（ 続紙 1 ）

京都大学	博士（情報学）	氏名	塩田　光司
論文題目	Machine Learning-Based Methods for Predicting the Most Stable Conformation and Binding Affinity of Protein-Drug Complexes （機械学習に基づくタンパク質-薬剤分子の最安定配座および結合親和性の予測手法）		

（論文内容の要旨）

創薬においては多数の化合物のリストから対象疾患に対する新薬の候補となる化合物（薬剤分子）を計算機の支援により選択するバーチャルスクリーニングとよばれる過程が重要である。バーチャルスクリーニングにはいくつかの部分課題があるが、本論文では計算機による処理の重要度が高いと考えられる二種類の課題を対象としている。具体的には、指定されたタンパク質と薬剤分子間の結合における最安定配座の予測、および、複数の薬剤分子候補が与えられた際のタンパク質と各薬剤分子間の結合親和性の予測を対象としている。それぞれの課題に対し新規の機械学習手法を開発し、計算機実験によりその有効性を評価している。

第1章は序論であり、バーチャルスクリーニングの重要性、バーチャルスクリーニングにおける４個の部分課題について説明している。特に本論文で対象とする最安定配座計算および結合親和性の予測手法について、最近の手法の問題点について説明し、それを解決するために本論文において提案した二種類の手法の概要を説明している。その後で、本論文の構成を示している。

第2章では、本研究において使用するデータ、特にPDBbindというデータベース、および、タンパク質-薬剤分子の結合の配置や親和性の評価に広く用いられているCASF-2016というデータ、および、LIT-PCBAというデータについて詳細に説明している。さらに、本論文で用いる機械学習手法である深層学習および勾配ブースティング木について背景と概要を簡潔に説明している。一方、本論文では量子化学計算を模倣する深層学習手法を開発しているため、量子化学計算手法についても概観している。

第3章では、タンパク質-薬剤分子が結合する際の最安定配座を予測するための新規手法ADQnetについて述べている。最安定配座の予測のためには、結合に関するエネルギーを量子化学計算により計算することが必要になるが、多数の候補配座についてその計算を実行するのは困難である。そこで、SQM/COSMOという量子化学計算に広く用いられる計算を深層学習により模倣するためのACSFという計算法が開発されていた。しかし、ACSFでは2分子系であるタンパク質-薬剤分子間の結合予測には十分に対応できないという問題があるため、それを解決するための新たな特徴量を提案している。さらに、効果的に深層学習手法を適用するためには大量のデータが必要となるが、既存のデータベースには十分な量のデータがないという問題点もあったため、最安定ではないがそれに近い配座を大量に発生させる手法も提案している。これらを統合して得られた予測手法がADQnetであり、CASF-2016データおよびLIT-PCBAデータを用いて、多数の既存手法と比較することにより、その有効性を示している。

第4章では、タンパク質-薬剤分子間の配座データから、その親和性を予測するための新規手法について述べている。結合親和性の予測のために、従来、各原子を原子の種類や環境などに基づく6種類の特徴量で表現し、それに勾配ブースティング木という機械学習モデルを適用したECIFという手法が提案されていた。しかし、ECIFでは、結合親和性を評価する際に重要となる原子間距離が考慮されていないという問題点がある。そこで原子間距離を特徴量に組み込んだMulti-Shelled ECIF、および、元素の組み合わせごとに原子間距離で重みづけしたWeighted ECIFという特徴量を新規に提案している。そして、CASF-2016データ、LIT-PCBAデータ、さらには、その他のCASFデータを用いたECIFを含む他手法との計算機実験による比較などを行い、 二種類の提案手法、とりわけMulti-Shelled ECIFの有効性を示している。

第5章は結論であり、提案手法についてのまとめと意義、および、今後の課題が述べられている。

（続紙 ２ ）

（論文審査の結果の要旨）

本論文は、薬剤設計におけるバーチャルスクリーニングという過程におけるタンパク質-薬剤分子の最安定配座予測および結合親和性予測それぞれについての機械学習に基づく手法について述べたものであり、得られた成果は以下のとおりである。

（１）タンパク質-薬剤分子の最安定配座の予測精度向上のために、結晶構造からの多数の配座生成によるデータ増強と、原子間の相互作用を考慮した新たな特徴量を用いた深層学習手法による量子化学計算の模倣を組み合わせたADQnetという手法を開発した。実験データからの結合親和性を用いた標準的なベンチマークデータであるCASF-2016を用いた計算機実験によりADQnetと約30種類の既存手法と比較した結果、標的タンパク質に対するリガンドの最安定配座を予測する性能を測るCASF-2016 Docking power testにおいて92.6%という最良の正解率を達成した。さらに、標的タンパク質と薬剤との結合活性を実験的に評価したベンチマークであるLIT-PCBAデータを用いた計算機実験によりEFスコア（活性化合物の割合に基づくスコア）をもとにした複数の指標により5種類の最先端手法との比較を行ったところ、15種類の標的タンパク質のうち、3種類においてほかの手法よりも優れた性能を示した。これらの結果は、最安定配座予測におけるADQnetの有効性を示すものとなっている。

（２）タンパク質-薬剤分子の結合親和性の予測精度向上のために、既存の特徴量であるECIFをもとに、原子間距離を考慮した特徴量を導入したMulti-Shelled ECIFおよびWeighted ECIFという新規特徴量を設計し、これと勾配ブースティング木という機械学習手法を組み合わせた予測手法を開発した。Multi-Shelled ECIFをベンチマークデータであるCASF-2016 Scoring power testで評価を行った結果、結合親和性予測に特化した既存手法との比較においてPeirson's R 0.877という最も優れた性能を示した。また、Orhoborらが定義したCASF-2007、CASF-2013、CASF-2016、CASF-2019を用いてECIFを含む他の特徴量と比較したところ、CASF-2013以外においては最良の相関係数を達成した。また、前述のLIT-PCBAデータにおいてEFスコアに基づく指標を用いてMulti-Shelled ECIFとECIFを比較した結果、Multi-Shelled ECIFは15種類の標的タンパク質中9種類においてECIFより優れた結果を得た。Weighted ECIFについては、CASF-2016データを用いた場合はMulti-Shelled ECIFより低い相関係数となったが、LIT-PCBAデータを用いた場合は15種類中11種類でECIFより優れた結果を得た。これらの結果は、結合親和性予測におけるMulti-Shelled ECIFおよびWeighted ECIFの両方の有効性を示すものとなっている。

以上、本論文ではタンパク質-薬剤分子の最安定配座および結合親和性の予測というバイオインフォマティクスおよび情報学における重要な研究課題に取り組み、いずれにおいても新たな手法を開発し、それぞれの手法を実際のベンチマークデータを用いた計算機実験により評価した。提案手法のいずれもが新規性、有用性が高く、当該分野の発展のために十分な寄与をしている。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和６年７月２３日、論文内容とそれに関連した事項について試問を行った結果、合格と認めた。なお、本論文のインターネットでの全文公表についても支障がないことを確認した。