Graduate School of Science and Technology Doctoral Thesis Abstract

Laboratory name (Supervisor))	Data Driven Chemistry Laboratory (Prof. Yukiharu Uraoka)		
Student ID	2121438	Submission date	2024 / 12 / 17 (yyyy/mm/dd)
Name (Surname)(given)	Syahdi, Rezi Riadhi		
Title	Hybrid structure- and ligand-based approach to virtual screening for rational drug discovery		

Drug discovery (DD) has always been a significant topic. Virtual screening (VS), is a process based on computer simulation to select/filter the bioactive compounds, and it has been integrated as an essential step to minimize the DD cost and resources.1) VS can be classified into two major types: ligand- and structure-based VS. DD protocols frequently combine these two VS approaches to enhance the result.2)

We have investigated a hybrid of the ligand- and structure-based approaches to enhance VS performances. The results of our research were described in a five-chapter thesis. In **Chapter 1**, the background of the research is described, including the connection of previous research to our research, the layout of our research, the research question, and the research aim.

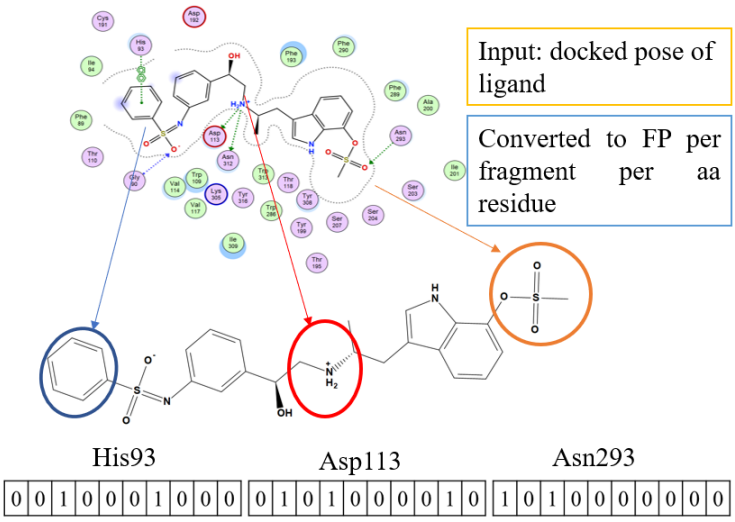

Fig.1. Schematic of the proposed hybrid fingerprint

In **Chapter 2**, several machine learning (ML) using LBVS and SBVS fingerprints were explored, including a parallel virtual screening method through consensus scoring of these fingerprints and their results on the Structure Activity Relationship Matrix (SARM)s data sets of six diverse targets. The SARMs data sets used in our research are implementations of limited information data sets that might occur during the early stage of drug discoveries.3) From the result of work in this part, we conclude that even using consensus scoring, which is a form of a meta-level combination of LBVS and SBVS strategies, has not enhanced the performance of a single VS result. Therefore, a more powerful combination strategy needs to be developed and tested.

In **Chapter 3**, we delve deeper into the combination of VS at a methodological level, termed as hybrid

VS approach. The hybrid approach is expected to utilize more information than singular information.4) We compared three types of hybrid approaches: (i) simple concatenation model; (ii) our new proposed approach, fragmented interaction fingerprint (FIFI)5); and (iii) a state-of-the-art example of a hybrid VS approach, protein-ligand extended connectivity (PLEC).6) In the case of PLEC, it lacks of consideration of the amino acid residue order which is vital for some targets. In the other hand, our FIFI approach retains this kind of information.

FIFI (Figure 1) captures fragmented molecular structures within a defined distance from each amino acid residue, representing them as a bit array (FIFI bit array / FIFI BA) or as one-hot encoded bits for unique substructures (FIFI unique SMARTs / FIFI US). Then we evaluated machine learning (ML) performance using FIFI as input and compared it with other various VS techniques to previously made restricted and structurally similar data sets in six distinct target receptors.3) In general, FIFI demonstrated improved performance over PLEC and several other VS approaches in the SARMs data sets.5)

In **Chapter 4**, we also tried to develop the binary vectors contained by FIFI into other representations to see whether a more descriptive feature would help the performance. This is owing to the vast sparse matrix generated in FIFI that could potentially cause the computation to suffer from inefficiency. We investigated alternative embedding representation using two different deep learning models: (1) bidirectional encoded representations from transformers (BERT), and (2) graph neural networks (GNN). The GNN models were made from two previously published research: (i) GNN-DTI7) which uses graph attention network (GAT) through the aggregation of atom pairs and (ii) LGN8) which is itself a fusion model of complex protein-ligand GNN, ligand graph isomorphism network (GIN) and interaction fingerprint. In this chapter, we compared the performance of various VS fingerprints and the DL models in 33 selected targets of Docking Score ML as external data sets.9) However, all three deep learning models generally have not provided better results than FIFI. Nonetheless, more viable embedding and architecture could be explored and compared in the future to get better representation.

In **Chapter 5**, the conclusion and future research direction on top of the current work i.e., perspective are provided. FIFI is a novel hybrid approach to combine the information from both SBVS and LBVS parts. It shows better performance than the previous state-of-the-art hybrid approach in the SARMs and external data sets, and better than the deep learning model representations in the external data sets. However, due to its dependence on both SBVS and LBVS information, inaccurate data from either approach could deter hybrid VS performance, including FIFI. Therefore, high-quality data, and a precaution are needed to make the utmost result from hybrid VS practice in the future.

References: [1] Gimeno et al, Int. J. Mol. Sci., 2019. 20(6):1375. [2] Drwal & Griffith, Drug Discov. Today, 2013. 10(3):e395-e401. [3] Maeda et al, J. Comput Aided Mol Des, 2022. 36:237-252. [4] Vázquez et al, Molecules, 2020. 25(20):472. [5] Syahdi et al, ACS Omega, 2024. 9(37):38957-38969. [6] Wójcikowski et al. Bioinformatics, 2019. 15;35(8):1334-1341.[7] Lim et al, J. Chem. Inf. Model, 2019. 59(9):3981-3988. [8] Jia G, PloS One, 2024. 19(1):e0296676. [9] Liu et al, J. Chem. Inf. Model, 2024. 64(14):5413-54.

氏　名	REZI RIADHI SYAHDI

（論文審査結果の要旨）

効率的な低分子創薬のためにコンピュータを利用した化合物仮想スクリーニングが行われる。仮想スクリーニングでは、活性化合物を多数の不活性化合物から識別し実験対象とする化合物を選定することで、創薬開発のコスト削減に貢献する。スクリーニング手法は、標的マクロ分子の情報を利用せず化合物情報のみを利用する「リガンドベース」と、ドッキングスタディなどに代表される標的マクロ分子との相互作用情報を利用する「ストラクチャーベース」のアプローチに分類される。本論文は、スクリーニング精度向上を目的として、両アプローチを組み合わせる新規手法を提案している。得られた成果は下記の通りである。

１．「リガンドの構造情報」と「標的マクロ分子とリガンド分子との相互作用情報」を組み合わせた表現として fragmented interaction fingerprint (FIFI) が考案された。FIFI は、従来の interaction fingerprint と異なりアミノ酸残基毎の相互作用を表現することができる。FIFI による相互作用表現を入力とした活性予測モデル（機械学習モデル）による仮想スクリーニングが提案され、複数の標的マクロ分子と活性化合物、不活性化合物を利用したレトロスペクティブな検証が実施された。「リガンドベース」や「ストラクチャーベース」の様々な手法と比較して、提案手法は総じて精度高く活性化合物を選定することができた。

２．「リガンドベース」と「ストラクチャーベース」それぞれのアプローチから算出されるスコアを組み合わせる様々な方法が評価され、望ましい方法が明らかになった。例えば「リガンドベース」としての extended connectivity fingerprint (ECFP)による識別モデルの出力値と「ストラクチャーベース」としてのドッキングスタディによる相互作用エネルギーとを組み合わせることで新規スコアを算出する。しかし、このようなメタレベルでの組み合わせ方法は、単純な「リガンドベース」手法である ECFP を利用した機械学習モデルを識別精度の観点で上回ることができず、ECFP を表現とした機械学習モデルが手法として望ましいことを実証した。

３．開発された FIFI はスパースなベクトルである。このスパース性のためデータ容量は大きくなりコンピュータ上での扱いが難しくなる課題がある。そこで、大量の化合物を効率的に評価できる深層学習に基づく新規手法が提案された。提案モデルは FIFI を利用した結果より識別能力が若干劣る結果となったが、従前考案された類似の深層学習モデルよりは優れていた。

以上、本論文では、「リガンドベース」と「ストラクチャーベース」の仮想スクリーニング手法を統合する新規手法開発と、公平な性能評価、提案手法の弱点を克服するための深層学習モデル考案を行なっている。これらの結果は、インシリコ低分子創薬に新しい知見を与えるものであり、学術的にも大きな意義がある。よって、審査委員一同は本論文が博士（理学）の学位論文として価値あるものと認めた。