`SafeDrug`: Dual Molecular Graph Encoders for Recommending Effective and Safe Drug Combinations

Chaoqi Yang[1] , **Cao Xiao**[2] , **Fenglong Ma**[3] , **Lucas Glass**[2] and **Jimeng Sun**[1*]

[1]Department of Computer Science, University of Illinois Urbana-Champaign
[2]Analytics Center of Excellence, IQVIA
[3]College of Information Sciences and Technology, Pennsylvania State University
[1]{chaoqiy2, jimeng}@illinois.edu, [2]{cao.xiao, Lucas.Glass}@iqvia.com, [3]fenglong@psu.edu

Abstract

Medication recommendation is an essential task of AI for healthcare. Existing works focused on recommending drug combinations for patients with complex health conditions solely based on their electronic health records. Thus, they have the following limitations: (1) some important data such as *drug molecule structures* have not been utilized in the recommendation process. (2) drug-drug interactions (DDI) are modeled implicitly, which can lead to sub-optimal results. To address these limitations, we propose a DDI-controllable drug recommendation model named `SafeDrug` to leverage drugs' molecule structures and model DDIs explicitly. `SafeDrug` is equipped with a global message passing neural network (MPNN) module and a local bipartite learning module to fully encode the connectivity and functionality of drug molecules. `SafeDrug` also has a controllable loss function to control DDI level in the recommended drug combinations effectively. On a benchmark dataset, our `SafeDrug` is relatively shown to reduce DDI by 19.43% and improves 2.88% on Jaccard similarity between recommended and actually prescribed drug combinations over previous approaches. Moreover, `SafeDrug` also requires much fewer parameters than previous deep learning based approaches, leading to faster training by about 14% and around 2× speed-up in inference.

1 Introduction

Today abundant health data such as longitudinal electronic health records (EHR) and massive biomedical data available on the web enable researchers and doctors to build better predictive models for clinical decision making [Choi *et al.*, 2016a; Xiao *et al.*, 2018]. Among others, recommending effective and safe medication combinations is an important task, especially for helping patients with complex health conditions [Shang *et al.*, 2019; Shang *et al.*, 2018]. The primary goal of recommending medication combination is to customize a safe combination of drugs for a particular patient based on the patient's health conditions. Earlier medication recommendation models are *instance-based*, which are

only based on the current hospital visit [Zhang *et al.*, 2017; Wang *et al.*, 2017]. As a result, a patient with newly diagnosed hypertension will likely be treated the same as another patient who has suffered chronic uncontrolled hypertension. Such a limitation affects the safety and utility of the recommendations. To overcome this issue, *Longitudinal* methods such as [Le *et al.*, 2018; Shang *et al.*, 2019] are proposed. They leverage the temporal dependencies within longitudinal patient history to provide a more personalized recommendation. However, they still suffer from the following limitations.

- **Inadequate Medication Encoding.** Existing works [Zhang *et al.*, 2017; Shang *et al.*, 2019] often represent medications using one-hot encoding. Each drug is considered a (binary) unit, ignoring that drugs in their meaningful molecular graph representation encode important drug properties, such as efficacy and safety profiles. Also, molecule substructures are correlated with functionalities. Such knowledge can be useful for improving accuracy and safety in medication recommendations.

- **Implicit and Non-controllable DDI Modeling.** Some existing works model drug-drug interactions (DDI) via soft or indirect constraints, like knowledge graphs (KGs) [Wang *et al.*, 2017; Mao *et al.*, 2019] and reinforcement post-processing [Zhang *et al.*, 2017]. These implicit handling of DDIs results in non-controllable rates in the final recommendation or sub-optimal recommendation accuracy.

To address these, we propose a DDI-controllable drug recommendation model, named `SafeDrug`, to leverage drugs' molecule structures and explicitly model DDIs. We argue it is beneficial to incorporate molecule structures in medication recommendation. Our `SafeDrug` has the following contributions:

- **Dual Molecular Encoders to Capture Global and Local Molecule Patterns.** `SafeDrug` model firstly learns *patient representation*, which is fed into dual molecule encoders to capture the global pharmacological properties [Brown and Fraser, 1868] and the local sub-structural patterns of a drug [Huang *et al.*, 2020b]. Globally, a message-passing neural network (MPNN) encoder is constructed to pass molecular information messages for the whole drug layer by layer. Drug connectivity information is well captured in multiple hops. Locally, a bipartite encoder segments drug molecules into substructures, each possibly associated with

small functional groups. In this work, the substructure representation is fed into an effective masked neural network. The final output of the model is obtained by element-wise integration of the global and local encoded embeddings.

- **DDI Controllable Loss Function.** Inspired by proportional-integral-derivative (PID) controller [Astrom and Hagglund, 1995], we design a new technique to adaptively combine supervised loss and unsupervised DDI constraints with considerable flexibility. During the training, the negative DDI signal would be emphasized and backpropagated if the DDI rate of individual samples is above a certain threshold/target. In the experiment, the adaptive gradient descending can balance the model accuracy and final DDI level. With a preset target, `SafeDrug` model can provide reliable drug combinations to meet different levels of DDI requirements.

- **Comprehensive Evaluations.** We follow previous works [Zhang *et al.*, 2017; Shang *et al.*, 2019; Le *et al.*, 2018] and evaluate `SafeDrug` on a benchmark medication recommendation dataset extracted from the *MIMIC-III* database [Johnson *et al.*, 2016]. `SafeDrug` outperforms the best baselines relatively by 19.43% improvement in DDI reduction, 2.88% in Jaccard similarity, and 2.14% in F1 measure. Besides, `SafeDrug` requires much fewer parameters than previous deep learning based drug recommendation approaches with 14% reduction in training time and around 2× speed-up during inference.

Data processing files and codes are released in Github [1]. A full version of the paper can be found in arXiv [2].

2 Related Works

2.1 Medication Recommendation

Existing medication recommendation algorithms can be categorized into instance-based and longitudinal recommendation methods. Instance-based approaches focus on the patient's current health status. For example, LEAP [Zhang *et al.*, 2017] extracts feature information from the current encounter and adopts a multi-instance multi-label drug recommendation setting. Longitudinal approaches are proposed to leverage the temporal dependencies within the clinical history, see [Choi *et al.*, 2016a; Xiao *et al.*, 2018]. Among them, RETAIN [Choi *et al.*, 2016b] utilizes a two-level RNN with reverse time attention to model the longitudinal information. GAMENet [Shang *et al.*, 2019] adopts memory neural networks and stores historical drug information as references for further prediction. These methods either model the final drug combination as multi-label binary classification [Choi *et al.*, 2016b; Shang *et al.*, 2019; Shang *et al.*, 2018] or by sequential decision making [Zhang *et al.*, 2017]. Despite their initial success, these existing works still have the following limitations. Other important data such as *drug molecule structures* can help augment drug recommendation but have not been utilized by existing models. Also, existing works model DDI inadequately, either via soft or indirect constraints.

---

[1] https://github.com/ycq091044/SafeDrug
[2] https://arxiv.org/abs/2105.02711

This paper proposes a DDI-controllable drug recommendation model that leverages drugs' molecule structures and models DDIs effectively.

2.2 Molecule Representation

Molecule graph representation remains an important topic due to the association between molecule structure and properties, e.g., efficacy and safety. Early days, molecular descriptors [Mauri *et al.*, 2006] and drug fingerprint [Rogers and Hahn, 2010; Duvenaud *et al.*, 2015] were commonly used to represent drug molecules. Deep learning models were developed in recent years to learn molecule representation and model molecule substructures (a set of connected atoms). For example, [Huang *et al.*, 2020b] developed a sequential pattern mining method on SMILES strings and modeled a pair of drugs as a set of their substring representation. Also, [Huang *et al.*, 2020a] proposed to directly model molecule graphs using graph-based neural network models.

In this paper, inspired by [Huang *et al.*, 2020a], `SafeDrug` is proposed to capture both the molecular global and local information during the recommendation.

3 Problem Formulation

Electrical Health Records (EHR). Patient EHR data captures comprehensive medical histories of patients in the format of longitudinal vectors of medical codes (e.g., diagnoses, procedures, and drugs). Formally, EHR for patient j can be represented as a sequence $\mathbf{X}_j = [\mathbf{x}_j^{(1)}, \mathbf{x}_j^{(2)}, \ldots, \mathbf{x}_j^{(V_j)}]$, where V_j is the total number of visits for patient j. The i-th entry of $\mathbf{X}_j$ documents the i-th visit of patient j, specified by $\mathbf{x}_j^{(i)} = [\mathbf{d}_j^{(i)}, \mathbf{p}_j^{(i)}, \mathbf{m}_j^{(i)}]$, where $\mathbf{d}_j^{(i)} \in \{0,1\}^{|\mathcal{D}|}, \mathbf{p}_j^{(i)} \in \{0,1\}^{|\mathcal{P}|}, \mathbf{m}_j^{(i)} \in \{0,1\}^{|\mathcal{M}|}$, are multi-hot diagnosis, procedure and medication vectors, respectively. $\mathcal{D}, \mathcal{P}, \mathcal{M}$ are the corresponding element sets, and $|\cdot|$ denotes the cardinality.

Safe Drug Combination Recommendation. We focus on drug recommendation while reducing/controlling the drug-drug interaction (DDI) rates in the prediction. We use a symmetric binary adjacency matrix $\mathbf{D} \in \{0,1\}^{|\mathcal{M}|\times|\mathcal{M}|}$ to denote DDI relation. $\mathbf{D}_{ij} = 1$ means that the interaction between drug i and j has been reported while $\mathbf{D}_{ij} = 0$ indicates a safe co-prescription. The paper aims at learning a drug recommendation function, $f(\cdot)$, such that for each visit, e.g., the t-th visit, given the patient's up-to-now diagnosis and procedure sequences $[\mathbf{d}^{(1)}, \mathbf{d}^{(2)}, \ldots, \mathbf{d}^{(t)}]$ and $[\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \ldots, \mathbf{p}^{(t)}]$, along with the DDI matrix $\mathbf{D}$, $f(\cdot)$ can generate a drug recommendation,

$$\hat{\mathbf{m}}^{(t)} = f\left([\mathbf{d}^{(i)}]_{i=1}^t, [\mathbf{p}^{(i)}]_{i=1}^t\right) \in \{0,1\}^{|\mathcal{M}|}. \quad (1)$$

The objective function consists of two parts: (i) extracting real drug combinations, $\mathbf{m}^{(t)}$, as supervision to penalize $\hat{\mathbf{m}}^{(t)}$; (ii) using matrix $\mathbf{D}$ to derive unsupervised DDI constraints on $\hat{\mathbf{m}}^{(t)}$.

4 The `SafeDrug` Model

As illustrated in Figure 1, our `SafeDrug` model comprises of four components: (1) a **longitudinal patient representation module** that learns patient representation from their
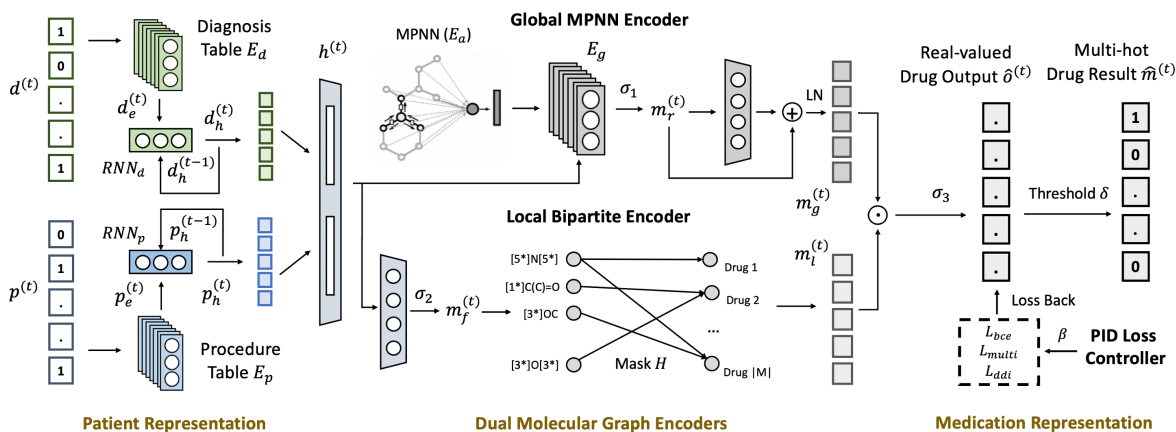
Figure 1: The `SafeDrug` Model. We first encode diagnosis and procedure sequences by RNNs to generate a patient health representation, $\mathbf{h}^{(t)}$; this representation then passes through dual molecular graph encoders for global and local molecular structural embeddings, $\mathbf{m}_g^{(t)}$ and $\mathbf{m}_l^{(t)}$; two embedding vectors are finally combined and thresholded to obtain the recommended drug combination, $\hat{\mathbf{m}}^{(t)}$.

EHR data; (2) a **global message passing neural network (MPNN) encoder** that takes patient representation as input and outputs a global drug vector with each entry quantifying the similarity between the patient representation and each of the drug representations; (3) in parallel, a **bipartite encoder** also takes the same patient representation and outputs a local drug vector, which encodes drugs' molecular substructural functionalities; (4) finally, the global and local drug representation vectors are element-wise combined in the **medication representation module**, where the final drug output is obtained from element-wise thresholding.

4.1 Longitudinal Patient Representation

From longitudinal EHR data, patient health can be encoded by their diagnosis and procedure information.

Diagnosis Embedding. To fully utilize rich diagnosis information, we design an embedding table, $\mathbf{E}_d \in \mathbb{R}^{|\mathcal{D}| \times dim}$, where each row stores the embedding vector for a particular diagnosis and dim is the dimension of the embedding space. Given a multi-hot diagnosis vector $\mathbf{d}^{(t)} \in \{0, 1\}^{|\mathcal{D}|}$, we project the corresponding diagnosis codes into the embedding space by vector-matrix dot product, which essentially takes the summation of each diagnosis embedding,

$$\mathbf{d}_e^{(t)} = \mathbf{d}^{(t)} \mathbf{E}_d. \qquad (2)$$

During the training, $\mathbf{E}_d$ is learnable and shared between each visit and every patient.

Procedure Embedding. Similarly, a shared procedure table $\mathbf{E}_p \in \mathbb{R}^{|\mathcal{P}| \times dim}$ is also designed to encode the associated procedure vector $\mathbf{p}^{(t)}$ (also a multi-hot vector),

$$\mathbf{p}_e^{(t)} = \mathbf{p}^{(t)} \mathbf{E}_p. \qquad (3)$$

Two *embedding vectors*, $\mathbf{d}_e^{(t)}, \mathbf{p}_e^{(t)} \in \mathbb{R}^{dim}$, encode the patient's current health condition. In fact, a health snapshot might be insufficient for treatment decisions. For example, it is inappropriate to prescribe the same medicine for one patient with a newly diagnosed hypertension and another patient who has chronic hypertension for more than two years.

Here we use RNN models to model dynamic patient history. Specifically, we use two separate RNNs to obtain the *hidden diagnosis and procedure vectors*, $\mathbf{d}_h^{(t)}$ and $\mathbf{p}_h^{(t)}$, in case that one of the sequences might be inaccessible in practice,

$$\mathbf{d}_h^{(t)} = \text{RNN}_d\left(\mathbf{d}_e^{(t)}, \mathbf{d}_h^{(t-1)}\right) = \text{RNN}_d\left(\mathbf{d}_e^{(t)}, \dots, \mathbf{d}_e^{(1)}\right), \quad (4)$$

$$\mathbf{p}_h^{(t)} = \text{RNN}_p\left(\mathbf{p}_e^{(t)}, \mathbf{p}_h^{(t-1)}\right) = \text{RNN}_p\left(\mathbf{p}_e^{(t)}, \dots, \mathbf{p}_e^{(1)}\right), \quad (5)$$

where $\mathbf{d}_h^{(t)}, \mathbf{p}_h^{(t)} \in \mathbb{R}^{dim}$. And $\mathbf{d}_h^{(0)}, \mathbf{p}_h^{(0)}$ are all-zero vectors.

Patient Representation. We then concatenate the diagnosis embedding $\mathbf{d}_h^{(t)}$ and procedure embedding $\mathbf{p}_h^{(t)}$ into a more compact *patient representation* $\mathbf{h}^{(t)}$. We follow a common and effective approach to first *concatenate* two vectors as a double-long vector, and then apply a feed-forward neural network, $\text{NN}_1(\cdot) : \mathbb{R}^{2dim} \mapsto \mathbb{R}^{dim}$, ($\mathbf{W}_1$ is the parameter, and $\#$ is the concatenation operation),

$$\mathbf{h}^{(t)} = \text{NN}_1([\mathbf{d}_h^{(t)} \# \mathbf{p}_h^{(t)}]; \mathbf{W}_1). \qquad (6)$$

To this end, we obtain an overall patient representation $\mathbf{h}^{(t)} \in \mathbb{R}^{dim}$, from which, we next generate a safe drug recommendation by comprehensively modeling the drug molecule databases.

4.2 Dual Molecular Graph Encoders

To investigate drug properties and their dependencies, we exploit the global molecule connectivity features and the local molecule functionalities. In the following sections, we introduce two new molecular graph encoders, in parallel, to comprehensively encode the drug molecules.

(I) Global MPNN Encoder

We encode drug molecule data using a message passing neural network (MPNN) operator with learnable fingerprints, aiming at convolving and pooling atom information across single molecule graphs into vector representations.

To begin with, we collect a set of all appeared atoms, $\mathcal{C} = \{a_i\}$, and design a learnable atom embedding table,

$\mathbf{E}_a \in \mathbb{R}^{|\mathcal{C}| \times dim}$, where each row looks up the initial embedding/fingerprint for a particular atom. Our design of the MPNN model is very different from previous works [Coley *et al.*, 2017; Yang *et al.*, 2019], which mostly use atom descriptors, e.g., atom mass and chirality, as the initial features. However, for molecule structure modeling, atom-atom connectivity is more important than individual atoms.

Given a drug molecule graph (atoms $a_0, \ldots, a_n$ as the vertices and atom-atom bonds as edges) with adjacency matrix $\mathbf{A}$ and the initial atom fingerprints from $\mathbf{E}_a$, i.e., $\{\mathbf{y}_0^{(0)}, \mathbf{y}_1^{(0)}, \ldots, \mathbf{y}_n^{(0)}\}$, we formulate the layer-wise message passing over the graph as,

$$\mathbf{z}_i^{(l+1)} = \sum_{j:\, \mathbf{A}_{ij}=1} \text{MESSAGE}_l\, (\mathbf{y}_i^{(l)}, \mathbf{y}_j^{(l)}; \mathbf{W}^{(l)}), \quad (7)$$

$$\mathbf{y}_i^{(l+1)} = \text{UPDATE}_l\, (\mathbf{y}_i^{(l)}, \mathbf{z}_i^{(l+1)}), \; i = 0, \ldots, n \quad (8)$$

where l is the layer index, $\mathbf{z}_i^{(l+1)}$ is the encoded messages from the neighbors of atom i at the l-th iteration, $\mathbf{y}_i^{(l+1)}$ is the hidden state of atom i. During message passing, $\mathbf{z}_i^{(l+1)}$ and $\mathbf{y}_i^{(l+1)}$ with each vertex atom i are updated using *message function*, $\text{MESSAGE}_l(\cdot)$, and vertex *update function*, $\text{UPDATE}_l(\cdot)$, and $\mathbf{W}^{(l)}$ is the layer-wise parameter matrix. After applying message passing for L layers, the global representation of a drug molecule is pooled by a *readout function*, which calculates the average of all atom fingerprints,

$$\mathbf{y} = \text{READOUT}\, (\{\mathbf{y}_i^{(L)} \mid i = 0, \ldots, n\}). \quad (9)$$

We apply the same MPNN encoder with shared parameters for each single drug moledule (there are $|\mathcal{M}|$ different drugs in total). We collect the MPNN embeddings of all drug molecules into a drug memory $\mathbf{E}_g \in \mathbb{R}^{|\mathcal{M}| \times dim}$, where each row is the embedding of a drug.

Patient-to-drug Matching. When treating patient representation $\mathbf{h}^{(t)}$ as a query, we aim at searching the most *relevant* drugs from the memory $\mathbf{E}_g$. The health-to-drug matching score is calculated by the dot product of the MPNN representation and the patient representation $\mathbf{h}^{(t)}$. For all drugs, we operate each row of $\mathbf{E}_g$ with $\mathbf{h}^{(t)}$, followed by a sigmoid function $\sigma_1(\cdot)$,

$$\mathbf{m}_r^{(t)} = \sigma_1\, (\mathbf{E}_g \mathbf{h}^{(t)}). \quad (10)$$

Every element of $\mathbf{m}_r^{(t)}$ stores a matching score for one drug. In light of the recent successful post-LN architecture [Wang *et al.*, 2019] in Transformer [Vaswani *et al.*, 2017], the matching score is further parameterized by a feed-forward neural network, $\text{NN}_2(\cdot) : \mathbb{R}^{dim} \mapsto \mathbb{R}^{dim}$, followed by layer-normalization (LN) step with residual connection,

$$\mathbf{m}_g^{(t)} = \text{LN}\, (\mathbf{m}_r^{(t)} + \text{NN}_2\, (\mathbf{m}_r^{(t)}; \mathbf{W}_2)). \quad (11)$$

(II) Local Bipartite Encoder

Molecules with similar structures shall be mapped into nearby space using the MPNN encoder. However, a pair of drugs that have significant overlap on structure domain might behave differently in terms of DDI interaction or other functional activities [Ryu *et al.*, 2018]. In fact, a drug's functionality is better reflected by molecule substructures, which are defined as a set of connected atoms that may have associated functional groups. This section focuses on the substructure composition of molecules and builds a substructure-to-drug bipartite architecture for capturing drug local functionality and their dependencies.

Molecule Segmentation. To capture the relations between drug molecules and their substructures, we first adopt the well-known molecule segmentation method, breaking retrosynthetically interesting chemical substructures (BRICS) [Degen *et al.*, 2008], which preserves the most critical drug functional groups and more crucial bonds. `SafeDrug` can easily support other segmentation methods such as RECAP, ECFP. After the segmentation, each molecule is decomposed into a set of chemical substructures. In this setting, a pair of drugs could intersect on a mutual set of substructures. We denote the overall substructure set as $\mathcal{S}$. Based on the many-to-many relations, it is ready to construct a bipartite architecture, represented by a *mask matrix* $\mathbf{H} \in \{0,1\}^{|\mathcal{S}| \times |\mathcal{M}|}$: $\mathbf{H}_{ij} = 1$ indicates that drug j contains substructure i.

Bipartite Learning. We hope that $\mathbf{H}$ can help to derive a drug representation on its functionality information, given the input, $\mathbf{h}^{(t)}$. To do so, we first apply a feed-forward neural network, $\text{NN}_3(\cdot) : \mathbb{R}^{dim} \mapsto \mathbb{R}^{|\mathcal{S}|}$, for dimensionality transformation, followed by a sigmoid function, $\sigma_2(\cdot)$. In fact, $\text{NN}_3(\cdot)$ transforms the patient representation $\mathbf{h}^{(t)}$ into a local functional vector,

$$\mathbf{m}_f^{(t)} = \sigma_2\, (\text{NN}_3\, (\mathbf{h}^{(t)}; \mathbf{W}_3)). \quad (12)$$

Each entry of $\mathbf{m}_f^{(t)}$ quantifies the significance of the corresponding functionality. Therefore, $\mathbf{m}_f^{(t)}$ can be viewed as an anti-disease functionality combination for the current patient. Next, we are motivated to generate the recommendation of drugs that can cover all these anti-disease functionalities while considering to prevent DDIs.

We therefore perform network pruning and design a masked 1-layer neural network, $\text{NN}_4(\cdot) : \mathbb{R}^{|\mathcal{S}|} \mapsto \mathbb{R}^{|\mathcal{M}|}$, where the parameter matrix is masked by the bipartite architecture $\mathbf{H}$, essentially through matrix element-wise product $\odot$. During the training process, $\text{NN}_4(\cdot)$ will learn to map the functionality vector $\mathbf{m}_f^{(t)}$ into the corresponding local drug representation,

$$\mathbf{m}_l^{(t)} = \text{NN}_4\, (\mathbf{m}_f^{(t)}; \mathbf{W}_4 \odot \mathbf{H}). \quad (13)$$

We show that the mask $\mathbf{H}$ would enable the model to (i) have far fewer parameters (due to the sparsity of $\mathbf{H}$); (ii) prevent DDIs by avoiding a co-prescription of interacted drugs (we demonstrate it in Appendix).

4.3 Medication Representation

The final medication representation is given by treating the global drug matching vector as attention signals to further adjust the local drug functionality representation. We finally use

a sigmoid function $\sigma_3(\cdot)$ to scale the output,

$$\hat{\mathbf{o}}^{(t)} = \sigma_3\,(\mathbf{m}_g^{(t)} \odot \mathbf{m}_l^{(t)}), \qquad (14)$$

where $\odot$ represents the element-wise product. By a threshold value δ, we can achieve the recommended drug combination, i.e., a multi-hot vector $\hat{\mathbf{m}}^{(t)}$, by picking those entries where the value is greater than δ.

4.4 Model Training and Inference

`SafeDrug` is trained end-to-end. We simultaneously learn $\mathbf{E}_d$, $\mathbf{E}_p$, parameters in RNN_d and RNN_p, network parameters $\mathbf{W}_1$, $\mathbf{E}_a$, layer-wise parameters in MPNN $\{\mathbf{W}^{(i)}\}$, parameters in layer normalization (LN), and $\mathbf{W}_2$, $\mathbf{W}_3$ and $\mathbf{W}_4$.

Multiple Loss. In this paper, the recommendation task is formulated as multi-label binary classification. Suppose $|\mathcal{M}|$ is the total number of drugs. We use $\mathbf{m}^{(t)}$ to denote the target multi-hot drug recommendation, $\hat{\mathbf{o}}^{(t)}$ for the output real-valued drug representation (opposed to the predicted multi-hot vector $\hat{\mathbf{m}}^{(t)}$ after applying threshold δ). We consider the prediction of each drug as a sub-problem and use binary cross-entropy (BCE) as a loss,

$$L_{bce} = -\sum_{i=1}^{|\mathcal{M}|} \mathbf{m}_i^{(t)} log(\hat{\mathbf{o}}_i^{(t)}) + (1 - \mathbf{m}_i^{(t)}) log(1 - \hat{\mathbf{o}}_i^{(t)}), \quad (15)$$

where $\mathbf{m}_i^{(t)}$ and $\hat{\mathbf{o}}_i^{(t)}$ are the i-th entries. To make the result more robust, we also adopt multi-label hinge loss to ensure that the truth labels have at least 1 margin larger than others,

$$L_{multi} = \sum_{i,j:\ \mathbf{m}_i^{(t)}=1, \mathbf{m}_j^{(t)}=0} \frac{\max(0, 1 - (\hat{\mathbf{o}}_i^{(t)} - \hat{\mathbf{o}}_j^{(t)}))}{|\mathcal{M}|}. \quad (16)$$

Based on the DDI adjacency matrix $\mathbf{D}$, we also design the adverse DDI loss over $\hat{\mathbf{o}}^{(t)}$ as,

$$L_{ddi} = \sum_{i=1}^{|\mathcal{M}|}\sum_{j=1}^{|\mathcal{M}|} \mathbf{D}_{ij} \cdot \hat{\mathbf{o}}_i^{(t)} \cdot \hat{\mathbf{o}}_j^{(t)}, \qquad (17)$$

where $\cdot$ is the product between scalars. Note that the above loss functions are defined for a single visit. During the training, loss back propagation will be conducted at patient level by the averaged losses across all visits.

Controllable Loss Function. A standard approach to training with multiple loss functions is by the weighted sum of the loss measuring terms [Dosovitskiy and Djolonga, 2019],

$$L = \beta\,(\alpha L_{bce} + (1-\alpha)L_{multi}) + (1-\beta)\,L_{ddi}, \quad (18)$$

where α and β are usually pre-defined hyperparameters.

In this scenario, two prediction loss, L_{bce} and L_{multi}, are compatible, and thus we select α from the validation set. We observe that adverse DDIs also exist in the dataset (doctors might mistakenly prescribe the interacted drugs). Therefore, by training with ground truth labels, the correct learning process might conversely increase DDI. Thus, we consider adjusting β during the training process by a proportional–integral–derivative (PID) controller [An et al., 2018]

to balance the prediction loss and DDI loss. For simplicity, we only utilize the proportional error signal as negative feedback when the DDI rate of the recommended drugs is above certain thresholds.

In safe drug recommendation, the key is to maintain a low DDI, so we set a DDI acceptance rate $\gamma \in (0, 1)$ for the loss function. Our motivation is that if the patient-level DDI is lower than the threshold, γ, then we consider only maximizing the prediction accuracy; otherwise, β will adjust adaptively to reduce DDI as well.

The controllable factor β follows the piece-wise strategy,

$$\beta = \begin{cases} 1, & \text{DDI} \leq \gamma \\ \max\,\{0,\ 1 - \frac{DDI-\gamma}{K_p}\}, & \text{DDI} > \gamma \end{cases} \quad (19)$$

where K_p is a correcting factor for the proportional signal.

We show in the experiment that γ can be viewed as an upper bound of the output DDI rate. By presetting γ, our `SafeDrug` can meet different level of DDI requirements.

The model inference phase basically follows the same pipeline as training. We apply a threshold $\delta = 0.5$ on the output drug representation in Eqn. (14) and select the drugs, corresponding to entries where the value is greater than δ, as the final recommendations. We structure the overall algorithm in Appendix.

5 Experiments

Dataset and Metrics. The experiments are carried out on *MIMIC-III* [Johnson et al., 2016]. The statistics of the post-processed data is reported in Table 1. We use five efficacy metrics: DDI rate, Jaccard similarity, F1 score, PRAUC, and # of medications, to evaluate the recommendation efficacy and three metrics: # of parameters, training time, and inference time, to evaluate the complexity of the models. Due to space limitation, we push other details to Appendix.

Items	Size
# of visits / # of patients	14,995 / 6,350
diag. / prod. / med. space size	1,958 / 1430 / 131
avg. / max # of visits	2.37 / 29
avg. / max # of diagnoses per visit	10.51 / 128
avg. / max # of procedures per visit	3.84 / 50
avg. / max # of medicines per visit	11.44 / 65
total # of DDI pairs	448
total # of substructures	491

Table 1: Data Statistics

Baselines. We compare `SafeDrug` with the following baselines from different perspectives: standard Logistic Regression (LR), multi-label classification approach: Ensemble Classifier Chain (ECC), [Read et al., 2009], RNN-based approach: RETAIN [Choi et al., 2016b], instance-based approach: LEAP [Zhang et al., 2017], longitudinal memory-based approaches: DMNC [Le et al., 2018] and GAMENet [Shang et al., 2019]. [Shang et al., 2018] is not considered a baseline since it requires extra ontology data. We also compared `SafeDrug` with its two model variants: `SafeDrug`$_L$ is for the model with only bipartite encoder (local, L) and `SafeDrug`$_G$ for the model with only the MPNN encoder

Model	DDI	Jaccard	F1-score	PRAUC	Avg. # of Drugs
LR	0.0829 ± 0.0009 (0.0)	0.4865 ± 0.0021 (0.0)	0.6434 ± 0.0019 (0.0)	0.7509 ± 0.0018 (3e-11)	16.1773 ± 0.0942
ECC	0.0846 ± 0.0018 (0.0)	0.4996 ± 0.0049 (5e-10)	0.6569 ± 0.0044 (4e-10)	0.6844 ± 0.0038 (0.0)	18.0722 ± 0.1914
RETAIN	0.0835 ± 0.0020 (0.0)	0.4887 ± 0.0028 (2e-15)	0.6481 ± 0.0027 (5e-15)	0.7556 ± 0.0033 (2e-06)	20.4051 ± 0.2832
LEAP	0.0731 ± 0.0008 (0.0)	0.4521 ± 0.0024 (0.0)	0.6138 ± 0.0026 (0.0)	0.6549 ± 0.0033 (0.0)	18.7138 ± 0.0666
DMNC	0.0842 ± 0.0011 (0.0)	0.4864 ± 0.0025 (2e-16)	0.6529 ± 0.0030 (3e-13)	0.7580 ± 0.0039 (2e-04)	20.0000 ± 0.0000
GAMENet	0.0864 ± 0.0006 (0.0)	0.5067 ± 0.0025 (6e-10)	0.6626 ± 0.0025 (4e-10)	0.7631 ± 0.0030 (0.21)	27.2145 ± 0.1141
SafeDrug_L	**0.0580 ± 0.0004** (−−)	0.5166 ± 0.0026 (1e-03)	0.6670 ± 0.0024 (9e-08)	0.7632 ± 0.0026 (0.20)	18.6663 ± 0.1417
SafeDrug_G	0.0606 ± 0.0007 (7e-06)	0.4862 ± 0.0027 (4e-16)	0.6442 ± 0.0024 (2e-16)	0.7423 ± 0.0029 (4e-13)	19.0724 ± 0.0971
SafeDrug	0.0589 ± 0.0005	**0.5213 ± 0.0030**	**0.6768 ± 0.0027**	**0.7647 ± 0.0025**	19.9178 ± 0.1604

Table 2: Performance Comparison on *MIMIC-III* (ground truth DDI rate is 0.0808)

(global, G). In implementation, we use $\mathbf{m}_l^{(t)}$ and $\mathbf{m}_g^{(t)}$ as the final drug representation for them, respectively.

Performance Comparison. From Table 2, it is easy to see the drug sequence generation models LEAP, DMNC, and ECC yield poor results among the baselines. We , therefore, conclude that formulating the drug recommendation task as multi-label prediction might be more straightforward and effective. LR and RETAIN do not consider the DDI information, so they output undesirable DDI rate. The recommendations given by GAMENet are based on historical combinations, however, real medication records usually contain high DDI rate, e.g., around 8% in *MIMIC-III*, so it also provides a high DDI rate. Moreover, the performance of two variants, SafeDrug_L and SafeDrug_G, show that both of the molecular encoders do help with the prediction. Overall, our SafeDrug also outperforms the other baselines with significantly lower DDI rate and better accuracy.

Significance Test. The above results are given by 10 rounds of bootstrapping sampling. Next, we conduct T-test on each metric for two-tailed hypothesis. The p-values are shown in the parenthesis. Commonly, a p-value smaller than 0.05 would be considered as significant. We also use (−−) for cases where the result of SafeDrug is worse than the baseline. To conclude, our model can beat all compared baselines in almost every metric with p=0.001 significance level.

Model	# of Param.	Training Time	Inference
RETAIN	287,940	36.35s / Epoch	3.98s
LEAP	433,286	336.14s / Epoch	32.31s
DMNC	525,533	4697.40s / Epoch	262.19s
GAMENet	449,092	162.10s / Epoch	26.85s
SafeDrug_L	299,404	133.43s / Epoch	9.93s
SafeDrug_G	324,577	129.66s / Epoch	9.79s
SafeDrug	325,473	138.77s / Epoch	10.64s
Δ (improve.)	↓ 24.88%	↓ 14.39%	↓ 44.21%

Table 3: DL Model Complexity Comparison

Comparison of Model Efficiency. Model complexity of SafeDrug and other DL baselines are evaluated in Table 3. Since RETAIN is not specialized for drug recommendation, improvements in Table 3 are over LEAP, DMNC, and GAMENet. Efficiency wise, our model has relatively lower space and time complexity over LEAP, DMNC, and GAMENet. Also, it is much more efficient during inference. Leap and DMNC adopt sequential modeling and recommend

γ	DDI	# of Med.	Jaccard	F1-score	PRAUC
0.00	0.0073	13.82	0.4198	0.5743	0.6902
0.01	0.0152	15.35	0.4391	0.5924	0.6788
0.02	0.0254	17.64	0.4741	0.6255	0.7196
0.03	0.0301	18.33	0.4959	0.6540	0.7492
0.04	0.0381	18.87	0.5083	0.6651	0.7556
0.05	0.0479	19.08	0.5164	0.6720	0.7604
0.06	0.0589	19.91	0.5213	0.6768	0.7647
0.07	0.0672	20.18	0.5228	0.6776	0.7634
0.08	0.0722	20.75	0.5250	0.6794	0.7691

Table 4: Performance under Acceptance DDI Rate γ

drugs one by one, which is time-consuming. GAMENet stores a large memory bank, and thus it requires larger space. By comparison, we conclude that SafeDrug is efficient and flexible for real deployment.

Analysis of DDI Controllability. We evaluate DDI controllability and show that the average DDI rate of our output can be controlled by threshold γ. The ground truth DDI rate in *MIMIC-III* is 0.0808. We consider that learning-based drug recommendation approaches should optimize the final recommendations by processing EHR and drug-drug relation databases, so the output DDI rate is supposed to be lower than human experts'. In this section, we test for a series of targeted DDI γ ranging from 0 to 0.08. For each γ, we train a separate model. The experiment is conducted for 5 times. After convergence, we show the average metrics in Table 4.

From the result, the DDI rates are controlled and upper-bounded by γ in most cases, which is consistent with the design. When γ becomes larger, more drugs are allowed in one combination, and SafeDrug is more accurate. When γ is too small (< 0.02), model accuracy will drop dramatically. The parameter γ provides a way for doctors to control the tradeoff between DDI rates and accuracy in recommendations.

6 Conclusion

In this paper, we propose SafeDrug accurate and safe drug recommendation. Our model extracts and encode rich molecule structure information to augment the drug recommendation task. We design an adaptive loss controller to provide additional flexibility. We evaluated SafeDrug using benchmark data and showed better accuracy and efficiency.

Acknowledgements

References

[An *et al.*, 2018] Wangpeng An, Haoqian Wang, Qingyun Sun, Jun Xu, Qionghai Dai, and Lei Zhang. A pid controller approach for stochastic optimization of deep networks. In *CVPR*, 2018.

[Astrom and Hagglund, 1995] Karl Johan Astrom and Tore Hagglund. *PID controllers: theory, design, and tuning*, volume 2. 1995.

[Brown and Fraser, 1868] Alexander Crum Brown and Thomas R Fraser. On the connection between chemical constitution and physiological action. *JAP*, 1868.

[Choi *et al.*, 2016a] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *MLHC*, 2016.

[Choi *et al.*, 2016b] Edward Choi, Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NIPS*, 2016.

[Coley *et al.*, 2017] Connor W Coley, Regina Barzilay, William H Green, Tommi S Jaakkola, and Klavs F Jensen. Convolutional embedding of attributed molecular graphs for physical property prediction. *JCIM*, 2017.

[Degen *et al.*, 2008] Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem*, 2008.

[Dosovitskiy and Djolonga, 2019] Alexey Dosovitskiy and Josip Djolonga. You only train once: Loss-conditional training of deep networks. In *ICLR*, 2019.

[Duvenaud *et al.*, 2015] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *NIPS*, 2015.

[Huang *et al.*, 2020a] Kexin Huang, Tianfan Fu, Cao Xiao, Lucas Glass, and Jimeng Sun. Deeppurpose: a deep learning based drug repurposing toolkit. *Bioinformatics*, 2020.

[Huang *et al.*, 2020b] Kexin Huang, Cao Xiao, Trong Hoang, Lucas Glass, and Jimeng Sun. Caster: Predicting drug interactions with chemical substructure representation. In *AAAI*, 2020.

[Johnson *et al.*, 2016] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[Le *et al.*, 2018] Hung Le, Truyen Tran, and Svetha Venkatesh. Dual memory neural computer for asynchronous two-view sequential learning. In *KDD*, 2018.

[Mao *et al.*, 2019] Chengsheng Mao, Liang Yao, and Yuan Luo. Medgcn: Graph convolutional networks for multiple medical tasks. *arXiv preprint arXiv:1904.00326*, 2019.

[Mauri *et al.*, 2006] Andrea Mauri, Viviana Consonni, Manuela Pavan, and Roberto Todeschini. Dragon software: An easy approach to molecular descriptor calculations. *Communications in MCC*, 2006.

[Read *et al.*, 2009] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. In *ECML PKDD*, 2009.

[Rogers and Hahn, 2010] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *JCIM*, 2010.

[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*, 2018.

[Shang *et al.*, 2018] Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of graph augmented transformers for medication recommendation. *IJCAI*, 2018.

[Shang *et al.*, 2019] Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. Gamenet: Graph augmented memory networks for recommending medication combination. In *AAAI*, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[Wang *et al.*, 2017] Meng Wang, Mengyue Liu, Jun Liu, Sen Wang, Guodong Long, and Buyue Qian. Safe medicine recommendation via medical knowledge graph embedding. *arXiv:1710.05980*, 2017.

[Wang *et al.*, 2019] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787*, 2019.

[Xiao *et al.*, 2018] Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *JAMIA*, 2018.

[Yang *et al.*, 2019] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *JCIM*, 2019.

[Zhang *et al.*, 2017] Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart, and Jimeng Sun. Leap: Learning to prescribe effective and safe treatment combinations for multimorbidity. In *KDD*, 2017.