京都大学	博　士（　薬科学　）	氏 名	NGUYEN DUC ANH
論文題目	Establishing advanced deep learning models for predicting drug side effects （薬物の副作用を予測するための高度なディープラーニングモデルの構築）		

A drug side effect or an adverse drug reaction is a response to a medicine that is noxious and unintended occurring at doses, which can be a single drug or a drug combination (drug-drug interactions), normally used in humans. Drug side effects are responsible for significant patient morbidity and mortality, costing billions of dollars annually. Hence, determining drug side effects is an important task in pharmacology to guide drug safety. Traditionally, drug side effects are obtained from clinical trials or surveillance reports of released drugs on the market, which are time-consuming and costly. To deal with these disadvantages, machine learning models integrating various kinds of drug data sources have been applied to obtain fast, inexpensive, and highly accurate predictions of drug side effects. The prediction results provide not only potential side effects but also the mechanisms which can support further clinical verification to improve drug side effect studies.

In this thesis, we explore machine learning models used in predicting drug side effects with a focus on deep learning models with the highest prediction performances. Basically, deep learning models aim to learn latent vector representations of drugs in low dimensional spaces which reflect drug properties causing side effects. We analyze the remaining problems in learning latent representations of drugs of the current cutting-edge methods and then propose new advanced models. The contributions of the thesis include 1) we present a comprehensive survey on data resources, tasks, and machine learning models used in drug side effect studies; 2) we present CentSmoothie, a central-smoothing hypergraph neural network for predicting drug-drug interactions, that not only learns representations of drugs but also latent representations of side effects to improve the prediction performances; 3) we present SPARSE for further improving CentSmoothie in terms of prediction accuracy and explaining the biological interpretation of the drug-drug interactions. We summarize the organization of the thesis as follows.

In Chapter 1, we introduce the predicting drug side effect problems with relevant background and terminologies.

In Chapter 2, we survey and classify data resources in drug side effects and machine learning models used on them. Data resources related to drug side effects consist of two types: (i) clinical data and (ii) non-clinical data. The clinical data contains observations of side effects in clinical treatments, which are often electronic health records or records from adverse report systems. The non-clinical data contains information on the chemical, physical, and biological properties of drugs and biological systems. The results showed that the deep learning models integrating both types of data achieved the highest prediction performance on the side effects of each drug, showing the prominence of the deep learning models.

In Chapter 3, we present CentSmoothie, a central-smoothing hypergraph neural network for predicting drug-drug interactions (DDI). DDI is usually represented as a graph in that nodes are drugs and edges are interacting drug pairs with side effects as labels. The task is to predict the labels of all pairs of nodes in the DDI graph. Existing work often uses graph neural networks to learn vector representations of drug nodes on the DDI graph and uses them to predict interactions. One drawback of this method is the lack of learning side effect representations. Side effects have complex relationships, for example, co-occurrences. Previous methods often represent each side effect as a one-hot vector indicating the presence of the side effect. This representation considers that side effects are independent, potentially under-utilizing the side effect relationships. Hence, it is necessary to learn representations of both side effects and drugs altogether. To address the above drawback, we propose to encode DDI data with a hypergraph that a node in the hypergraph can be either a drug or a side effect and each hyperedge is a triple of two drugs and a side effect that they cause. CentSmoothie, with the core idea that the side effect is caused by a single combination of the properties of two corresponding drugs, was proposed to learn on the new DDI hypergraph. The experimental results on the largest DDI benchmark dataset showed that CentSmoothie outperformed existing methods with 0.9348 and 0.8749 in AUC (area under the ROC curve) and AUPR (area under the precision-recall curve) while the second-best method was only 0.9044 and 0.8410, respectively.

In Chapter 4, we present SPARSE, a model for learning multiple types of latent combinations of drug-drug interactions. In CentSmoothie, we assumed that the side effect is caused by a single combination of the properties of two corresponding drugs. However, in reality, a side effect might have multiple, different mechanisms that cannot be represented by a single combination of latent representations of drugs. Furthermore, DDI data is sparse, suggesting that using a sparsity regularization would help to learn better latent representations to improve prediction performances. To solve these remaining problems, we propose SPARSE, which encodes the DDI hypergraph and drug features to latent spaces to learn multiple types of combinations of latent features of drugs and side effects, controlling the model sparsity by a sparse prior. The experimental results on the largest DDI benchmark data showed that SPARSE achieved an AUC of 0.9524 and AUPR of 0.882, which was higher than CentSmoothie with 0.9348 and 0.8749. We also validated the prediction results by analyzing the biological properties such as target proteins of the top prediction obtained by the learned latent interactions of SPARSE. For the top 10 cases, we could find relevant references for all cases, suggesting the prominence of prediction and the usefulness of SPARSE in practice.

In Chapter 5, we conclude our work in establishing advanced deep learning models for predicting drug side effects and give some possible future directions to enhance the models.

（論文審査の結果の要旨）

人工知能は、計算機の登場とともに考えられ始めた歴史のある技術である。当初の人工知能技術は、医療診断システムのように、エキスパートの知識を規則として書き下し、それらを集めたものであった。もちろん、規則をうまく使いこなすことにより、計算機が医師と同様に診断することが可能となる。しかし、このような「書き下し」の問題点は、そもそも書き下すことが可能な知識に限界があること、すなわち、計算機が既存の知識を超えることができないことである。加えて、知識が増えれば増えるほど大量な知識を書き下すことは容易ではない。一方、知識が得られる経験、より普遍的に言えば「データ」は時代の進展とともに、豊富に蓄積されつつある。従って、知識をデータから自動的に計算機が取得するという「データ駆動型」技術、すなわち、データから規則・パタン・仮説等を計算機が自動的に獲得する「機械学習」という考え方が生まれるのはごく自然な流れである。社会の様々な側面で大量のデータが得られる現在、機械学習は、人工知能の根幹技術であり、さらに、機械学習は現実社会のみならず、理学、工学等、科学においても、その進展のために重要な技術となっている。本研究は、薬学・薬科学において重要な問題の一つである、薬の副作用に着目し、薬の組み合わせを入力とし、入力の組み合わせが各副作用を起こすかどうかを予測する、オリジナルな機械学習手法を構築し、実データへの適用から実際に副作用を起こすような薬の組み合わせの予測が計算機で可能なことを示す試みである。

薬は、もちろん１つの薬でも副作用を起こし得るが、組み合わせ、すなわち２つの薬により副作用を起こし得るか否かがより興味深く、また人間の知識のみでは予測が難しい。一方、組み合わせの薬を３つ以上に増やした場合には、十分なデータがこれまでには得られていない。従って、２つの薬の組み合わせにより、副作用が起こるか、という問題設定が、現在のデータ、計算機資源等を考慮した際に、最も標的とすべき問題設定である。この問題設定の下、これまで様々な計算機による手法、特に機械学習に基づく手法が提案されてきた。その多くは、既存の機械学習手法を単純にあてはめるものだが、これらについては、本論文の第一章にまとめられている。最も標準的な手法は、２つの薬のそれぞれを特徴ベクトルとして、特徴ベクトルの連結によるベクトルを１つの事例として機械学習の入力とする方法だが、この方法では、薬が２つの組み合わせ（ペア）であることを必ずしも明示的に入力できない。そのため、薬をノード、予測すべき副作用をエッジ（のラベル）としたグラフを入力とし、グラフの機械学習を行うことがこれまでの最先端技術であった。

ここで、副作用は複数あるため、上記グラフのエッジには複数の副作用がラベルとなる。さらに、これら複数の副作用には相互に関係があり、すなわち、マルチタスク学習をしなければならない。しかし、通常のグラフの学習では、エッジ上の複数のラベル間（すなわち副作用間）の相互作用を考慮したマルチタスク学習はなされない。従って、この副作用データに即したグラフの学習が必要となる。そこで、以下のようにグラフのデータを拡張することを考えた。すなわち、副作用もノードとし、２つの薬と副作用の計３つのノードを一つのエッジでつなぐ、すなわち、これはノードを３つ持つハイパーエッジであり、グラフ全体はハイパーグラフとなる。本論文では、薬の組み合わせの副作用データをハイパーグラフとみなし（副作用ハイパーグラフ）、副作用ハイパーグラフ上の独自の機械学習手法の構築から、副作用予測問題の解決を試みる。特に、以下の２つの異なる手法を構築した。

１，
通常のハイパーグラフの学習では、ハイパーグラフ内のノード（に添えられる特徴ベクトル）がなるたけ同じになるように学習する。しかし、副作用ハイパーグラフでは、この学習方針は正しくない。そこで、２つの薬のノード（に添えられる特徴ベクトル）の中点が、副作用のノード（に添えられる特徴ベクトル）となるように学習する手法を構築した。この手法が通常のハイパーグラフ学習手法を凌駕することを実験的に確かめた。

２，
１と同様に、ハイパーエッジ内のノード（に添えられる特徴ベクトル）が同じになる必要はない。また、１の方針である中点は、すべてのデータに対して達成することが困難という問題点がある。ここで、予測への重要性を考えた場合には、ハイパーエッジ（内のノードに添えられる特徴ベクトル）同士が類似しているか否かが予測に重要となる。そこで、いわばハイパーエッジをクラスタリングするようなモデル、いわゆるブロックモデル（各ブロックは各クラスタのようなもの）をハイパーグラフに対して構築した。これにより、副作用に複数の原因があっても対応可能である（各クラスタはいわば一つの原因（理由）に対応すると考えられる）さらに、副作用は稀な出来事であり、副作用ハイパーグラフのハイパーエッジは非常に疎である。従って、疎であることを仮定し、ハイパーグラフ用のブロックモデルを構築した。この方法は、上記１を含めたほぼすべての既存の、組み合わせの副作用予測手法を凌駕することを実験的に確かめた。

上記、２つのモデルは、本論文の第二章、第三章にあたり、各々論文にまとめられている。特に、上記２の第三章のモデルは、2022 年に、バイオインフォマティクス（生命情報科学）のトップ国際会議である ISMB（Intelligent Systems for Molecular Biology）に採択され（採択率 19.8%（242 件投稿,48 件採択））、同国際会議の予稿集である Bioinformatics 誌の特別号に掲載された。この方法は、バイオインフォマティクスまた機械学習分野の研究者には、薬の組み合わせによる副作用を予測する、最も精度の高い手法と認知され、高く評価されている。また、この分野の調査を行った、本論文の第一章の内容は、サーベイ論文として、Briefings in Bioinformatics 誌に掲載されている。

これら３つの論文の内容は、生命情報科学及び機械学習の研究成果として、一定の基準を十分に満たしているとみなすことができる。

よって、本論文は、博士（薬科学）の学位論文として価値あるものと認める。また、令和５年２月１０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。

要旨公表可能日：　　年　　月　　日以降