4SDrug: Symptom-based Set-to-set Small and Safe Drug Recommendation

Yanchao Tan
College of Computer and Data Science, Fuzhou University, China
yanchaotan@gmail.com

Chengjun Kong
Faculty of Science, National University of Singapore, Singapore
e0732699@u.nus.edu

Leisheng Yu
Department of Computer Science, Emory University, USA
leisheng.yu@emory.edu

Pan Li
Department of Computer Science, Purdue University, USA
panli@purdue.edu

Chaochao Chen
College of Computer Science, Zhejiang University, China
zjuccc@zju.edu.cn

Xiaolin Zheng
College of Computer Science, Zhejiang University, China
xlzheng@zju.edu.cn

Vicki S Hertzberg
Nell Hodgson Woodruff School of Nursing, Emory University, USA
vhertzb@emory.edu

Carl Yang*
Department of Computer Science, Emory University, USA
j.carlyang@emory.edu

ABSTRACT

Drug recommendation is an important task of AI for healthcare. To recommend proper drugs, existing methods rely on various clinical records (e.g., diagnosis and procedures), which are commonly found in data such as electronic health records (EHRs). However, detailed records as such are often not available and the inputs might merely include a set of symptoms provided by doctors. Moreover, existing drug recommender systems usually treat drugs as individual items, ignoring the unique requirements that drug recommendation has to be done on a set of items (drugs), which should be as small as possible and safe without harmful drug-drug interactions (DDIs).

To deal with the challenges above, in this paper, we propose a novel framework of Symptom-based Set-to-set Small and Safe drug recommendation (4SDrug). To enable set-to-set comparison, we design set-oriented representation and similarity measurement for both symptoms and drugs. Further, towards the symptom sets, we devise importance-based set aggregation to enhance the accuracy of symptom set representation; towards the drug sets, we devise intersection-based set augmentation to ensure smaller drug sets, and apply knowledge-based and data-driven penalties to ensure safer drug sets. Extensive experiments on two real-world EHR datasets, i.e., the public benchmark one of MIMIC-III and the industrial large-scale one of NELL, show drastic performance gains brought by 4SDrug, which outperforms all baselines in most effectiveness measures, while yielding the smallest sets of recommended drugs and 26.83% DDI rate reduction from the ground-truth data.

*Carl Yang is the corresponding author.

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539089

CCS CONCEPTS

• **Applied computing → Life and medical sciences**.

KEYWORDS

Drug recommendation, Symptom-based, Set-to-set comparison, Small and safe drug sets

ACM Reference Format:
Yanchao Tan, Chengjun Kong, Leisheng Yu, Pan Li, Chaochao Chen, Xiaolin Zheng, Vicki S Hertzberg, and Carl Yang. 2022. 4SDrug: Symptom-based Set-to-set Small and Safe Drug Recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3534678.3539089

1 INTRODUCTION

Today abundant healthcare data such as electronic health records (EHRs) enable researchers and doctors to build better predictive models for clinical decision making [30, 47]. Among them, drug recommendation is an important task, which provides candidate drug sets for doctors to work from with improved prescription efficiency [30, 47, 52]. Instead of replacing the effort of doctors, many industrial platforms have successfully assisted doctors with drug recommendation (e.g., Medical Brain of Baidu[1], Waston-health of IBM[2], and Medical AI of Tencent[3]). To recommend proper drugs, existing methods rely on various clinical records from actual hospital visits (e.g., diagnoses [2, 10, 29, 51], lab tests [52], and procedures [9, 30, 40, 43, 47]). Moreover, to provide more personalized drug recommendations, some methods also require historical health records [9, 30, 40, 43, 47]. Such reliance on complicated clinical records and personal information largely limits the use cases of existing drug recommendation methods.

In this work, we consider a simple yet realistic scenario of symptom-based drug recommendation, which provides efficient

[1]https://ai.baidu.com/industry/healthcare
[2]https://www.ibm.com/watson-health
[3]https://healthcare.tencent.com/

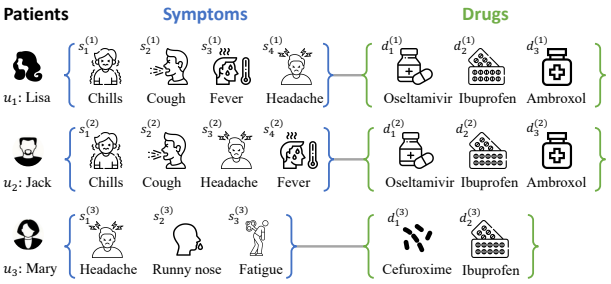


Figure 1: A toy example of the symptom-based set-to-set drug recommendation.

references for doctors during the actual diagnoses. Instead of using complicated historical clinical records, we propose to only require *a set of symptoms*, which can be actively entered by the doctor (e.g., the NELL dataset[4]) or automatically extracted from the medical records (e.g., the MIMIC-III dataset [15]). Since symptoms can reflect a patient's physical status [16, 28, 36] without exposing his/her personal information, the symptom-based drug recommendation system is secure from privacy issues and can be widely applied to assist doctors in prescribing proper drugs. As shown in Figure 1, both Lisa and Jack who show the same set of symptoms (i.e., $\{Chills, Cough, Fever, Headache\}$) are likely to be diagnosed with the same disease by doctors (i.e., viral influenza), and thus be prescribed with the same set of drugs (i.e., $\{Oseltamivir, Ibuprofen, Ambroxol\}$).

Since not all doctors can always avoid prescription errors (e.g., prescribing drug sets that include harmful drug-drug interactions (DDIs)) due to the limited expert experience and possible human negligence [3, 5], such a symptom-based drug recommendation system can be of great help. Specifically, the system can recommend a small set of safe drugs given a set of symptoms, which is convenient and accessible to doctors. However, this novel task of set-to-set recommendation between symptoms and drugs (as shown in Figure 1) poses several unique challenges for us.

Challenge I: *How to effectively model the set-to-set relations among symptoms and drugs?* Different from general recommendation which aims at modeling single users on one side and single items on the other, we need to consider sets of symptoms on one side and sets of drugs on the other. Since the sets have multiple elements and variable sizes, how to properly represent the sets and effectively optimize their relations is unknown.

Challenge II: *How to enhance the accuracy of symptom set representation?* To accurately represent a symptom set, it is practical to consider the varying importance of individual symptoms. For example, as shown in Figure 1, although Headache appears frequently in different sets, Fever plays a much more important role than Headache in the symptom set of Lisa, since it dominantly leads to the diagnosis of viral influenza. Without specific consideration of the importance of Fever, the model may fail to provide an accurate symptom representation towards effective drug recommendation.

Challenge III: *How to properly recommend small and safe sets of drugs?* Unlike general recommendation, the output of drug recommendation is a set of drugs. Firstly, the set needs to be small because too many drugs will increase the patient's financial burden and reluctancy of taking all of them. Moreover, the set of drugs needs to be safe, which should not include harmful drug-drug interactions (DDIs). Some existing studies model DDIs based on the external drug knowledge base, but this is not always applicable.

To address these challenges, we propose Symptom-based Set-to-set Small and Safe drug recommendation (4SDrug), which consists of three pivotal technical modules: (i) a *set-to-set comparison module*, which is introduced to effectively model the relations among symptom sets and drug sets; (ii) a *symptom set module*, which is presented to enhance the accuracy of symptom set representation by considering the importance of individual symptoms; and (iii) a *drug set module*, which recommends sets of drugs by ensuring the small and safe drug set principles.

Our overall contributions in this work are summarized as follows:

- *Formulation of symptom-based set-to-set drug recommendation.* 4SDrug is the first drug recommendation framework solely based on symptoms, which can provide convenient assistance to doctors while protecting the privacy of the patients. (Section 3.1).
- *Effective model designs.* In the *set-to-set comparison module*, we introduce set-oriented representation and similarity measurement to effectively model the set-to-set relations among symptoms and drugs (Section 3.2). In the *symptom set module*, importance-based set aggregation is devised to enhance the accuracy of symptom set representation (Section 3.3). In the *drug set module*, we devise intersection-based set augmentation, knowledge-based, and data-driven penalties to ensure small and safe drug sets recommendations (Section 3.4).
- *Extensive experiments on real EHR datasets.* We conduct comprehensive experimental evaluations on drug recommendation tasks against state-of-the-art approaches over both public benchmark and industrial large-scale EHR datasets. Extensive experimental results demonstrate the superiority of 4SDrug (Section 5).

2 RELATED WORK

General Recommendation. Recently, matrix factorization (MF) has become the *de facto* method, which uses inner products to model the similarity of the user-item relations [21, 27]. To model complex relations in real-world applications, recently, metric learning for recommendations [12, 35, 48] and graph learning for recommendations [33, 41, 42, 44–46] have attracted significant research attention. However, the above methods aim at modeling single users on one side and single items on the other, which fail to recommend sets of drugs in drug recommendation scenario.

Drug Set Recommendation. In the setting of drug set recommendation, many existing works model patient representations as sequences of hospital visits, where each visit consists of diagnoses [2, 10, 29, 34, 51], lab tests [52], and procedures [9, 30, 40, 43, 47]. For example, RETAIN [9] employed an attention model to identify the most meaningful historical visit, so as to make the model interpretable. [2] modeled a sequence of ICD-9 [1, 7, 26] codes based on GRU. However, these methods fail to model patients whose historical visits are not available.

To alleviate the reliance on the historical records, existing approaches recommend drugs based on clinical information of patients' current visits. For example, G-BERT [29] pretrained the records of patients with a single hospital visit. LEAP [51] extracted

[4]https://www.nursing.emory.edu/pages/project-nell

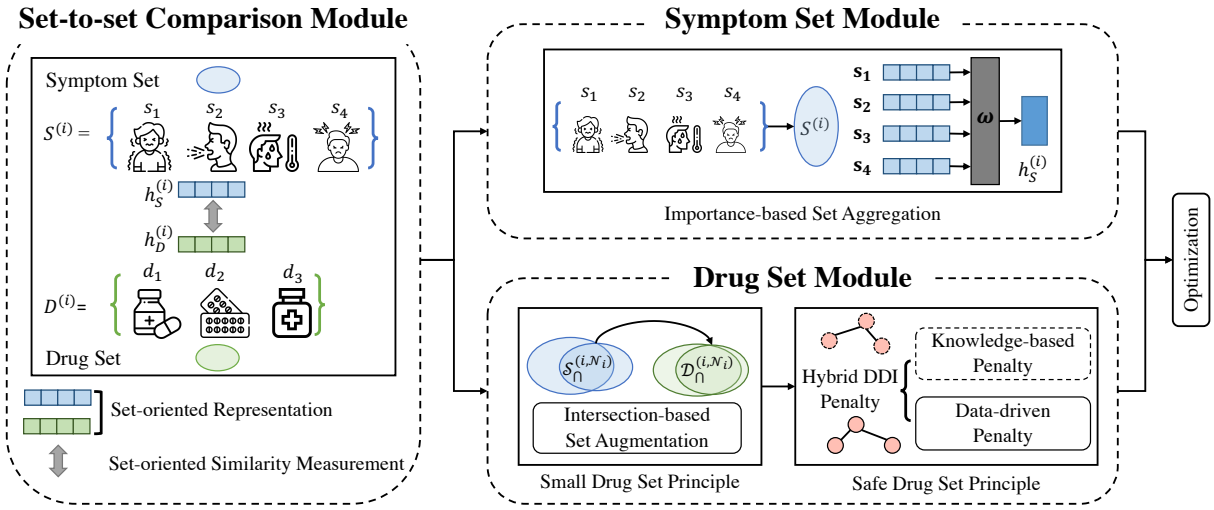


Figure 2: Overview of our proposed Symptom-based Set-to-set Small and Safe Drug Recommendation (4SDrug) framework.

information only from patients' current visit. However, these models still require access to patients' various personal information beyond symptoms, which may lead to severe privacy concerns.

Furthermore, considering drug-drug interactions (DDIs) [20, 38] is necessary in recommending drug sets, since the mixture of drugs may undermine the therapeutic effect and cause side effects. Existing studies model DDIs implicitly either via knowledge graphs (KGs) [10, 40], reinforcement post-processing [51], or acquiring probability distributions of safe drug sets from raw EHR records for adversarial regularization [43]. However, such additional data like KGs are not always applicable. There are also studies that model DDIs explicitly with a controllable loss function [30, 47]. Although these works have considered DDIs, they fail to control the number of recommended drugs, which may increase patient's financial burden and make patient reluctant of taking all of them.

Deep Learning with Sets. Since there are many domains where the data can be treated as unordered sets, recent years have witnessed a growth of interest in utilizing neural networks to learn set representations. Specifically, learning set representations has been widely studied in computer vision (CV) [13, 31, 49], natural language processing (NLP) [17], information retrieval [24], and product bundle recommendation [4]. For example, PointNet [25] applied multi-layer perceptron (MLP) and feature transformations on the elements in the set and used max-pooling to aggregate information. [32] measured the similarity between the input set and each one of the hidden sets by bipartite matching. However, the above set-oriented representation methods do not consider the unique properties of symptoms and drugs, and thus cannot be directly applied to our setting.

3 THE 4SDRUG FRAMEWORK

3.1 Problem Statement and 4SDrug Overview

Our task of symptom-to-drug recommendation aims to generate a drug set as the treatment to a specific symptom set as shown in Figure 1. Let $\mathcal{S} = \{s_1, s_2, \ldots, s_M\}$ and $\mathcal{D} = \{d_1, d_2, \ldots, h_N\}$ denote all symptoms and drugs, respectively. Each query consists of a symptom set and a drug set, e.g., $\mathcal{S}^{(i)} = \{s_1, s_2, \ldots\}$ and $\mathcal{D}^{(i)} = \{d_1, d_2, \ldots\}$. We denote $\mathbf{h}_{\mathcal{S}}^{(i)}$ for i-th symptom set and $\mathbf{h}_{\mathcal{D}}^{(i)}$ for i-th drug set. Given a symptom set $\mathcal{S}^{(i)}$, an N-dimensional probability vector is computed, where the value of dimension k represents the probability that drug k can treat some symptoms inside $\mathcal{S}^{(i)}$. This is achieved by a learned set-oriented similarity measurement $g\{\mathbf{h}_{\mathcal{S}}^{(i)}, \mathbf{d}_j\}$ between the representation of symptom set $\mathbf{h}_{\mathcal{S}}^{(i)}$ and drug $\mathbf{d}_j$, which also represents the probability of recommending drug d_j to treat $\mathcal{S}^{(i)}$. The input and output are defined as follows:

- Input: Symptom sets $\{\mathcal{S}^{(1)}, \mathcal{S}^{(2)}, \ldots, \mathcal{S}^{(Q)}\}$ and the drug sets $\{\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, \ldots, \mathcal{D}^{(Q)}\}$ that treat $\{\mathcal{S}^{(1)}, \mathcal{S}^{(2)}, \ldots, \mathcal{S}^{(Q)}\}$, where Q denote the total number of queries.
- Output: A learned set-oriented similarity measurement function $g\{\mathbf{h}_{\mathcal{S}}^{(i)}, \mathbf{d}_j\}$, which generates the probability vector for all drugs from $\mathcal{D}$ given the symptom set $\mathcal{S}^{(i)}$.

We summarize the main modules of the 4SDrug framework in Figure 2 to provide an overview. 4SDrug takes symptom sets $\mathcal{S}^{(i)}$ and drug sets $\mathcal{D}^{(i)}$ as inputs, and performs symptom-to-drug recommendation with the help of three technical modules. In the *set-to-set comparison module*, we use $\mathbf{h}_{\mathcal{S}}^{(i)}$ and $\mathbf{h}_{\mathcal{D}}^{(i)}$ to represent $\mathcal{S}^{(i)}$ and $\mathcal{D}^{(i)}$ via the proposed set-oriented representation method, and measure the relations between $\mathcal{S}^{(i)}$ and $\mathcal{D}^{(i)}$ via the set-oriented similarity measurement $g\{\cdot, \cdot\}$. In the *symptom set module*, 4SDrug reformulate $\mathbf{h}_{\mathcal{S}}^{(i)}$ via importance-based set aggregation. Finally, in the *drug set module*, we recommend proper sets of drugs by intersection-based set augmentation as well as a hybrid DDI penalty mechanism to ensure the principles of small and safe drug set.

3.2 Set-to-set Comparison Module

Different from general recommendation which aims at modeling the relations between single users on one side and single items on the other, in drug recommendation, we need to consider sets of symptoms and sets of drugs. To model such set-to-set relations, a straightforward way is to represent symptom sets and drug sets by one-hot encoding. This is obviously inappropriate because there are many possible sets which appear for very small numbers of times.

Table 1: The occurrence times of symptom sets on MIMIC-III.

# of set occurrence	1	2	3	4	5	6	7	≥8
# of set	23106	27	10	2	3	1	2	9

As shown in Table 1, over 99% cold-start sets only appear once on MIMIC-III. A model that ignores the intersections of similar sets can easily fail to properly model such cold-start sets. To address this, in this section, we first devise a set-oriented representation method for both symptoms and drugs, and then present set-oriented similarity measurement between symptom sets and drug sets.

3.2.1 Set-oriented Representation. To leverage the information of symptoms and drugs in sets and alleviate the cold-start problem brought by the set-oriented data, we represent a set based on its elements. Different from the data formats like sequences and lists, a set has two main properties: 1) *Permutation invariance:* a set keeps same no matter how its elements are permutated; 2) *Variable cardinality:* the set can have different sizes (shown in Figure 3a). These properties pose challenges in set representation, which invalidates existing methods using sequence models for drug sets [18, 40].

To satisfy these properties of sets, we use average pooling for the representation of symptom sets and drug sets, which is guaranteed to be permutation invariant and can handle variable cardinality. Let $\mathbf{h}_{\mathcal{S}}^{(i)}$ be the embedding for the symptom set $\mathcal{S}^{(i)}$ and $\mathbf{h}_{\mathcal{D}}^{(i)}$ be the embedding for drug set $\mathcal{D}^{(i)}$, we have the representation as:

$$\mathbf{h}_{\mathcal{S}}^{(i)} = \bigoplus_{s_{i'} \in \mathcal{S}^{(i)}} \frac{1}{|\mathcal{S}^{(i)}|} \mathbf{s}_{i'}, \quad \mathbf{h}_{\mathcal{D}}^{(i)} = \bigoplus_{d_{i'} \in \mathcal{D}^{(i)}} \frac{1}{|\mathcal{D}^{(i)}|} \mathbf{d}_{i'}, \tag{1}$$

where both $\mathbf{s}_{i'}$ and $\mathbf{d}_{i'}$ are the learnable embeddings of symptom $s_{i'}$ and drug $d_{i'}$.

3.2.2 Set-oriented Similarity Measurement. To conduct symptom-based set-to-set recommendation, we need to ensure that the representation of the symptom set $\mathcal{S}^{(i)}$ is more similar to the drugs in $\mathcal{D}^{(i)}$ than the drugs in $\mathcal{D} - \mathcal{D}^{(i)}$, which can be formulated as:

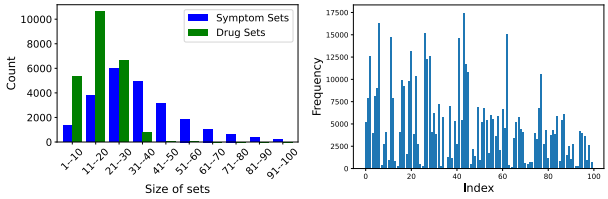
$$sim\left(\mathbf{h}_{\mathcal{S}}^{(i)}, \mathbf{h}_{\mathcal{D}}^{(i)}\right) >_{\mathcal{S}} sim\left(\mathbf{h}_{\mathcal{S}}^{(i)}, \mathbf{h}_{-\mathcal{D}}^{(i)}\right), \tag{2}$$

where $>_{\mathcal{S}}$ between sets denotes that the similarity of the former one is higher than that of the latter one. $\mathbf{h}_{\mathcal{S}}^{(i)}$, $\mathbf{h}_{\mathcal{D}}^{(i)}$, and $\mathbf{h}_{-\mathcal{D}}^{(i)}$ are the representations of the symptom set $\mathcal{S}^{(i)}$, the drug set $\mathcal{D}^{(i)}$ that treats $\mathcal{S}^{(i)}$, and the drugs that are in $\mathcal{D} - \mathcal{D}^{(i)}$, respectively.

Note that, during training, all symptom sets and drug sets are given, but during testing, only symptom sets are given. To make the training phase and testing phase consistent, we should first generate drug sets, which are subsets of the drugs $\mathcal{D}$ with varying sizes. However, it is impossible to traverse 2^N candidate drug sets for such set-oriented drug recommendation.

To capture the latent relations between symptoms and drugs effectively and efficiently, we take one step back and devise a set-oriented similarity measurement g between symptom sets and individual drugs, to optimize the probability of recommending a drug based on a set of symptoms. Note that, modeling all elements individually is a common practice to approximate the modeling of the set as a whole to avoid the combinatorial complexity with approximation guarantee [6, 19], and we will revisit the set properties of drugs in Section 3.4. The measurement is:

$$g\left(\mathbf{h}_{\mathcal{S}}^{(i)}, \mathbf{d}_j\right) = \sigma\left(\mathbf{h}_{\mathcal{S}}^{(i)} \odot \mathbf{d}_j\right), \tag{3}$$



(a) Histogram of size of sets. **(b) The frequency of symptoms.**

Figure 3: Typical data analysis on MIMIC-III.

where σ is a sigmoid function to scale the output to a probability measure between 0 and 1. $\odot$ represents the element-wise product. In this way, by summing up the similarity between $\mathcal{S}^{(i)}$ and each drug d_j in the training phase, we can approximately calculate the similarity between symptom sets and drug sets as:

$$sim\left(\mathbf{h}_{\mathcal{S}}^{(i)}, \mathbf{h}_{\mathcal{D}}^{(i)}\right) = \frac{1}{|\mathcal{D}^{(i)}|} \sum_{j=1}^{|\mathcal{D}^{(i)}|} g\left(\mathbf{h}_{\mathcal{S}}^{(i)}, \mathbf{d}_j\right). \tag{4}$$

Following abundant recent studies on drug set recommendation [30, 47, 51], we formulate the drug set recommendation as a multi-label binary classification task. Towards the recommended drug, we have $g\{\mathbf{h}_{\mathcal{S}}^{(i)}, \mathbf{d}_j\} \to 1$; towards the other drugs, we have $g\{\mathbf{h}_{\mathcal{S}}^{(i)}, \mathbf{d}_j\} \to 0$. Thereby, the objective function of set-to-set drug recommendation is:

$$\mathcal{L}_{rec}^{(i)} = \bigoplus_{d_j \in \mathcal{D}^{(i)}} \log g\left(\mathbf{h}_{\mathcal{S}}^{(i)}, \mathbf{d}_j\right) + \tilde{\bigoplus}_{d_j \in (\mathcal{D} - \mathcal{D}^{(i)})} \log\left(1 - g\left(\mathbf{h}_{\mathcal{S}}^{(i)}, \mathbf{d}_j\right)\right), \tag{5}$$

where $\mathcal{D}^{(i)}$ is the drug set for treating symptom set $\mathcal{S}^{(i)}$.

3.3 Symptom Set Module

Although the pooling strategy in Eq. 1 can be applied to convert a set of symptom embeddings to one unified set representation, it cannot capture the different importance of symptoms, which should be considered carefully when doctors perform diagnosis. For example, as shown in Figure 1, Fever is more important than other symptoms (e.g., Chills) of Lisa since it directly leads to the diagnosis of viral influenza. Moreover, since symptoms have different frequencies (shown in Figure 3b), the averaging strategy that is prone to the ignorance of unusual symptoms will lead to the failure of disease detection. For example, Dark neck skin is a rare symptom of diabetes, which deserves more attention when representing a symptom set so as to recommend diabetes-oriented drugs.

To capture the aforementioned different importance of symptoms in a symptom set, we design an importance-based set aggregation mechanism, which is inspired by the attention mechanism in neural networks [17]. Specifically, we rewrite $\mathbf{h}_{\mathcal{S}}^{(i)}$ in Eq. 1 to explicitly take the importance of symptoms into consideration as follows:

$$\mathbf{h}_{\mathcal{S}}^{(i)} = \bigoplus_{s_{i'} \in \mathcal{S}^{(i)}} \frac{\omega_{i'}}{\sum_{s_z \in \mathcal{S}^{(i)}} \omega_z} \mathbf{s}_{i'}, \tag{6}$$

where $\omega_{i'}$ is a learnable weight of symptom $s_{i'}$ and $\mathbf{s}_{i'}$ denotes the embedding of $s_{i'}$. Furthermore, we add L_1 normalization on the weights to eliminate the impact of varied sizes of different symptom sets (as shown in Figure 3a).

3.4 Drug Set Module

In Section 3.2, for the efficiency of testing phase, we reduce the modeling of drug sets into individual drugs as a multi-label binary classification task. However, this does not take the unique requirements of recommending drug sets into consideration. Moreover, unlike set recommendation in E-commerce, where irrelevant items have little impact on customers and items in the set do not interact, extra drugs can be harmful and drugs can interact. Specifically, recommending more drugs than necessary can increase the patient's financial burden and reluctancy of taking all of them, and more drugs can also lead to more potential side effects and harmful drug-drug interactions (DDIs). Therefore, it is important to produce sets of drugs that are small and safe, besides high accuracy.

In light of this, we propose a small drug set principle and a safe drug set principle, which explicitly stress the effective drugs and model DDIs based on the set-oriented data.

3.4.1 Small Drug Set Principle. The small set principle aims to treat symptoms as much as possible by recommending as small number of drugs as possible. However, directly limiting the recommended number of drugs may sacrifice the treatment towards really severe and complicated diseases. With the only knowledge about symptom sets $\mathcal{S}^{(i)}$ and drug sets $\mathcal{D}^{(i)}$, it is challenging to reduce the size of recommended drug sets while still ensuring the effectiveness.

Inspired by the way how experienced doctors prescribe drugs [28] and the popular voting mechanism [11] that regards the overlaps of decisions as more confident, we propose intersection-based set augmentation by paying attention to the overlapping symptoms and overlapping drugs. For example, as shown in Figure 1, the symptom sets of u_1 and u_3 overlap on Headache while the drug sets of them overlap on Ibuprofen, so Ibuprofen is more likely effective for Headache, and we augment a new positive relation between this pair of symptoms and drugs.

To efficiently augment the data, we aim to find pairs of symptom sets with more overlapping elements. Specifically, given a symptom set $\mathcal{S}^{(i)}$, we first rank all other symptom sets based on their Jaccard coefficients with $\mathcal{S}^{(i)}$, and select the candidate $\mathcal{S}^{(\mathcal{N}_i)}$ with the largest score, which is formulated as:

$$\mathcal{N}_i = \arg\max |\mathcal{S}^{(i)} \cap \mathcal{S}^{(\mathcal{N}_i)}| \;/\; |\mathcal{S}^{(i)} \cup \mathcal{S}^{(\mathcal{N}_i)}|. \tag{7}$$

Then, the intersection of the two symptom sets (i.e., $\mathcal{S}^{(i)}$ and $\mathcal{S}^{(\mathcal{N}_i)}$) and the intersection of the two drug sets (i.e., $\mathcal{D}^{(i)}$ and $\mathcal{D}^{(\mathcal{N}_i)}$) are obtained as follows:

$$\mathcal{S}_{\cap}^{(i,\mathcal{N}_i)} = \mathcal{S}^{(i)} \cap \mathcal{S}^{(\mathcal{N}_i)}, \quad \mathcal{D}_{\cap}^{(i,\mathcal{N}_i)} = \mathcal{D}^{(i)} \cap \mathcal{D}^{(\mathcal{N}_i)}. \tag{8}$$

According to the representation of symptom set in Eq. 6, we obtain the representation for the intersection symptom set $\mathcal{S}_{\cap}^{(i,\mathcal{N}_i)}$:

$$\mathbf{h}_{\mathcal{S}\cap}^{(i,\mathcal{N}_i)} = \sum_{s_{i'} \in \mathcal{S}_{\cap}^{(i,\mathcal{N}_i)}} \frac{\omega_{i'}}{\sum_{s_z \in \mathcal{S}_{\cap}^{(i,\mathcal{N}_i)}} \omega_z} \mathbf{s}_{i'}, \tag{9}$$

Finally, to highlight the effective relations between symptom sets and drug sets, we add a new objective function based on the intersection as follows:

$$\begin{aligned}\mathcal{L}_{inter}^{(i)} = &\sum_{d_j \in \mathcal{D}_{\cap}^{(i,\mathcal{N}_i)}} \log g\left(\mathbf{h}_{\mathcal{S}\cap}^{(i,\mathcal{N}_i)}, \mathbf{d}_j\right) \\ &+ \sum_{d_j \in (\mathcal{D} - \mathcal{D}_{\cap}^{(i,\mathcal{N}_i)})} \log(1 - g\left(\mathbf{h}_{\mathcal{S}\cap}^{(i,\mathcal{N}_i)}, \mathbf{d}_j\right)).\end{aligned} \tag{10}$$

In this case, we encourage the drugs in $\mathcal{D}_{\cap}^{(i,\mathcal{N}_i)}$ to be recommended while decreasing the probability of recommending the drugs in $\mathcal{D} - \mathcal{D}_{\cap}^{(i,\mathcal{N}_i)}$, so as to keep the effective drugs only and reduce the number of recommended drugs.

3.4.2 Safe Drug Set Principle. The safe drug set principle aims to recommend a set of drugs that can avoid the drug-drug interactions (DDIs). Existing works model DDIs via soft or indirect constraints, like knowledge graphs (KGs) [22, 39] and reinforcement post-processing [51]. However, the implicit handling of DDIs results in non-controllable rates in the final recommendation or sub-optimal recommendation accuracy.

To address these limitations, the existing studies leverage the drug knowledge base (DKB) [37] to explicitly model DDIs, which is not always applicable. Specifically, they are only applicable to datasets whose drug codes can be converted to ATC Third Level [30] (e.g., MIMIC-III). For other datasets, such as the industrial one of NELL, such external DKB cannot be leveraged since the drugs are encoded with American Hospital Formulary System (AHFS) drug encoding [8, 23] and cannot be converted to ATC Third Level.

To prevent DDIs on all kinds of datasets, we introduce a hybrid penalty mechanism, which includes: 1) the knowledge-based penalty, and 2) the data-driven penalty, where the data-driven one provide side signals as completments of the ground-truth DDIs.

Firstly, on datasets whose drug encodings can be converted to ATC Third Level, we are able to leverage the ground-truth DDIs from the DKB. Specifically, we design a knowledge-based penalty $\mathcal{L}_{K-DDI}^{(i)}$ on the predicted similarity based on external DDI adjacency matrix $\boldsymbol{A}^d$, which is computed from TWOSIDES dataset [37] via the ATC Third Level drug codes. For the representations of two durgs $\mathbf{d}_k$ and $\mathbf{d}_l$, if the combination of drug d_k and drug d_l induce a DDI, then $\boldsymbol{A}_{kl}^d = 1$. Intuitively, we want the probability of recommending a pair of drugs to be penalized if the two drugs induce DDI, which we enforce with the following objective:

$$\mathcal{L}_{K-DDI}^{(i)} = \sum_{d_k \in \mathcal{D}} \sum_{d_l \in \mathcal{D}} (\boldsymbol{A}_{kl}^d \cdot (g\left(\mathbf{h}_{\mathcal{S}}^{(i)}, \mathbf{d}_k\right) \cdot g\left(\mathbf{h}_{\mathcal{S}}^{(i)}, \mathbf{d}_l\right))), \tag{11}$$

where $\cdot$ is the product between scalars. $g\{\mathbf{h}_{\mathcal{S}}^{(i)}, \mathbf{d}_k\} \cdot g\{\mathbf{h}_{\mathcal{S}}^{(i)}, \mathbf{d}_l\}$ denotes a pair-wise probability of recommending d_k and d_k together.

Secondly, on top of the knowledge-based penalty, and in cases where the knowledge-based penalty cannot be applied, we design a data-driven penalty as side signals for the safe drug set principle.

Since we stress the relations between the intersections of symptom sets and drug sets in Eq. 10, the drugs in the difference set of two similar drug sets are seldom used together and may have DDIs. For example, {Compound Liquorice Tablets, Ibuprofen} and {Azithromycin, Ibuprofen} can be used to treat cough with headache. However, taking Compound Liquorice Tablets with Azithromycin may lead to cardiac arrhythmia. Compound Liquorice Tablets and Azithromycin do not appear in the ground-truth DDI table but show up in the difference set of two drug sets for similar symptoms for more than 150K times in the dataset. Based on this intuition, we propose to punish the relations between $\mathcal{S}_{\cap}^{(i,\mathcal{N}_i)}$ and the difference sets of the two drug sets, where the relative complement of $\mathcal{D}^{(i)}$ in $\mathcal{D}^{(\mathcal{N}_i)}$ and that of $\mathcal{D}^{(\mathcal{N}_i)}$ in $\mathcal{D}^{(i)}$ are calculated as:

$$\mathcal{D}_{-}^{(i)} = \mathcal{D}^{(i)} - \mathcal{D}_{\cap}^{(i,\mathcal{N}_i)}, \quad \mathcal{D}_{-}^{(\mathcal{N}_i)} = \mathcal{D}^{(\mathcal{N}_i)} - \mathcal{D}_{\cap}^{(i,\mathcal{N}_i)}. \tag{12}$$

Table 2: Statistics of the datasets used in our experiments.

Items	MIMIC-III	NELL
# of visits	27,869	278,388
# symptom	1,113	17,898
# drugs	131	230
avg # of symptoms per symptom set	31.81	11.02
avg # of drugs per drug set	14.36	7.62
total # of DDI pairs	448	-

With such infrequently recommended pairs, we enforce an additional data-driven penalty objective by yielding large $\mathcal{L}_{D-DDI}^{(i)}$ if d_k appear in $\mathcal{D}_{-}^{(i)}$ and d_l appear in $\mathcal{D}_{-}^{((N_i))}$, respectively:

$$\mathcal{L}_{D-DDI}^{(i)} = \sum_{d_k \in \mathcal{D}_{-}^{(i)}} \sum_{d_l \in \mathcal{D}_{-}^{(N_i)}} g\left(\mathbf{h}_{\mathcal{S}\cap}^{(i,\mathcal{N}_i)}, \mathbf{d}_k\right) \cdot g\left(\mathbf{h}_{\mathcal{S}\cap}^{(i,\mathcal{N}_i)}, \mathbf{d}_l\right). \tag{13}$$

Finally, we apply weighted sum strategy over the loss for training the final proposed objective as follows:

$$\mathcal{L} = \mathcal{L}_{rec} + \alpha \mathcal{L}_{inter} + \beta(\mathcal{L}_{K-DDI} + \mathcal{L}_{D-DDI}), \tag{14}$$

where α controls the weight of the loss $\mathcal{L}_{inter}$ from intersection-based set augmentation and β controls the weight of the hybrid DDI loss, respectively.

4 INSIGHTS AND DISCUSSIONS

Our development and experiments of 4SDrug have led to several unique insights. (1) Modeling sets as sets instead of sequences is important: in this way, we can get rid of the irrelevant signals from the orders of the elements by ensuring permutation-invariant set representation; (2) differentiating the importance of elements in a set is important: by discovering and highlighting rare but significant elements, the overall representations of sets can be largely improved; (3) combining external knowledge and data-driven knowledge is important: they compliment each other and allow us to achieve drug sets with more controllable sizes and less harmful drug-drug interactions.

This work intends to study and develop a comprehensive and automatic drug recommendation pipeline, to help doctors quickly locate possible drugs and avoid harmful drug-drug interactions. The set-up of 4SDrug is designed so that it can be easily implemented into a deployable system for the real-world application of drug recommendation: 4SDrug focuses on the input of general symptoms without other patient-specific information, allowing the model to be efficient for large-scale real-time drug recommendation, and at the same time, secure from compromising patients' privacy. Note that, the goal of 4SDrug is to provide complementary assistance to doctors for efficient and safe drug prescription, and its recommendations should be further considered by the doctors given the more complicated and private information of individual patients. It is not suggested to be directly used by individuals without necessary medical expertise at its current stage.

5 EXPERIMENT

In this section, we evaluate our proposed 4SDrug framework focusing on the following four research questions:

- **RQ1:** How does 4SDrug perform in comparison to state-of-the-art recommendation methods?
- **RQ2:** What are the effects of different model components?
- **RQ3:** How do the hyperparameters affect the recommendation performance and how to choose optimal values?
- **RQ4:** What real drugs are recommended by 4SDrug and how are they accurate, small, and safe?

5.1 Experimental Setup

5.1.1 Datasets and Evaluation Protocols. We use two real-world EHR datasets to verify the effectiveness of compared methods, i.e., the public benchmark one of MIMIC-III [15] and the industrial large-scale one of NELL. NELL is provided by the Nell Hodgson Woodruff School of Nursing at Emory University. Both datasets are fully anonymized and carefully sanitized before our access. The statistics are summarized in Table 2.

The symptoms here are extracted differently on two datasets. Specifically, the "query" field in NELL directly describes the symptoms of patients while MIMIC-III does not have this. Therefore, we adopt the symptom extraction from the clinical texts following [50] in MIMIC-III. Recall that the DKB from the TWOSIDES dataset [37] is applicable to datasets whose drug codes can be converted to ATC Third Level [30]. We apply the knowledge-based penalty for DDI on MIMIC-III, and skip it on NELL, whose drugs are encoded with AHFS [8, 23]. This actually emphasizes the value of our novel data-driven penalty for DDIs. Following [30, 47], we use two standard effectiveness metrics (i.e., Jaccard coefficient and F1 score) and two specific drug-set metrics (i.e., Avg # of drug and DDI Rate) to evaluate the results of drug recommendation. The detailed evaluation protocols can be found in Appendix. A.

5.1.2 Baselines. We compare 4SDrug with the following baselines from two perspectives: 1) traditional set-oriented models: K-freq [40], K-near [30], ECC [47], and MLP [40]; 2) existing drug recommendation methods: LEAP [51], RETAIN [9], GAMENet [30], and SafeDrug [47]. Following the recent works in drug recommendation [30, 47], we do not compare with general recommendation methods because they cannot provide drug set recommendations. The details of the compared baselines can be referred in Appendix B.

5.1.3 Implementation Details. The full code for our 4SDrug is available[5]. Implementations of the compared baselines are from GAMENet[6] and SafeDrug[7]. We follow the same setting as [30, 47] and split the dataset into training, validation, and testing with a ratio of 4:1:1. We tune all hyperparameters on the validation set through grid search, in particular, α in α in $\{0, 0.25, 0.50, 0.75, 1.00\}$, β in $\{0, 0.25, 0.50, 0.75, 1.00, 1.25\}$. We use 64 as the embedding size for all compared methods on both MIMIC-III and NELL. The batch size is set to 50. We also carefully tune the hyperparameters of baselines on the validation set as suggested in the original papers to achieve their best performance.

5.2 Overall Performance Comparison (RQ1)

We compare the recommendation results of the proposed 4SDrug framework to those of the baseline models. Table 3 shows the Jaccard, F1, Avg # of drug, and DDI Rate on MIMIC-III. Since the

[5] https://github.com/Melinda315/4SDrug
[6] https://github.com/sjy1203/GAMENet
[7] https://github.com/ycq091044/SafeDrug

Table 3: Experimental results on MIMIC-III. Ground-truth Avg # Drug is 14.3600. Ground-truth DDI Rate is 0.0850.

Method	Jaccard	F1	Avg # of Drug	DDI Rate	|Δ| Avg # of Drug	Δ% DDI Rate
K-freq	0.4048 ± 0.0011	0.5681 ± 0.0013	18.7622 ± 0.0584	0.0659 ± 0.0003	4.4022	-19.63%
K-near	0.4041 ± 0.0030	0.5593 ± 0.0030	19.2018 ± 0.1569	0.0815 ± 0.0007	4.8418	-0.61%
ECC	0.4499 ± 0.0030	0.5977 ± 0.0030	17.9707 ± 0.1125	0.0808 ± 0.0008	3.6107	-1.46%
MLP	0.4788 ± 0.0011	0.6317 ± 0.0011	18.0724 ± 0.0853	0.0821 ± 0.0005	3.7124	+0.12%
LEAP	0.4677 ± 0.0011	0.6081 ± 0.0013	18.5374 ± 0.0646	0.0645 ± 0.0001	4.1774	-21.34%
RETAIN	0.4717 ± 0.0024	0.6290 ± 0.0023	18.9957 ± 0.0391	0.0817 ± 0.0003	4.6357	-0.36%
GAMENet	0.4848 ± 0.0022	0.6393 ± 0.0021	26.3139 ± 0.0668	0.0975 ± 0.0003	11.9539	+18.90%
SafeDrug	0.4894 ± 0.0020	0.6454 ± 0.0018	19.7909 ± 0.0531	0.0649 ± 0.0002	5.4309	-20.85%
4SDrug	0.5041 ± 0.0016	0.6581 ± 0.0016	17.5040 ± 0.0533	0.0600 ± 0.0004	3.1440	-26.83%

Table 4: Ablation analysis of our proposed 4SDrug on MIMIC-III.

Submodels	Jaccard	F1	Avg # of Drug	DDI Rate	Δ Avg # of Drug	Δ DDI Rate
MLP with one-hot set encodings	0.4788	0.6317	18.0724	0.0821	3.7124	+0.12%
+ set-to-set comparison module (1SDrug)	0.4873	0.6427	18.1452	0.0827	3.7852	+0.85%
+ symptom set module (2SDrug)	0.5071	0.6608	19.7129	0.0787	5.3529	-4.02%
+ small drug set principle (3SDrug)	0.5078	0.6614	18.3079	0.0770	3.9479	-6.10%
+ safe drug set principle (4SDrug)	0.5041	0.6581	17.5040	0.0600	3.1440	-26.83%

Table 5: Experimental results on NELL. Ground-truth Avg # Drug is 7.6200.

Method	Jaccard	F1	Avg # of Drug
K-freq	0.1495 ± 0.0009	0.2435 ± 0.0014	9.1854 ± 0.0284
K-near	0.1423 ± 0.0017	0.2362 ± 0.0020	9.2026 ± 0.0894
ECC	0.1985 ± 0.0024	0.2770 ± 0.0013	9.5066 ± 0.1005
MLP	0.2371 ± 0.0015	0.3040 ± 0.0012	9.6521 ± 0.0498
LEAP	0.2359 ± 0.0010	0.2980 ± 0.0011	9.6191 ± 0.0250
RETAIN	0.2441 ± 0.0013	0.3098 ± 0.0023	9.8006 ± 0.0198
4SDrug	0.2618 ± 0.0015	0.3485 ± 0.0016	9.1380 ± 0.0278

DKB is not applicable to NELL, we only verify the performance on NELL with the first three metrics except DDI Rate in Table 5.

In general, 4SDrug outperforms all baselines across all evaluation metrics on both datasets. This answers RQ1, showing that our proposed symptom-based set-to-set recommendation framework is capable of effective drug set recommendation. Note that, limited by the deep reliance on DDI knowledge, some baselines (e.g., GAMENet and SafeDrug) are not available on NELL, and the second best performance scattered among different models like SafeDrug and RETAIN. Compared with the second best performance, the performance gains of 4SDrug in terms of Jaccard and F1 range from reasonably large (1.79% achieved with F1 on the MIMIC-III dataset) to significantly large (12.49% achieved with F1 on the NELL dataset).

Moreover, the proposed 4SDrug method can achieve best Jaccard and F1 scores with the smallest number of recommended drugs on both MIMIC-III and NELL, which is particularly evident for its effectiveness towards the small drug set principle. Note that, real drug records usually contain high DDI Rate, e.g., around 0.0850 on MIMIC-III. Although GAMENet already consider DDI based on DKG, it does not consider the number of recommended drug and outputs undesirable DDI Rate, which is consistent with the results in the recent work [47]. The proposed 4SDrug framework can achieve the lowest value on DDI Rate (i.e., 0.0600 on MIMIC-III).

Table 6: Model Complexity Comparison.

Dataset	Model	# of Param.	Training(s)	Testing(s)
MIMIC-III	LEAP	252,675	243.79	18.65
	RETAIN	255,947	100.23	15.64
	GAMENet	365,258	150.47	18.03
	SafeDrug	285,612	134.68	18.49
	4SDrug	79,681	18.00	14.93
NELL	LEAP	252,675	859.65	69.12
	RETAIN	255,947	438.76	47.97
	4SDrug	79,681	17.43	45.26

Table 6 shows the runtimes of 4SDrug and four baselines for drug recommendation, where 4SDrug is more efficient with lower space and time complexity than others. Specifically, our set-to-set recommendation does not involve complex neural architectures and is trivially compatible with efficient mini-batch training. Among the four compared methods, LEAP adopts sequential modeling and recommend drugs one by one, and thus is the most time-consuming. GAMENet stores a large memory bank, and thus requires the largest space. By comparison, we conclude that 4SDrug is efficient and flexible, friendly towards real industrial deployment.

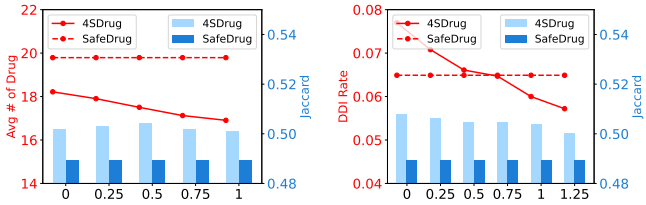
5.3 Model Ablation Study (RQ2)

To better understand 4 "S" in 4SDrug, we closely study our framework by adding the components one by one, i.e., set-to-to comparison module (1SDrug), symptom sets module (2SDrug), small drug set principle (3SDrug), and safe drug set principle (4SDrug). From Table 4, we have the following observations:

- 1SDrug outperforms MLP not only on the effectiveness metrics by achieving 3.24% improvement in Jaccard and 2.18% improvement in F1, but also on the drug-set metrics by achieving a 1.15 decrease in drug number and a 0.97% DDI Rate reduction. Such results are consistent with those in Table 3 and Table 5, showing the effectiveness of mining complex relations among sets.

Table 7: Example recommended drug set for a given symptom set on MIMIC-III. Here "FN" refers to the drugs that are in the ground-truth drug sets but are not predicted, while "FP" indicates the drugs predicted but are not in ground-truth drug sets. The drug pairs denoted by the same non-black colors have harmful DDIs. Best viewed in color.

Method	Recommended Drug Set
Ground-Truth Num:7	7 **TP** (Antithrombotic Agents; Other Mineral Supplements; Anesthetics, General; Irrigating Solutions; Stomatological Preparations; Vitamin B1, Plain and in Combination with Vitamin B6 and B12; Vitamin B12 and Folic Acid)
GAMENet Num: 19 Recall=6/7 Precision=6/19	6 **TP** (Antithrombotic Agents; Other Mineral Supplements; Anesthetics, General; Irrigating Solutions; Stomatological Preparations; Vitamin B1, Plain and in Combination with Vitamin B6 and B12) 1 **FN** (Vitamin B12 and Folic Acid) 13 **FP** (Antiepileptics; Potassium; Other Abesics and Antipyretics; Drugs for Constipation; Drugs for Peptic Ulcer and Gastro-Oesophageal Reflux Disease (GORD); Beta Blocking Agents; Adrenergics, Inhalants; Calcium; Opioids; Quinolone Antibacterials; Arteriolar Smooth Muscle, Agents Acting on; Anxiolytics; Hypnotics and Sedatives)
SafeDrug Num: 5 Recall=3/7 Precision=3/5	3 **TP** (Antithrombotic Agents; Other Mineral Supplements; Vitamin B12 and Folic Acid) 4 **FN** (Anesthetics, General; Irrigating Solutions; Stomatological Preparation; Vitamin B1, Plain and in Combination with Vitamin B6 and B12) 2 **FP** (Drugs for Peptic Ulcer and Gastro-Oesophageal Reflux Disease (GORD); Calcium)
4SDrug Num: 8 Recall=6/7 Precision=3/4	6 **TP** (Antithrombotic Agents; Other Mineral Supplements; Anesthetics, General; Irrigating Solutions; Stomatological Preparations; Vitamin B1, Plain and in Combination with Vitamin B6 and B12) 1 **FN** (Vitamin B12 and Folic Acid) 2 **FP** (Other Analgesics and Antipyretics; Drugs for Constipation)



(a) Varying α **(b) Varying β**

Figure 4: Performance for varying the weight of $\mathcal{L}_{inter}$ (i.e., α) and the weight of $(\mathcal{L}_{K-DDI} + \mathcal{L}_{D-DDI})$ (i.e., β) on MIMIC-III.

- Compared with 1SDrug, 2SDrug achieves performance gains on the effectiveness metrics by achieving 2.59% improvement in Jaccard and 2.37% improvement in F1. In the drug-set metrics, 2SDrug bring about 3.2% DDI Rate reduction and 1.57 increase in drug number, since 2SDrug does not consider the unique requirements of drug sets.
- Moreover, the performance gain of 3SDrug over 2SDrug includes both effectiveness metrics and drug-set metrics, where 3SDrug can achieve 1.41 decrease in drug number and 2.16% DDI Rate reduction based on 2SDrug. These results show the effectiveness of applying our novel intersection-based set augmentation.
- 4SDrug can achieve the smallest number of drugs and the lowest DDI Rate with significant 20.73% improvement over the DDI of 3SDrug. Although 4SDrug brings a slight decrease on the effectiveness metrics, it can finally achieve a satisfactory trade-off among the multiple objectives of drug recommendation.

5.4 Major Hyperparameter Study (RQ3)

Our proposed 4SDrug framework mainly introduces two hyperparameters, i.e., α, and β, which control the weight of the loss $\mathcal{L}_{inter}$ from intersection-based set augmentation and the hybrid DDI loss, respectively. Here we show how these two hyperparameters impact the performance and clarify how to set them.

Firstly, we show the model performance with varying $\mathcal{L}_{inter}$. The loss $\mathcal{L}_{inter}$ can control the number of drug. If α is too small, the interactions between the intersection symptom set and drugs will likely be weakened. However, too large α will likely cause the model to overfit. The results are shown in Figure 4a. We found that the optimal α values on MIMIC-III to be about 0.5. Note that, when $\alpha \in [0, 1]$, 4SDrug is always better than the best baseline. In the range of [0,1], the optimal α can be obtained by slight tuning.

Secondly, for hyperparameter β, the optimal β on MIMIC-III is 1.0, as shown in Figure 4b. In particular, we observe the effectiveness of $\mathcal{L}_{K-DDI}$ and $\mathcal{L}_{D-DDI}$ as increasing β always leads to the reduction of drug number. However, further increasing it beyond the optimal value makes the accuracy performance worse. In practice, β = 1.0 seems to be the rule-of-thumb.

5.5 Case Studies (RQ4)

To demonstrate the advantages of 4SDrug over the two drug recommendation baselines methods, we demonstrate the recommended drugs learned by GAMENet, SafeDrug, and the proposed 4SDrug on MIMIC-III. Some example results are presented in Table 7 and more results can be found in Table C.1 in Appendix C. According to the listed metrics, we have the following observations.

In general, 4SDrug achieves the highest value on Recall and Precision under the given symptom set {Sputum, Ulcer, Cool, Cough, Bleed, ...}. The only one **FN** drug of 4SDrug is Vitamin B12 and Folic Acid, and by checking DKB, we find the reason of not recommending this drug as to avoid the harmful DDIs between Vitamin B12 and Folic Acid and Stomatological Preparations. Note that, compared with GAMENet that recommend too many drugs with many DDIs, 4SDrug can automatically avoid DDIs, which is consistent with its overall advantageous performance in Table 3.

Since different doctors might give different drug sets for a certain symptom set [14], and thus no gold-standard ground truth exists. In this case, 4SDrug, which provides small and safe drug sets learned from various doctors actual prescriptions, may even serve as a better option than the provided ground-truth drug sets.

6 CONCLUSION

In this paper, we propose a symptom set-based drug recommendation framework, towards the prescription assistance for doctors and privacy protection for patients. Specifically, we propose a novel framework of Symptom-based Set-to-set Small and Safe drug recommendation (4SDrug), including a set-to-set comparison module, a symptom set module, and a drug set module. Extensive quantitative experiments demonstrate the clear advantages of our 4SDrug over the state-of-the-art baselines towards the recommendation of accurate, small and safe drug sets, which is further consolidated with our real case study results.

ACKNOWLEDGMENTS

Xiaolin Zheng was supported by the National Natural Science Foundation of China (No.62172362 and No.72192823).

REFERENCES

[1] A. Avati, K. Jung, S. Harman, L. Downing, A. Ng, and N. H. Shah. Improving palliative care with deep learning. *BMC medical informatics and decision making*, 18(4):55–64, 2018.
[2] J. M. Bajor and T. A. Lasko. Predicting medications from diagnostic codes with recurrent neural networks. In *ICLR*, 2017.
[3] Y. Bao and X. Jiang. An intelligent medicine recommender system framework. In *ICIEA*, pages 1383–1388, 2016.
[4] L. Chen, Y. Liu, X. He, L. Gao, and Z. Zheng. Matching user with item set: Collaborative bundle recommendation with deep attention network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 2019.
[5] M. Chen and H. Wang. The reason and prevention of hospital medication errors. *Practical Journal of Clinical Medicine*, 4, 2013.
[6] U. Chitra and B. Raphael. Random walks on hypergraphs with edge-dependent vertex weights. In *International Conference on Machine Learning*, 2019.
[7] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318, 2016.
[8] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun. Multi-layer representation learning for medical concepts. In *proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.
[9] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *NIPS*, 29, 2016.
[10] F. Gong, M. Wang, H. Wang, S. Wang, and M. Liu. Smr: Medical knowledge graph embedding for safe medicine recommendation. *Big Data Research*, 23:100174, 2021.
[11] L. Guo, H. Yin, Q. Wang, B. Cui, Z. Huang, and L. Cui. Group recommendation with latent voting mechanism. In *ICDE*, pages 121–132, 2020.
[12] C.-K. Hsieh, L. Yang, Y. Cui, T.-Y. Lin, S. Belongie, and D. Estrin. Collaborative metric learning. In *WWW*, pages 193–201, 2017.
[13] M. Jia, X. Cheng, Y. Zhai, S. Lu, S. Ma, Y. Tian, and J. Zhang. Matching on sets: Conquer occluded person re-identification without alignment. In *AAAI*, pages 1673–1681, 2021.
[14] Y. Jin, W. Zhang, X. He, X. Wang, and X. Wang. Syndrome-aware herb recommendation with multi-graph convolution network. In *ICDE*, pages 145–156, 2020.
[15] A. E. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, R. G. Mark, and et al. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1), 2016.
[16] H.-C. Kao, K.-F. Tang, and E. Chang. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In *AAAI*, 2018.
[17] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*, pages 3744–3753, 2019.
[18] C. Li, B. Wang, V. Pavlu, and J. Aslam. Conditional bernoulli mixtures for multi-label classification. In *ICML*, pages 2482–2491, 2016.
[19] P. Li and O. Milenkovic. Inhomogoenous hypergraph clustering with applications. In *Advances in Neural Information Processing Systems*, 2017.
[20] X. Lin, Z. Quan, Z.-J. Wang, T. Ma, and X. Zeng. Kgnn: Knowledge graph neural network for drug-drug interaction prediction. In *IJCAI*, pages 2739–2745, 2020.
[21] Y. Liu, X. Xia, L. Chen, X. He, C. Yang, and Z. Zheng. Certifiable robustness to discrete adversarial perturbations for factorization machines. In *SIGIR*, pages 419–428, 2020.
[22] C. Mao, L. Yao, and Y. Luo. Medgcn: Graph convolutional networks for multiple medical tasks. *arXiv preprint arXiv:1904.00326*, 2019.
[23] G. K. McEvoy. Ahfs drug information. *Oncology Issues*, 9(5):12–13, 1994.
[24] L. Pang, J. Xu, Q. Ai, Y. Lan, X. Cheng, and J. Wen. Setrank: Learning a permutation-invariant ranking model for information retrieval. In *SIGIR*, 2020.
[25] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017.
[26] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13, 2021.
[27] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *UAI*, pages 452—-461, 2009.
[28] C. L. A.-P. Rutter and D. Newby. *Community Pharmacy-E-Book, livre ebook*. 2011.
[29] J. Shang, T. Ma, C. Xiao, and J. Sun. Pre-training of graph augmented transformers for medication recommendation. In *IJCAI*, pages 5953–5959, 2019.
[30] J. Shang, C. Xiao, T. Ma, H. Li, and J. Sun. Gamenet: Graph augmented memory networks for recommending medication combination. In *AAAI*, pages 1126–1133, 2019.
[31] Y. Shi, J. Oliva, and M. Niethammer. Deep message passing on sets. In *AAAI*, pages 5750–5757, 2020.
[32] K. Skianis, G. Nikolentzos, S. Limnios, and M. Vazirgiannis. Rep the set: Neural networks for learning set representations. In *AISTATS*, pages 1410–1420, 2020.
[33] Y. Tan, C. Yang, X. Wei, C. Chen, L. Li, and X. Zheng. Enhancing recommendation with automated tag taxonomy construction in hyperbolic space. In *ICDE*, 2022.
[34] Y. Tan, C. Yang, X. Wei, C. Chen, W. Liu, L. Li, J. Zhou, and X. Zheng. Metacare++: Meta-learning with hierarchical subtyping for cold-start diagnosis prediction in healthcare data. In *SIGIR*, 2022.
[35] Y. Tan, C. Yang, X. Wei, Y. Ma, and X. Zheng. Multi-facet recommender networks with spherical optimization. In *ICDE*, pages 1524–1535, 2021.
[36] K.-F. Tang, H.-C. Kao, C.-N. Chou, and E. Y. Chang. Inquire and diagnose: Neural symptom checking ensemble using deep reinforcement learning. In *NIPS Workshop on Deep Reinforcement Learning*, 2016.
[37] N. P. Tatonetti, P. P. Ye, R. Daneshjou, and R. B. Altman. Data-driven prediction of drug effects and interactions. *Science translational medicine*, 4(125):125ra31–125ra31, 2012.
[38] 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*, 9(9):2147–2163, 2014.
[39] M. Wang, M. Liu, J. Liu, S. Wang, G. Long, and B. Qian. Safe medicine recommendation via medical knowledge graph embedding. *ArXiv e-prints*, pages arXiv–1710, 2017.
[40] S. Wang, P. Ren, Z. Chen, Z. Ren, J. Ma, and M. de Rijke. Order-free medicine combination prediction with graph convolutional reinforcement learning. In *CIKM*, pages 1623–1632, 2019.
[41] X. Wang, T. Huang, D. Wang, Y. Yuan, Z. Liu, X. He, and T.-S. Chua. Learning intents behind interactions with knowledge graph for recommendation. In *WWW*, pages 878–887, 2021.
[42] X. Wang, H. Jin, A. Zhang, X. He, T. Xu, and T.-S. Chua. Disentangled graph collaborative filtering. In *SIGIR*, pages 1001–1010, 2020.
[43] Y. Wang, W. Chen, D. PI, L. Yue, S. Wang, and M. Xu. Self-supervised adversarial distribution regularization for medication recommendation. In *IJCAI*, pages 3134–3140, 2021.
[44] Y. Xie, Z. Wang, C. Yang, Y. Li, B. Ding, H. Deng, and J. Han. Komen: Domain knowledge guided interaction recommendation for emerging scenarios. In *WWW*, pages 1301–1310, 2022.
[45] C. Yang, L. Bai, C. Zhang, Q. Yuan, and J. Han. Bridging collaborative filtering and semi-supervised learning: a neural approach for poi recommendation. In *KDD*, pages 1245–1254, 2017.
[46] C. Yang, A. Pal, A. Zhai, N. Pancha, J. Han, C. Rosenberg, and J. Leskovec. Multisage: Empowering gcn with contextualized multi-embeddings on web-scale multipartite networks. In *KDD*, pages 2434–2443, 2020.
[47] C. Yang, C. Xiao, F. Ma, L. Glass, and J. Sun. Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. In *IJCAI*, pages 3735–3741, 2021.
[48] P. Zadeh, R. Hosseini, and S. Sra. Geometric mean metric learning. In *ICML*, pages 2464–2471, 2016.
[49] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola. Deep sets. In *NIPS*, volume 30, 2017.
[50] X. Zeng, G. Yu, Y. Lu, L. Tan, X. Wu, S. Shi, H. Duan, Q. Shu, and H. Li. Pic, a paediatric-specific intensive care database. *Scientific data*, 7(1):1–8, 2020.
[51] Y. Zhang, R. Chen, J. Tang, W. F. Stewart, and J. Sun. Leap: learning to prescribe effective and safe treatment combinations for multimorbidity. In *KDD*, pages 1315–1324, 2017.
[52] Z. Zheng, C. Wang, T. Xu, D. Shen, P. Qin, B. Huai, T. Liu, and E. Chen. Drug package recommendation via interaction-aware graph induction. In *WWW*, pages 1284–1295, 2021.

APPENDIX

A EVALUATION PROTOCOLS DETAILS

For a particular symptom set $S^{(i)}$ and the recommended drug set is $D^{(i)}$. $\hat{D}^{(i)}$ is the ground truth drug set prescribed by doctors. The mean jaccard coefficient is defined as the size of the intersection divided by the size of the union of predicted drugs and ground truth drugs. Recall measures the completeness of predicted drugs and Precision measures the correctness of predicted drugs. F1 score is the harmonic mean of Precision and Recall, and is often used as a comprehensive evaluation metric of prediction models:

$$\text{Jaccard} = \frac{1}{Q}\sum_i \frac{D^{(i)} \cap \hat{D}^{(i)}}{D^{(i)} \cup \hat{D}^{(i)}}, \quad \text{Recall} = \frac{1}{Q}\sum_i \frac{D^{(i)} \cap \hat{D}^{(i)}}{\hat{D}^{(i)}}$$

$$\text{Precision} = \frac{1}{Q}\sum_i \frac{D^{(i)} \cap \hat{D}^{(i)}}{D^{(i)}}$$

$$F1 = \frac{1}{Q}\sum_i \frac{2 * \text{Precision}_i * \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

where i is a query index in the test set.

To measure drug safety, we define DDI Rate as the percentage of drug recommendation that contain DDIs.

$$\text{DDI Rate} = \frac{1}{Q}\sum_i \mathbb{1}\left[(d_a, d_b) \in D^{(i)} \;\&\; (d_a, d_b) \in \mathcal{E}_{ddi}\right]$$

where the set will count each drug pair (d_a, d_b) in recommendation set $D^{(i)}$ if the pair belongs to drug knowledge base (DKB) in $\mathcal{E}_{ddi}$.

B BASELINE DETAILS

- K-freq [40]. K-frequent predicts drugs by counting the the top K most frequently occurring drugs of each symptom. We tried K from 1 to 8 and finally set K to 5 according to the validation.
- K-near [30]. To prescribe drugs for a symptom set S_i, K-nearest selects the drugs prescribed for patient S_j that has the most similar symptom set embedding with S_i. Similarity between two sets is measured by Jaccard measurement. Here, we set K to 1.
- Ensemble Classifier Chain (ECC) [47]. Classifier chain (CC) is a popular multi-label classification approach, which feeds previous classification results into the latter classifiers. We implement a 10-member ensemble of CCs also by scikit-learn, where each CC consists of a dependent series of logistic regression classifiers.
- Multi-layer Perceptron (MLP) [40]. MLPs are conventional methods to solve multi-label classification problem, where we use a three-layer perceptron and sigmoid as activation function to predict the probability of each drug.
- LEAP [51] treats drug recommendation as a sentence generation task and recommend drugs one at a time.
- RETAIN [9] is a temporal based method. It utilizes a two-level RNN with reverse time attention to model the symptom information.
- GAMENet [30]. GAMENet adopts memory augmented neural networks and stores historical drug memory records as references for future prediction.
- SafeDrug [47]. SafeDrug propose dual molecular encoders to capture global and local molecule patterns and explicitly design DDI controllable loss function.

C FULL CASE STUDIES

Given a symptom set { Poor mental status, Mental Status Change, Secretions, sputum, ulcer,Cool, Productive Cough, Cough, Apnea, Bleed}, the recommended drugs learned by the compared methods are listed in Table C.1.

Table C.1: Example recommended drug set for a given symptom set on MIMIC-III. Here "FN" refers to the drugs that are in the ground-truth drug sets but are not predicted, while "FP" indicates the drugs predicted but are not in ground-truth drug sets. The drug pairs denoted by the same non-black colors have harmful DDIs. Best viewed in color.

Method	Recommended Drug Set
Ground Truth	7 **TP** (Antithrombotic Agents; Other Mineral Supplements; Anesthetics, General; Irrigating Solutions; Stomatological Preparations; Vitamin B1, Plain and in Combination with Vitamin B6 and B12; Vitamin B12 and Folic Acid)
K-frequent Num: 21 Recall=4/7 Precision=4/21	4 **TP** (Antithrombotic Agents; Other Mineral Supplements; Irrigating Solutions; Stomatological Preparations) 3 **FN** (Anesthetics, General; Vitamin B1, Plain and in Combination with Vitamin B6 and B12; Vitamin B12 and Folic Acid) 17 **FP** (Antimalarials; Potassium; Other Analgesics and Antipyretics; Drugs for Constipation; Drugs for Peptic Ulcer and Gastro-Oesophageal Reflux Disease (GORD); Intestinal Antiinfectives; Direct Acting Antivirals; Immunosuppressants; Beta Blocking Agents; Adrenergics, Inhalants; High-Ceiling Diuretics; Calcium; Opioids; Beta-Lactam Antibacterials, Penicillins; Quinolone Antibacterials; Blood Glucose Lowering Drugs, excl. Insulins; Antivaricose Therapy)
K-nearest Num: 17 Recall=6/7 Precision=3/17	6 **TP** (Antithrombotic Agents; Other Mineral Supplements; Anesthetics, General; Stomatological Preparations; Vitamin B1, Plain and in Combination with Vitamin B6 and B12; Vitamin B12 and Folic Acid) 1 **FN** (Irrigating Solutions) 11 **FP** (Potassium; Other Analgesics and Antipyretics; Drugs for Constipation; Drugs for Peptic Ulcer and Gastro-Oesophageal Reflux Disease (GORD); Adrenergics, Inhalants; Quinolone Antibacterials; Anxiolytics; Decongestants and Other Nasal Preparations for Topical Use; Hypnotics and Sedatives; Drugs Used in Addictive Disorders; Expectorants, excl. Combinations with Cough Suppressants)
ECC Num: 3 Recall=2/7 Precision=2/3	2 **TP** (Antithrombotic Agents; Other Mineral Supplements) 5 **FN** (Anesthetics, General; Irrigating Solutions; Stomatological Preparations; Vitamin B1, Plain and in Combination with Vitamin B6 and B12; Vitamin B12 and Folic Acid) 1 **FP** (Drugs for peptic ulcer and gastroesophageal reflux disease (GORD))
MLP Num: 7 Recall=4/7 Precision=4/7	4 **TP** (Antithrombotic Agents; Other Mineral Supplements; Irrigating Solutions; Anesthetics, General) 3 **FN** (Stomatological Preparation; Vitamin B1, Plain and in Combination with Vitamin B6 and B12; Vitamin B12 and Folic Acid) 3 **FP** (Drugs for Constipation; Drugs for Peptic Ulcer and Gastro-Oesophageal Reflux Disease (GORD); Anxiolytics; Hypnotics and Sedatives)
LEAP Num: 4 Recall=1/7 Precision=1/4	1 **TP** (Antithrombotic Agents) 6 **TN** (Other Mineral Supplements; N01A; Irrigating Solutions; Stomatological Preparation; Vitamin B1, Plain and in Combination with Vitamin B6 and B12; Vitamin B12 and Folic Acid) 3 **FP** (Intestinal Antiinfectives; Anxiolytics; Beta Blocking Agents)
RETAIN Num: 2 Recall=2/7 Precision=1	2 **TP** (Antithrombotic Agents; Other Mineral Supplements) 5 **FN** (Anesthetics, General; Irrigating Solutions; Stomatological Preparation; Vitamin B1, Plain and in Combination with Vitamin B6 and B12; Vitamin B12 and Folic Acid) 0 **FP** (None)
GAMENet Num: 19 Recall=6/7 Precision=6/19	6 **TP** (Antithrombotic Agents; Other Mineral Supplements; Anesthetics, General; Irrigating Solutions; Stomatological Preparations; Vitamin B1, Plain and in Combination with Vitamin B6 and B12) 1 **FN** (Vitamin B12 and Folic Acid) 13 **FP** (Antiepileptics; Potassium; Other Analgesics and Antipyretics; Drugs for Constipation; Drugs for Peptic Ulcer and Gastro-Oesophageal Reflux Disease (GORD); Beta Blocking Agents; Adrenergics, Inhalants; Calcium; Opioids; Quinolone Antibacterials; Arteriolar Smooth Muscle, Agents Acting on; Anxiolytics; Hypnotics and Sedatives)
SafeDrug Num: 5 Recall=3/7 Precision=3/5	3 **TP** (Antithrombotic Agents; Other Mineral Supplements; Vitamin B12 and Folic Acid) 4 **FN** (Anesthetics, General; Irrigating Solutions; Stomatological Preparation; Vitamin B1, Plain and in Combination with Vitamin B6 and B12) 2 **FP** (Drugs for Peptic Ulcer and Gastro-Oesophageal Reflux Disease (GORD); Calcium)
4SDrug Num: 8 Recall=6/7 Precision=3/4	6 **TP** (Antithrombotic Agents; Other Mineral Supplements; Anesthetics, General; Irrigating Solutions; Stomatological Preparations; Vitamin B1, Plain and in Combination with Vitamin B6 and B12) 1 **FN** (Vitamin B12 and Folic Acid) 2 **FP** (Other Analgesics and Antipyretics; Drugs for Constipation)