Artif. Intell. Med.*, vol. 63, no. 3, pp. 135–152, Mar. 2015.

[33] D.-L. Ma, D. Shiu-Hin Chan, and C.-H. Leung, "Drug repositioning by structure-based virtual screening," *Chem. Soc. Rev.*, vol. 42, no. 5, pp. 2130–2141, 2013.

[34] J. D. Durrant and J. A. McCammon, "NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein−Ligand Complexes," *J. Chem. Inf. Model.*, vol. 50, no. 10, pp. 1865–1871, Oct. 2010.

[35] H. M. Ashtawy and N. R. Mahapatra, "Machine-learning scoring functions for identifying native poses of ligands docked to known and novel proteins," *BMC Bioinformatics*, vol. 16, no. 6, p. S3, Apr. 2015.

[36] W. Wang, W. He, X. Zhou, and X. Chen, "Optimization of molecular docking scores with support vector rank regression," *Proteins Struct. Funct. Bioinforma.*, vol. 81, no. 8, pp. 1386–1398, 2013.

[37] I. Wallach, M. Dzamba, and A. Heifets, "AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery," *ArXiv151002855 Cs*, 2015.

[38] J. Gomes, B. Ramsundar, E. N. Feinberg, and V. S. Pande, "Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity," *ArXiv170310603 Cs*, 2017.

[39] F. Imrie, A. R. Bradley, M. van der Schaar, and C. M. Deane, "Protein Family-Specific Models Using Deep Neural Networks and Transfer Learning Improve Virtual Screening and Highlight the Need for More Data," *J. Chem. Inf. Model.*, vol. 58, no. 11, pp. 2319–2330, Nov. 2018.

[40] J. Jiménez, M. Škalič, G. Martínez-Rosell, and G. De Fabritiis, "KDEEP : Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks," *J. Chem. Inf. Model.*, vol. 58, no. 2, pp. 287–296, Feb. 2018.

[41] T. Lau and R. Dror, “Brendan - A Deep Convolutional Network for Representing Latent Features of Protein-Ligand Binding Poses,” p. 9.

[42] M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, and P. Siedlecki, “Development and evaluation of a deep learning model for protein–ligand binding affinity prediction,” *Bioinformatics*, vol. 34, no. 21, pp. 3666–3674, Nov. 2018.

[43] H. Li, K.-S. Leung, M.-H. Wong, and P. J. Ballester, “Low-Quality Structural and Interaction Data Improves Binding Affinity Prediction via Random Forest,” *Molecules*, vol. 20, no. 6, pp. 10947–10962, Jun. 2015.

[44] B. K. Shoichet and I. D. Kuntz, “Protein docking and complementarity,” *J. Mol. Biol.*, vol. 221, no. 1, pp. 327–346, Sep. 1991.

[45] R. L. DesJarlais, R. P. Sheridan, G. L. Seibel, J. S. Dixon, I. D. Kuntz, and R. Venkataraghavan, “Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure,” *J. Med. Chem.*, vol. 31, no. 4, pp. 722–729, Apr. 1988.

[46] E. C. Meng, B. K. Shoichet, and I. D. Kuntz, “Automated docking with grid-based energy evaluation,” *J. Comput. Chem.*, vol. 13, no. 4, pp. 505–524, 1992.

[47] D. E. Koshland, “Correlation of Structure and Function in Enzyme Action,” *Science*, vol. 142, no. 3599, pp. 1533–1541, 1963.

[48] G. G. Hammes, “Multiple Conformational Changes in Enzyme Catalysis,” *Biochemistry*, vol. 41, no. 26, pp. 8221–8228, Jul. 2002.

[49] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, “A Fast Flexible Docking Method using an Incremental Construction Algorithm,” *J. Mol. Biol.*, vol. 261, no. 3, pp. 470–489, Aug. 1996.

[50] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor, “Development and validation of a genetic algorithm for flexible docking11Edited by F. E. Cohen,” *J. Mol. Biol.*, vol. 267, no. 3, pp. 727–748, Apr. 1997.

[51] T. J. A. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz, “DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases,” *J. Comput. Aided Mol. Des.*, vol. 15, no. 5, pp. 411–428, May 2001.

[52] Z. Zsoldos, D. Reid, A. Simon, S. B. Sadjad, and A. P. Johnson, “eHiTS: A new fast, exhaustive flexible ligand docking system,” *J. Mol. Graph. Model.*, vol. 26, no. 1, pp. 198–212, Jul. 2007.

[53] C. M. Venkatachalam, X. Jiang, T. Oldfield, and M. Waldman, “LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites,” *J. Mol. Graph. Model.*, vol. 21, no. 4, pp. 289–307, Jan. 2003.

[54] F. Österberg, G. M. Morris, M. F. Sanner, A. J. Olson, and D. S. Goodsell, “Automated docking to multiple target structures: Incorporation of protein mobility and structural water

heterogeneity in AutoDock," *Proteins Struct. Funct. Bioinforma.*, vol. 46, no. 1, pp. 34–40, 2002.

[55] A. N. Jain, "Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine," *J. Med. Chem.*, vol. 46, no. 4, pp. 499–511, Feb. 2003.

[56] R. A. Friesner *et al.*, "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy," *J. Med. Chem.*, vol. 47, no. 7, pp. 1739–1749, Mar. 2004.

[57] M. R. McGann, H. R. Almond, A. Nicholls, J. A. Grant, and F. K. Brown, "Gaussian docking functions," *Biopolymers*, vol. 68, no. 1, pp. 76–90, Jan. 2003.

[58] C. R. Corbeil, C. I. Williams, and P. Labute, "Variability in docking success rates due to dataset preparation," *J. Comput. Aided Mol. Des.*, vol. 26, no. 6, pp. 775–786, Jun. 2012.

[59] H. Zhao and A. Caflisch, "Discovery of ZAP70 inhibitors by high-throughput docking into a conformation of its kinase domain generated by molecular dynamics," *Bioorg. Med. Chem. Lett.*, vol. 23, no. 20, pp. 5721–5726, Oct. 2013.

[60] O. Trott and A. J. Olson, "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *J. Comput. Chem.*, vol. 31, no. 2, pp. 455–461, 2010.

[61] S. Ruiz-Carmona *et al.*, "rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids," *PLoS Comput. Biol.*, vol. 10, no. 4, Apr. 2014.

[62] W. J. Allen *et al.*, "DOCK 6: Impact of New Features and Current Docking Performance," *J. Comput. Chem.*, vol. 36, no. 15, pp. 1132–1156, Jun. 2015.

[63] N. S. Pagadala, K. Syed, and J. Tuszynski, "Software for molecular docking: a review," *Biophys. Rev.*, vol. 9, no. 2, pp. 91–102, Jan. 2017.

[64] B. K. Shoichet and I. D. Kuntz, "Matching chemistry and shape in molecular docking," *Protein Eng.*, vol. 6, no. 7, pp. 723–732, Sep. 1993.

[65] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. Cambridge, MA, USA: MIT Press, 1992.

[66] N. Metropolis and S. Ulam, "The Monte Carlo method," *J. Am. Stat. Assoc.*, vol. 44, no. 247, pp. 335–341, Sep. 1949.

[67] Z. Wang *et al.*, "Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power," *Phys. Chem. Chem. Phys.*, vol. 18, no. 18, pp. 12964–12975, May 2016.

[68] X. Li, Y. Li, T. Cheng, Z. Liu, and R. Wang, "Evaluation of the performance of four molecular docking programs on a diverse set of protein-ligand complexes," *J. Comput. Chem.*, vol. 31, no. 11, pp. 2109–2125, 2010.

[69] K. Onodera, K. Satou, and H. Hirota, “Evaluations of Molecular Docking Programs for Virtual Screening,” *J. Chem. Inf. Model.*, vol. 47, no. 4, pp. 1609–1618, Jul. 2007.

[70] M. Kontoyianni, L. M. McClellan, and G. S. Sokol, “Evaluation of Docking Performance: Comparative Data on Docking Algorithms,” *J. Med. Chem.*, vol. 47, no. 3, pp. 558–565, Jan. 2004.

[71] B. D. Bursulaya, M. Totrov, R. Abagyan, and C. L. Brooks, “Comparative study of several algorithms for flexible ligand docking,” *J. Comput. Aided Mol. Des.*, vol. 17, no. 11, pp. 755–763, Nov. 2003.

[72] C. Bissantz, G. Folkers, and D. Rognan, “Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations,” *J. Med. Chem.*, vol. 43, no. 25, pp. 4759–4767, Dec. 2000.

[73] B. Kramer, M. Rarey, and T. Lengauer, “Evaluation of the FLEXX incremental construction algorithm for protein–ligand docking,” *Proteins Struct. Funct. Bioinforma.*, vol. 37, no. 2, pp. 228–241, 1999.

[74] H.-M. Chen, B.-F. Liu, H.-L. Huang, S.-F. Hwang, and S.-Y. Ho, “SODOCK: Swarm optimization for highly flexible protein–ligand docking,” *J. Comput. Chem.*, vol. 28, no. 2, pp. 612–623, 2007.

[75] V. Namasivayam and R. Günther, “A Fast Flexible Molecular Docking Program Based on Swarm Intelligence,” *Chem. Biol. Drug Des.*, vol. 70, no. 6, pp. 475–484, 2007.

[76] Y. Liu, L. Zhao, W. Li, D. Zhao, M. Song, and Y. Yang, “FIPSDock: A new molecular docking technique driven by fully informed swarm optimization algorithm,” *J. Comput. Chem.*, vol. 34, no. 1, pp. 67–75, 2013.

[77] R. E. Amaro *et al.*, “Ensemble Docking in Drug Discovery,” *Biophys. J.*, vol. 114, no. 10, pp. 2271–2278, May 2018.

[78] S. B. Nabuurs, M. Wagener, and J. de Vlieg, “A Flexible Approach to Induced Fit Docking,” *J. Med. Chem.*, vol. 50, no. 26, pp. 6507–6518, Dec. 2007.

[79] Clark J. J., Benson M. L., Smith R. D., and Carlson H. A., “Inherent versus induced protein flexibility: Comparisons within and between apo and holo structures,” *PLOS Comput. Biol.*, vol. 15, no. 1, p. e1006705, Jan. 2019.

[80] H. A. Carlson, K. M. Masukawa, and J. A. McCammon, “Method for Including the Dynamic Fluctuations of a Protein in Computer-Aided Drug Design,” *J. Phys. Chem. A*, vol. 103, no. 49, pp. 10213–10219, Dec. 1999.

[81] H. A. Carlson *et al.*, “Developing a Dynamic Pharmacophore Model for HIV-1 Integrase,” *J. Med. Chem.*, vol. 43, no. 11, pp. 2100–2114, Jun. 2000.

[82] K. Kapoor, N. McGill, C. B. Peterson, H. V. Meyers, M. N. Blackburn, and J. Baudry, “Discovery of Novel Nonactive Site Inhibitors of the Prothrombinase Enzyme Complex,” *J. Chem. Inf. Model.*, vol. 56, no. 3, pp. 535–547, Mar. 2016.

[83] Z. Xiao *et al.*, “A computationally identified compound antagonizes excess FGF-23 signaling in renal tubules and a mouse model of hypophosphatemia,” *Sci. Signal.*, vol. 9, no. 455, pp. ra113–ra113, Nov. 2016.

[84] R. E. Amaro, R. Baron, and J. A. McCammon, “An improved relaxed complex scheme for receptor flexibility in computer-aided drug design,” *J. Comput. Aided Mol. Des.*, vol. 22, no. 9, pp. 693–705, Sep. 2008.

[85] D. E. Shaw *et al.*, “Millisecond-scale Molecular Dynamics Simulations on Anton,” in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, New York, NY, USA, 2009, pp. 39:1–39:11.

[86] X. Hu, L. Hong, M. Dean Smith, T. Neusius, X. Cheng, and J. C. Smith, “The dynamics of single protein molecules is non-equilibrium and self-similar over thirteen decades in time,” *Nat. Phys.*, vol. 12, no. 2, pp. 171–174, Feb. 2016.

[87] S.-Y. Huang, S. Z. Grinter, and X. Zou, “Scoring functions and their evaluation methods for protein –ligand docking: recent advances and future directions,” *Phys. Chem. Chem. Phys.*, vol. 12, no. 40, pp. 12899–12908, 2010.

[88] J. Liu and R. Wang, “Classification of Current Scoring Functions,” *J. Chem. Inf. Model.*, vol. 55, no. 3, pp. 475–482, Mar. 2015.

[89] J. Li, A. Fu, and L. Zhang, “An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking,” *Interdiscip. Sci. Comput. Life Sci.*, Mar. 2019.

[90] J. Michel, J. Tirado-Rives, and W. L. Jorgensen, “Prediction of the Water Content in Protein Binding Sites,” *J. Phys. Chem. B*, vol. 113, no. 40, pp. 13337–13346, Oct. 2009.

[91] Ross G. A., Morris G. M., and Biggin P. C., “Rapid and Accurate Prediction and Scoring of Water Molecules in Protein Binding Sites,” *PLOS ONE*, vol. 7, no. 3, p. e32036, Mar. 2012.

[92] Y. Yang, F. C. Lightstone, and S. E. Wong, “Approaches to efficiently estimate solvation and explicit water energetics in ligand binding: the use of WaterMap,” *Expert Opin. Drug Discov.*, vol. 8, no. 3, pp. 277–287, Mar. 2013.

[93] A. Kumar and K. Y. J. Zhang, “Investigation on the Effect of Key Water Molecules on Docking Performance in CSARdock Exercise,” *J. Chem. Inf. Model.*, vol. 53, no. 8, pp. 1880–1892, Aug. 2013.

[94] S. Uehara and S. Tanaka, “AutoDock-GIST: Incorporating Thermodynamics of Active-Site Water into Scoring Function for Accurate Protein-Ligand Docking,” *Molecules*, vol. 21, no. 11, p. 1604, Nov. 2016.

[95] H. Sun, Y. Li, D. Li, and T. Hou, “Insight into Crizotinib Resistance Mechanisms Caused by Three Mutations in ALK Tyrosine Kinase using Free Energy Calculation Approaches,” *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 2376–2389, Sep. 2013.

[96] P. Chaskar, V. Zoete, and U. F. Röhrig, “On-the-Fly QM/MM Docking with Attracting

Cavities," *J. Chem. Inf. Model.*, vol. 57, no. 1, pp. 73–84, Jan. 2017.

[97] H. J. Kulik, "Large-scale QM/MM free energy simulations of enzyme catalysis reveal the influence of charge transfer," *Phys. Chem. Chem. Phys.*, vol. 20, no. 31, pp. 20650–20660, Aug. 2018.

[98] Y. Orozco-Gonzalez *et al.*, "An Average Solvent Electrostatic Configuration Protocol for QM/MM Free Energy Optimization: Implementation and Application to Rhodopsin Systems," *J. Chem. Theory Comput.*, vol. 13, no. 12, pp. 6391–6404, Dec. 2017.

[99] M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini, and R. P. Mee, "Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes," *J. Comput. Aided Mol. Des.*, vol. 11, no. 5, pp. 425–445, Sep. 1997.

[100] C. W. Murray, T. R. Auton, and M. D. Eldridge, "Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of Bayesian regression to improve the quality of the model," *J. Comput. Aided Mol. Des.*, vol. 12, no. 5, pp. 503–519, Sep. 1998.

[101] R. A. Friesner *et al.*, "Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes," *J. Med. Chem.*, vol. 49, no. 21, pp. 6177–6196, Oct. 2006.

[102] Z. Zheng and K. M. Merz, "Ligand Identification Scoring Algorithm (LISA)," *J. Chem. Inf. Model.*, vol. 51, no. 6, pp. 1296–1306, Jun. 2011.

[103] I. Muegge and Y. C. Martin, "A General and Fast Scoring Function for Protein−Ligand Interactions: A Simplified Potential Approach," *J. Med. Chem.*, vol. 42, no. 5, pp. 791–804, Mar. 1999.

[104] H. Gohlke, M. Hendlich, and G. Klebe, "Knowledge-based scoring function to predict protein-ligand interactions11Edited by R. Huber," *J. Mol. Biol.*, vol. 295, no. 2, pp. 337–356, Jan. 2000.

[105] L. Zhang *et al.*, "Virtual screening approach to identifying influenza virus neuraminidase inhibitors using molecular docking combined with machine-learning-based scoring function," *Oncotarget*, vol. 8, no. 47, pp. 83142–83154, Sep. 2017.

[106] M. Karplus and J. A. McCammon, "Molecular dynamics simulations of biomolecules," *Nat. Struct. Biol.*, vol. 9, no. 9, p. 646, Sep. 2002.

[107] J. A. McCammon, B. R. Gelin, and M. Karplus, "Dynamics of folded proteins," *Nature*, vol. 267, no. 5612, p. 585, Jun. 1977.

[108] S. Lifson and A. Warshel, "Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and n-Alkane Molecules," *J. Chem. Phys.*, vol. 49, no. 11, pp. 5116–5129, Dec. 1968.

[109] M. Levitt and S. Lifson, "Refinement of protein conformations using a

macromolecular energy minimization procedure," *J. Mol. Biol.*, vol. 46, no. 2, pp. 269–279, Dec. 1969.

[110] S. A. Hollingsworth and R. O. Dror, "Molecular Dynamics Simulation for All," *Neuron*, vol. 99, no. 6, pp. 1129–1143, Sep. 2018.

[111] D. E. Shaw *et al.*, "Anton, a Special-purpose Machine for Molecular Dynamics Simulation," *Commun ACM*, vol. 51, no. 7, pp. 91–97, Jul. 2008.

[112] R. Salomon-Ferrer, A. W. Götz, D. Poole, S. Le Grand, and R. C. Walker, "Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald," *J. Chem. Theory Comput.*, vol. 9, no. 9, pp. 3878–3888, Sep. 2013.

[113] S. Bernèche and B. Roux, "Energetics of ion conduction through the K + channel," *Nature*, vol. 414, no. 6859, p. 73, Nov. 2001.

[114] J. Li, S. A. Shaikh, G. Enkavi, P.-C. Wen, Z. Huang, and E. Tajkhorshid, "Transient formation of water-conducting states in membrane transporters," *Proc. Natl. Acad. Sci.*, vol. 110, no. 19, pp. 7696–7701, May 2013.

[115] K. Khafizov *et al.*, "Investigation of the sodium-binding sites in the sodium-coupled betaine transporter BetP," *Proc. Natl. Acad. Sci.*, vol. 109, no. 44, pp. E3035–E3044, Oct. 2012.

[116] A. J. Clark *et al.*, "Prediction of Protein–Ligand Binding Poses via a Combination of Induced Fit Docking and Metadynamics Simulations," *J. Chem. Theory Comput.*, vol. 12, no. 6, pp. 2990–2998, Jun. 2016.

[117] A. Koehl *et al.*, "Structure of the μ-opioid receptor–G i protein complex," *Nature*, vol. 558, no. 7711, p. 547, Jun. 2018.

[118] J. D. Durrant and J. A. McCammon, "Molecular dynamics simulations and drug discovery," *BMC Biol.*, vol. 9, no. 1, p. 71, Oct. 2011.

[119] D. W. Borhani and D. E. Shaw, "The future of molecular dynamics simulations in drug discovery," *J. Comput. Aided Mol. Des.*, vol. 26, no. 1, pp. 15–26, Jan. 2012.

[120] M. Udier-Blagović, J. Tirado-Rives, and W. L. Jorgensen, "Validation of a Model for the Complex of HIV-1 Reverse Transcriptase with Nonnucleoside Inhibitor TMC125," *J. Am. Chem. Soc.*, vol. 125, no. 20, pp. 6016–6017, May 2003.

[121] V. Spahn *et al.*, "A nontoxic pain killer designed by modeling of pathological receptor conformations," *Science*, vol. 355, no. 6328, pp. 966–969, Mar. 2017.

[122] K. Kappel, Y. Miao, and J. A. McCammon, "Accelerated molecular dynamics simulations of ligand binding to a muscarinic G-protein-coupled receptor," *Q. Rev. Biophys.*, vol. 48, no. 4, pp. 479–487, Nov. 2015.

[123] R. O. Dror *et al.*, "Structural basis for modulation of a G-protein-coupled receptor by allosteric drugs," *Nature*, vol. 503, no. 7475, pp. 295–299, Nov. 2013.

[124] A. Perez, J. A. Morrone, C. Simmerling, and K. A. Dill, "Advances in free-energy-

based simulations of protein folding and ligand binding," *Curr. Opin. Struct. Biol.*, vol. 36, pp. 25–31, Feb. 2016.

[125] L. Wang *et al.*, "Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field," *J. Am. Chem. Soc.*, vol. 137, no. 7, pp. 2695–2703, Feb. 2015.

[126] D. L. Mobley and K. A. Dill, "Binding of Small-Molecule Ligands to Proteins: 'What You See' Is Not Always 'What You Get,'" *Structure*, vol. 17, no. 4, pp. 489–498, Apr. 2009.

[127] H. Fujitani *et al.*, "Direct calculation of the binding free energies of FKBP ligands," *J. Chem. Phys.*, vol. 123, no. 8, p. 084108, 2005.

[128] T. Hou, J. Wang, Y. Li, and W. Wang, "Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations," *J. Chem. Inf. Model.*, vol. 51, no. 1, pp. 69–82, Jan. 2011.

[129] J. Schlitter, M. Engels, and P. Krüger, "Targeted molecular dynamics: A new approach for searching pathways of conformational transitions," *J. Mol. Graph.*, vol. 12, no. 2, pp. 84–89, Jun. 1994.

[130] M. P. Luitz and M. Zacharias, "Protein–Ligand Docking Using Hamiltonian Replica Exchange Simulations with Soft Core Potentials," *J. Chem. Inf. Model.*, vol. 54, no. 6, pp. 1669–1675, Jun. 2014.

[131] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chem. Phys. Lett.*, vol. 314, no. 1, pp. 141–151, Nov. 1999.

[132] L. Maragliano and E. Vanden-Eijnden, "A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations," *Chem. Phys. Lett.*, vol. 426, no. 1, pp. 168–175, Jul. 2006.

[133] D. Sabbadin and S. Moro, "Supervised Molecular Dynamics (SuMD) as a Helpful Tool To Depict GPCR–Ligand Recognition Pathway in a Nanosecond Time Scale," *J. Chem. Inf. Model.*, vol. 54, no. 2, pp. 372–376, Feb. 2014.

[134] D. Hamelberg, J. Mongan, and J. A. McCammon, "Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules," *J. Chem. Phys.*, vol. 120, no. 24, pp. 11919–11929, Jun. 2004.

[135] C. A. F. de Oliveira, D. Hamelberg, and J. A. McCammon, "On the Application of Accelerated Molecular Dynamics to Liquid Water Simulations," *J. Phys. Chem. B*, vol. 110, no. 45, pp. 22695–22701, Nov. 2006.

[136] D. R. Koes, M. P. Baumgartner, and C. J. Camacho, "Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise," *J. Chem. Inf. Model.*, vol. 53, no. 8, pp. 1893–1904, Aug. 2013.

[137] J. Hochuli, A. Helbling, T. Skaist, M. Ragoza, and D. R. Koes, “Visualizing convolutional neural network protein-ligand scoring,” *J. Mol. Graph. Model.*, vol. 84, pp. 96–108, Sep. 2018.

[138] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” *ArXiv13112524 Cs*, Nov. 2013.

[139] J. Donahue *et al.*, “DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition,” *ArXiv13101531 Cs*, Oct. 2013.

[140] Razavian A. S., Azizpour H., Sullivan J., and Carlsson S., “CNN Features off-the-shelf: an Astounding Baseline for Recognition,” *ArXiv14036382 Cs*, Mar. 2014.

[141] Agrawal P., Girshick R., and Malik J., “Analyzing the Performance of Multilayer Neural Networks for Object Recognition,” *ArXiv14071610 Cs*, Jul. 2014.

[142] J. B. Dunbar *et al.*, “CSAR Benchmark Exercise of 2010: Selection of the Protein–Ligand Complexes,” *J. Chem. Inf. Model.*, vol. 51, no. 9, pp. 2036–2046, Sep. 2011.

[143] J.-I. Ito, Y. Tabei, K. Shimizu, K. Tsuda, and K. Tomii, “PoSSuM: a database of similar protein-ligand binding and putative pockets,” *Nucleic Acids Res.*, vol. 40, no. D1, pp. D541–D548, Jan. 2012.

[144] H. M. Berman *et al.*, “The Protein Data Bank,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, Jan. 2000.

[145] *Molecular Operating Environment (MOE), 2013.01; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2013.* .

[146] “Pfam.” [Online]. Available: http://pfam.xfam.org/. [Accessed: 15-Dec-2019].

[147] A. Sato, H. Yuki, C. Watanabe, J. Saito, A. Konagaya, and T. Honma, “Prediction of the site of CYP3A4 metabolism of tolterodine by molecular dynamics simulation from multiple initial structures of the CYP3A4-tolterodine complex,” *Chem-Bio Inform. J.*, vol. 17, pp. 38–52, May 2017.

[148] Jia Y. *et al.*, “Caffe: Convolutional Architecture for Fast Feature Embedding,” *ArXiv14085093 Cs*, Jun. 2014.

[149] J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve.,” *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982.

[150] “RCSB PDB: Structure Classification and Analysis.” [Online]. Available: https://www.rcsb.org/pages/thirdparty/structure_classification. [Accessed: 15-Apr-2019].

[151] Y. Lin *et al.*, “Substrate Inhibition Kinetics for Cytochrome P450-Catalyzed Reactions,” *Drug Metab. Dispos.*, vol. 29, no. 4, pp. 368–374, Jan. 2001.

[152] U. M. Zanger and M. Schwab, “Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation,” *Pharmacol. Ther.*, vol. 138, no. 1, pp. 103–141, Apr. 2013.

[153] L. Di and E. H. Kerns, *Drug-like Properties: Concepts, Structure Design and Methods:*

from ADME to Toxicity Optimization. Academic Press, 2010.

[154] M. B. Fisher, K. R. Henne, and J. Boer, “The complexities inherent in attempts to decrease drug clearance by blocking sites of CYP-mediated metabolism,” *Curr. Opin. Drug Discov. Devel.*, vol. 9, no. 1, pp. 101–109, Jan. 2006.

[155] H.-J. Böhm *et al.*, “Fluorine in Medicinal Chemistry,” *ChemBioChem*, vol. 5, no. 5, pp. 637–643, May 2004.

[156] L. Olsen, C. Oostenbrink, and F. S. Jørgensen, “Prediction of cytochrome P450 mediated metabolism,” *Adv. Drug Deliv. Rev.*, vol. 86, pp. 61–71, Jun. 2015.

[157] S. Shaik, D. Kumar, S. P. de Visser, A. Altun, and W. Thiel, “Theoretical Perspective on the Structure and Mechanism of Cytochrome P450 Enzymes,” *Chem. Rev.*, vol. 105, no. 6, pp. 2279–2328, Jun. 2005.

[158] B. Meunier, S. P. de Visser, and S. Shaik, “Mechanism of Oxidation Reactions Catalyzed by Cytochrome P450 Enzymes,” *Chem. Rev.*, vol. 104, no. 9, pp. 3947–3980, Sep. 2004.

[159] S. Shaik, S. Cohen, Y. Wang, H. Chen, D. Kumar, and W. Thiel, “P450 Enzymes: Their Structure, Reactivity, and Selectivity—Modeled by QM/MM Calculations,” *Chem. Rev.*, vol. 110, no. 2, pp. 949–1017, Feb. 2010.

[160] R. Liu, J. Liu, G. Tawa, and A. Wallqvist, “2D SMARTCyp Reactivity-Based Site of Metabolism Prediction for Major Drug-Metabolizing Cytochrome P450 Enzymes,” *J. Chem. Inf. Model.*, vol. 52, no. 6, pp. 1698–1712, Jun. 2012.

[161] P. Vasanthanathan *et al.*, “Virtual Screening and Prediction of Site of Metabolism for Cytochrome P450 1A2 Ligands,” *J. Chem. Inf. Model.*, vol. 49, no. 1, pp. 43–52, Jan. 2009.

[162] T. Huang, J. Zaretzki, C. Bergeron, K. P. Bennett, and C. M. Breneman, “DR-Predictor: Incorporating Flexible Docking with Specialized Electronic Reactivity and Machine Learning Techniques to Predict CYP-Mediated Sites of Metabolism,” *J. Chem. Inf. Model.*, vol. 53, no. 12, pp. 3352–3366, Dec. 2013.

[163] J. D. Tyzack, M. J. Williamson, R. Torella, and R. C. Glen, “Prediction of Cytochrome P450 Xenobiotic Metabolism: Tethered Docking and Reactivity Derived from Ligand Molecular Orbital Analysis,” *J. Chem. Inf. Model.*, vol. 53, no. 6, pp. 1294–1305, Jun. 2013.

[164] H. Yuki, T. Honma, M. Hata, and T. Hoshino, “Prediction of sites of metabolism in a substrate molecule, instanced by carbamazepine oxidation by CYP3A4,” *Bioorg. Med. Chem.*, vol. 20, no. 2, pp. 775–783, Jan. 2012.

[165] P. Van Kerrebroeck, K. Kreder, U. Jonas, N. Zinner, and A. Wein, “Tolterodine once-daily: superior efficacy and tolerability in the treatment of the overactive bladder1,” *Urology*, vol. 57, no. 3, pp. 414–421, Mar. 2001.

[166] N. Brynne *et al.*, “Pharmacokinetics and pharmacodynamics of tolterodine in man: a new drug for the treatment of urinary bladder overactivity,” *Int. J. Clin. Pharmacol. Ther.*,

vol. 35, no. 7, pp. 287–295, Jul. 1997.

[167] Z. Deng, C. Chuaqui, and J. Singh, “Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein−Ligand Binding Interactions,” *J. Med. Chem.*, vol. 47, no. 2, pp. 337–344, Jan. 2004.

[168] S. Soga, H. Shirai, M. Kobori, and N. Hirayama, “Use of Amino Acid Composition to Predict Ligand-Binding Sites,” *J. Chem. Inf. Model.*, vol. 47, no. 2, pp. 400–406, Mar. 2007.

[169] “MacQueen, J. B, Berkeley, University of California Press. 1967, 1, 281-297.” [Online]. Available: http://www-m9.ma.tum.de/foswiki/pub/WS2010/CombOptSem/kMeans.pdf. [Accessed: 06-Jun-2016].

[170] E. Sano *et al.*, “Mechanism of the decrease in catalytic activity of human cytochrome P450 2C9 polymorphic variants investigated by computational analysis,” *J. Comput. Chem.*, vol. 31, no. 15, pp. 2746–2758, Nov. 2010.

[171] “Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. Gaussian 03; Gaussian, Inc., Wallingford, CT, 2004.,” 2004.

[172] P. Cieplak, W. D. Cornell, C. Bayly, and P. A. Kollman, “Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins,” *J. Comput. Chem.*, vol. 16, no. 11, pp. 1357–1377, Nov. 1995.

[173] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, “Comparison of simple potential functions for simulating liquid water,” *J. Chem. Phys.*, vol. 79, no. 2, p. 926, 1983.

[174] Y. Duan *et al.*, “A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations,” *J. Comput. Chem.*, vol. 24, no. 16, pp. 1999–2012, 2003.

[175] “D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke,

R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A.W. Götz, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, R.M. Wolf, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P.A. Kollman (2012), AMBER 12, University of California, San Francisco.," 2012.

[176] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes," *J. Comput. Phys.*, vol. 23, no. 3, pp. 327–341, 1977.

[177] E. G. Hrycay and S. M. Bandiera, "Monooxygenase, Peroxidase and Peroxygenase Properties and Reaction Mechanisms of Cytochrome P450 Enzymes," in *Monooxygenase, Peroxidase and Peroxygenase Properties and Mechanisms of Cytochrome P450*, vol. 851, E. G. Hrycay and S. M. Bandiera, Eds. Cham: Springer International Publishing, 2015, pp. 1–61.

[178] J. K. Yano, M. R. Wester, G. A. Schoch, K. J. Griffin, C. D. Stout, and E. F. Johnson, "The Structure of Human Microsomal Cytochrome P450 3A4 Determined by X-ray Crystallography to 2.05-A Resolution," *J. Biol. Chem.*, vol. 279, no. 37, pp. 38091–38094, Sep. 2004.

[179] P. A. Williams, "Crystal Structures of Human Cytochrome P450 3A4 Bound to Metyrapone and Progesterone," *Science*, vol. 305, no. 5684, pp. 683–686, Jul. 2004.

[180] M. Ekroos and T. Sjögren, "Structural basis for ligand promiscuity in cytochrome P450 3A4," *Proc. Natl. Acad. Sci.*, vol. 103, no. 37, pp. 13682–13687, Sep. 2006.

[181] I. F. Sevrioukova and T. L. Poulos, "Structure and mechanism of the complex between cytochrome P4503A4 and ritonavir," *Proc. Natl. Acad. Sci.*, vol. 107, no. 43, pp. 18422–18427, 2010.

[182] I. F. Sevrioukova and T. L. Poulos, "Structural and Mechanistic Insights into the Interaction of Cytochrome P4503A4 with Bromoergocryptine, a Type I Ligand," *J. Biol. Chem.*, vol. 287, no. 5, pp. 3510–3517, Jan. 2012.

[183] "Dassault Systèmes BIOVIA, *PipelinePilot*, Release 9.2, San Diego: Dassault Systèmes, 2014."

[184] M. Awale, R. van Deursen, and J.-L. Reymond, "MQN-Mapplet: Visualization of Chemical Space with Interactive Maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13," *J. Chem. Inf. Model.*, vol. 53, no. 2, pp. 509–518, Feb. 2013.

[185] L. Ruddigkeit, L. C. Blum, and J.-L. Reymond, "Visualization and Virtual Screening of the Chemical Universe Database GDB-17," *J. Chem. Inf. Model.*, vol. 53, no. 1, pp. 56–65, Jan. 2013.

[186] Bushati N., Smith J., Briscoe J., and Watkins C., "An intuitive graphical visualization

technique for the interrogation of transcriptome data," *Nucleic Acids Res.*, vol. 39, no. 17, pp. 7380–7389, Sep. 2011.

[187] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[188] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *ArXiv160806993 Cs*, Aug. 2016.