MLRDA: A Multi-Task Semi-Supervised Learning Framework for Drug-Drug Interaction Prediction

Xu Chu[1,3], **Yang Lin**[1,3], **Yasha Wang**[1,2*], **Leye Wang**[1,3], **Jiangtao Wang**[4], **Jingyue Gao**[1,3]
[1]Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing, China
[2]National Engineering Research Center of Software Engineering, Peking University, Beijing, China
[3]School of Electronics Engineering and Computer Science, Peking University, Beijing, China
[4]School of Computing and Communications, Lancaster University, Lancashire, UK
{chu_xu, bdly, wangyasha, leyewang, gaojingyue1997}@pku.edu.cn, jiangtao.wang@lancaster.ac.uk

Abstract

Drug-drug interactions (DDIs) are a major cause of preventable hospitalizations and deaths. Recently, researchers in the AI community try to improve DDI prediction in two directions, *incorporating multiple drug features* to better model the pharmacodynamics and *adopting multi-task learning* to exploit associations among DDI types. However, these two directions are challenging to reconcile due to the sparse nature of the DDI labels which inflates the risk of overfitting of multi-task learning models when incorporating multiple drug features. In this paper, we propose a multi-task semi-supervised learning framework MLRDA for DDI prediction. MLRDA effectively exploits information that is beneficial for DDI prediction in unlabeled drug data by leveraging a novel unsupervised disentangling loss CuXCov. The CuXCov loss cooperates with the classification loss to disentangle the DDI prediction relevant part from the irrelevant part in a representation learnt by an autoencoder, which helps to ease the difficulty in mining useful information for DDI prediction in both labeled and unlabeled drug data. Moreover, MLRDA adopts a multi-task learning framework to exploit associations among DDI types. Experimental results on real-world datasets demonstrate that MLRDA significantly outperforms state-of-the-art DDI prediction methods by up to 10.3% in AUPR.

1 Introduction

Drug-drug interactions (DDIs) are modification effects of a drug when administered with another drug, resulting in many adverse drug reactions (ADRs) that may cause injuries or deaths [Qato *et al.*, 2016]. Most DDIs are discovered by accident once a drug is already on the market [Percha and Altman, 2013]. Early detection of DDIs based on drug features helps drug safety professionals allocate investigative resources and take regulatory action [Zhang *et al.*, 2015]. Since ADRs (e.g., Nausea, Asthenia, Kidney Failure, etc.) associated with DDIs are important for both clinical and pharmaceutical decisions [Vilar *et al.*, 2017], researchers classify DDIs into different types according to different ADRs in DDI prediction [Jin *et al.*, 2017; Ryu *et al.*, 2018; Ma *et al.*, 2018]. The DDI prediction studied in this paper is defined as: For a pair of drugs with some known drug features, predicting the occurrence of all different DDI types based on the drug features.

The pioneering work for DDI prediction [Vilar *et al.*, 2012] computes similarities between drug pairs with drug chemical structures and predicts DDIs with a nearest neighbor method. The research works done to improve DDI prediction in recent years can be summarized in two directions: (a) Integrating multiple drug features (e.g., chemical structures, indications, etc) to calculate the similarities among drugs more comprehensively (compared to a single feature), and then predict DDIs more accurately based on the fused similarity [Zhang *et al.*, 2015; Abdelaziz *et al.*, 2017; Ma *et al.*, 2018]. (b) Adopting multi-task learning to exploit associations among DDI types [Jin *et al.*, 2017; Ryu *et al.*, 2018; Zitnik *et al.*, 2018]. Considering the prediction of a certain type of DDI as a task, the correlation between DDI types can be utilized in a multi-task learning framework to improve the overall predictive accuracy.

However, the methods of the two directions are challenging to reconcile. Not all drugs have complete data of all features [Ryu *et al.*, 2018]. Considering more features would lead to fewer drugs with known data of all considered features as well as a sparser labeled dataset. For example, in the widely utilized DDI label resource Twosides database [Tatonetti *et al.*, 2012], considering one single feature the drug chemical structures, there are 645 drugs and 63473 drug pairs with positive DDI labels. Nevertheless, considering two more features the drug indications and drug side effects, there would be only 318 drugs with complete feature data of three features[1] and corresponding 16775 drug pairs with positive DDI labels left. Thus a multi-task learning model with an excessive quantity of parameters would tend to overfit.

Actually, there are a large number of unlabeled drug pairs that contain not only information of multiple drug features explicitly, but also instructive information for DDI prediction implicitly, yet existing multi-task learning methods fail to incorporate the unlabeled DDI data. Indeed, it is not trivial

[*] Corresponding author

[1] Please see Section 5.1 for drug feature data sources.

to effectively take advantage of both labeled and unlabeled drug data simultaneously for better DDI prediction. On one hand, the state-of-the-art deep semi-supervised learning models (such as Ladder Network [Rasmus *et al.*, 2015], Mean Teacher [Tarvainen and Valpola, 2017], and Virtual Adversarial Training [Miyato *et al.*, 2017], etc.) construct unsupervised consistency regularization losses that rely on an underlying assumption: the classes are well-separated. While for DDI prediction, the classes, i.e., the DDI types, are strongly associated, resulting in nebulous boundaries for each class. On the other hand, although unsupervised algorithms such as autoencoders allow us to exploit information hidden in the unlabeled data by learning a compact representation that explains the variations of drug features, the representations learnt by unsupervised methods would, in general, entangle the factors relevant to DDI predictions with factors accounting for remaining variations of drug features [Cheung *et al.*, 2015]. The irrelevant factors for DDI prediction in the complex representation would introduce undesired bias that depresses the performance of DDI prediction. If we manage to disentangle the DDI prediction relevant part from the irrelevant part in the representation, then we may exclude the disturbance from irrelevant part and leverage beneficial information from both unlabeled and labeled data in the DDI prediction relevant part, and thereby enhance the generalization of a model by restraining overfitting.

Inspired by the above insights, we propose **M**ulti-**L**abel **R**obust **D**isentangling **A**utoencoders (MLRDA) for DDI prediction. Our contributions are summarized as follows:

- We propose a multi-task semi-supervised learning framework MLRDA for DDI prediction. MLRDA reconciles integrating multiple drug features and multi-task learning, thus is able to better describe pharmacodynamics and exploit the associations among DDI types.
- The MLRDA framework has three distinct technical highlights. (a) MLRDA exploits information in unlabeled data that is beneficial for DDI prediction. (b) MLRDA leverages a novel robust unsupervised loss CuXCov to help disentangling the DDI prediction relevant part from the irrelevant part in the representation, lessening the feature complexity of the representation and reducing the risk of overfitting. (c) MLRDA adopts a multi-task learning framework compatible with (a) and (b) to exploit associations among DDI types.
- Experimental results on real-world datasets demonstrate that MLRDA significantly outperforms state-of-the-art DDI prediction methods by up to 10.3% in AUPR.

2 Related Work

The pioneering computational work for DDI prediction [Vilar *et al.*, 2012] is based on a simple but effective idea. They first calculate the similarities between drug pairs with drug chemical structure fingerprints, then they predict DDIs based on a similarity based idea, i.e., if drug A is similar to drug B, then the drugs that have the i-th type of DDI with drug A are likely to have the i-th type of DDI with drug B.

Lately, researchers show that the prediction could be improved by incorporating more drug features to better describe pharmacodynamics[Cheng and Zhao, 2014; Takeda *et al.*, 2017; Zhang *et al.*, 2017]. [Zhang *et al.*, 2015] propose an integrative framework to fuse the similarities of different views of drug features with proper weights and predict DDIs by label propagation method. [Abdelaziz *et al.*, 2017; Kastrin *et al.*, 2018] model the DDIs and drug features as edges and nodes and predict DDIs with link prediction method. A semi-supervised Graph autoencoder method is proposed by [Ma *et al.*, 2018] that employs graph convolutional networks (GCN)[Kipf and Welling, 2017]. Compared with traditional graph methods, the GCN model not only encodes the graph structure, but also encode the node features. However, those methods fail to exploit the task associations among DDI types. Actually, different types of DDI events are related. For example, if a specific drug pair causes Nausea/High blood pressure, then the specific drug pair is likely to cause Emesis/Difficulty breathing.

Recently, efforts leveraging multi-task learning attempts to exploit the associations among DDI types. Researchers take pairwise chemical structure features as input to model interactions between drug pairs and predict different DDIs simultaneously [Jin *et al.*, 2017; Ryu *et al.*, 2018]. [Zitnik *et al.*, 2018] proposes a multi-modal graph of drug-protein target interaction and protein-protein interaction to predict DDIs simultaneously. However, it is challenging to integrate multiple drug features in multi-task learning models for better predictive performance because of the trade-off between information gain from additional features and better immunity against overfitting. In this paper, we propose a multi-task semi-supervised learning DDI prediction framework MLRDA that reconciles integrating multiple drug features and multi-task learning.

3 Preliminaries

We define some notations to prepare our method in Table 1.

Symbol	Definition
v	Number of different DDI types
n	Number of different drug features
r_i	The i-th type of DDI
$\mathbf{e}_j$	The j-th single drug feature vector
$\mathbf{d}_j$	The j-th pairwise drug feature vector
$\mathcal{B}$	Set of drug pairs with known features
$\mathcal{P}_i$	Set of **p**ositive samples of the ith DDI
$\mathcal{N}_i$	Set of **n**egative samples of the ith DDI
$\mathcal{L}_{Cls}$	Classification loss
$\mathcal{L}_{Rcnst}$	Reconstruction loss
$\mathcal{L}_{CuXCov}$	Cumulative cross-covariance loss

Table 1: Notation.

For the j-th drug feature, $j = 1, \cdots, n$, the pairwise drug feature vector $\mathbf{d}_j$s are associated with drug pairs, and the single drug feature vectors $\mathbf{e}_j$s are associated with single drugs. The $\mathbf{d}_j$s are constructed by simply concatenating a pair of $\mathbf{e}_j$s. For example, omitting the subscript j for simplicity, suppose $\mathbf{e}_A$ denotes the single feature vector of drug A, so does $\mathbf{e}_B$ of drug B. The pairwise drug feature vectors $\mathbf{d}_{AB}$ and $\mathbf{d}_{BA}$ associated with drug pair $\{$A, B$\}$ are concatenations of $\mathbf{e}_A$ and $\mathbf{e}_B$, i.e., $\mathbf{d}_{AB} = (\mathbf{e}_A, \mathbf{e}_B)$ and $\mathbf{d}_{BA} = (\mathbf{e}_B, \mathbf{e}_A)$.

Problem Statement

The problem of DDI prediction is formulated as follows. Suppose we have a set of drug pairs $\mathcal{B}$. For each drug pair in $\mathcal{B}$, we know n different pairwise drug feature vector $\mathbf{d}_j, j = 1, \cdots, n$. Considering v DDI types, for each DDI type r_i, suppose we know the positive sample set $\mathcal{P}_i$ (containing drug pairs that causing r_i) and the negative sample set $\mathcal{N}_i$ (containing drug pairs that not causing r_i empirically). Our goal is to estimate the occurrence probabilities of DDI type r_i of drug pairs in $\mathcal{B} - \mathcal{P}_i \cup \mathcal{N}_i$ for all DDI types $r_i, i = 1, \cdots, v$.

4 The Proposed MLRDA

4.1 Framework Overview

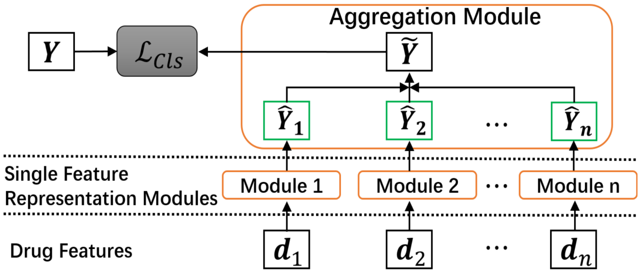

Figure 1: The framework of proposed MLRDA.

The framework of our proposed MLRDA is shown in Figure 1. Suppose we want to predict v DDI types based on n drug features, then MLRDA consists of n *Single Feature Representation Modules* (See more details in section 4.2) and one *Aggregation Module* (See more details in section 4.4). For each drug pair, the j-th Single Feature Representation Module takes in the j-th pairwise feature $\mathbf{d}_j$ as input and outputs a binary vector $\hat{\mathbf{Y}}_j \in \mathbb{R}^v$, with each bit encoding the occurrence probability of each DDI type r_i predicted by j-th feature. The Aggregation module collects prediction vectors $\hat{\mathbf{Y}}_j$s given by different features and calculates a weighted prediction vector $\tilde{\mathbf{Y}}$ with attention mechanism. In MLRDA, all representation modules and aggregation module are jointly optimized in an end-to-end way (See training objective in section 4.5). In the following subsections, we give more technical details.

4.2 Single Feature Representation Module

To exploit the beneficial information for DDI prediction and to reduce the undesired bias caused by irrelevant information in both labeled and unlabeled data, we design the Single Feature Representation Module (SFRM), of which the architecture is illustrated in Figure 2. For ease of illustration, we omit the subscript j indexing different drug features. The overall neural network is built with an autoencoder structure. The network consists of $H + 1$ layers where H is an even number. The first $H/2 + 1$ hidden layers are encoders to learn a representation of each input and the last $H/2$ hidden layers are decoders to reconstruct the input.

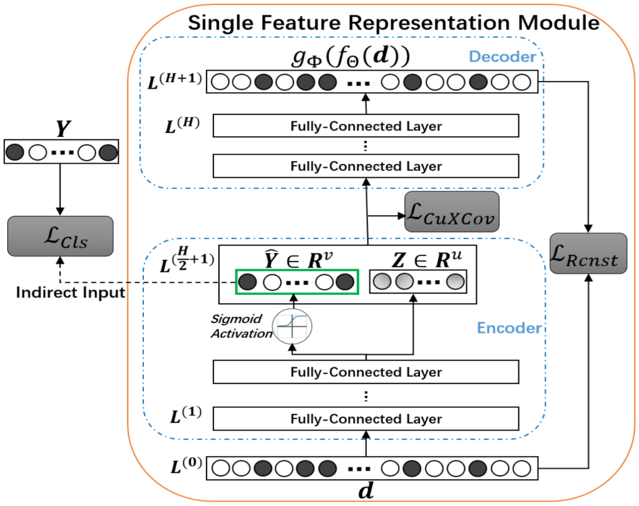

Figure 2: The architecture of SFRM.

Let $\mathbf{L}^{(0)} = \mathbf{d}$ denote an input to the first layer and let

$$\mathbf{L}^{(h)} = t^{(h)}((\mathbf{W}^{(h)})^T \mathbf{L}^{(h-1)} + \mathbf{b}^{(h)}) \in \mathbb{R}^{b_h} \tag{1}$$

be the output of the h-th layer, $h = 1, \cdots, H, H+1$. b_h denotes the dimension of the output at the h-th layer and $t^{(h)}$s are activation functions, which we take Rectified Linear Unit (ReLU) for all hidden layers except the *high-level representation layer* (the last layer of the encoder), i.e., $h = H/2 + 1$,

$$\begin{cases} \mathbf{L}^{(\frac{H}{2}+1)} = \begin{pmatrix} \hat{\mathbf{Y}} \\ \mathbf{Z} \end{pmatrix} = ((\hat{y}_1, \cdots, \hat{y}_v)^T, (z_1, \cdots, z_u)^T)^T. \\ \hat{y}_i = Sigmoid((\mathbf{W}_i^{(\frac{H}{2}+1)})^T \mathbf{L}^{(\frac{H}{2})} + b_i^{(\frac{H}{2}+1)}), i = 1, \cdots, v; \\ z_k = (\mathbf{W}_{v+k}^{(\frac{H}{2}+1)})^T \mathbf{L}^{(\frac{H}{2})} + b_{v+k}^{(\frac{H}{2}+1)}, k = 1, \cdots, u. \end{cases} \tag{2}$$

We define the function of the encoder as f_Θ and the function of the decoder as g_Φ (Θ and Φ denote the parameter space of encoder and decoder respectively), i.e.,

$$f_\Theta(\mathbf{d}) = \mathbf{L}^{(\frac{H}{2}+1)} = \begin{pmatrix} \hat{\mathbf{Y}} \\ \mathbf{Z} \end{pmatrix}, \quad g_\Phi(f_\Theta(\mathbf{d})) = \mathbf{L}^{(H+1)}.$$

Given inputs $\mathbf{d}$s, $f_\Theta(\mathbf{d})$s are abstract representations learnt by the encoder. In the learnt representation, without some explicit means, the latent factors relevant for DDI predictions would be entangled with other latent factors explaining the remaining variations of $\mathcal{B}$. Our goal is partitioning $f_\Theta(\mathbf{d})$ into two part $\hat{\mathbf{Y}} \in \mathbb{R}^v$ and $\mathbf{Z} \in \mathbb{R}^u$. $\hat{\mathbf{Y}} = (\hat{y}_1, \cdots, \hat{y}_v)^T$ encodes the DDI-prediction-relevant factors with $\hat{y}_i$ *the occurrence of the i-th event* r_i and $\mathbf{Z}$ encodes the factors that are irrelevant with DDI prediction. To achieve the above idea, (a) we introduce a robust unsupervised disentangling loss CuXCov (See more details in section 4.3) that make $\hat{\mathbf{Y}}$ as uncorrelated with $\mathbf{Z}$ as possible, (b) at the same time endowing $\hat{y}_i$ with semantic meaning of the occurrence probability by DDI label vectors $\mathbf{Y}$s that provide indirect supervised learning signals.

The learning of $\hat{y}_i$ can be viewed as a binary classification task r_i. In DDI prediction, each $\hat{y}_i$ is corresponding to DDI type. Actually, each SFRM reformulates the DDI prediction problem as a multi-label classification problem, and the class label vector $\hat{\mathbf{Y}}$ is given by learning the dimerous representation $f_\Theta(\mathbf{d})$ of pairwise drug feature $\mathbf{d}$. The class labels in $\hat{\mathbf{Y}}$s and $\tilde{\mathbf{Y}}$s are learnt simultaneously with shared parameters in layers $\mathbf{L}^{(h)}, h = 1, \cdots, H/2$, thus MLRDA is a multi-task learning framework [Caruana, 1997] in nature.

In addition, when drug pair {A,B} is in validation/test set, its predicted DDI label is the arithmetic average of two predicted label vectors w.r.t. inputs $\mathbf{d}_{AB}$ and $\mathbf{d}_{BA}$ since $\mathbf{d}_{AB}$ and $\mathbf{d}_{BA}$ should share the same DDI label.

4.3 CuXCov Loss

To identify the DDI-prediction-relevant part $\hat{\mathbf{Y}}$ from the irrelevant part $\mathbf{Z}$ in the representation $f_\Theta(\mathbf{d})$ of each SFRM, we wish $\hat{\mathbf{Y}}$ as uncorrelated with $\mathbf{Z}$ as possible. [Cheung *et al.*, 2015] proposed an unsupervised loss called XCov loss aimed at decorrelating $\hat{\mathbf{Y}}$ with $\mathbf{Z}$ by minimizing the covariance of each dimension in $\hat{\mathbf{Y}}$ and each dimension in $\mathbf{Z}$.

Let $\mathbb{Y} = (\hat{\mathbf{Y}}^1, \hat{\mathbf{Y}}^2, \cdots, \hat{\mathbf{Y}}^N)$, and $\mathbb{Z} = (\mathbf{Z}^1, \mathbf{Z}^2, \cdots, \mathbf{Z}^N)$ denote the matrices of high-level representations over a mini-batch with size N. The XCov loss takes the form:

$$L_{XCov} = 1/N(\mathbb{Y}(\mathbf{I} - \mathbf{ee}^T/N))(\mathbb{Z}(\mathbf{I} - \mathbf{ee}^T/N))^T. \quad (3)$$

Where $\mathbf{e} \in \mathbb{R}^N$ is a column vector with all entries being 1, $\mathbf{I}$ is the identity matrix.

However, XCov loss estimates the cross-covariance matrix in each mini-batch separately. The mini-batch limits the sample size involved in estimating the cross-covariance matrix, and could result in inaccurate estimation and gradient descent directions with a large variance, which may disturb decorrelating $\hat{\mathbf{Y}}$ with $\mathbf{Z}$ and bring about undesired bias for DDI prediction. Better estimation of the cross-covariance matrix is possible by enlarging sample size, yet simply enlarging the mini-batch size would not achieve a monotone increasing gain in predictive performance. Large batch size method tends to converge to sharp minimizers of the training function and result in worse generalizability [Keskar *et al.*, 2017].

Utilizing all samples in $\mathcal{B}$, i.e., the full batch to estimate the cross-covariance would result in a lower estimation variance. But it is challenging to calculate a full-batch estimation in mini-batch based deep learning models. Inspired by [Chang *et al.*, 2018], we propose a robust cumulative loss CuXCov that approximates the full-batch cross-covariance matrix. The approximation is achieved by stochastic incremental learning. Let $\Sigma_f^k, \Sigma_a^k, \Sigma_c^k, \Sigma_m^k$ denote the *full, approximate, cumulative, mini-batch* cross-covariance estimator at the k-th training step respectively. The approximation works as follows:

$$\begin{cases} \Sigma_c^k = \alpha\Sigma_c^{k-1} + \Sigma_m^k, & with\ \Sigma_c^0 = \mathbf{0}, \\ p^k = \alpha p^{k-1} + 1, & with\ p^0 = 0. \end{cases} \quad (4)$$

Where $\alpha \in [0, 1]$ is the decay rate, and p is a normalizing factor computed accumulatively to estimate the cross-covariance matrix more accurately. Then let

$$\Sigma_a^k = \Sigma_c^k/p^k, \quad (5)$$

and Σ_a^k approximates Σ_f^k better than Σ_m^k does as k gets larger for accumulating statistics collected for previous mini-batches. To calculate The CuXCov loss, we rewrite (3) as:

$$\Sigma_m^k = 1/N(\mathbb{Y}(\mathbf{I}-\mathbf{ee}^T/N))(\mathbb{Z}(\mathbf{I}-\mathbf{ee}^T/N))^T = 1/N\mathbb{Y}\mathbf{H}\mathbb{Z}^T, \quad (6)$$

where $\mathbf{H} = (\mathbf{I} - \mathbf{ee}^T/N)(\mathbf{I} - \mathbf{ee}^T/N)^T \in \mathbb{R}^{N\times N}$.

From (4), (5) and (6), we have

$$\Sigma_a^k = 1/p^k(\alpha\Sigma_c^{k-1} + 1/N\mathbb{Y}\mathbf{H}\mathbb{Z}^T) \in \mathbb{R}^{v\times u}. \quad (7)$$

Our goal is to minimize all entries in Σ_a^k. We **define CuXCov loss** as:

$$\mathcal{L}_{CuXCov} = trace((\Sigma_a^k)^T\Sigma_a^k)/2. \quad (8)$$

4.4 Aggregation Module

We follow the common paradigm of multi-view learning [Sun, 2013; Gao *et al.*, 2019] to take advantages of complementary information in multiple drug features. We seek help from the well-known attention mechanism [Bahdanau *et al.*, 2015] to aggregate the results from single views and learn a weighted prediction vector $\tilde{\mathbf{Y}} = (\tilde{y}_1, \cdots, \tilde{y}_v)$ adaptively. The details of the attention mechanism work as follows. Suppose we have n drug features and the SFRM give predictions $\hat{\mathbf{Y}}_j \in \mathbb{R}^v, j = 1, 2, \cdots, n$. The final prediction vector $\tilde{\mathbf{Y}}$ is the weighted average of $\hat{\mathbf{Y}}_j$s, where the softmax weights a_js are determined by parameters $\mathbf{B}_j \in \mathbb{R}^v$ and $c_j \in \mathbb{R}$. Specifically,

$$\begin{cases} \tilde{\mathbf{Y}} = \sum\limits_{j=1}^{n} a_j\hat{\mathbf{Y}}_j, a_j = e^{A_j} / \sum\limits_{j=1}^{n} e^{A_j}, \;\; j = 1, \cdots, n, \\ A_j = \mathbf{B}_j^T\hat{\mathbf{Y}}_j + c_j. \end{cases} \quad (9)$$

The parameters $\mathbf{B}_j$, c_j, as well as all other parameters in MLRDA are learnt in the end-to-end optimization process.

4.5 Loss Functions

Classification Loss and Reconstruction Loss

The Classification loss is defined as:

$$\mathcal{L}_{Cls} = -\sum_{i=1}^{v} \sum_{\mathbf{d}_s \in \mathcal{P}_i \cup \mathcal{N}_i} (y_i^s log(\tilde{y}_i^s) + (1 - y_i^s)log(1 - \tilde{y}_i^s)) \quad (10)$$

Where s indexes examples and i indexes DDI tasks. y_i^s is the label of s-th sample in labeled set $\mathcal{P}_i \cup \mathcal{N}_i$ with 1 coding the occurrence of DDI r_i and 0 otherwise.

The architecture of autoencoder allows introducing the unsupervised reconstruction loss to leverage information in both labeled and unlabeled data. For j-th drug feature, we define reconstruction loss over all drug pairs in $\mathcal{B}$:

$$\mathcal{L}_{Rcnst}^j = \sum_{\mathcal{B}} ||\mathbf{d}_j - g_{\Phi_j}(f_{\Theta_j}(\mathbf{d}_j))||^2. \quad (11)$$

Training Objective

The training objective of the MLRDA network is to minimize the weighted integration of three losses:

$$\min \mathcal{L}_{Cls} + \beta\sum_{j=1}^{n} \mathcal{L}_{CuXCov}^j + \gamma\sum_{j=1}^{n} \mathcal{L}_{Rcnst}^j. \quad (12)$$

Where hyperparameters $\beta > 0$ and $\gamma > 0$ control relative weights of $\sum\limits_{j=1}^{n} \mathcal{L}_{CuXCov}^j$ and $\sum\limits_{j=1}^{n} \mathcal{L}_{Rcnst}^j$ over $\mathcal{L}_{Cls}$, j indexes for the j-th drug feature.

5 Experiments

5.1 Datasets

Drug Features

The drug features considered are from the following sources. *Drug chemical structure data I*: The first chemical structure data are extracted from Pubchem[2] substructure fingerprint. The chemical structure of each drug are binary coded as an 881-bit feature vector, each bit representing a Boolean determination of the presence of a substructure in a drug molecule. *Drug chemical structure data II*: The second chemical structure features are extended-connectivity fingerprints with diameter 6 (ECFP6) generated by R package "rcdk"[3], and are hashed binary coded as a 1024-bit feature vector.
Drug indication data: The drug indication data is from SIDER[4], and are binary coded as a 2714-bit feature vector, each bit representing a Boolean determination if the drug is clinically significant for an indication.
Drug targets data: The drug target data is from Therapeutic Target Database[5], and are binary coded as a 2150-bit feature vector, each bit representing a Boolean determination if a protein or a nucleic acid is a target of the drug.
Drug side effect data: The drug indication data is also from SIDER, and are binary coded as a 5868-bit feature vector, each bit representing a Boolean determination if the drug is clinically significant for an side effect.

DDI Data

The labeled DDI data is from Twosides database[6] [Tatonetti *et al.*, 2012]. It contains 645 drugs and 1318 types of DDIs, and in total 63473 drug pairs associated with DDI reports.

Datasets for DDI Prediction

We consider two sets of drug features for DDI prediction.

- **C1IT**: Drug chemical structure data I, Drug indication data, Drug targets data;
- **C2IS**: Drug chemical structure data II, Drug indication data, Drug side effect data.

The sets of drugs associated with various drug features in two datasets are different, we take the two intersections of drug sets in two datasets as the considered drug sets. In the two datasets, a simple filtering process is operated by keeping only informational bits of drug features, for example, if a bit in the original 2714-bit indication vector is 0 for all considered drugs, then this bit is omitted. In our study, in agreement with [Jin *et al.*, 2017], we consider *the top 100 frequent types of DDIs* associated with drugs in **C1IT** and **C2IS** respectively. The DDI interactions from TWOSIDES are used as positive drug pair samples. The complement set of positive samples in TWOSIDES are utilized as negative samples. The drug pairs not included in TWOSIDES are used as the unlabeled dataset. Some basic statistics are shown in Table 2.

Dataset	C1IT	C2IS
Number of considered drugs	309	317
Bits number of chemical structures	563	1024
Bits number of indications	1457	1510
Bits number of targets	190	Null
Bits number of side effects	Null	3908
Average positive samples count	4757.6	4756.0
Average negative samples count	11946.4	12019.0
Unlabeled samples count	30882	33311

Table 2: Basic statistics of datasets **C1IT** and **C2IS**. The average counts are computed over v=100 DDI types.

5.2 Methods for Comparison

Baselines

MLRDA is compared with following DDI prediction models.

- Nearest Neighbor method in [Vilar *et al.*, 2012].
- Label Propagation method in [Zhang *et al.*, 2015].
- Dyadic Prediction method in [Jin *et al.*, 2017].
- Graph AutoEncoders method in [Ma *et al.*, 2018].
- DeepDDI method in [Ryu *et al.*, 2018].
- Other than above 5 state-of-the-art methods designed for DDI prediction. We also studied the well known semi-supervised Ladder Network [Rasmus *et al.*, 2015] as a baseline, regarding DDI prediction as a multi-label classification problem in a multi-task learning framework.

Ablation Study

We studied the effects of different components in MLRDA.

- MLRDA-Un: The MLRDA model trained with only labeled data.
- MLRDA-X: The MLRDA model without considering the cross-covariance penalty.
- MLRDA-X^+: The MLRDA model without **Z** in high-level representation layer, leading to no cross-covariance penalty as well.
- concateMLRDA: A simplified version of MLRDA that concatenates multiple drug features into one input vector **d** and predicts DDIs with one SFRM.

Evaluations and Implementations

We randomly select 10% of drugs and mask all DDIs associated with these drugs for testing. DDIs associated with drugs not in the testing set are used for training all models and we use 10-fold cross-validation to tune all hyperparameters of different methods. We optimized the hyperparameters for MLRDA and fix them for all MLRDA variants. The hyperparameters are shown in Table 4. The models are trained by Adam optimizer[Kingma and Ba, 2014] with a learning rate of 0.0001 and a dropout rate of 0.1. We consider decaying learning rate after 10 epochs.

For testing data, we evaluate all methods on different collections of DDIs. For a given collection of DDIs, we randomly select 50% of the testing set for evaluation and repeated the selection-evaluation process for 50 times. We report the mean and standard deviation of the Area Under Precision-Recall Curve(AUPR) and the area under Receiver Operating Characteristic curve (AUROC) over 50 repetitions. The AUPR and AUROC are averaged over 100 DDI tasks.

[2] https://pubchem.ncbi.nlm.nih.gov/

[3] https://cran.r-project.org/web/packages/rcdk/index.html

[4] http://sideeffects.embl.de/download/

[5] https://db.idrblab.org/ttd/full-data-download

[6] http://tatonettilab.org/resources/tatonetti-stm.html

Method		**C1IT**		**C2IS**	
		AUPR	AUROC	AUPR	AUROC
Baselines	Nearest Neighbor	0.320 ± 0.0042	0.564 ± 0.0035	0.389 ± 0.0048	0.635 ± 0.0037
	Label Propagation	0.399 ± 0.0057	0.649 ± 0.0027	0.434 ± 0.0050	0.659 ± 0.0035
	Dyadic Prediction	0.352 ± 0.0043	0.604 ± 0.0031	0.438 ± 0.0057	0.655 ± 0.0035
	Graph AutoEncoder	0.394 ± 0.0044	0.651 ± 0.0027	0.426 ± 0.0046	0.676 ± 0.0024
	Deep DDI	0.390 ± 0.0055	0.639 ± 0.0032	0.454 ± 0.0051	0.683 ± 0.0028
	Ladder Network	0.378 ± 0.0046	0.629 ± 0.0030	0.462 ± 0.0048	0.688 ± 0.0026
Proposed	MLRDA	$\mathbf{0.440 \pm 0.0058}$	$\mathbf{0.667 \pm 0.0027}$	$\mathbf{0.483 \pm 0.0053}$	$\mathbf{0.697 \pm 0.0029}$
Ablation Study	MLRDA-Un	0.397 ± 0.0053	0.620 ± 0.0033	0.429 ± 0.0048	0.656 ± 0.0031
	MLRDA-X	0.377 ± 0.0049	0.622 ± 0.0029	0.433 ± 0.0042	0.663 ± 0.0027
	MLRDA-X^+	0.354 ± 0.0046	0.606 ± 0.0033	0.395 ± 0.0038	0.631 ± 0.0026
	concateMLRDA	0.419 ± 0.0048	0.653 ± 0.0026	0.471 ± 0.0047	0.683 ± 0.0025

Table 3: Perfomance of MLRDA against comparative approaches.

Hyperparameters	Values
Number of layers $H+1$	5 (2+1+2)
Number of neurons in encoders	2048, 1024
Number of neurons in decoders	1024, 2048
Dimension of **Z**s	256
Activation function $t(\cdot)$	ReLU
(β, γ) of **C1IT**	(125,0.25)
(β, γ) of **C2IS**	(50,0.25)
Decay rate αs of **C1IT**	(0.2, 0.6, 0.2)
Decay rate αs of **C2IS**	(0.3, 0.2, 0.6)
Mini-batch size	1024

Table 4: Hyperparameters of MLRDA.

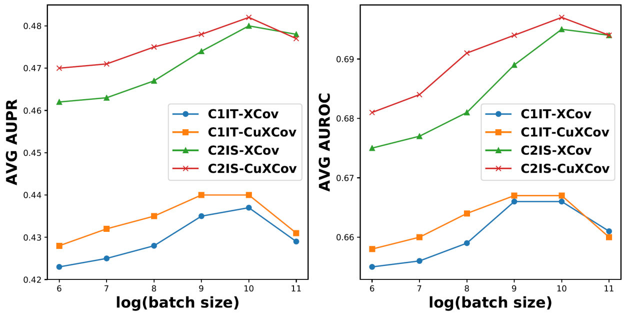

Figure 3: AUPRs and AUROCs of models leveraging CuXCov loss and XCov loss with variant batch sizes ranging from 64 to 2048.

5.3 Results and Analysis

Table 3 compares the performance of the proposed MLRDA against its variants and competing baseline methods. The table shows that MLRDA and concateMLRDA consistently achieve higher AUPRs and AUROCs for better exploiting the associations among DDI events and leveraging hidden information in unlabeled DDI data. MLRDA-Un achieves worse performance for failing to leverage hidden information in the unlabeled data. Though MLRDA-X and MLRDA-X^+ consider unlabeled data in reconstruction loss, they suffer from overfitting and achieve even worse results for failing to disentangle the class-relevant factors, and as a result, incorporating the unlabeled data will introduce undesired bias for DDI prediction instead of revealing hidden information. MLRDA-X outperforms MLRDA-X^+ by reserving **Z** for better reconstruction of drug features, and the reconstruction losses regularize classification. The better performance of MLRDA over concateMLRDA demonstrates the advantage of isolating features first and learn adaptive weights to fuse features by adopting multiple SFRM.

5.4 CuXCov Loss vs XCov Loss

In this subsection we study the performances of CuXCov loss against XCov loss [Cheung *et al.*, 2015]. Note that the cumulative CuXCov loss degenerates to XCov loss when the decay rate parameter $\alpha = 0$. For CuXCov loss, αs are set as in Table 4. Fixing other hyperparameters as in Table 4, we study the performances of models leveraging two losses with variant batch sizes ranging from 64 to 2048. The results in Figure 3 demonstrates that CuXCov model outperforms XCov model consistently for better approximating the full-batch statistics. The performances of two models are increasing as batch size getting larger at first for better estimation of cross-covariance with smaller estimation variance. As the batch size continually getting larger, performances decay because of the degradation of generalization caused by convergence to sharp minimizers of the training function [Keskar *et al.*, 2017].

6 Conclusion

In this paper, we propose a multi-task semi-supervised learning framework MLRDA for DDI prediction. MLRDA effectively exploits information that is beneficial for DDI prediction in both labeled and unlabeled drug data by leveraging a novel unsupervised disentangling loss CuXCov. Moreover, MLRDA adopts a multi-task learning framework to exploit associations among DDI types. Experimental results on real-world datasets demonstrate that MLRDA significantly outperforms state-of-the-art DDI prediction methods by up to 10.3% in AUPR.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.61772045). We thank the anonymous reviewers for their helpful comments and suggestions.

References

[Abdelaziz *et al.*, 2017] Ibrahim Abdelaziz, Achille Fokoue, Oktie Hassanzadeh, Ping Zhang, and Mohammad Sadoghi. Large-scale structural and textual similarity-based mining of knowledge graph to predict drugdrug interactions. *Web Semant.*, 44(C):104–117, May 2017.

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[Caruana, 1997] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[Chang *et al.*, 2018] Xiaobin Chang, Tao Xiang, and Timothy M Hospedales. Scalable and effective deep cca via soft decorrelation. In *CVPR*, 2018.

[Cheng and Zhao, 2014] Feixiong Cheng and Zhongming 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 Jamia*, 21(2), 2014.

[Cheung *et al.*, 2015] Brian Cheung, Jesse A Livezey, Arjun K Bansal, and Bruno A Olshausen. Discovering hidden factors of variation in deep networks. *ICLR workshop*, 2015.

[Gao *et al.*, 2019] Jingyue Gao, Xiting Wang, Yasha Wang, and Xing Xie. Explainable recommendation through attentive multi-view learning. In *AAAI*, 2019.

[Jin *et al.*, 2017] Bo Jin, Haoyu Yang, Cao Xiao, Ping Zhang, Xiaopeng Wei, and Fei Wang. Multitask dyadic prediction and its application in prediction of adverse drug-drug interaction. In *AAAI*, 2017.

[Kastrin *et al.*, 2018] A. Kastrin, P. Ferk, and B. Leskošek. Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning. *Plos One*, 13(5):e0196865, 2018.

[Keskar *et al.*, 2017] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *ICLR*, 2017.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[Ma *et al.*, 2018] Tengfei Ma, Cao Xiao, Jiayu Zhou, and Fei Wang. Drug similarity integration through attentive multi-view graph auto-encoders. In *IJCAI*, 2018.

[Miyato *et al.*, 2017] Takeru Miyato, Shin Ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *arXiv preprint arXiv:1704.03976*, 2017.

[Percha and Altman, 2013] B Percha and R. B. Altman. Informatics confronts drug–drug interactions. *Trends in Pharmacological Sciences*, 34(3):178–184, 2013.

[Qato *et al.*, 2016] D. M. Qato, J Wilder, L. P. Schumm, V Gillet, and G. C. Alexander. Changes in prescription and over-the-counter medication and dietary supplement use among older adults in the united states, 2005 vs 2011. *Jama Internal Medicine*, 176(4):473, 2016.

[Rasmus *et al.*, 2015] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *NeurlPS*, 2015.

[Ryu *et al.*, 2018] Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. Deep learning improves prediction of drug–drug and drug–food interactions. *PNAS*, 115(18):E4304–E4311, 2018.

[Sun, 2013] Shiliang Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038, 2013.

[Takeda *et al.*, 2017] T Takeda, M. Hao, T. Cheng, S. H. Bryant, and Y. Wang. Predicting drug-drug interactions through drug structural similarities and interaction networks incorporating pharmacokinetics and pharmacodynamics knowledge. *Journal of Cheminformatics*, 9(1):16, 2017.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurlPS*, 2017.

[Tatonetti *et al.*, 2012] Nicholas P Tatonetti, P Ye Patrick, Roxana Daneshjou, and Russ B Altman. Data-driven prediction of drug effects and interactions. *Science translational medicine*, 4(125):125ra31–125ra31, 2012.

[Vilar *et al.*, 2012] Santiago Vilar, Rave Harpaz, Eugenio Uriarte, Lourdes Santana, Raul Rabadan, and Carol Friedman. Drug—drug interaction through molecular structure similarity analysis. *Journal of the American Medical Informatics Association Jamia*, 19(6):1066, 2012.

[Vilar *et al.*, 2017] 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*, 2017.

[Zhang *et al.*, 2015] P. Zhang, F. Wang, J. Hu, and R Sorrentino. Label propagation prediction of drug-drug interactions based on clinical side effects. *Scientific Reports*, 5:12339, 2015.

[Zhang *et al.*, 2017] Wen Zhang, Yanlin Chen, Feng Liu, Fei Luo, Gang Tian, and Xiaohong Li. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. *Bmc Bioinformatics*, 18(1):18, 2017.

[Zitnik *et al.*, 2018] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018.