APPLICATIONS OF DATA MINING IN HEALTHCARE

A Thesis

Submitted to the Faculty

of

Purdue University

by

Bo Peng

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

May 2019

Purdue University

Indianapolis, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. George Mohler, Chair

Department of Computer & Information Science

Dr. Murat Dundar

Department of Computer & Information Science

Dr. Jiang Yu Zheng

Department of Computer & Information Science

Approved by:

Dr. Shiaofen Fang

Head of the Graduate Program

To my parents and lovely girlfriend.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. George Mohler and all my committee members for their help in overcoming numerous obstacles I have been facing through my research. I would also like to thank Dr. Xia Ning and Dr. Martin Renqiang Min for all their help both academically and personally. Their patience and support are critical to steer my research in the right direction. It's a really enjoyable and fruitful experience to work with them. I am also grateful to all my lab mates and friends for their support when I get troubles in my research or daily life.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Peng, Bo. M.S., Purdue University, May 2019. Applications of Data Mining in Healthcare. Major Professor: George Mohler.

With increases in the quantity and quality of healthcare related data, data mining tools have the potential to improve people's standard of living through personalized and predictive medicine. In this thesis we improve the state-of-the-art in data mining for several problems in the healthcare domain. In problems such as drug-drug interaction prediction and Alzheimer's Disease (AD) biomarkers discovery and prioritization, current methods either require tedious feature engineering or have unsatisfactory performance. New effective computational tools are needed that can tackle these complex problems.

In this dissertation, we develop new algorithms for two healthcare problems: high-order drug-drug interaction prediction and amyloid imaging biomarker prioritization in Alzheimer's Disease. Drug-drug interactions (DDIs) and their associated adverse drug reactions (ADRs) represent a significant detriment to the public health. Existing research on DDIs primarily focuses on pairwise DDI detection and prediction. Effective computational methods for high-order DDI prediction are desired. In this dissertation, I present a deep learning based model $\texttt{D}^3\texttt{I}$ for cardinality-invariant and order-invariant high-order DDI prediction. The proposed models achieve 0.740 F1 value and 0.847 AUC value on high-order DDI prediction, and outperform classical methods on order-2 DDI prediction. These results demonstrate the strong potential of $\texttt{D}^3\texttt{I}$ and deep learning based models in tackling the prediction problems of high-order DDIs and their induced ADRs.

The second problem I consider in this thesis is amyloid imaging biomarkers discovery, for which I propose an innovative machine learning paradigm enabling precision medicine in this domain. The paradigm tailors the imaging biomarker discovery process to individual characteristics of a given patient. I implement this paradigm using a newly developed

learning-to-rank method PLTR. The PLTR model seamlessly integrates two objectives for joint optimization: pushing up relevant biomarkers and ranking among relevant biomarkers. The empirical study of PLTR conducted on the ADNI data yields promising results to identify and prioritize individual-specific amyloid imaging biomarkers based on the individual's structural MRI data. The resulting top ranked imaging biomarkers have the potential to aid personalized diagnosis and disease subtyping.

1. INTRODUCTION

Adverse drug reactions and Alzheimer's Disease (AD) are two significant public health problems [1, 2] and a number of research efforts in the data mining community have been dedicated to solve these problems. Zhang *et al.* [3] applied multiple classical methods such as neighbor-based recommendation, random walk and matrix perturbation to predict and rank DDI on drug pairs. Wang *et al.* [4] incorporated different types of drug features to learn drug embedding of single drugs. Then they employed deep neural networks to predict the probability of single drugs in inducing side-effects. There are also several methods proposed to extract mentioned drug pairs via text mining from medical literature and electronic medical records [5–9]. For AD detection, there is a large body of neuroimaging studies that develop image-based predictive models for early detection of AD as well as identification of relevant biomarkers [10–12]. However, the existing studies have several limitations and unsolved problems remain. For DDI prediction, few studies, to the best of our knowledge, have addressed representing, quantifying, discovering and visualizing relations among high order DDIs. However, high order DDIs comprise a significant portion of real life adverse drug interaction cases [6, 13–15]. In most AD detection studies, proposed models are limited to identifying imaging biomarkers that are at the population level, but not specific to individuals.

In recent years, new computational methods such as deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs) and attention mechanisms have shown promise in solving complicated problems that were previously considered infeasible. These deep models achieve impressive results in image-based diagnosis [16] and electronic health records (EHRs) based prediction tasks [5–9]. However, there remain many problems in the healthcare domain that are yet to be solved by deep learning.

In this dissertation, I study two challenging problems:

- high-order drug-drug interaction prediction; and
- amyloid imaging biomarker prioritization in Alzheimer's Disease.

I propose data-driven approaches (i.e. deep learning, learning to rank) to the two problems and conduct comprehensive experiments to test the proposed approaches on public datasets. The experimental results demonstrate the strong potential of the proposed methods in solving DDI and AD prediction tasks. The performance of the proposed methods also demonstrates the potential of applying newly developed data-driven approaches to solve complex healthcare related problems more generally.

1.1 Thesis Outline

The remaining sections of the dissertation are organized as follows. In Chapter 2, I will describe a deep learning method for high-order drug-drug interaction prediction. In Chapter 3, I will describe amyloid imaging biomarker prioritization in Alzheimer's Disease via learning to rank in detail. Finally, I will summarize the dissertation in Chapter 4. Supplementary materials are provided in the appendix.

2. DEEP LEARNING FOR HIGH-ORDER DRUG-DRUG INTERACTION PREDICTION

2.1 Introduction

Drug-drug interactions (DDIs) and their associated adverse drug reactions (ADRs) represent a significant detriment to the public health. Approximately 195,000 hospitalizations and 74,000 emergency room visits are resulted out of DDIs in the United States [1]. The increasing rates of polypharmacy, particularly among aging populations [1], will further deteriorate this situation [13]. Consequent upon these facts, significant research efforts have been dedicated to detecting DDIs, including DDI extraction from medical literature [6, 7] or social media [17–19], and biochemical and molecular information integration for DDI scoring [20–22, 22, 23], etc. However, most of the existing DDI studies are limited to interactions between pairs of drugs (i.e., order-2 DDIs), while DDIs among multiple drugs (i.e., high-order DDIs) occupy a significant portion in real-life cases. It is reported that more than 76% of the elderly Americans take two or more drugs daily [1]. Another study [6] estimates that about 29.4% of elderly American patients taking six or more drugs every day. Therefore, understanding high-order DDIs and their associated ADRs becomes urgent and critical [13–15].

Unfortunately, very limited efforts, to the best of our knowledge, have been dedicated to representing, quantifying, discovering and visualizing relations among high-order DDIs. Emerging methods on high-order DDI studies are only focused on the discovery of high-order DDIs through mining frequent drug combinations efficiently. Meanwhile, as the cardinality of drug combinations (i.e., the number of drugs in drug combinations; also refereed to as the order of drug combinations) increases, modeling of DDI relations, particularly of arbitrary cardinalities/orders in a unified framework, becomes increasingly non-trivial.

In this chapter, we present a new deep model to conduct cardinality- and order-invariant high-order DDI prediction, refereed to as **D**eep **DDI** model and denoted as $\texttt{D}^3\texttt{I}$. $\texttt{D}^3\texttt{I}$ is invariant of drug combination cardinalities and the order in which the drugs are considered in the model, that is, $\texttt{D}^3\texttt{I}$ is able to predict ADR labels for combinations of arbitrary numbers of drugs in arbitrary input orders. Meanwhile, $\texttt{D}^3\texttt{I}$ is able to generate embeddings for single drugs and aggregate single drug embeddings into drug-combination embeddings. Thus, these drug-combination embeddings are able to capture the synergistic latent signals that are related to ADRs among the constituent single drugs. We conducted extensive experiments on two public datasets of high-order DDIs, and tested multiple $\texttt{D}^3\texttt{I}$ variations on the datasets. Our experimental results demonstrate that $\texttt{D}^3\texttt{I}$ is able to achieve 0.740 F1 value and 0.847 AUC value on balanced high-order DDI prediction, and outperform other models on order-2 DDI prediction. The experiments also show that by integrating DDIs of high orders, $\texttt{D}^3\texttt{I}$ models are even able to further improve prediction performance on order-2 DDIs. In addition, the single drug embeddings produced from $\texttt{D}^3\texttt{I}$ models also represent clustering structures that conform to domain knowledge. To the best of our knowledge, $\texttt{D}^3\texttt{I}$ is the first deep model for high-order DDI prediction.

The rest of this chapter is organized as follows. Section 2.2 presents the literature review. Section 2.3 presents the definitions and notations used in this chapter. Section 2.4 presents the $\texttt{D}^3\texttt{I}$ method. Section 2.5 presents the datasets used for the experiments. Section 2.6 presents the experimental protocol. Section 2.7 presents the experimental results. Section 2.8 presents conclusions and future research.

2.2 Literature Review

2.2.1 DDI Detection and Prediction

Current research on detecting DDIs can be broadly classified into four categories [13, 21, 24, 25]. The first category of methods focus on text mining from medical literature and electronic medical records, and they extract mentioned drug pairs [5–9]. A second category of methods integrate various biochemical and molecular drug/target data to measure drug-

drug similarities and score/predict DDIs using the similarities. These data include chemical structures [20, 26], target information [21, 27], compound-target docking results [28], phenotypic and genomic information [22], and drug side effects [23], etc. The collected data are used in various data-driven computational methods such as classification [22], regression [23], statistical testing [29] to detect DDIs. For example, Zhang *et al.* [3] applies multiple methods such as neighbor-based recommendation, random walk and matrix perturbation for pairwise DDI ranking and prediction. The third category of methods leverage healthcare information on social media and online communities to detect DDIs that have been mentioned/inferred in online discussions and posts [17–19]. The last category of methods predict the probability of ADR event counts due to high-order DDIs [30, 31] and use either electronic medical records or pharmacokinetic modeling to validate potential DDIs. A notable shortcoming of these methods is that they work for low-order or fixed-order DDIs but do not scale well to arbitrary orders.

2.2.2 Deep Learning based DDI Detection and Prediction

The interactions between drugs are very complex and may go far beyond simple or linear relations. Thus, it inspires the use of Deep Learning (DL) in this field due to the strong capability of DL in approximating complex relations. High-order DDIs prediction has some analogies to multi-instance learning [32] over bags of instances. Wang *et al.* [32] proposed a deep framework for multi-instance learning, which first learns an embedding for each of the instances in the bag, and then applies an aggregator to combine these embeddings into a bag-level representation for classification. Ilse *et al.* [33] proposed an attention-based deep model to integrate instance embeddings into bag embeddings. One drawback of this method is that it combines instances linearly, which might not always be optimal. Zaheer *et al.* [34] introduced constraints on the weight matrix of the deep model to learn over sets, and enforced symmetry of the learned weight matrix to enable order-invariant property into the model. Wang *et al.* [4] incorporated different types of drug features to

learn a drug embedding for single drugs, and used a deep neural network architecture to predict potential side-effects of single drugs.

Deep learning technologies are also used in detecting and predicting DDIs. Segura-Bedmar *et al.* [35] proposed to use convolutional neural networks (CNNs) to extract DDIs from biomedical text. Text information is represented as a matrix, in which each column or row is a word vector [36]. Then CNN layers are applied to the matrix to extract features and do the prediction. This work achieves the second place in the 2013 ranking of the DDIs extraction challenge. In Sahu *et al.* [37], instead of using CNNs, Long Short-Term Memory (LSTM) model is used to extract features from text and then do the prediction. Graph Convolutional neural Network (GCN) is also introduced to predict pairwise DDIs. Zitnik *et al.* [38] views pairwise DDI prediction as a link prediction task over drug-drug graphs. They applied GCN on the constructed DDI graph to learn embeddings for each drug, and calculated link probabilities (i.e., DDI probabilities) based on learned embeddings.

2.3 Definitions and Notations

Table 2.1.

Notations

notation	meaning
d	a drug
D	a drug combination
$\mathbf{f}$	a vector of drug features
$\mathbf{e}$	a vector of drug embedding
$\mathbf{E}$	a vector of drug combination embedding

The key notations used in this chapter are listed in Table 2.1. In this cgapter, all the vectors are by default row vectors and represented using lower-case bold letters (e.g., $\mathbf{e}$);

all the matrices are represented using upper-cases letters (e.g., X). The key definitions used in this chapter are listed as follows:

- *Drug combination*: a set of drugs that are prescribed/taken together, denoted as D.
- *Cardinality of drug combinations*: the number of drugs in a drug combination D is the cardinality of the drug combination, denoted as $\|D\|$. Drug combination cardinality is also refereed to as drug order of the combination.
- *Cardinality invariance*: the model is able to predict for drug combinations of arbitrary cardinalities; the prediction mechanism is invariant of the cardinalities of input drug combinations.
- *Order invariance*: the model is able to produce a same result for a same drug combination, regardless of the order in which the drugs in the combination are input to the model. Note that in order invariance, the term "order" does not refer to cardinality but to a notion of ordering.

The problem that we try to solve is defined as follows:

Problem definition: Given a set of drug combinations and their ADR labels, build a classification model of cardinality invariance and order invariance that is able to predict the ADR labels for new drug combinations of arbitrary cardinalities.

In this chapter, we are only concerned with one specific ADR, that is, myopathy [39]. Therefore, the classification model is a binary classifier. However, multi-class classifier can be extended from our models, and will be investigated in our future research. In this study, we a use feature vector to represent each drug. The feature vector is consisted by the pairwise similarities between the profile of different drugs.

2.4 Methods

We develop a new deep model to conduct cardinality- and order-invariant high-order DDI prediction. This model is refereed to as **D**eep **DDI** model and denoted as $\texttt{D}^3\texttt{I}$. $\texttt{D}^3\texttt{I}$ has the following three key components:

- An encoder, which encodes each of the drugs in an input drug combination into a latent representation (i.e., embedding);
- An aggregator, which learns a single, high-level representation/embedding for the drug combination from the representations/embeddings of its component drugs; and
- A predictor, which predicts the likelihood of ADR labels using the drug-combination representation/embedding.

Figure 2.1 presents the architecture of D^{3}I. The novelty of D^{3}I is that its aggregator is able to deal with arbitrary number of drug embeddings in drug combinations regardless of drug input orders. Meanwhile, the single drug embeddings and the drug-combination embeddings could enable additional insights on the drug properties and relations in inducing ADRs.

Note that in this chapter, only myopathy is considered as the ADR of interest. That is, we predict if a drug combination will induce myopathy or not. The reason for myopathy as the interested ADR is that it has been better studied [40] than other side effects, particularly in terms of the underlying mechanisms and the ground-truth myopathy-inducing single drugs. Even though, our D^{3}I is effortlessly applicable to other specific, single ADRs, and can be easily extended to the prediction of multiple, specific ADRs (by learning multiple outputs) and to the prediction of general ADRs (i.e., whether there will be ADRs or not; not specific to a certain type of ADR).

2.4.1 D^{3}I Encoder

The D^{3}I encoder learns and represents signals that could be pertinent to ADR prediction from each drug in the input drug combination. For a drug combination $D = \{d_1, d_2, \cdots, d_n\}$ of n drugs, the encoder g_e learns an embedding $\mathbf{e}_i$ for each drug d_i from its feature vector $\mathbf{f}_i$ as follows:

$$\mathbf{e}_i = g_e(\mathbf{f}_i), \tag{2.1}$$

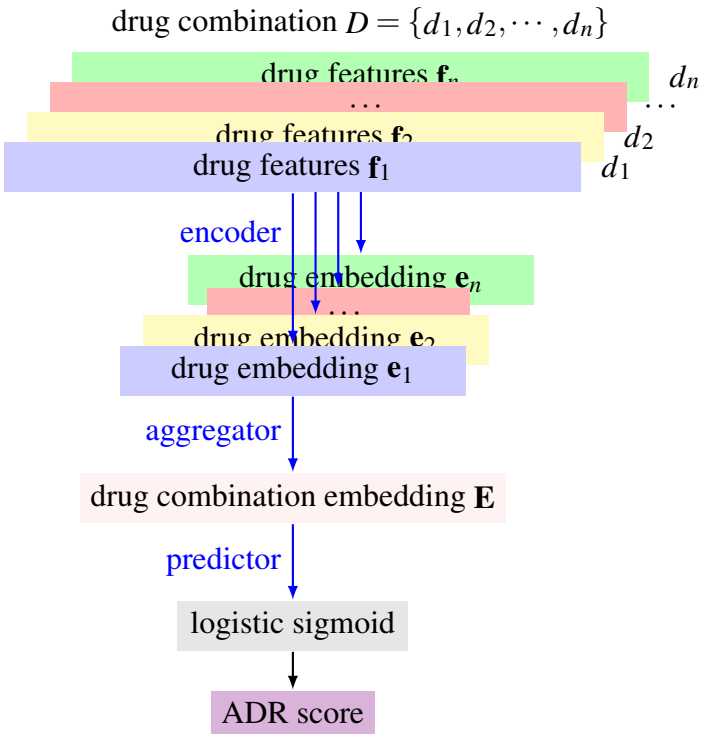

Fig. 2.1. $\mathtt{D^3I}$ Architecture

where $\mathbf{e}_i \in \mathbb{R}^{1\times k}$, $\mathbf{f}_i \in \mathbb{R}^{1\times m}$ and typically $k < m$. We use an n_e-layer neural network (NN) as g_e, that is,

$$g_e(\mathbf{f}) = g_{n_e}(\cdots(g_2(g_1(\mathbf{f}))), \tag{2.2}$$

with each layer parameterized by a weighting matrix W_j^e $(j = 1,\cdots,n_e)$ of appropriate dimensions. The input drug features will be discussed later in Section 2.5. Note that the encoder applies on each individual drug in the input drug combination independently, and thus it is order invariant. For input $D = \{d_1, d_2, \cdots, d_n\}$, the output from the encoder is denoted as $\mathbf{e}(D) = [\mathbf{e}_1; \mathbf{e}_2; \cdots; \mathbf{e}_n]$, that is, $\mathbf{e}(D)$ is a $n \times k$ matrix.

2.4.2 D^{3}I Aggregator

The D^{3}I aggregator learns one embedding for the input drug combination from its individual drug embeddings out of the encoder. We adopt three aggregation strategies: 1). max pooling, 2). mean pooling and 3). aggregation with attentions, respectively, in the D^{3}I aggregator.

Max Pooling

In the max pooling strategy, we calculate the drug-combination embedding, denoted as $\mathbf{E}_D$, for $D = \{d_1, d_2, \cdots, d_n\}$ as follows:

$$\mathbf{E}_D = \max(\mathbf{e}(D)) = [\max_{\forall i}\{\mathbf{e}_{i,1}\}, \max_{\forall i}\{\mathbf{e}_{i,2}\}, \cdots, \max_{\forall i}\{\mathbf{e}_{i,k}\}], \tag{2.3}$$

where max is an element-wise operator that selects the maximum value in each dimension in all the drug embeddings $\mathbf{e}(D) = [\mathbf{e}_1; \mathbf{e}_2; \cdots; \mathbf{e}_n]$. The max pooling is trivially cardinality invariant and order invariant due to the max function used. It is expected that drugs contribute differently in their interactions and induced ADRs, and their respective contributions could be represented in their maximum values in their embeddings through learning and the max pooling. D^{3}I with the max pooling strategy is denoted as D^{3}I$_{\texttt{max}}$.

Mean Pooling

In the mean pooling strategy, we calculate the drug-combination embedding $\mathbf{E}_D$ as follows:

$$\mathbf{E}_D = \text{mean}(\mathbf{e}(D)) = [\text{avg}_{\forall i}\{\mathbf{e}_{i,1}\}, \text{avg}_{\forall i}\{\mathbf{e}_{i,2}\}, \cdots, \text{avg}_{\forall i}\{\mathbf{e}_{i,k}\}], \tag{2.4}$$

where the avg operator calculates the average value in each dimension in all the drug embeddings $\mathbf{e}(D) = [\mathbf{e}_1; \mathbf{e}_2; \cdots; \mathbf{e}_n]$. The mean pooling is also trivially cardinality invariant and order invariant. It intends to average the information from each involved drug in representing a drug combination. $\texttt{D}^3\texttt{I}$ with the mean pool strategy is denoted as $\texttt{D}^3\texttt{I}_{\texttt{mean}}$.

Self-Attention

Inspired by the recent work in deep multi-instance learning [33], we propose to use a weighted sum of drug embeddings to learn a single embedding of a drug combination. For a drug combination $D = \{d_1, d_2, \cdots, d_n\}$, the embedding of D is calculated as follows:

$$\mathbf{E}_D = \sum_{i=1}^{n} a_i \mathbf{e}_i, \tag{2.5}$$

where a_i is a weight on $\mathbf{e}_i$. To allow the drug embeddings to determine their own importance in the drug-combination embedding, a_i is also calculated as a function of $\mathbf{e}_i$ as follows,

$$a_i = \text{softmax}(\mathbf{w} \tanh(V \mathbf{e}_i^{\mathsf{T}})),$$

where $V \in \mathbb{R}^{l \times k}$ and $\mathbf{w} \in \mathbb{R}^{1 \times l}$ are two parameters that will be learned, and $\text{softmax}(x)$ is the softmax function defined as follows:

$$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}, \tag{2.6}$$

and thus $\sum_i a_i = \sum_i (\text{softmax}(\mathbf{w}^{\mathsf{T}} \tanh(V \mathbf{e}_i^{\mathsf{T}}))) = 1$; and the hyperbolic tangent function $\tanh(\cdot)$ is used to introduce element-wise non-linearity. The attention mechanism as in Equation 2.5 is order invariant simply because the sum operation in Equation 2.5 is order invariant. It is also cardinality invariant because of the normalization in softmax in Equation 2.6. $\texttt{D}^3\texttt{I}$ with the self-attention pooling strategy is denoted as $\texttt{D}^3\texttt{I}_{\texttt{Att}}$.

2.4.3 D^{3}I Predictor

The D^{3}I predictor predicts the probability of a drug combination in inducing ADRs. For a drug combination D, its embedding $\mathbf{E}_D$ is first converted through n_p fully-connected layers with tanh as the activation function, that is,

$$h_e(\mathbf{E}_D) = h_{n_p}(\cdots(h_2(h_1(\mathbf{E}_D)))), \tag{2.7}$$

with each layer parameterized by a weighting matrix W_j^E $(j = 1, \cdots, n_p)$ of appropriate dimensions. Then a sigmoid function is used to do the prediction as follows:

$$p(D) = \frac{1}{1 + \exp(-h_e(\mathbf{E}_D)^\top)} \tag{2.8}$$

where $\mathbf{E}_D$ is the drug-combination embedding of D out of D^{3}I aggregator, and $p(D)$ is the probability of D in inducing ADRs ($p(D) \in [0,1]$).

2.4.4 Learning Algorithm

In D^{3}I, we formulate the DDI-induced ADR prediction as a binary classification problem, and learn the D^{3}I models by solving the following optimization problem, in which the cross entropy loss is used as the objective:

$$\min_{\Theta} -\sum_{i=1}^{n} y_i \log p_i + (1 - y_i) \log (1 - p_i), \tag{2.9}$$

where y_i is the label of the i-th drug combination (positive for ADR inducing and negative otherwise), p_i as calculated in Equation 2.8 is the probability of i-th drug combination in inducing ADRs, and Θ is the set of parameters of the D^{3}I model, including the weighting matrices $\{W^e\}$ (in D^{3}I encoder as in Section 2.4.1) and $\{W^E\}$ (in D^{3}I predictor as in Section 2.4.3) among the fully-connected layers. We use the Adam gradient descent algorithm [41] to solve the problem 2.9. We use batch training, described in Section A.2.1 in the supplementary materials, to train D^{3}I models. All the hyper-parameters are reported in Section A.2.2 in the supplementary materials.

2.4.5 Data Availability

The data and the code will be made publicly available upon the acceptance of this chapter.

2.5 Materials

We use two datasets in our experiments to test the performance of $\texttt{D}^3\texttt{I}$. The first dataset is derived from Chiang *et al.* [42], denoted as FEARS. The second dataset is derived from Zhang *et al.* [3], denoted as BMC. The dataset statistics is presented in Table 2.2.

Table 2.2.
Dataset Statistics

dataset	$\#d$	$\#D$	$\|\|D\|\|$	$\|\|\bar{D}\|\|$	features
FEARS	826	6,338	2-52	3.6	substructures (FP), targets (TG), side effects (SE), indications (TI)
BMC	548	48,584	2	2	substructures (FP), targets (TG), off-side effects (OSE), indications (TI), enzymes (EM), pathways (PW), transporters (TP)

The columns corresponding to $\#d$, $\#D$, $\|\|D\|\|$, $\|\|\bar{D}\|\|$ and "features" have the number of drugs, the number of drug combinations, the cardinalities of drug combinations, average cardinaltiy of drug combinations and the drug features in the dataset, respectively.

2.5.1 FEARS Dataset

The FEARS dataset has 6,338 drug combinations from 826 drugs, including 2,981 2-drug combinations, 1,555 3-drug combinations, 652 4-drug combinations, 323 5-drug combinations, 220 6-drug combinations, 157 7-drug combinations and 450 combinations with more than 7 drugs. The maximum number of drugs in a combination is 52, and the

average is 3.6. The drug combinations are selected based on their odds ratios [43] of inducing myopathy among a large collection of spontaneous reports to FDA[1] The detailed description of drug combination selection and dataset construction is available in Section 8 "Data Preparation" in Chiang *et al.* [42]. We collected 4 types of information for the drugs, including chemical substructure fingerprints (FP), side-effect profiles (SE), therapeutic-indication profiles (TI) and target profiles (TG). Unfortunately, we cannot find all the 4 types of features for each drug. As a result, the size of used data when using different features as input is different, that is, 6,638 combinations with FP, 3,330 combinations with SE, 3,088 combinations with TI, and 5,621 combinations with TG. The detailed description on such features is available in Section A.1 in the supplementary materials. Please note that in FEARS, half of the drug combinations induce myopathy (i.e., positive drug combinations) and the rest do not (i.e., negative drug combinations).

2.5.2 BMC Dataset

The BMC has 48,584 drug pairs from 548 drugs with 9 different types of drug features, including chemical substructures denoted as FP, drug target profiles denoted as TG, transporter profiles denoted as TP, enzymes denoted as EM, pathways denoted as PW, drug indications denoted as TI, side effects denoted as SE, off-side effects denoted as OSE and the drug-drug interaction profiles. We download the drug similarity profiles calculated from the 7 types of features[2]. For more details about the drug features, drug similarity profiles and BMC, please refer to Zhang *et al.* [3]. Note that in BMC, the drug combinations all induce side effects (i.e., positive drug combinations).

2.5.3 Generation of Drug Feature Vectors

For each dataset, we calculate the pairwise Jaccard similarity coefficients for all the drugs in the dataset using each of the drug features (e.g., TG, TI), and use each row of the

[1] `https://www.fda.gov/drugs/informationondrugs/ucm135151.htm`

[2] `https://github.com/zw9977129/drug-drug-interaction/`

similarity matrix as the corresponding feature vector representation of the corresponding drug. Intuitively, the feature vector $\mathbf{f}$ of a drug d presents the similarities between d and all drugs in the same dataset using the corresponding drug features. This feature representation scheme is inspired by the idea in Que and Belkin [44]. It provides an easy framework to mitigate high-dimension features with missing values and integrate multiple types of features.

2.6 Experimental Protocols

2.6.1 Positive and Negative Data Generation

We conduct the experiments under two settings, denoted as TPTN and TPRN, respectively. In TPTN, we use the true positive and true negative drug combinations from the datasets to train and test our models. That is, the positive and negative samples are fixed from the datasets. In TPRN, we only use the positive drug combinations in the datasets and sample corresponding equal-size negative drug combinations for training and testing.

Negative Data Sampling in TPRN

The negative sample generation process is only conducted in the TPRN setting, that is, for a cardinality-k positive drug combination $D = \{d_1, ...d_k\}$, we sample k drugs and construct a corresponding negative drug combination $D' = \{d'_1, ...d'_k\}$ such that D' is not in the positive drug combinations. Drug d' is selected according to the following distribution P,

$$P(d') = \frac{f(d')}{\sum_{i=1}^{n}(f(d'_i))}, \tag{2.10}$$

where $f(d')$ denotes the frequency of drug d' in training and validation set (see Section 2.6.2 for details on cross validation). Please note that sampled drug combinations could be false negative, and thus we need to check the sampled combinations against the training and validation set to remove false negative samples.

The reason why we do negative sampling, even though there could be labeled negative drug combinations, is to avoid the situation in which the classification is biased by a confounder from the cardinalities of drug combinations. We noticed that combinations of high cardinalities are more likely to induce side effects, but true negative drug combinations tend to have low cardinalities (will be discussed later in Section 2.7.3). Therefore, a model trained from such negative drug combinations could be biased by the signals in high-cardinality, true positive drug combinations, and the signals in low-cardinality, true negative drug combinations. By doing the negative sampling as above, we introduce negative training instances of high cardinality, and thus force the model to learn non-trivial signals from drug combinations.

2.6.2 Cross Validation

We conduct 5-fold cross validation in both TPTN and TPRN settings. In the TPTN setting, we randomly split the original datasets into 5 folds of equal size, with all the folds having relatively same number of true positive/true negative drug combinations. We use 3 folds for model training, 1 fold for validation and 1 fold for testing each time. In the TPRN setting, we randomly split the positive drug combinations in the datasets into 5 folds of equal size. Similar to the first setting, 3 folds are used for training, and the rest 2 folds are used for testing and validation each time. Before training, we sample negative drug combinations for testing and validation sets and fix them (i.e., the negative drug combinations will not change during and after training for the testing and validation sets). The negative drug combinations of training set are sampled during training on the fly. That is, in each training batch (Section A.2.1 in the supplementary materials), we sample negative drug combinations of the same size and order distribution for the positive drug combinations in that batch. The positive drug combinations and sampled negative drug combinations are together used as training data in the batch to train the model. In both settings, we run experiments for 5 times, with 1 fold as the testing set each time, and report results that are averaged out of the five experiments.

2.6.3 Evaluation Metrics

We use accuracy, precision, recall, F1 and Area Under the ROC Curve (AUC) to evaluate the performance of the various methods. We use TP, FN, TN and FP to denote the number of true positive drug combinations, false negative drug combinations, true negative drug combinations and false positive drug combinations in the testing set, respectively. We also use P to denote the number of positive drug combinations (i.e.,P = TP + FN) and N to denote the number of negative drug combinations (i.e., N = FP + TN). Thus, accuracy (acc) is defined as follows,

$$\text{acc} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}, \tag{2.11}$$

that is, acc is the fraction of all correctly classified drug combinations over all the drug combinations. Precision (pre) is defined as follows,

$$\text{pre} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{2.12}$$

that is, pre is the fraction of correctly classified positive drug combinations over all the drug combinations that are classified as positive. Recall (rec) is defined as follows,

$$\text{rec} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2.13}$$

that is, it's the fraction of correctly classified positive drug combinations over all the positive drug combinations. F1 is defined as follows,

$$\text{F1} = 2 \times \frac{\text{rec} \times \text{pre}}{\text{rec} + \text{pre}}, \tag{2.14}$$

that is, it's the harmonic mean of the precision and recall. Area Under the ROC Curve (AUC) [45] is the normalized area under the curve of the true-positive rate against the false positive rate over different classification thresholds. For all the 5 metrics, the larger value indicates better classification performance.

2.7 Experimental Results

We present the experimental results in this section. Additional experimental results including comparison on drug features and model architectures are available in Section A.3 in the supplementary materials..

2.7.1 Overall Performance

Overall performance on FEARS

Table 2.3 presents the best performance of the three methods $\texttt{D}^3\texttt{I}_{\texttt{max}}$, $\texttt{D}^3\texttt{I}_{\texttt{mean}}$ and $\texttt{D}^3\texttt{I}_{\texttt{Att}}$ on FEARS under the two experimental settings TPTN and TPRN. Note that all the results in the table are selected according to the best F1 values, and the other evaluation measurements according to the best F1 values are also presented. Overall, $\texttt{D}^3\texttt{I}_{\texttt{max}}$ achieves the best performance compared to the other two methods with the best F1 0.815 and AUC 0.892 in TPTN, and the best F1 0.740 and AUC 0.847 in TPRN. $\texttt{D}^3\texttt{I}_{\texttt{mean}}$ ranks as the second with the best F1 0.766 and AUC 0.842 in TPTN, and the best F1 0.704 and corresponding AUC 0.767 (best AUC 0.770) in TPRN. $\texttt{D}^3\texttt{I}_{\texttt{Att}}$ performs the worst with best F1 0.756 and AUC 0.834 in TPTN, and best F1 0.672 and AUC 0.760 in TPRN. These results demonstrate the strong capability of $\texttt{D}^3\texttt{I}_{\texttt{max}}$ in predicting ADRs of drug combinations of various orders.

The primary difference among $\texttt{D}^3\texttt{I}_{\texttt{max}}$, $\texttt{D}^3\texttt{I}_{\texttt{mean}}$ and $\texttt{D}^3\texttt{I}_{\texttt{Att}}$ relies on their aggregators. $\texttt{D}^3\texttt{I}_{\texttt{max}}$ utilizes max pooling as in Equation 2.3 to construct a combination embedding that consists of the strong signals from each dimension of individual drug embeddings. It is very likely that in the combination embedding, different dimensions selected via $\max()$ operator are from different drugs, and therefore, non-linearity in aggregation is realized. More importantly, such combination of embedding dimensions from different drugs corresponds to the notation of drug-drug interaction – intuitively, drugs contribute different aspects all together to introduce ADRs.

The TPTN and TPRN settings are different in the negative drug combinations in both training and testing sets. In TPTN, the negative drug combinations typically have different cardinalities compared to those of the positive drug combinations. However, in TPRN, our sampling method as described in Section 2.6.1 guarantees same cardinalities for each positive drug combination and its paired, sampled negative drug combination. The overall worse performance in TPRN compared to that in TPTN indicates the difficulty in learning from same-dimensionality positive and negative drug combinations, and the difficulty in learning synergistic interaction signals from high-cardinality drug combinations. However, our methods are still able to achieve F1 0.740 and AUC 0.847 in TPRN, indicating its strong potential in predicting high-order DDIs and induced ADRs. Compared to TPRN, the TPTN setting is closer to the real application scenario (e.g., different cardinality distributions in positive drug combinations and negative drug combinations), and the good performance of our methods demonstrates their strong potential in high-order DDI prediction in real applications.

Overall performance on BMC

Table 2.4 presents the overall performance of $\texttt{D}^3\texttt{I}$ methods and the comparison with other methods on the BMC dataset. Please note that BMC dataset has only true positive drug combinations of cardinality 2. The results reported in the original paper [3] correspond to very unbalanced testing data (i.e., 9,716 positives, 101,294 negatives). Therefore, the performance of neighbor-based recommender, random walk and matrix perturbation as used in the paper [3] is good in accuracy and AUC, but not in other metrics. In $\texttt{D}^3\texttt{I}$ methods, we conducted negative sampling and thus the testing data are balanced. In terms of precision, recall and F1, $\texttt{D}^3\texttt{I}$ methods significantly outperform others. In particular, in terms recall, $\texttt{D}^3\texttt{I}_{\texttt{Att}}$ is 4.6% better than random walk (recall 0.803 vs 0.768), which is the best non-$\texttt{D}^3\texttt{I}$ method. Also, in terms of F1, $\texttt{D}^3\texttt{I}_{\texttt{max}}$ is also better than matrix perturbation that achieves the best F1 among all the non-$\texttt{D}^3\texttt{I}$ methods (F1 0.720 vs 0.707).

Clustering Analysis

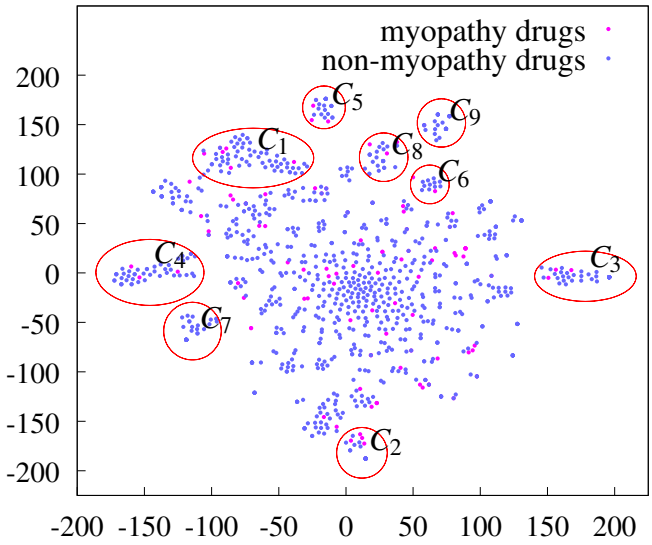

Fig. 2.2. Single Drug Embeddings from $\mathtt{D^3I_{max}}$ (FEARS, TPRN)

Figure 2.2, generated using t-SNE [49] method, presents the single drug embeddings generated from $\mathtt{D^3I_{max}}$ (TPRN) on the FEARS dataset. In this figure, there are some well-formed clusters (e.g., C_1, C_3 and C_4). Cluster C_1 primarily includes antipsychotic drugs (e.g., amisulpride, aripiprazole, droperidol, perphenazine, pimozide, pipotiazine, risperidone), antidepressants (e.g., amitriptyline, desipramine, trazodone) and drugs for Parkinson treatment (e.g., isuride, ropinirole) and Huntington treatment (e.g., tetrabenazine). Cluster C_3 includes many anti-inflammatory drugs (e.g., acetylsalicylic acid, flurbiprofen, ibuprofen, loxoprofen, naproxen, rofecoxib, salicyclic acid, tenoxicam). In cluster C_4, most of the drugs (e.g., butabarbital, clonazepam, clotiazepam, etizolam, oxazepam, pentobarbital, thiopental) are used to treat tension, anxiety, nervousness, insomnia, seizures and panic disorders. Cluster C_5 represents a group of drugs (e.g., codeine, heroin, oxycodone, propoxyphene, sufentanil, tramadol) that are used to treat pains. The above clustering structures among single drug embeddings demonstrate that $\mathtt{D^3I}$ methods learn latent representations from single drugs that may conform to domain knowledge.

2.7.2 Case Study

Table 2.5 presents some examples that $\texttt{D}^3\texttt{I}_{\texttt{max}}$ is able to correctly predict on FEARS dataset. The first 5 drug combinations have single drugs that induce myopathy on their own (bold), and have various cardinalities. The last 5 drug combinations do not involve any myopathy-inducing single drugs, but all the drugs together in a combination still induce myopathy based on their odds ratios. These results show that $\texttt{D}^3\texttt{I}$ models do not trivially learn from single drugs that induce myopathy, but learn from the synergistic signals from multiple drugs in drug combinations for ADR prediction.

2.7.3 Comparison over Drug Combination Cardinalities

Performance Comparison over Drug Combinations of Various Cardinalities

Table 2.6 presents the cardinality distribution of drug combinations in the FEARS dataset. The majority of the drug combinations with ADRs is of order/cardinality 2 or 3. Please note that the total number of drug combinations for different features may be different due to the availability of different features on the drug combinations (Section 2.5.1).

Table 2.7 and 2.8 presents the model performance over drug combinations of each cardinality using TG as drug features. In the experiments, all the drug combinations of various cardinalities are used for model training and only drug combinations of each respective cardinality are tested. Table 2.7 shows that in TPTN setting, interestingly, all the methods share a similar trend in their performance over cardinalities, that is, the F1 values in general increase as the cardinalities increase. However, Table 2.8 presents that in TPRN, all the methods tend to achieve their best performance in F1 at drug combination cardinality 3 or 4, and the performance tends to remain similar even when the cardinality increases. In TPTN, as cardinality increases, the true positive drug combinations become more than the true negative drug combinations (Table 2.6). Therefore, $\texttt{D}^3\texttt{I}$ model training in TPTN is biased by the true positive drug combinations of higher cardinalities, and the true negative drug combinations of lower cardinalities. Consequentially, all $\texttt{D}^3\texttt{I}$ methods in TPTN tend to

have better precision and recall performance on drug combinations of higher cardinalities. Please note that $\texttt{D}^3\texttt{I}$ methods are cardinality-invariant and they do not use the cardinality information in prediction. The biased performance in TPTN, although not preferable, actually demonstrates that $\texttt{D}^3\texttt{I}$ methods do learn signals from the multiple drugs in drug combinations.

The strong ability of $\texttt{D}^3\texttt{I}$ methods in learning from multiple drugs in drug combinations is also demonstrated by their performance in TPRN in Table 2.8. In TPRN, each true positive drug combination will have a corresponding negative drug combinations of same cardinality, and thus the learning of $\texttt{D}^3\texttt{I}$ models will not be biased by the unbalanced distribution between positive and negative drug combinations. In TPRN, $\texttt{D}^3\texttt{I}_{\texttt{max}}$ is able to achieve F1 values above 0.760 for cardinalities higher than 3. In particular, for higher cardinalities, $\texttt{D}^3\texttt{I}_{\texttt{max}}$ achieves even better performance, for example, for cardinality higher than or equal to 8, $\texttt{D}^3\texttt{I}_{\texttt{max}}$ achieves F1 value 0.811.

Performance Comparison over Order-2 Drug Combinations

Table 2.9 presents the testing results on drug pairs (i.e., drug combinations of cardinality 2) using drug combinations of only cardinality 2 for model training, and using all cardinalities for model training, in $\texttt{D}^3\texttt{I}$ methods. All the experiments are conducted in TPRN setting to avoid biases from imbalanced training data distributions. Table 2.9 shows that when only drug pairs are used for training (i.e., the first column block in Table 2.9), the best F1 performance is 0.680, achieved by $\texttt{D}^3\texttt{I}_{\texttt{mean}}$ (with FP as the drug features), and the best AUC performance is 0.765, achieved by $\texttt{D}^3\texttt{I}_{\texttt{max}}$ (with TG as the drug features). However, when drug combinations of all cardinalities are used for training (i.e., the second column block in Table 2.9), the best F1 performance is 0.685, achieved by $\texttt{D}^3\texttt{I}_{\texttt{max}}$ (with FP as the drug features), and the best AUC performance is 0.786, achieved by $\texttt{D}^3\texttt{I}_{\texttt{max}}$ (with TG as the drug features). The better performance using all-cardinality drug combinations for training demonstrates that $\texttt{D}^3\texttt{I}$ methods do not trivially consider drug combination cardinalities in learning and prediction, but do learn the signals from all drug combinations.

In addition, when all-cardinality drug combinations are used for training, D^3I methods are able to capture the more and richer information carried by those drug combinations, and thus better predict drug pairs.

2.8 Conclusions and Future Research

In this chapter, we presented our deep learning model D^3I for predicting adverse drug reactions induced by high-order drug-drug interactions. D^3I is able to predict for drug combinations of arbitrary numbers of drugs, and generate meaningful embeddings for single drugs and drug combinations. We tested D^3I on two real datasets, one involving pairwise drug-drug interactions and the other involving high-order drug-drug interactions. Our experimental results demonstrate that D^3I is able to achieve superior performance on high-order drug-drug interaction prediction.

In D^3I, different drug features (e.g., target profiles, side effect profiles) are used independently. Effective integration of such features together may better represent drugs and their properties, and thus enable better performance of deep learning models. In our future work, we will explore feature integration and fusion in D^3I models. In addition, other information may be also highly related to drug-drug interactions and their induced adverse reactions, such as protein pathways and evidences from electronic medical records. We also plan to explore effective methods to integrate such information in D^3I models to further improve D^3I performance. Interpretability and evidence support are important for prediction methods in biomedical applications. A known issue in deep learning is its lack of interpretability by design, and thus it is worthwhile to address the interpretability issues of D^3I (e.g., what each layer learns, what the embeddings represent) in our future research. Mining evidences to support high-order drug-drug interactions and their adverse reactions from literature and electronic medical records is a challenging, related task that we would like to explore in the future.

Table 2.3.
Overall Performance on FEARS Dataset

method	TPTN					
	feature	acc	pre	rec	F1	AUC
D^3I_{max}	TG	**<u>0.823</u>**	**<u>0.862</u>**	0.773	**<u>0.815</u>**	**<u>0.892</u>**
	TG	0.815	0.834	**<u>0.790</u>**	0.811	0.889
D^3I_{mean}	FP	**0.773**	**0.790**	0.744	**0.766**	**0.842**
	FP	0.742	0.734	**0.762**	0.747	0.823
	TG	0.761	0.768	0.750	0.759	0.833
D^3I_{Att}	TG	**0.758**	0.768	0.744	**0.756**	**0.834**
	FP	0.753	**0.772**	0.720	0.745	0.819
	FP	0.753	0.756	**0.749**	0.752	0.828
method	TPRN					
	feature	acc	pre	rec	F1	AUC
D^3I_{max}	TG	**<u>0.762</u>**	**<u>0.813</u>**	0.680	**<u>0.740</u>**	**<u>0.847</u>**
	SE	0.700	0.689	**<u>0.748</u>**	0.714	0.784
D^3I_{mean}	TG	0.706	0.708	0.702	**0.704**	0.767
	TG	**0.707**	**0.717**	0.683	0.699	**0.770**
	SE	0.665	0.650	**0.721**	0.683	0.738
D^3I_{Att}	TI	**0.703**	**0.750**	0.609	**0.672**	**0.760**
	FP	0.649	0.647	**0.661**	0.653	0.719
	SE	0.668	0.675	0.647	0.661	0.737

Columns “acc”, “pre”, “rec”, “F1” and “AUC” correspond to accuracy, precision, recall, F1 and AUC. The best performance for each method under each evaluation metric is **fold**. The best performance over all the methods is <u>underlined</u>.

Table 2.4.
Overall Performance on BMC Dataset (TPRN)

method	feature	acc	pre	rec	F1	AUC
D^3I_{max}	OSE	**0.693**	**0.663**	**0.788**	**<u>0.720</u>**	**0.744**
D^3I_{mean}	OSE	**0.687**	**<u>0.669</u>**	0.742	**0.703**	**0.743**
	TI	0.681	0.659	**0.752**	0.702	0.734
D^3I_{Att}	OSE	**0.670**	0.635	**<u>0.803</u>**	**0.709**	**0.710**
	TI	**0.670**	**0.640**	0.779	0.702	0.707
neighbor recommender [46]	OSE	**0.951**	**0.629**	**0.765**	**0.691**	**0.940**
random walk [47]	TI	**<u>0.952</u>**	**0.641**	**0.768**	**0.699**	**0.941**
matrix perturbation [48]	-	**<u>0.952</u>**	**0.666**	**0.755**	**0.707**	**<u>0.948</u>**

Columns "acc", "pre", "rec", "F1" and "AUC" correspond to accuracy, precision, recall, F1 and AUC. The original paper [3] did not report drug features used in matrix perturbation method. The best performance for each method under each evaluation metric is **fold**. The best performance over all the methods is <u>underlined</u>.

Table 2.5.
Examples of Correctly Predicted Drug Combinations by $\mathtt{D^3I_{max}}$ (FEARS, TG, TPRN)

idx	drug combination
1	acetaminophen, alprazolam, **amitriptyline**, amlodipine, anastrozole, azithromycin, baclofen, buprenorphine, calcium, cevimeline, ciprofloxacin, doxycycline, duloxetine, **escitalopram**, estradiol, **fentanyl**, fondaparinux sodium, fulvestrant, furosemide, gabapentin, glucosamine, hydrochlorothiazide, hydrocodone, hydromorphone, ibuprofen, levofloxacin, lidocaine, **methadone**, methocarbamol, metoprolol, montelukast, morphine, moxifloxacin, omeprazole, oxycodone, pamidronate, pantoprazole, pentosan polysulfate, potassium, **pregabalin**, rabeprazole, triamterene, valaciclovir, **valdecoxib**, vitamin c, zoledronate, zolpidem
2	alprazolam, pioglitazone, **rosuvastatin**, **sunitinib**, tamsulosin, valsartan
3	alendronate, cetaminophen, chlorpheniramine, codeine, naproxen, **prednisolone**, **zopiclone**
4	atenolol, **pravastatin**
5	diphenhydramine, hydromorphone, montelukast, omeprazole, razepam, **triamcinolone**
6	gabapentin, haloperidol, morphine, propofol
7	alprazolam, diazepam, diclofenac, dicyclomine, etizolam, losartan, sulpiride
8	atenolol levofloxacin
9	amikacin, amiodarone
10	acetaminophen, alendronate, oxycodone

The drugs that induce myopathy on their own are **bold**.

Table 2.6.
Cardinality Distribution in FEARS (TPTN)

feature		total	2	3	4	5	6	7	$\geq$8
	all	6,338	2,981	1,555	652	323	220	157	450
FP	#pos	3,169	865	841	442	263	195	138	425
	#neg	3,169	2,116	714	210	60	25	19	25
	total	5,621	2,809	1,395	544	252	169	132	320
TG	#pos	2,821	822	795	402	222	158	121	301
	#neg	2,800	1,987	600	142	30	11	11	19

The columns corresponding to "2", "3", ..., "$\geq$8" represent the numbers of drug combinations of cardinality 2, 3, ..., greater than 8. The row of "all" has the total number of drug combinations. The row of "#pos" has the numbers of positive drug combinations. The row of "#neg" has the numbers of negative drug combinations.

Table 2.7.
Performance Comparison of TPTN on Different Cardinalities in FEARS Dataset (TG)

method	cardinality	TPTN				
		acc	pre	rec	F1	AUC
D^3I_{max}	2	0.810	0.749	0.532	0.621	0.821
	3	0.789	0.853	0.763	0.804	0.870
	4	0.819	0.879	0.875	0.877	0.848
	5	0.913	0.954	0.949	0.950	0.925
	6	0.918	0.944	0.971	0.956	0.563
	7	0.909	0.913	0.994	0.951	0.716
	$\geq$8	0.938	0.941	0.997	0.968	0.709
D^3I_{mean}	2	0.754	0.562	0.719	0.631	0.811
	3	0.763	0.836	0.726	0.776	0.836
	4	0.718	0.901	0.695	0.784	0.818
	5	0.777	0.972	0.768	0.857	0.880
	6	0.726	0.940	0.758	0.834	0.479
	7	0.791	0.928	0.838	0.877	0.716
	$\geq$8	0.877	0.951	0.914	0.931	0.735
D^3I_{Att}	2	0.744	0.550	0.696	0.614	0.806
	3	0.750	0.838	0.698	0.760	0.838
	4	0.741	0.913	0.719	0.803	0.816
	5	0.786	0.972	0.778	0.864	0.867
	6	0.756	0.942	0.789	0.857	0.394
	7	0.833	0.930	0.883	0.904	0.701
	$\geq$8	0.897	0.956	0.933	0.944	0.677

Columns “acc”, “pre”, “rec”, “F1” and “AUC” correspond to accuracy, precision, recall, F1 and AUC. Target profiles (TG) are used as drug features. The best results presented for each drug combination cardinality are selected based on F1.

Table 2.8.
Performance Comparison of TPRN on Different Cardinalities in FEARS Dataset (TG)

method	cardinality	TPRN				
		acc	pre	rec	F1	AUC
D^3I_{max}	2	0.697	0.752	0.587	0.659	0.786
	3	0.774	0.808	0.719	0.760	0.847
	4	0.775	0.798	0.737	0.766	0.862
	5	0.793	0.883	0.676	0.765	0.885
	6	0.799	0.900	0.679	0.770	0.887
	7	0.802	0.879	0.704	0.780	0.882
	$\geq$8	0.828	0.895	0.743	0.811	0.913
D^3I_{mean}	2	0.664	0.642	0.744	0.688	0.732
	3	0.723	0.707	0.766	0.734	0.786
	4	0.748	0.760	0.724	0.740	0.806
	5	0.750	0.812	0.653	0.722	0.837
	6	0.724	0.824	0.570	0.672	0.803
	7	0.760	0.873	0.611	0.711	0.827
	$\geq$8	0.670	0.940	0.364	0.523	0.824
D^3I_{Att}	2	0.665	0.674	0.644	0.658	0.726
	3	0.651	0.652	0.645	0.648	0.735
	4	0.680	0.683	0.670	0.675	0.753
	5	0.662	0.693	0.583	0.631	0.761
	6	0.651	0.671	0.587	0.622	0.718
	7	0.689	0.731	0.595	0.653	0.773
	$\geq$8	0.632	0.687	0.508	0.576	0.733

Columns “acc”, “pre”, “rec”, “F1” and “AUC” correspond to accuracy, precision, recall, F1 and AUC. Target profiles (TG) are used as drug features. The best results presented for each drug combination cardinality are selected based on F1.

Table 2.9.
Best Performance on Cardinality-2 Drug Combinations (FEARS, TPRN)

method	training with cardinality-2 drug combinations					
	feature	acc	pre	rec	F1	AUC
D^3I_{max}	FP	0.655	0.637	**<u>0.724</u>**	**0.677**	0.710
	TG	**<u>0.697</u>**	**<u>0.742</u>**	0.610	0.668	**<u>0.765</u>**
D^3I_{mean}	FP	0.666	0.652	**0.714**	**<u>0.680</u>**	0.725
	TG	**0.695**	**0.725**	0.633	0.675	**0.747**
D^3I_{Att}	TG	**0.689**	0.721	**0.620**	**0.665**	**0.749**
	TG	0.687	**0.736**	0.589	0.653	0.736
method	training with all-cardinality drug combinations					
	feature	acc	pre	rec	F1	AUC
D^3I_{max}	FP	0.672	0.659	**0.715**	**<u>0.685</u>**	0.729
	TG	**<u>0.697</u>**	**<u>0.752</u>**	0.587	0.659	**<u>0.786</u>**
D^3I_{mean}	TG	**0.651**	0.630	**<u>0.742</u>**	**0.680**	**0.732**
	FP	0.635	**0.634**	0.646	0.638	0.697
D^3I_{Att}	TG	**0.665**	**0.674**	0.644	**0.658**	**0.726**
	FP	0.617	0.610	**0.656**	0.629	0.687

Columns "acc", "pre", "rec", "F1" and "AUC" correspond to accuracy, precision, recall, F1 and AUC. The best performance for each method under each evaluation metric is **fold**. The best performance over all the methods is <u>underlined</u>.

3. PRIORITIZING AMYLOID IMAGING BIOMARKERS IN ALZHEIMER'S DISEASE VIA LEARNING TO RANK

3.1 Introduction

Alzheimer's disease (AD) is a national priority, with 5.5 million Americans affected at an annual cost of $259 billion in 2017 and no available cure [2]. Brain characteristics related to AD progression may be captured by multimodal magnetic resonance imaging (MRI) [50] and positron emission tomography (PET) [51] scans. Thus, there is a large body of neuroimaging studies in AD, aiming to develop image-based predictive machine learning models for early detection of AD as well as identification of relevant imaging biomarkers (e.g., [10–12, 52]). These models are typically designed to accomplish learning tasks such as classification [53], regression [12, 54, 55] or both [56]. The identified imaging biomarkers are at the population level and not specific to an individual subject. Although such studies can improve the mechanistic understanding of AD, they are not designed to directly impact clinical practice.

In this work, we propose a novel learning paradigm that embraces the concept of precision medicine and tailors the imaging biomarker discovery process to the individual characteristics of a given patient. Specifically, we perform an innovative application of a newly developed learning-to-rank method, denoted as `PLTR` [57], to the structural MRI and amyloid PET data of the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort [58]. Using structural MRI as the individual characteristics, our goal is to not only predict individual-specific amyloid imaging biomarkers but also prioritizes them according to AD-specific abnormality. Note that amyloid imaging is more expensive and more invasive than structural MRI. Compared with traditional biomarker studies at the population level, the uniqueness of our study is twofold: (1) the identified biomarkers are tailored to each individual patient; and (2) the identified biomarkers are prioritized based on the

individual's characteristics, which has the potential to enable personalized diagnosis (e.g., determining whether or not the corresponding test is needed) and disease subtyping.

3.2 Materials and Data Processing

To demonstrate the effectiveness of the learning-to-rank method for personalized prioritization of the amyloid imaging biomarkers, we applied it to the ADNI cohort [58]. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI, a prodromal stage of AD) and early AD. For up-to-date information, see www.adni-info.org.

Data used in the preparation of this article were obtained from the 2017 ADNI TADPOLE grand challenge (tadpole.grand-challenge.org/), and was downloaded from the ADNI website (adni.loni.usc.edu). The TADPOLE data used in this study consists of structural MRI and AV45-PET (amyloid) imaging data as well as diagnostic information. Both MRI and amyloid imaging data have been pre-processed with standard ADNI pipelines as described previously in [59].

In this study, we included all the regional MRI measures with field name containing "UCSFFSX" in the TADPOLE D1 and D2 data sets. Specifically, these are FreeSurfer regional volume and cortical thickness measures processed by the ADNI UCSF team. We also included all the regional amyloid measures with field name containing "UCBERKELEYAV45" in the TADPOLE D1 and D2 data sets. These are cortical and subcortical amyloid deposition measures processed by the ADNI UC Berkeley team.

Originally, there are totally 12,741 participant visit records with 103 amyloid features, 125 FreeSurfer features and diagnostic information corresponding to each visit. To convert this longitudinal data set into a cross-sectional one as well as handle the missing data issue, we use the following procedure to generate a clean set of cross-sectional data: (1) remove visit records that have more than 50 percent of null values across 103 amyloid features, with

10,623 records removed; (2) extract the earliest AV45-PET visit for each participant, with 1,091 records kept; (3) remove visit records that have more than 50 percent of null values across 125 FreeSurfer features, with 58 records removed; (4) remove features that have more than 50 percent of null values across records, with 16 FreeSurfer features removed; (5) remove 3 participants with no diagnostic information. Finally, 1,030 participants with 103 amyloid and 109 FreeSurfer measures are studied, including 351 health control (HC), 501 MCI and 178 AD participants. We treat both MCI and AD subjects as patients, and so have a total of 679 cases and 351 controls.

3.3 Methods

We use the joint push and learning-to-rank method as developed in He *et al.* [57], denoted as PLTR, for personalized patient feature prioritization. Our goal is to prioritize amyloid features for each patient that are most relevant to his/her disease diagnosis using patients' existing information (i.e., FreeSurfer measures extracted from MRI scans). The underlying hypothesis is that patients with similar FreeSurfer feature profiles would have similar ranking structures among their amyloid features. In the context of AD feature prioritization, PLTR learns and uses latent vectors of patients and amyloid features to score each amyloid feature for each patient, and ranks the features based on their scores; patients with similar FreeSurfer feature profiles will have similar latent vectors. During the learning process, PLTR explicitly pushes the most relevant amyloid features on top of the less relevant ones for each patient, and thus optimizes the latent patient and amyloid feature vectors so they will reproduce the pushed ranking structures.

3.3.1 Overview of PLTR

In PLTR, the ranking of features in terms of their relatedness to MCI/AD in a patient is determined by their latent scores on the patient. For a feature f_i and a patient $\mathscr{P}_p$, f_i's

latent score on $\mathscr{P}_p$, denoted as $s_p(f_i)$, is calculated as the dot product of f_i's latent vector $\mathbf{v}_i \in \mathbb{R}^{l\times 1}$ and $\mathscr{P}_p$'s latent vector $\mathbf{u}_p \in \mathbb{R}^{l\times 1}$, where l is the latent dimension, as follows,

$$s_p(f_i) = \mathbf{u}_p^\top \mathbf{v}_i, \tag{3.1}$$

where the latent vectors $\mathbf{u}_p$ and $\mathbf{v}_i$ will be learned. All the features are then sorted based on their scores on $\mathscr{P}_p$, with the most relevant features having the highest scores and ranked higher than irrelevant features.

PLTR leverages ranking with push [60] to enforce the high rank of relevant features. In PLTR, the height of an irrelevant feature f_i^- in $\mathscr{P}_p$, denoted as $h_s(f_i^-, \mathscr{P}_p)$, is used to measure the ranking position of f_i^- in $\mathscr{P}_p$ [60], and is determined as the number of relevant features that are ranked below f_i^-, that is,

$$h_s(f_i^-, \mathscr{P}_p) = \sum_{f_j^+ \in \mathscr{P}_p^+} \mathbb{I}(s_p(f_j^+) \leq s_p(f_i^-)), \tag{3.2}$$

where $\mathscr{P}_p^+$ is the set of relevant features in patient $\mathscr{P}_p$, $s_p(f_j^+)$ and $s_p(f_i^-)$ are the scores of f_j^+ and f_i^- in $\mathscr{P}_p$, respectively, and $\mathbb{I}(x)$ is the indicator function ($\mathbb{I}(x) = 1$ if x is true, otherwise 0). To rank relevant features higher in a patient, PLTR minimizes the total height of all irrelevant features in that patient (i.e., minimize the total number of relevant features that are ranked below irrelevant features). For all the patients, PLTR minimizes the total heights, denoted as $P_s^\uparrow$, defined as,

$$P_s^\uparrow = \sum_{p=1}^{m} \frac{1}{n_p^+ n_p^-} \sum_{f_i^- \in \mathscr{P}_p} h_s(f_i^-, \mathscr{P}_p), \tag{3.3}$$

where m is the number of patients, and n_p^+ and n_p^- are the numbers of relevant and irrelevant features in patient $\mathscr{P}_p$, respectively. The normalization by n_p^+ and n_p^- is to eliminate the effects due to different numbers of relevant and irrelevant features across the patients.

In addition to pushing relevant features on top of irrelevant features, PLTR uses $f_i \succ_R f_j$ to represent that f_i is ranked higher than f_j under the relation R. The concordance index (CI) [61] is used to measure feature ranking structures compared to the ground truth, which is defined as follows,

$$\mathtt{CI}(\{f_i\}, \mathscr{P}, s) = \frac{1}{|\{f_i \succ_{\mathscr{P}} f_j\}|} \sum_{f_i \succ_{\mathscr{P}} f_j} \mathbb{I}(f_i \succ_s f_j), \tag{3.4}$$

where $\{f_i\}$ is the set of features in patient $\mathscr{P}$, $\{f_i \succ_{\mathscr{P}} f_j\}$ is the set of ordered pairs of features in patient $\mathscr{P}$ ($f_i \succ_{\mathscr{P}} f_j$ represents that f_i is more relevant, and thus ranked higher, than f_j in $\mathscr{P}$), s is the scoring function (Equation (3.1)) that produces an estimated feature ranking, $f_i \succ_s f_j$ represents that f_i is ranked higher than f_j by the scoring function s, and $\mathbb{I}$ is the indicator function. Essentially, `CI` measures the ratio of correctly ordered feature pairs by s among all possible pairs. Higher `CI` values indicate better ranking structures.

To produce correct ranking orders among relevant features in all the patients, `PLTR` minimizes O_s^+ as part of its objective, which is defined as the sum of $1 - \texttt{CI}$ values (i.e., the ratio of mis-ordered feature pairs among all pairs) over the relevant features of all the patients as follows,

$$O_s^+ = \sum_{p=1}^{m} [1 - \texttt{CI}(\{f_i^+\}, \mathscr{P}_p, s_p)] = \sum_{p=1}^{m} \frac{1}{|\{f_i^+ \succ_{\mathscr{P}_p} f_j^+\}|} \sum_{f_i^+ \succ_{\mathscr{P}_p} f_j^+} \mathbb{I}(f_i^+ \prec_{s_p} f_j^+). \tag{3.5}$$

Overall, `PLTR` seeks the patient latent vectors and feature latent vectors that will be used in feature scoring function s (Equation (3.1)) such that for each patient, the relevant features will be ranked on top and in right orders using the latent vectors. In `PLTR`, such latent vectors are learned by solving the following optimization problem:

$$\min_{U,V} \mathscr{L}_s = (1-\alpha) P_s^{\uparrow} + \alpha O_s^+ + \frac{\beta}{2} R_{uv} + \frac{\gamma}{2} R_{\text{csim}}, \tag{3.6}$$

where $\mathscr{L}_s$ is the overall loss function; $P_s^{\uparrow}$ and O_s^+ are defined in Equation (3.3) and Equation (3.5), respectively; $U = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m]$ and $V = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n]$ are the latent vector matrices for patients and features, respectively ($U \in \mathbb{R}^{l \times m}$, $V \in \mathbb{R}^{l \times n}$, where l is the latent dimension); α ($\alpha \in [0,1]$) is a weighting parameter to control the contribution from push (i.e., $P_s^{\uparrow}$) and ranking (i.e., O_s^+); β and γ are regularization parameters ($\beta \geq 0$, $\gamma \geq 0$) on the two regularizers R_{uv} and R_{csim}, respectively. Note that in `PLTR`, only AD/MCI patients are used for model training, as the feature prioritization is for AD/MCI patients, and makes little sense for HCs.

In Problem (3.6), R_{uv} is a regularizer on U and V to prevent overfitting, defined as

$$R_{uv} = \frac{1}{m} \|U\|_F^2 + \frac{1}{n} \|V\|_F^2, \tag{3.7}$$

where $\|X\|_F$ is the Frobenius norm of matrix X. R_{csim} is a regularizer on patients to constrain patient latent vectors, defined as

$$R_{\text{csim}} = \frac{1}{m^2} \sum_{p=1}^{m} \sum_{q=1}^{m} w_{pq} \|\mathbf{u}_p - \mathbf{u}_q\|_2^2, \tag{3.8}$$

where w_{pq} is the similarity between $\mathscr{P}_p$ and $\mathscr{P}_q$ that is calculated using FreeSurfer features of the patients.

3.3.2 Patient Similarities from FreeSurfer Features

We consider 109 FreeSurfer features and represent each patient as a FreeSurfer feature vector, denoted as $\mathbf{r}_p = [r_{p1}, r_{p2}, \cdots, r_{pn_r}]$, where r_{pi} $(i = 1, \cdots, n_r)$ is a FreeSurfer feature for patient p. Thus, for all the patients, we construct a FreeSurfer feature matrix $R_{\text{AD}} = [\mathbf{r}_1^+; \mathbf{r}_2^+; \cdots; \mathbf{r}_{m^+}^+] \in \mathbb{R}^{m^+ \times n_r}$ and for all the health control subjects (HCs), a FreeSurfer feature matrix $R_{\text{HC}} = [\mathbf{r}_1^-; \mathbf{r}_2^-; \cdots; \mathbf{r}_{m^-}^-] \in \mathbb{R}^{m^- \times n_r}$, where m^+ and m^- are the numbers of AD/MCI patients and HCs, respectively, and n_r is the number of FreeSurfer features. We scale R_{AD} values into the unit interval by dividing each column of R_{AD} (i.e., each FreeSurfer feature) its maximum value. The normalized R_{AD} matrix is denoted as $\bar{R}_{\text{AD}}$, and the similarities between patients are calculated over $\bar{R}_{\text{AD}}$ using the radial basis function (RBF) kernel:

$$w_{pq} = \exp\left(-\frac{\|\bar{R}_{\text{AD}}(p,:) - \bar{R}_{\text{AD}}(q,:)\|^2}{2\sigma^2}\right), \tag{3.9}$$

where w_{pq} is the patient similarity used in Equation (3.8). This patient similarity measurement is denoted as `simU`.

3.3.3 Patient Amyloid Features in Ground Truth

Similarly, each patient is also represented by an amyloid feature vector, denoted as $\mathbf{c}_p = [c_{p1}, c_{p2}, \cdots, c_{pn_c}]$, where c_{pi} $(i = 1, \cdots, n_c)$ is an amyloid feature for patient p. Thus, we construct an amyloid feature matrix $C_{\text{AD}} = [\mathbf{c}_1^+; \mathbf{c}_2^+; \cdots, \mathbf{c}_{m^+}^+]$ for AD/MCI patients, and an amyloid feature matrix $C_{\text{HC}} = [\mathbf{c}_1^-; \mathbf{c}_2^-; \cdots, \mathbf{c}_{m^-}^-]$ for HC subjects. We normalize C_{AD} by

dividing each column of C_{AD} (i.e., each amyloid feature) by the mean value of the corresponding column in C_{HC}. Thus, the normalization results in C_{AD} measure the extent to which an amyloid feature in patients deviates from that in HCs. The normalized matrix, denoted as $\bar{C}_{\mathrm{AD}}$, is used as the ground truth of amyloid feature ranking. That is, the optimization problem (3.6) tries to learn the latent vectors that reconstruct the ordering structures in $\bar{C}_{\mathrm{AD}}$, and through such reconstruction prioritize amyloid features that are most relevant to patients. The reason why we use FreeSurfer features to quantitatively measure patients and prioritize amyloid features correspondingly is that MRI imaging is non-invasive and relatively low-cost as compared to PET imaging.

3.4 Experimental Protocol

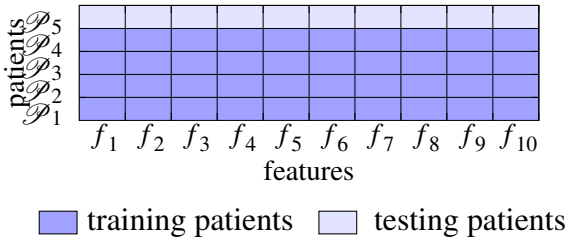

Fig. 3.1. Data split for testing new patients

We split patients into training and testing set, such that a certain patient and all his/her features will be either in the training set or in the testing set. We train the PLTR model using training patients and test its performance on the testing patients. This corresponds to the use scenario in which we want to identify the most potentially useful AD biomarkers for new patients, based on the existing information of the patients, when such biomarkers have not been tested on the new patients. Figure 3.1 demonstrates the data split process.

We define average hit at k, denoted as AH@k, to evaluate the ranking performance. AH@k is defined as follows:

$$\text{AH@}k(\tau, \tilde{\tau}) = \sum_{i=1}^{k} \mathbb{I}(\tilde{\tau}_i \in \tau), \tag{3.10}$$

where τ is the ground-truth ranking list, $\tilde{\tau}$ is the predicted ranking list, and $\tilde{\tau}_i$ is the i-th ranked item in $\tilde{\tau}$. That is, AH@k calculates the number of items among top k in the predicted lists that are also in the ground truth (i.e., hits). Higher AH@k values indicate better prioritization performance.

We define a second evaluation metric weighted average high at k, denoted as WAH@k as follows:

$$\text{WAH@}k(\tau, \tilde{\tau}) = \sum_{j=1}^{k} AH@j(\tau, \tilde{\tau})/k, \tag{3.11}$$

that is, WAH@k is a weighted version of AH@k that calculates the average of AH@ over top k. Higher WAH@k indicate more hits and those hits are ranked on top in the ranking list. By default, the ground-truth τ has k items (i.e., the top-k items among all the sorted items) in Equation (3.10) and Equation (3.11).

3.5 Experimental Results

3.5.1 Overall Performance

We first hold out 35 and 163 patients as testing patients, respectively. These testing patients are determined such that they have more than 10 similar patients in the training set, and the corresponding patient similarities are higher than 0.75 and 0.65, respectively. Patient latent vectors and feature latent vectors are learned on the training patients. The feature scores for the testing patients are calculated as the weighted sum of the predicted feature scores from their top-10 most similar training patients, where the weights are the corresponding patient similarities. The patient similarities are calculated using `simU` (Equation (3.9), $\sigma = 1$). The patient amyloid features are normalized as described in Section 3.3.3. Please note that we only use patients (i.e., MCI and AD subjects) for model

training and testing, and only use controls (i.e., HC subjects) to set the standard for patient data normalization, as feature prioritization for healthy controls has limited clinical interests.

Table 3.1 presents the testing performance of PLTR in terms of AH@5 for each latent dimension. When 35 patients are hold out for testing, the best AH@5 is 1.886 when latent dimension $d = 20$, and the corresponding AH@10 is 3.286. This performance is significantly better than random, which has a theoretical AH@5 value lower than 0.25 for more than 100 features. Note that we use predicted feature scores to prioritize features for the testing patients. A baseline is to use the weighted sum of the ground-truth feature values from the similar training patients, which does not require any model training. This baseline method has an AH@5 1.714 in our data, whereas the learning-based PLTR achieves 10.0% better performance (i.e., 1.886) than the baseline. When 163 patients are hold out for testing, the best performance of PLTR (i.e., AH@5 1.429 when $d = 20$) is 5.9% better than its baseline (i.e., AH@5 1.350). This indicates that PLTR is able to capture the signals that lead to accurate feature rankings among training data, potentially correct the noise in the data and use the signals to prioritize features for new patients.

Table 3.1 also shows that the best testing performance for the 35 testing patients is better than that for the 163 testing patients. In addition, the performance improvement over the baseline for the 35 testing patients is also better than that for the 163 testing patients. This indicates that as long as there are sufficiently similar patients for modeling training, PLTR is able to achieve stronger performance than the baseline.

Feature Prioritization on Population Level

We also investigate which features are frequently prioritized for all the testing patients. We sort all the top-5 ranked features from all the testing patients, weighted by their aggregated ranking positions among the patients, so that features that are frequently ranked high among many patients will be sorted on top. Table 3.2 lists the top 10 of such frequently prioritized features by PLTR among the 163 testing patients. Among these 10 features, 8 of

Table 3.1.
Overall Performance of PLTR (simU, $\sigma = 1$)

n	α	β	γ	d	AH@5	WAH@5	AH@10	WAH@10
	0.3	0.5	1.0	10	1.857	1.545	**3.371**	**2.249**
35	0.3	0.5	1.0	20	**1.886**	**1.632**	3.286	1.987
	0.3	0.5	1.0	50	1.857	1.560	3.314	2.007
	0.5	1.0	1.0	10	1.343	0.930	3.080	**2.497**
163	0.5	1.0	1.0	20	**1.429**	**1.067**	3.074	2.402
	0.5	1.0	1.0	50	**1.429**	1.012	**3.110**	2.437

The column "n" corresponds to the number of hold-out testing patients. Best performance under each evaluation metric is in **bold**. Baseline AH@5 performance for $n = 35$ is 1.714, and for $n = 163$ 1.350.

them are among the top 10 identified from the ground truth. Similarly, for the 35 testing patients, 7 of the top-10 most frequently prioritized features are among the top 10 identified from the ground truth. This indicates the capability of PLTR to find common AD biomarkers on a population level.

Most of the above top ranked amyloid features are related to AD or its biomarkers. For example, frontal lobe, the region where frontal pole, rostral middle frontal gyrus and medial orbitofrontal cortex are located, shows significantly higher amyloid deposition in AD/MCI patients than in MCI [62]. Furthermore, Huang *et al.* [63] report that both frontal lobe and precuneus show significantly higher amyloid deposition in both MCI and AD compared to HC. Additionally, they report the negative correlation between MiniMental State Examination (MMSE) score with amyloid deposition in frontal lobe and precuneus, which further validates increased amyloid deposition in these regions of MCI and AD patients.

Table 3.2.
Top-10 frequent features by PLTR (simU, $\sigma = 1$)

rank	features	p-value	GT
1	ctx-lh-frontal pole	8.67e-20	Y
2	ctx-rh-frontal pole	5.68e-20	Y
3	right-lateral ventricle	4.34e-04	Y
4	ctx-rh-medial orbitofrontal	4.79e-23	Y
5	left-lateral ventricle	1.09e-04	Y
6	ctx-lh-rostral middle frontal	5.12e-21	Y
7	right-choroid plexus	4.41e-05	N
8	ctx-rh-rostral middle frontal	3.68e-20	N
9	ctx-lh-precuneus	3.19e-19	Y
10	non-wm-hypointensities	8.75e-01	Y

The p-value measures whether the feature means are statistically different between controls and patients. Column "GT" indicates if the feature is in ground truth (Y) or not (N). These features are frequently prioritized by PLTR when 163 patients are hold out for testing.

3.5.2 Study on Patient-Patient Similarities

Table 3.3 presents the testing performance when a different patient similarity is applied. In this case, the patient similarities are calculated using a RBF kernel ($\sigma = 5$) on the FreeSurfer features of the patients, after the FreeSurfer features are divided by the corresponding feature mean from normal patients. This feature normalization measures how much the FreeSurfer features in patients deviate from those in HCs. This similarity measurement is denoted as simN. 62 patients are hold out for testing, who have at least 10 training patients each with patient similarities higher than 0.65. The feature ranking is done in a same way as in Section 3.5.1. The corresponding baseline performance in terms of AH@5 is 1.081. Table 3.3 shows that the PLTR outperforms the baseline at 29.8%. Table 3.3 and Table 3.1 together demonstrate that regardless of similar functions used to measure patient

similarities in FreeSurfer features, PLTR is robust in outperforming baseline given that the testing patients have sufficient similar training patients.

Table 3.3.
Overall Performance of PLTR (simN, $\sigma = 5$)

n	α	β	γ	d	AH@5	WAH@5	AH@10	WAH@10
	0.5	1.0	1.0	10	1.371	1.161	3.129	2.295
62	0.5	1.0	1.0	20	1.387	1.186	3.081	2.162
	0.5	1.0	1.0	50	**1.403**	1.165	3.113	2.117

The column "n" corresponds to the number of hold-out testing patients. Best performance under each evaluation metric is in **bold**. Baseline AH@5 performance for $n = 62$ is 1.081.

3.6 Conclusions

We have proposed an innovative machine learning paradigm enabling precision medicine for AD imaging biomarker discovery. The paradigm tailors the imaging biomarker discovery process to individual characteristics of a given patient, and has been implemented based on a newly developed learning-to-rank method PLTR. To the best of our knowledge, this learning-to-rank method has never been applied to the AD imaging biomarker studies. It is a paradigm shifting strategy to facilitate precision medicine research in brain imaging study of AD. The PLTR model seamlessly integrates two objectives for joint optimization: pushing up relevant biomarkers and ranking among relevant biomarkers. The empirical study of PLTR has been performed on the ADNI data and yielded promising results to identify and prioritize individual-specific amyloid imaging biomarkers based on the individual's structural MRI data.

Our overarching goal is to enable precision medicine for AD imaging biomarker discovery. The proposed paradigm not only identifies individual-specific imaging biomarkers but also prioritizes them according to AD-specific abnormality. The resulting top ranked imaging biomarkers have the potential to aid personalized diagnosis and disease subtyping.

With this observation, work is in progress to expand the proposed method into a new model that can not only identify top ranked imaging biomarkers in a subject specific manner but also use this reduced set of biomarkers for accurate prediction of outcome of interest such as diagnostic status or conversion to AD.

4. SUMMARY

In this dissertation, I have developed a new deep model $\texttt{D}^3\texttt{I}$ for high-order drug-drug interaction prediction. To my best knowledge, it's the first model that is able to conduct cardinality-invariant and order-invariant high-order DDI prediction. Moreover, I have proposed a novel machine learning paradigm enabling amyloid imaging biomarker discovery and prioritization. The proposed paradigm can find the most informative amyloid features of each patient. Therefore, the paradigm can help significantly save the diagnosis time and reduce costs, while maintain similar diagnostic power.

Studies showed that above half of the elderly Americans take two or more drugs daily and about one-third of elderly American patients take more than 5 drugs daily. It makes detecting drug-drug interactions (DDI) an urgent and crucial task for keeping the patients from adverse drug reactions. Most of the existing DDI studies facus on pairwise DDI prediction. However, considerable amount of patients take more then 2 drugs daily. It's highly desired to develop efficient computational tools for arbitrary-order DDI prediction. As a possible solution to this problem, I developed deep learning based methods, denoted as $\texttt{D}^3\texttt{I}$. The developed methods contain an encoder, an aggregator and a predictor. The encoder encodes each of the drugs in an input drug combination into a latent representation. The aggregator takes the embedding of the drugs in a drug combination as input to learn a single high-level representation for the drug combination. The predictor then predicts the probability of this drug combination in inducing ADRs using the drug combination representation. In this study, I considered 3 aggregation strategies: max pooling, mean pooling and aggregation with attentions. Their performance is evaluated and compared on multiple public datasets. The experimental results demonstrate that the proposed methods $\texttt{D}^3\texttt{I}$ is able to achieve promising results and $\texttt{D}^3\texttt{I}$ outperforms other classic methods on order-2 DDI prediction.

Another healthcare problem I have studied is prioritizing amyloid imaging biomarkers in Alzheimer's Disease. Alzheimer's Disease (AD) is an irreversible brain disorder that will cause memory loss, think skills loss and, eventually, the loss of viability. Currently, no drugs are able to cure AD and there is no valid treatment for patients whose condition begins to deteriorate. Therefore, detecting AD in the early stage is crucial for the further treatment. In this dissertation, I have proposed a novel machine learning paradigm that enabling individual-specific amyloid imaging biomarkers discovery and prioritization that can help detect the progression of AD. I implemented the paradigm using a newly developed learning-to-rank method `PLTR`, which learns the latent representation of patients and amyloid features. The learned representations are used to score the relevancy of amyloid features to patients. We then rank the features based on their scores. I evaluated the paradigm on a subset of Alzheimers Disease Neuroimaging Initiative (ADNI) cohort, which includes 103 amyloid features and 109 FreeSurfer features. The experimental results are promising and demonstrate that the top ranked imaging biomarkers (i.e. amyloid features) have the potential to aid personalized diagnosis and disease subtyping.

REFERENCES

REFERENCES

[1] https://www.cdc.gov/nchs/nhanes/about_nhanes.htm, 2018.

[2] Alzheimer's Association, "2017 Alzheimer's disease facts and figures," 2017.

[3] W. Zhang, Y. Chen, F. Liu, F. Luo, G. Tian, and X. Li, "Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data," *BMC Bioinformatics*, vol. 18, no. 1, p. 18, Jan 2017. [Online]. Available: https://doi.org/10.1186/s12859-016-1415-9

[4] C.-S. Wang, P.-J. Lin, C.-L. Cheng, S.-H. Tai, Y.-H. K. Yang, and J.-H. Chiang, "Detecting potential adverse drug reactions using a deep neural network model," *Journal of medical Internet research*, vol. 21, no. 2, p. e11016, 2019.

[5] A. Kolchinsky, A. Lourenço, L. Li, and L. M. Rocha, "Evaluation of linear classifiers on articles containing pharmacokinetic evidence of drug-drug interactions," in *Biocomputing 2013*. World Scientific, 2013, pp. 409–420.

[6] S. V. Iyer, R. Harpaz, P. LePendu, A. Bauer-Mehren, and N. H. Shah, "Mining clinical text for signals of adverse drug-drug interactions," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 353–362, 2013.

[7] F. Hammann and J. Drewe, "Data mining for potential adverse drug–drug interactions," *Expert opinion on drug metabolism & toxicology*, vol. 10, no. 5, pp. 665–671, 2014.

[8] Y. Zhang and D.-Y. Yeung, "Learning high-order task relationships in multi-task learning," in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[9] S. Yan, X. Jiang, and Y. Chen, "Text mining driven drug-drug interaction detection," in *2013 IEEE International Conference on Bioinformatics and Biomedicine*. IEEE, 2013, pp. 349–354.

[10] J. Dukart, F. Sambataro, and A. Bertolino, "Accurate prediction of conversion to alzheimer's disease using imaging, genetic, and neuropsychological biomarkers," *J Alzheimers Dis*, vol. 49, no. 4, pp. 1143–59, 2015.

[11] M. Ten Kate, F. Barkhof, P. J. Visser, C. E. Teunissen, P. Scheltens, W. M. van der Flier, and B. M. Tijms, "Amyloid-independent atrophy patterns predict time to progression to dementia in mild cognitive impairment," *Alzheimers Res Ther*, vol. 9, no. 1, p. 73, 2017.

[12] J. Wan, Z. Zhang, B. D. Rao, S. Fang, J. Yan, A. J. Saykin, and L. Shen, "Identifying the neuroanatomical basis of cognitive impairment in Alzheimer's disease by correlation- and nonlinearity-aware sparse Bayesian learning," *IEEE Trans Med Imaging*, vol. 33, no. 7, pp. 1475–87, 2014.

[13] B. Percha and R. B. Altman, "Informatics confronts drug–drug interactions," *Trends in pharmacological sciences*, vol. 34, no. 3, pp. 178–184, 2013.

[14] R. Harpaz, W. DuMouchel, N. H. Shah, D. Madigan, P. Ryan, and C. Friedman, "Novel data-mining methodologies for adverse drug event discovery and analysis," *Clinical Pharmacology & Therapeutics*, vol. 91, no. 6, pp. 1010–1021, 2012.

[15] R. Harpaz, K. Haerian, H. S. Chase, and C. Friedman, "Statistical mining of potential drug interaction adverse effects in fda's spontaneous reporting system." *AMIA Annu Symp Proc*, vol. 2010, pp. 281–285, 2010.

[16] N. Hameed, A. Ruskin, K. A. Hassan, and M. A. Hossain, "A comprehensive survey on image-based computer aided diagnosis systems for skin cancer," in *2016 10th International Conference on Software, Knowledge, Information Management Applications (SKIMA)*, Dec 2016, pp. 205–214.

[17] H. Yang and C. C. Yang, "Harnessing social media for drug-drug interactions detection," in *2013 IEEE International Conference on Healthcare Informatics*. IEEE, 2013, pp. 22–29.

[18] S. Vilar, C. Friedman, and G. Hripcsak, "Detection of drug–drug interactions through data mining studies using clinical sources, scientific literature and social media," *Briefings in bioinformatics*, vol. 19, no. 5, pp. 863–877, 2017.

[19] H. Ibrahim, A. Saad, A. Abdo, and A. S. Eldin, "Mining association patterns of drug-interactions using post marketing fdas spontaneous reporting data," *Journal of biomedical informatics*, vol. 60, pp. 294–308, 2016.

[20] H. Luo, P. Zhang, H. Huang, J. Huang, E. Kao, L. Shi, L. He, and L. Yang, "Ddicpi, a server that predicts drug–drug interactions through implementing the chemical–protein interactome," *Nucleic acids research*, vol. 42, no. W1, pp. W46–W52, 2014.

[21] S. Vilar, E. Uriarte, L. Santana, T. Lorberbaum, G. Hripcsak, C. Friedman, and N. P. Tatonetti, "Similarity-based modeling in large-scale prediction of drug-drug interactions," *Nature protocols*, vol. 9, no. 9, p. 2147, 2014.

[22] F. Cheng and Z. Zhao, "Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties," *Journal of the American Medical Informatics Association*, vol. 21, no. e2, pp. e278–e286, 2014.

[23] N. P. Tatonetti, J. Denny, S. Murphy, G. Fernald, G. Krishnan, V. Castro, P. Yue, P. Tsau, I. Kohane, D. Roden *et al.*, "Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels," *Clinical Pharmacology & Therapeutics*, vol. 90, no. 1, pp. 133–142, 2011.

[24] L. C. Wienkers and T. G. Heath, "Predicting in vivo drug interactions from in vitro drug discovery data," *Nature reviews Drug discovery*, vol. 4, no. 10, p. 825, 2005.

[25] S. Ekins and S. A. Wrighton, "Application of in silico approaches to predicting drug–drug interactions," *Journal of pharmacological and toxicological methods*, vol. 45, no. 1, pp. 65–69, 2001.

[26] A. Gottlieb, G. Y. Stein, Y. Oron, E. Ruppin, and R. Sharan, "Indi: a computational framework for inferring drug interactions and their associated recommendations," *Molecular systems biology*, vol. 8, no. 1, p. 592, 2012.

[27] X.-M. Zhao, M. Iskar, G. Zeller, M. Kuhn, V. Van Noort, and P. Bork, "Prediction of drug combinations by integrating molecular and pharmacological data," *PLoS computational biology*, vol. 7, no. 12, p. e1002323, 2011.

[28] J. Huang, C. Niu, C. D. Green, L. Yang, H. Mei, and J.-D. J. Han, "Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network," *PLoS computational biology*, vol. 9, no. 3, p. e1002998, 2013.

[29] Y. Li and K. J. Hale, "Asymmetric total synthesis and formal total synthesis of the antitumor sesquiterpenoid (+)-eremantholide a," *Organic letters*, vol. 9, no. 7, pp. 1267–1270, 2007.

[30] C.-W. Chiang, P. Zhang, X. Wang, L. Wang, S. Zhang, X. Ning, L. Shen, S. K. Quinney, and L. Li, "Translational high-dimensional drug interaction discovery and validation using health record databases and pharmacokinetics models," *Clinical Pharmacology & Therapeutics*, vol. 103, no. 2, pp. 287–295, 2018.

[31] X. Wang, P. Zhang, C.-W. Chiang, H. Wu, L. Shen, X. Ning, D. Zeng, L. Wang, S. K. Quinney, W. Feng *et al.*, "Mixture drug-count response model for the high-dimensional drug combinatory effect on myopathy," *Statistics in medicine*, vol. 37, no. 4, pp. 673–686, 2018.

[32] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *arXiv preprint arXiv:1610.02501*, 2016.

[33] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *CoRR*, vol. abs/1802.04712, 2018. [Online]. Available: http://arxiv.org/abs/1802.04712

[34] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R. Salakhutdinov, and A. J. Smola, "Deep sets," *CoRR*, vol. abs/1703.06114, 2017. [Online]. Available: http://arxiv.org/abs/1703.06114

[35] I. Segura-Bedmar, P. Martínez, and M. H. Zazo, "Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)," in *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, vol. 2, 2013, pp. 341–350.

[36] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[37] S. K. Sahu and A. Anand, "Drug-drug interaction extraction from biomedical text using long short term memory network," *CoRR*, vol. abs/1701.08303, 2017. [Online]. Available: http://arxiv.org/abs/1701.08303

[38] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, pp. i457–i466, 2018.

[39] C. Castro and M. Gourley, "Diagnosis and treatment of inflammatory myopathy: issues and management," *Therapeutic advances in musculoskeletal disease*, vol. 4, no. 2, pp. 111–120, 2012.

[40] Q. Feng, R. A. Wilke, and T. M. Baye, "Individualized risk for statin-induced myopathy: current knowledge, emerging challenges and potential solutions," *Pharmacogenomics*, vol. 13, no. 5, pp. 579–594, 2012.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[42] W.-H. Chiang, S. Li, L. Lang, and N. Xia, "Drug-drug interaction prediction based on co-medication patterns and graph matching," *arXiv preprint arXiv:1902.08675*, 2019.

[43] M. Szumilas, "Explaining odds ratios," *Journal of the Canadian academy of child and adolescent psychiatry*, vol. 19, no. 3, p. 227, 2010.

[44] Q. Que and M. Belkin, "Back to the future: Radial basis function networks revisited," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016, pp. 1375–1383.

[45] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[46] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," *Knowledge-based systems*, vol. 46, pp. 109–132, 2013.

[47] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: statistical mechanics and its applications*, vol. 390, no. 6, pp. 1150–1170, 2011.

[48] L. Lü, L. Pan, T. Zhou, Y.-C. Zhang, and H. E. Stanley, "Toward link predictability of complex networks," *Proceedings of the National Academy of Sciences*, vol. 112, no. 8, pp. 2325–2330, 2015.

[49] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html

[50] J. Jack, C. R., J. Barnes, M. A. Bernstein, B. J. Borowski, J. Brewer, S. Clegg, A. M. Dale, O. Carmichael, C. Ching, C. DeCarli, R. S. Desikan, C. Fennema-Notestine, A. M. Fjell, E. Fletcher, N. C. Fox, J. Gunter, B. A. Gutman, D. Holland, X. Hua, P. Insel, K. Kantarci, R. J. Killiany, G. Krueger, K. K. Leung, S. Mackin, P. Maillard, I. B. Malone, N. Mattsson, L. McEvoy, M. Modat, S. Mueller, R. Nosheny, S. Ourselin, N. Schuff, M. L. Senjem, A. Simonson, P. M. Thompson, D. Rettmann, P. Vemuri, K. Walhovd, Y. Zhao, S. Zuk, and M. Weiner, "Magnetic resonance imaging in Alzheimer's Disease Neuroimaging Initiative 2," *Alzheimers Dement*, vol. 11, no. 7, pp. 740–56, 2015.

[51] W. J. Jagust, S. M. Landau, R. A. Koeppe, E. M. Reiman, K. Chen, C. A. Mathis, J. C. Price, N. L. Foster, and A. Y. Wang, "The Alzheimer's Disease Neuroimaging Initiative 2 PET Core: 2015," *Alzheimers Dement*, vol. 11, no. 7, pp. 757–71, 2015.

[52] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," *Radiographics*, vol. 37, no. 2, pp. 505–515, 2017.

[53] L. Shen, Y. Qi, S. Kim, K. Nho, J. Wan, S. L. Risacher, A. J. Saykin, and Adni, "Sparse bayesian learning for identifying imaging biomarkers in ad prediction," *Med Image Comput Comput Assist Interv*, vol. 13, no. Pt 3, pp. 611–8, 2010.

[54] C. Lange, P. Suppa, U. Pietrzyk, M. R. Makowski, L. Spies, O. Peters, R. Buchert, and I. Alzheimer's Disease Neuroimaging, "Prediction of alzheimer's dementia in patients with amnestic mild cognitive impairment in clinical routine: Incremental value of biomarkers of neurodegeneration and brain amyloidosis added stepwise to cognitive status," *J Alzheimers Dis*, vol. 61, no. 1, pp. 373–388, 2018.

[55] J. Yan, T. Li, H. Wang, H. Huang, J. Wan, K. Nho, S. Kim, S. L. Risacher, A. J. Saykin, L. Shen, and I. Alzheimer's Disease Neuroimaging, "Cortical surface biomarkers for predicting cognitive outcomes using group l2,1 norm," *Neurobiol Aging*, vol. 36 Suppl 1, pp. S185–93, 2015.

[56] L. Brand, H. Wang, H. Huang, S. Risacher, A. Saykin, and L. Shen, "Joint high-order multi-task feature learning to predict the progression of alzheimer's disease," in *Medical Image Computing and Computer Assisted Intervention (MICCAI 2018)*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-Lpez, and G. Fichtinger, Eds. Springer International Publishing, Conference Proceedings, pp. 555–562.

[57] Y. He, J. Liu, and X. Ning, "Drug selection via joint push and learning to rank," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–1, 2018.

[58] M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, D. Harvey, J. Jack, C. R., W. Jagust, J. C. Morris, R. C. Petersen, J. Salazar, A. J. Saykin, L. M. Shaw, A. W. Toga, J. Q. Trojanowski, and I. Alzheimer's Disease Neuroimaging, "The Alzheimer's Disease Neuroimaging Initiative 3: Continued innovation for clinical trial improvement," *Alzheimers Dement*, vol. 13, no. 5, pp. 561–571, 2017.

[59] R. V. Marinescu, N. P. Oxtoby *et al.*, "TADPOLE Challenge: Prediction of Longitudinal Evolution in Alzheimer's Disease," *arXiv e-prints*, p. arXiv:1805.03909, May 2018.

[60] S. Agarwal, *The Infinite Push: A new Support Vector Ranking Algorithm that Directly Optimizes Accuracy at the Absolute Top of the List*, 2011, pp. 839–850.

[61] F. E. Harrell, K. L. Lee, and D. B. Mark, "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in medicine*, vol. 15, no. 4, pp. 361–387, 1996.

[62] A. Forsberg, H. Engler, O. Almkvist, G. Blomquist, G. Hagman, A. Wall, A. Ringheim, B. Lngstrm, and A. Nordberg, "PET imaging of amyloid deposition in patients with mild cognitive impairment," *Neurobiology of Aging*, vol. 29, no. 10, pp. 1456 – 1465, 2008.

[63] K.-L. Huang, K.-J. Lin, I.-T. Hsiao, H.-C. Kuo, W.-C. Hsu, W.-L. Chuang, M.-P. Kung, S.-P. Wey, C.-J. Hsieh, Y.-Y. Wai, T.-C. Yen, and C.-C. Huang, "Regional amyloid deposition in amnestic mild cognitive impairment and alzheimer's disease evaluated by [18f]av-45 positron emission tomography in chinese population," *PLOS ONE*, vol. 8, no. 3, pp. 1–8, 03 2013.

[64] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The sider database of drugs and side effects," *Nucleic acids research*, vol. 44, no. D1, pp. D1075–D1079, 2015.

[65] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "Drugbank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic acids research*, vol. 36, no. suppl_1, pp. D901–D906, 2007.

[66] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[67] S. Ruder, "An overview of gradient descent optimization algorithms," *CoRR*, vol. abs/1609.04747, 2016. [Online]. Available: http://arxiv.org/abs/1609.04747

APPENDIX

A. SUPPLEMENTARY MATERIALS

A.1 Drug Features for FEARS

A.1.1 Chemical Substructure Fingerprints (FP)

Drug chemical structure fingerprints are commonly used as drug features. Chemical structures are highly related to drug physicochemical properties, which may correlate to the intrinsic reasons of drug-drug interactions. We extracted the substructure fingerprints for each drug from PubChem[1]. The substructure fingerprints are composed of 881 substructure-keys, each corresponding to a predefined substructure. The binary values on each substructure-key represent whether the drug has the corresponding substructure or not. This type of drug feature is denoted as FP.

A.1.2 Side-Effect Profiles (SE)

Drug side-effect profile is a high-level representation of drug properties. Two drugs with similar side effect profiles may have similar underlying mechanisms. We extracted drug side-effect profile from Side Effect Resource (SIDER) [64][2] and constructed binary drug side-effect profiles. Each dimension in the profiles corresponds to a specific drug side effect. The binary values on the dimensions represent whether the drug has the corresponding side effect or not. We found side effect information for 529 out of 826 FEARS drugs. These 529 drugs correspond to 3,330 drug combinations, with a maximum cardinality 18, minimum cardinality 2 and mean cardinality 2.9. In addition, these drug combinations include 1,896 2-drug combinations, 823 3-drug combinations, 296 4-drug combinations, 114

[1] `https://pubchem.ncbi.nlm.nih.gov`

[2] `http://sideeffects.embl.de/`

5-drug combinations, 78 6-drug combinations, 47 7-drug combinations and 76 over 7-drug combinations. This type of drug feature is denoted as SE.

A.1.3 Therapeutic-Indication Profiles (TI)

With a similar intuition as in side-effect profiles, we also consider drug theraputic-indication profiles in our experiments. We extracted drug theraputic-indication profiles also from SIDER, and constructed theraputic-indication profiles in a similar way as to construct SE. Still, we could not find theraputic-indication profiles for all the drugs in *Fears* dataset. We found therapeutic indication information for only 491 out of 826 FEARS drugs. These 491 drugs correspond to 3,088 drug combinations, with a maximum cardinality 14, minimum cardinality 2 and mean cardinality 2.8. In addition, these drug combinations include 1,812 2-drug combinations, 751 3-drug combinations, 243 4-drug combinations, 102 5-drug combinations, 72 6-drug combinations, 39 7-drug combinations and 69 over 7-drug combinations. This type of drug feature is denoted as TI.

A.1.4 Target Profiles (TG)

Drug target profile is a high-level representation of drug biological properties, and two drugs with similar target profiles may have similar biological properties. We extracted drug target information from DrugBank [65] and constructed binary drug target profiles. Each dimension in the profiles corresponds to a specific drug target. The binary values on the dimensions represent whether the drug has the corresponding target or not. We found target information for 704 out of 826 FEARS drugs. These 704 drugs correspond to 5621 drug combinations, with a maximum cardinality 47, minimum cardinality 2 and mean cardinality 3.3. In addition, these drug combinations include 2809 2-drug combinations, 1395 3-drug combinations, 544 4-drug combinations, 252 5-drug combinations, 169 6-drug combinations, 132 7-drug combinations and 320 over 7-drug combinations. This type of drug feature is denoted as TG.

A.2 Model Training

A.2.1 Batch Training and Roll-Back

We initialize all the parameters using the initialization method described in [66] to ensure that the gradient flow is smooth in the training. During the training, we also employ the mini-batch strategy [67] for regularization and efficiency purposes. After the training, we employ a roll-back strategy in order to avoid overfitting. That is, in the training, we record the loss value on the validation set every 20 epochs. After the training is done, we roll back to the model that has the minimal recorded loss value on the validation set. Please note that the roll-back strategy is conducted on the validation set, and we don't use any information of the testing set.

A.2.2 Parameters for $\texttt{D}^3\texttt{I}$ Experiments

We use tensorflow 1.9.0 to implement $\texttt{D}^3\texttt{I}$ methods. The $\texttt{D}^3\texttt{I}$ models are trained using Adam gradient descent algorithm. The learning rate in the the Adam gradient descent algorithm is initialized as 1e-3. The learning rate is decreased with a rate of 0.8 every 80 epochs of optimization. The parameter ε in Adam, which is used to prevent any division by zero, is set to 1e-4. The learning rate is the same for both of the datasets.

On the FEARS dataset, in the TPTN setting, the best performing (in terms of F1) $\texttt{D}^3\texttt{I}_{\texttt{max}}$ with TG has the following parameters: the dimension for single drug embeddings (k) is 128; the number of fully-connected layers before the aggregator (n_e) is 1; the number of fully-connected layers after the aggregator (n_p) is 3. In the TPRN setting, the best performing (in terms of F1) $\texttt{D}^3\texttt{I}_{\texttt{max}}$ with TG has the following parameters: the dimension for single drug embeddings (k) is 128; the number of fully-connected layers before the aggregator (n_e) is 1; the number of fully-connected layers after the aggregator (n_p) is 0. We set the batch size as 100 when training on all the drug combinations, and 50 when training on only the order-2 drug combinations. We set the number of the epochs as 600 in both TPTN and TPRN settings.

On the BMC dataset, the best performing (in terms of F1) $\texttt{D}^3\texttt{I}_{\texttt{max}}$ with OSE has the following parameters: the dimension for single drug embeddings (k) is 128; the number of fully-connected layers before the aggregator (n_e) is 1; the number of fully-connected layers after the aggregator (n_p) is 1. The batch size on BMC is 200, and the number of learning epochs is 400.

A.3 Additional Experimental Results

A.3.1 Comparison over Drug Features

Table A.1 presents the best performance in terms of F1 for each drug feature in FEARS dataset in the TPRN setting. Performance in other metrics corresponding to the best F1 is also presented. For high-cardinality drug combinations as in FEARS in TPRN, the best performing drug features are TG for $\texttt{D}^3\texttt{I}_{\texttt{max}}$ and $\texttt{D}^3\texttt{I}_{\texttt{mean}}$, and TI for $\texttt{D}^3\texttt{I}_{\texttt{Att}}$. The reason why TG has good performance might be that as more drugs are involved in a combination, it is likely that the interactions among their targets and secondary targets could induce the drug-drug interactions. Table A.2 presents the comparison over different drug features in $\texttt{D}^3\texttt{I}$ methods on the FEARS dataset in the TPTN setting. The best performing drug features in TPTN on FEARS is still TG. Table A.3 presents the best performance in terms of F1 for each drug feature in BMC dataset. The results in Table A.3 show that the off-side effect profiles OSE as drug features enable the best performance for all the three methods in terms of F1. The drug indication profiles TI as drug features also show promising results for $\texttt{D}^3\texttt{I}$ methods. This may be maybe because for drug pairs (all drug combines in BMC dataset are drug pairs), side effect profiles and drug indication profiles are the most direct sources of information that is related to pairwise drug-drug interactions.

Table A.1.
Comparison of Drug Features on FEARS Dataset (TPRN)

method	feature	acc	pre	rec	F1	AUC
D^3I_{max}	TG	**0.762**	**0.813**	0.680	**<u>0.740</u>**	0.845
	SE	0.738	0.744	**0.728**	0.735	0.820
	TI	0.755	0.807	0.672	0.732	**0.848**
	FP	0.699	0.693	0.718	0.704	0.767
D^3I_{mean}	TG	**0.706**	**0.708**	0.702	**0.704**	**0.767**
	SE	0.679	0.656	**0.760**	0.703	0.756
	TI	0.690	0.704	0.657	0.679	0.762
	FP	0.668	0.705	0.588	0.638	0.742
D^3I_{Att}	TI	**0.703**	**0.750**	0.609	**0.672**	**0.760**
	SE	0.668	0.675	0.647	0.661	0.737
	FP	0.649	0.649	**0.660**	0.653	0.703
	TG	0.659	0.673	0.623	0.646	0.734

Column "feature" corresponds to the drug features. Columns "acc", "pre", "rec", "F1" and "AUC" correspond to accuracy, precision, recall, F1 and AUC. The best results presented for each feature are selected based on F1. The best F1 and the other performance evaluation over each feature are **bold**. The best F1 over all the features is <u>underlined</u>.

A.3.2 Comparison over Model Architectures

Table A.4 and A.5 present the best performance of the three methods from different model architectures on FEARS in the TPTN and TPRN settings, respectively. The performance is presented in terms of F1.

Table A.2.
Comparison of Drug Features on FEARS Dataset (TPTN)

method	feature	acc	pre	rec	F1	AUC
D^3I_{max}	TG	**0.823**	**0.862**	0.773	**<u>0.815</u>**	**0.892**
	FP	0.817	0.838	**0.786**	0.811	0.882
	SE	0.807	0.793	0.752	0.771	0.881
	TI	0.807	0.820	0.721	0.767	0.877
D^3I_{mean}	TG	0.761	0.768	**0.750**	0.759	0.833
	FP	**0.773**	**0.790**	0.744	**0.766**	**0.842**
	SE	0.726	0.684	0.684	0.683	0.806
	TI	0.724	0.733	0.588	0.652	0.786
D^3I_{Att}	TG	**0.758**	**0.768**	0.744	**0.756**	**0.834**
	FP	0.753	0.756	**0.749**	0.752	0.828
	SE	0.742	0.728	0.644	0.683	0.805
	TI	0.752	0.763	0.636	0.693	0.816

Column "feature" corresponds to the drug features. Columns "acc", "pre", "rec", "F1" and "AUC" correspond to accuracy, precision, recall, F1 and AUC. The best results presented for each feature are selected based on F1. The best F1 and the other performance evaluation over each feature are **bold**. The best F1 over all the features is <u>underlined</u>.

Table A.3.
Comparison of Drug Features on BMC Dataset (TPRN)

method	feature	acc	pre	rec	F1	AUC
D^3I_{max}	OSE	**0.693**	**0.663**	**0.788**	**<u>0.720</u>**	**0.744**
	TI	0.672	0.643	0.777	0.703	0.717
	FP	0.668	0.645	0.750	0.693	0.713
	TG	0.643	0.621	0.739	0.674	0.684
	PW	0.633	0.613	0.731	0.666	0.676
	EM	0.616	0.621	0.601	0.611	0.650
	TP	0.592	0.608	0.524	0.562	0.620
D^3I_{mean}	OSE	**0.687**	**0.669**	0.742	**0.703**	**0.743**
	TI	0.681	0.659	**0.752**	0.702	0.734
	FP	0.670	0.657	0.714	0.684	0.721
	TG	0.654	0.643	0.698	0.669	0.707
	PW	0.650	0.637	0.704	0.667	0.702
	EM	0.624	0.633	0.590	0.610	0.666
	TP	0.605	0.624	0.531	0.573	0.637
D^3I_{Att}	OSE	**0.670**	0.635	**0.803**	**0.709**	**0.710**
	TI	**0.670**	**0.640**	0.779	0.702	0.707
	FP	0.661	0.626	0.801	0.703	0.696
	TG	0.659	0.639	0.735	0.683	0.698
	PW	0.638	0.611	0.761	0.678	0.681
	EM	0.631	0.634	0.623	0.629	0.669
	TP	0.603	0.622	0.530	0.573	0.635

Column "feature" corresponds to the drug features. Columns "acc", "pre", "rec", "F1" and "AUC" correspond to accuracy, precision, recall, F1 and AUC. The best F1 and the other performance evaluation over each feature are **bold**. The best F1 over all the features is <u>underlined</u>.

Table A.4.
F1 Comparison over Model Architectures (FEARS, TG, TPTN)

method	#layers	embedding dimensiosn				
		32	64	128	256	512
D^3I_{max}	0	0.809	0.808	0.812	0.812	0.812
	1	0.809	0.811	0.814	0.809	0.808
	3	0.811	0.809	0.815	0.813	0.808
	5	0.806	0.810	0.814	0.809	0.805
D^3I_{mean}	0	0.721	0.723	0.722	0.722	0.724
	1	0.724	0.726	0.725	0.723	0.727
	3	0.735	0.732	0.731	0.741	0.751
	5	0.735	0.759	0.749	0.729	0.735
D^3I_{Att}	16	0.750	0.753	0.754	0.753	0.750
	32	0.750	0.752	0.751	0.756	0.751
	64	0.750	0.752	0.754	0.755	0.750
	128	0.751	0.753	0.751	0.753	0.753

Column "#layers" corresponds to the number of fully-connected layers after the aggregator. Columns under "embedding dimension" correspond to the different numbers of embedding dimensions. The values in this Table are F1 values under the corresponding model architectures using TG as drug features.

Table A.5.
F1 Comparison over Model Architectures (FEARS, TG, TPRN)

method	#layers	embedding dimensiosn				
		32	64	128	256	512
D^3I_{max}	0	0.730	0.734	0.740	0.738	0.740
	1	0.731	0.732	0.718	0.727	0.718
	3	0.707	0.710	0.695	0.683	0.688
	5	0.702	0.698	0.667	0.696	0.651
D^3I_{mean}	0	0.610	0.606	0.611	0.615	0.614
	1	0.673	0.673	0.680	0.683	0.683
	3	0.688	0.683	0.691	0.690	0.699
	5	0.699	0.704	0.692	0.687	0.691
D^3I_{Att}	16	0.639	0.633	0.642	0.646	0.643
	32	0.621	0.635	0.636	0.646	0.642
	64	0.627	0.636	0.627	0.636	0.632
	128	0.627	0.632	0.629	0.641	0.632

Column "#layers" corresponds to the number of fully-connected layers after the aggregator. Columns under "embedding dimension" correspond to the different numbers of embedding dimensions. The values in this Table are F1 values under the corresponding model architectures using TG as drug features.