MoleRec: Combinatorial Drug Recommendation with Substructure-Aware Molecular Representation Learning

Nianzu Yang
Department of CSE and MoE Lab of Artificial Intelligence, Shanghai Jiao Tong University
800 Dongchuan RD. Minhang District, Shanghai, China
yangnianzu@sjtu.edu.cn

Kaipeng Zeng
Department of CSE and MoE Lab of Artificial Intelligence, Shanghai Jiao Tong University
800 Dongchuan RD. Minhang District, Shanghai, China
zengkaipeng@sjtu.edu.cn

Qitian Wu
Department of CSE and MoE Lab of Artificial Intelligence, Shanghai Jiao Tong University
800 Dongchuan RD. Minhang District, Shanghai, China
echo740@sjtu.edu.cn

Junchi Yan*
Department of CSE and MoE Lab of Artificial Intelligence, Shanghai Jiao Tong University
800 Dongchuan RD. Minhang District, Shanghai, China
yanjunchi@sjtu.edu.cn

ABSTRACT

Combinatorial drug recommendation involves recommending a personalized combination of medication (drugs) to a patient over his/her longitudinal history, which essentially aims at solving a combinatorial optimization problem that pursues high accuracy under the safety constraint. Among existing learning-based approaches, the association between drug substructures (i.e., a sub-graph of the molecule that contributes to certain chemical effect) and the target disease is largely overlooked, though the function of drugs in fact exhibits strong relevance with particular substructures. To address this issue, we propose a molecular substructure-aware encoding method entitled **MoleRec** that entails a hierarchical architecture aimed at modeling inter-substructure interactions and individual substructures' impact on patient's health condition, in order to identify those substructures that really contribute to healing patients. Specifically, MoleRec learns to attentively pooling over substructure representations which will be element-wisely re-scaled by the model's inferred relevancy with a patient's health condition to obtain a prior-knowledge-informed drug representation. We further design a weight annealing strategy for drug-drug-interaction (DDI) objective to adaptively control the balance between accuracy and safety criteria throughout training. Experiments on the MIMIC-III dataset demonstrate that our approach achieves new state-of-the-art performance w.r.t. four accuracy and safety metrics. Our source code is publicly available at: https://github.com/yangnianzu0515/MoleRec and MoleRec has been incorporated into the **PyHealth** package as a benchmark method for the combinatorial drug recommendation task: https://github.com/sunlabuiuc/PyHealth.

*Correspondence author is Junchi Yan who is also with Shanghai AI Laboratory. The work was in part supported by National Key Research and Development Program of China (2020AAA0107600), National Natural Science Foundation of China (62222607), and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583872

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → **Molecular structural biology**; • **Information systems** → **Recommender systems**; • **Social and professional topics** → **Medical records**.

KEYWORDS

Combinatorial Drug Recommendation, Recommender Systems, Molecule Representation Learning, Clinical Therapeutics

ACM Reference Format:
Nianzu Yang, Kaipeng Zeng, Qitian Wu, and Junchi Yan. 2023. MoleRec: Combinatorial Drug Recommendation with Substructure-Aware Molecular Representation Learning. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543507.3583872

1 INTRODUCTION

Learning-based predictive models have shown promising effects for improving the accuracy and safety of clinical decisions [31, 44, 49], which could be applied to web-based disease-diagnosis systems [2, 25, 47], with the increasing availability of individual medical data e.g., longitudinal electronic health records (EHR) [9, 14]. In particular, combinatorial drug recommendation[1] aims to provide a proper combination of drugs as final prescription for individual patient according to his/her health conditions.

The problem definition basically shares some similar spirits with sequential recommendation [4, 23, 37, 52, 60] that involves a sequential decision-making procedure over a patient's multiple visits. Still, the fundamental challenges lie in two-folds: 1) drug recommendation aims at returning a combination set of items (i.e., drugs) instead of a single item targeted by conventional recommendation tasks; 2) the recommended drug set needs to meet certain safety requirements, i.e., suppress the adverse chemical reactions among the final recommended drugs, namely *drug-drug-interaction* (DDI) [3, 6, 35, 42]. DDIs often occur when a drug is co-administered with another or multiple drugs, whose effect can be life-threatening [3, 6, 35, 42]. For instance, Tachycardia can often be triggered when the two drugs Adenosine and Ephedrine

[1] In this paper, we use "combinatorial drug recommendation" and "medication combination recommendation" interchangeably.

are taken together. The DDI rate [50, 67] is an important indicator referring to the rate of the number of conflict drug pairs against the total number of recommended drug pairs for a patient and both of them are accumulated over visits. The pairwise drug-to-drug conflict relation table is often given in advance by knowledge.

The starting works [55, 69] on medication combination recommendation ignore the sequential nature of the problem. They only focus on the current patient visit and ignore the history visits which can be informative for personalization. Some subsequent works [30, 50] start to consider the longitudinal patient history and the recommendation turns out to be more effective. Yang et al. [67] proposes to accommodate the information of an entire drug molecule, yet it fails to identify the association between substructures inside each drug molecule and the target disease, which can be informative as demonstrated by many of recent works on learning-based drug design [20, 21, 28, 58]. More recently, a symptom-based set-to-set small and safe drug recommendation method called 4SDrug [51] is developed. However, it tends to recommend a relatively smaller drug package to avoid DDIs and then meet the safety principle, which may result in lower recommendation accuracy.

In this paper, we observe that the biochemical activities of a drug are usually associated with a few privileged molecular substructures [24, 26, 45, 71], which is in fact a well-recognized agreement in literature especially for drug design. Hence we argue that patient's health condition can be related to the function of some particular substructures in drug molecules, and thus making prescriptions at a molecular substructure-aware may yield more desirable efficacy and explainability. Furthermore, it is shown in [43] that drug-drug-interaction essentially stems from substructure-substructure-interaction, suggesting that modeling the interactions among substructures would improve the recommendation safety.

To fill this gap, in this paper, we propose a novel molecular substructure-aware attentive learning method for sequential medication recommendation namely **MoleRec** to an individual patient. Our approach models the high-order interactions among molecular substructures, and learns the relevancy between the given patient's health condition and substructures. MoleRec is aimed at discriminating those substructures effectively contributing to the cure of patients. It learns substructures' representations and attentively aggregates these representations to obtain a substructure-aware representation for each drug, where the representations of substructures are scaled by their corresponding relevancy with patient's health condition before aggregation. These substructure-aware drug representations are then used for final prediction. To more explicitly improve the safety, inspired by [53], we further design an annealing-based re-weighting strategy for the DDI loss which can be derived from the drug-drug conflict table. The goal is to stabilize the training and better achieve trade-off between accuracy and safety of recommendation.

The highlights of this paper are as follows:

1) We develop a substructure-aware attentive method that models substructures' interactions and relevancy to patient's health condition. It extracts pairwise features based on the input of patient's health history and drug combination information. To our knowledge, this is the first work for explicitly modeling the substructure-level drug information in medication combination recommendation (or equivalently, combinatorial drug recommendation).

2) We devise a simple yet effective annealing-based weight adjusting approach that adaptively controls the importance of DDI loss to handle the constrained optimization problem of drug recommendation in consideration of both accuracy and safety criteria.

3) Experimental results on MIMIC-III show that our method outperforms state-of-the-art methods by a notable margin w.r.t. both accuracy and safety, whereby ablation studies further verify the effectiveness of our two new techniques.

2 RELATED WORKS

Medication Combination Recommendation. The majority of existing methods for medication combination recommendation fall into instance-based and longitudinal medication recommendation methods. Instanced-based methods focus on current patient's health condition but ignore the history. One early work refers to LEAP [69] that formulates the medication recommendation into a multi-instance multi-label sequential decision making process and adopts a variant of sequence-to-sequence model based on content-attention mechanism to make prescriptions.

In contrast, longitudinal methods leverage the temporal dependencies within clinical visits [7, 36]. Among them, RETAIN [8] is based on a two-level neural attention model that detects influential past visits and significant clinical variables within those visits. However, safety is little considered in RETAIN. Accordingly, GAMENet [50] considers DDI conflict relations by jointly modeling the longitudinal patient records as an EHR graph and drug knowledge base as a DDI graph. One step further, SafeDrug [67] extracts and encodes rich molecule structure information to improve the medication recommendation safety. The recent 4SDrug [51] recommends drugs with a small number to ensure safety.

This paper aims to explore the interactions among molecular substructures as well as the relevance between patient's health condition and substructures, which achieves notable improvement over the state-of-the-art models.

Learning Molecule Representation. Existing molecule representation learning methods can also be classified into two categories. The first is SMILES-based methods where SMILES refers to Simplified Molecular Input Line Entry System [1]. They use language models to process the textual representation (SMILES) of a molecule, for example, Transformer [54] or BERT [12]. SMILES is a linear encoding for molecules and highly depends on the traverse order of molecule graphs. Therefore its expressiveness is limited for problems like medication recommendation which we believe calls for fine-grained molecular structure extraction.

Beyond the above linear encoding protocol, structure-based methods are also developed, which can be further classified into fingerprint-based and graph neural networks (GNN)-based methods. The molecular fingerprint techniques date back to the Morgan fingerprints [40]. However, those fingerprint-based methods are often handcrafted and not trained in an end-to-end fashion [18]. Since molecules can be viewed as structured graphs, graph neural networks have been widely used to learn molecule representation, especially for drug design thanks to the high volume of available structure candidates [21].

A surge of works across bioinformatics, pharmacy and data mining [24, 26, 45, 71] have shown such a phenomenon that biochemical activities of a drug are usually associated with a few privileged molecular substructures. Thus, the significance of substructures

has been emphasized in a rich literature of molecule representation learning recently [5, 56, 57, 68, 70]. For example, based on this phenomenon and the causal invariance principle [41, 63], a recent work named MoleOOD [68] proposes to leverage those environment-invariant[2] substructures for learning robust molecular representations against distribution shifts. In this paper, also inspired by this widely-observed phenomenon, we aim to learn substructure-aware molecular representations for medication combination recommendation to provide safer and more accurate prescriptions.

3 PROBLEM FORMULATION

Given the patient's visit history, the model aims to predict his/her suitable medication combination for the current visit, in the sense of both accuracy compared with the ground truth provided by the doctor, as well as the safety as approximately measured here by the DDI rate which will increase if two related drugs (according to a prior DDI relation table) are recommended to the patient at the same time. Unless otherwise specified, all the vectors are row vectors in this paper.

Input Data (Electrical Health Records – EHR). For patient x, the corresponding EHR is represented as a sequence $\mathbf{V} = [\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(N_x)}]$, where $\mathbf{v}^{(i)}$ represents the i-th visit and N_x is the total number of visits for x. Specifically, $\mathbf{v}^{(i)}$ can be represented by a concatenation of multi-hot diagnosis, procedure (i.e. surgery) and medication vectors, i.e., $\mathbf{v}^{(i)} = [\mathbf{v}_d^{(i)}, \mathbf{v}_p^{(i)}, \mathbf{v}_m^{(i)}]$. $\mathbf{v}_d^{(i)} \in \{0,1\}^{|\mathcal{D}|}$, $\mathbf{v}_p^{(i)} \in \{0,1\}^{|\mathcal{P}|}$ and $\mathbf{v}_m^{(i)} \in \{0,1\}^{|\mathcal{M}|}$, where $\mathcal{D}$, $\mathcal{P}$ and $\mathcal{M}$ are diagnosis, procedure and medication code sets, respectively.

Known DDI Relation Matrix. We use a symmetric matrix $\mathbf{D} \in \{0,1\}^{|\mathcal{M}|\times|\mathcal{M}|}$ to denote the known DDI relation which is used as prior, where $\mathbf{D}_{ij} = 1$ if there exists interaction between drug i and j. As mentioned above, when the prescription contains two related drugs, its DDI score will decrease and become worse. These drug-drug-interation information is generated from Adverse Event Reporting Systems (AERS)[3].

Medication Combination Recommendation. Given the longitudinal diagnosis sequence and procedure sequence till time t $\mathbf{v}_d^{\mathbf{t}} = [\mathbf{v}_d^{(1)}, \mathbf{v}_d^{(2)}, \ldots, \mathbf{v}_d^{(t)}]$ and $\mathbf{v}_p^{\mathbf{t}} = [\mathbf{v}_p^{(1)}, \mathbf{v}_p^{(2)}, \ldots, \mathbf{v}_p^{(t)}]$, the DDI relation matrix $\mathbf{D}$, we aim to learn a medication combination recommendation function, $f(\cdot)$ that generates a multi-label output $\hat{\mathbf{o}}^{(t)} \in \{0,1\}^{|\mathcal{M}|}$. Note that $\hat{\mathbf{o}}^{(t)} = f(\mathbf{v}_d^{\mathbf{t}}, \mathbf{v}_p^{\mathbf{t}})$. During the training stage, we use the ground-truth recommendation $\mathbf{o}^{(t)}$ as supervision to penalize $\hat{\mathbf{o}}^{(t)}$.

4 THE PROPOSED APPROACH: MOLEREC

As shown in Fig. 1, MoleRec is composed of: 1) **patient representation module** encoding the longitudinal diagnosis and procedure information of patients. 2) **medication representation module** generating substructure-aware representations for drugs by patients' different condition 3) **prediction module** responsible for making prescription only using substructure-aware representations of drugs. In particular, in medication representation module, we model substructures' internal interactions and their relevancy by patient's health condition. It outputs a substructure-aware representation for each drug molecule as the input for prediction.

4.1 Patient Representation Module

The patient health is encoded by the diagnosis and procedure information. We define two learnable embedding tables, $\mathbf{E}_d \in \mathbb{R}^{|\mathcal{D}|\times h}$ and $\mathbf{E}_p \in \mathbb{R}^{|\mathcal{P}|\times h}$, corresponding to diagnosis and procedure, respectively, where h is the embedding size. Each row of $\mathbf{E}_d$ or $\mathbf{E}_p$ maintains a representation vector for a diagnosis or procedure. Given the multi-hot diagnosis and procedure vector at the i-th visit $\mathbf{v}_d^{(i)}$ and $\mathbf{v}_p^{(i)}$, we pick out the corresponding diagnosis and procedure embeddings and sum them up by vector-matrix dot product,

$$\mathbf{e}_d^{(i)} = \mathbf{v}_d^{(i)}\mathbf{E}_d, \quad \mathbf{e}_p^{(i)} = \mathbf{v}_p^{(i)}\mathbf{E}_p. \tag{1}$$

In line with [50, 67], we use a Dual-RNN to model history diagnosis and procedure:

$$\mathbf{h}_d^{(i)} = \text{RNN}_d(\mathbf{e}_d^{(i)}, \mathbf{h}_d^{(i-1)}), \quad \mathbf{h}_p^{(i)} = \text{RNN}_p(\mathbf{e}_p^{(i)}, \mathbf{h}_p^{(i-1)}) \tag{2}$$

where $\mathbf{h}_d^{(i)}, \mathbf{h}_p^{(i)} \in \mathbb{R}^h$ are hidden states of RNN_d and RNN_p. We set $\mathbf{h}_d^{(0)}$ and $\mathbf{h}_p^{(0)}$ as zero vectors. Next, we concatenate the hidden diagnosis state $\mathbf{h}_d^{(i)}$ and hidden procedure state $\mathbf{h}_p^{(i)}$ of the Dual-RNN to obtain a final patient representation,

$$\mathbf{e}^{(i)} = \text{CONCAT}[\mathbf{h}_d^{(i)}, \mathbf{h}_p^{(i)}]. \tag{3}$$

4.2 Medication Representation Module

Entirety-level Representation. We use Graph Neural Networks (GNNs) to encode the drug and obtain the entirety-level representation. GNNs updates node features iteratively using a neighbor aggregation strategy, i.e., updating a node's representation by aggregating representations of its neighbors. In general, we have:

$$\begin{aligned} \mathbf{a}^{(k)}(u) &= \text{AGG}^{(k)}\left(\left\{\mathbf{h}_n^{(k-1)}(v), \mathbf{h}_e(u,v) | v \in \mathcal{N}(u)\right\}\right), \\ \mathbf{h}_n^{(k)}(u) &= \text{COMBINE}^{(k)}\left(\mathbf{a}^{(k)}(u), \mathbf{h}_n^{(k-1)}(u)\right), \end{aligned} \tag{4}$$

where $\mathbf{h}_n^{(k)}(u)$ is the node feature of node u in the k-th layer, $\mathbf{h}_n^{(0)}(u)$ is the initial feature of node u, $\mathbf{h}_e(u,v)$ is the edge feature of edge (u,v), and $\mathcal{N}(u)$ denotes the neighbor of node u on the graph.

A drug molecule can be represented as a graph $G = (V, E)$, where V is the graph's node set corresponding to atoms and E denotes the graph's edge set corresponding to chemical bonds. Given a molecule G, the **entirety-level** representation is generated via a L-layer GNN followed by a *readout function* which aggregates all node features in the L-th layer,

$$\mathbf{h}_g(G) = \text{READOUT}\left(\left\{\mathbf{h}_n^{(L)}(u) | u \in V\right\}\right). \tag{5}$$

In this paper, we use GIN [65], which is widely used as backbone in recent works related to molecules [19, 39, 59], as our Graph Neural Network in practical implementation. We leave the detailed description of GIN in Appendix C due to limited space. Note that different from previous work, this entirety-level representation is not utilized directly for final recommendation in our method. Finally, all representations of drug molecules are collected together and compose an embedding table $\mathbf{E}_m \in \mathbb{R}^{|\mathcal{M}|\times h}$, of which each row corresponds to a drug.

[2]Environment-invariant substructures refer to the substructures that stably relate with the labels across environments (e.g. different scaffolds or sizes).

[3]The AERS is a database widely used to support post-marketing safety surveillance programs for all approved drug and therapeutic biologic products [16].

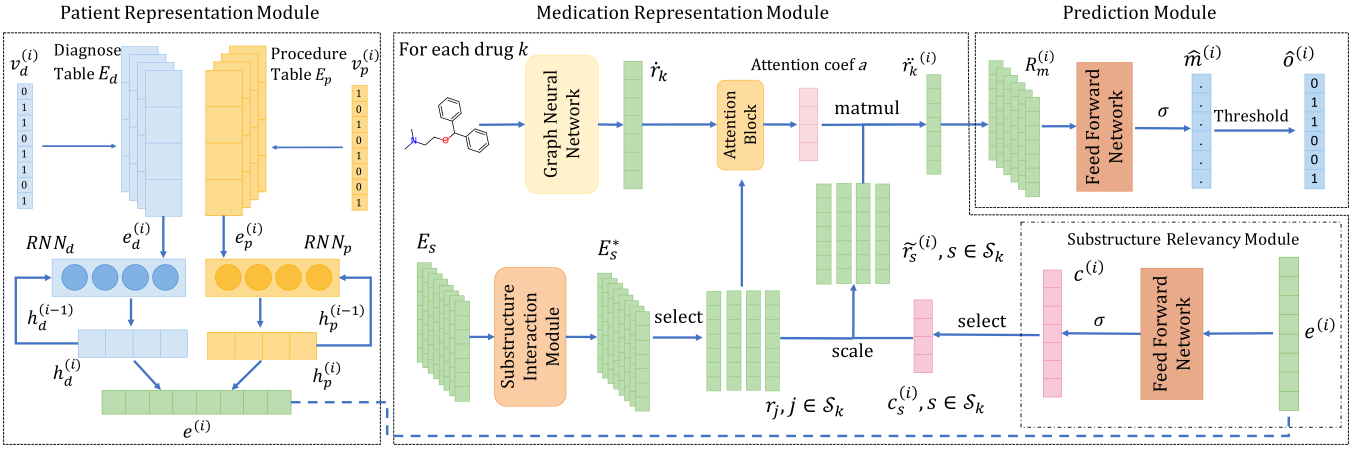

Figure 1: MoleRec: At the i-th visit of patient, we first use a Dual-RNN to encode longitudinal diagnosis and procedure (i.e. surgery) information and obtain a patient representation $\mathbf{e}^{(i)}$. Next, medication representation module models the interactions among all substructures and the relevancy between substructures and patient's health condition. Finally, the prediction module predicts each drug k's probability of appearing in the combination based on each corresponding substructure-aware representation $\ddot{\mathbf{r}}_k^{(i)}$. Those with probabilities larger than a given threshold are recommended in the final prescription.

Substructure-aware Representation. Assuming that the property of a drug molecule depends on its particular substructures and a treatment to a patient's disease also relies on some certain substructures' function, we aim to make full use of substructure-aware information of drugs. *Breaking retrosynthetically interesting chemical substructures* (BRICS) [11] method is adopted to decompose drug molecules into substructures, which is available as an API in RDKit [29] package. We decompose all drug molecules and obtain a substructure set $\mathcal{S}$. We define a embedding table $\mathbf{E}_s \in \mathbb{R}^{|\mathcal{S}|\times h}$, where each row denotes a embedding for a particular substructure.

Because substructures always function in a group instead of in isolation, we need to model high-order interactions among them. Thus, a **Substructure Interaction Module** (**SIM**) is designed to achieve this goal. We use the Set Attention Block (SAB) proposed in [32] as our Substructure Interaction Module here. SIM block takes the set of embeddings of all substructures, i.e., $\mathbf{E}_s$, and performs self-attention between the elements in the set, resulting in a set of equal size denoted as $\mathbf{E}_s^* \in \mathbb{R}^{|\mathcal{S}|\times h}$. Note that $\mathbf{E}_s^* \in \mathbb{R}^{|\mathcal{S}|\times h}$ is a new embedding table which encodes pairwise interactions among substructures. SIM is defined as below:

$$\begin{aligned} \mathbf{E}_s^* &= \mathrm{SIM}(\mathbf{E}_s) = \mathrm{LN}(\mathbf{H} + \mathrm{FF}_1(\mathbf{H})), \\ \text{where } \mathbf{H} &= \mathrm{LN}(\mathbf{E}_s + \mathcal{A}_{vanilla}(\mathbf{E}_s)), \end{aligned} \tag{6}$$

where FF_1 is a feed-forward network and LN is layer normalization. $\mathcal{A}_{vanilla}$ denotes the vanilla self-attention [54] defined by:

$$\mathcal{A}_{vanilla}(\mathbf{X}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}_x\mathbf{K}_x^\top}{\sqrt{d_k}}\right)\mathbf{V}_x, \tag{7}$$

where d_k is the dimension of key embedding, all $\mathbf{W}$ are projection matrices, $\mathbf{Q}_x$, $\mathbf{K}_x$ and $\mathbf{V}_x$ are separately the query, key and values matrices obtained by different transformations of the input $\mathbf{X}$. Next, we devise a module entitled **Substructure Relevancy Module** (**SRM**) responsible for explicitly modeling the degree at which the treatment is dependent on each substructure. Given the patient representation $\mathbf{e}^{(i)}$, we apply a feed-forward network $\mathrm{FF}_2 : \mathbb{R}^{2h} \to \mathbb{R}^{|\mathcal{S}|}$ with a sigmoid activation function $\sigma(\cdot)$,

$$\mathbf{c}^{(i)} = \mathrm{SRM}(\mathbf{e}^{(i)}) = \sigma(\mathrm{FF}_2(\mathbf{e}^{(i)})). \tag{8}$$

Each entry of $\mathbf{c}^{(i)} \in \mathbb{R}^{|\mathcal{S}|}$ depicts the contribution of each substructure to treating the patient's disease at the i-th visit.

Each drug molecule k is corresponding to a set of its substructures denoted as $\mathcal{S}_k \subseteq \mathcal{S}$. Given a drug molecule k, we pick out its entirety-level representation $\dot{\mathbf{r}}_k$ from $\mathbf{E}_m$ and its substructures' representations from $\mathbf{E}_s^*$. We denote the representation of substructure s as $\mathbf{r}_s$. Next, we can compute the attention coefficient for s that indicates the importance of substructure s's features to whole drug,

$$\mathbf{a}_{k,s} = \frac{\exp(\dot{\mathbf{r}}_k\mathbf{W}_{ent}\mathbf{W}_{sub}^\top\mathbf{r}_s^\top)}{\sum_{i\in\mathcal{S}_k}\exp(\dot{\mathbf{r}}_k\mathbf{W}_{ent}\mathbf{W}_{sub}^\top\mathbf{r}_i^\top)}, \tag{9}$$

where $\mathbf{W}_{ent}$ and $\mathbf{W}_{sub}$ are linear transformations for $\dot{\mathbf{r}}_k$ and $\mathbf{r}_s$, respectively. For each s, we scale its representation according to its contribution to curing the patient's disease:

$$\forall s \in \mathcal{S}_k, \tilde{\mathbf{r}}_s^{(i)} = \mathbf{c}_s^{(i)}\mathbf{r}_s, \tag{10}$$

where $\mathbf{c}_s^{(i)}$ is the entry in $\mathbf{c}^{(i)}$ corresponding to substructure s. Finally, we aggregate the scaled substructures' representation attentively and obtain a **substructure-aware** representation $\ddot{\mathbf{r}}_k^{(i)}$ for a given drug molecule k,

$$\ddot{\mathbf{r}}_k^{(i)} = \sum_{s\in\mathcal{S}_k}\mathbf{a}_{k,s}\tilde{\mathbf{r}}_s^{(i)}. \tag{11}$$

All drugs' substructure-aware representations make up $\mathbf{R}_m^{(i)}$.

4.3 Recommendation Prediction Module

Given all the substructure-aware representations $\mathbf{R}_m^{(i)}$ for drug set $\mathcal{M}$ as input, we apply a feed-forward network FF_3: $\mathbb{R}^h \to \mathbb{R}$ and a sigmoid activation function progressively,

$$\hat{\mathbf{m}}^{(i)} = \sigma(\mathrm{FF}_3(\mathbf{R}_m^{(i)})), \tag{12}$$

where $\hat{\mathbf{m}}^{(i)}$ represents appearance probability of each drug in the prescription. Then we can obtain a multi-hot prediction vector $\hat{\mathbf{o}}^{(i)}$ by picking out the entries of $\hat{\mathbf{m}}^{(i)}$ whose the value is greater than a predefined threshold value δ.

4.4 Training and Inference

In training, we optimize all the learnable parameters. In the inference phase, model works in the same pipeline as training. In this section, we introduce a combined loss used to find a better balance between safety and accuracy.

Multi-Label Prediction Loss. We treat the recommendation task as multi-label classification task. Thus, we adopt two common loss functions for multi-label classification, i.e., the binary cross-entropy loss $\mathcal{L}_{bce}$ and the multi-label margin loss $\mathcal{L}_{multi}$. Note that $\mathcal{L}_{multi}$ is necessary due to that it make the predicted probability of ground truth labels has at least 1 margin larger than others. Their definitions are:

$$\mathcal{L}_{bce} = -\sum_{j=1}^{|\mathcal{M}|} \mathbf{o}_j^{(i)} \log(\hat{\mathbf{m}}_j^{(i)}) + (1-\mathbf{o}_j^{(i)}) \log(1-\hat{\mathbf{m}}_j^{(i)}),$$

$$\mathcal{L}_{multi} = \sum_{p,q:\mathbf{o}_p^{(i)}=1,\mathbf{o}_q^{(i)}=0} \frac{\max(0, 1-(\hat{\mathbf{m}}_p^{(i)} - \hat{\mathbf{m}}_q^{(i)}))}{|\mathcal{M}|},$$

where the superscript i denotes the i-th entry of the vector.

DDI Loss. We define the DDI loss according to [50, 67]:

$$\mathcal{L}_{DDI} = \sum_i \sum_{p=1}^{|\mathcal{M}|} \sum_{q=1}^{|\mathcal{M}|} (\hat{\mathbf{m}}_p^{(i)} \cdot \hat{\mathbf{m}}_q^{(i)}) \cdot \mathbf{D}_{pq},$$

where $(\hat{\mathbf{m}}_p^{(i)} \cdot \hat{\mathbf{m}}_q^{(i)}) \cdot \mathbf{D}_{pq}$ gives the pairwise DDI probability.

Combined Controllable Loss Function. We adopt the weighted sum form of combined loss [13, 67],

$$\mathcal{L} = \alpha \cdot (\beta \cdot \mathcal{L}_{bce} + (1-\beta) \cdot \mathcal{L}_{multi}) + (1-\alpha) \cdot \mathcal{L}_{DDI}, \quad (13)$$

where α and β are hyperparameters. Because that DDI also exists in real EHR data, the correct predicted results may increase DDI rate as well as incorrect ones. Therefore, we can dynamically adjust α in the training phase to balance multi-label prediction loss and DDI loss, i.e., to find an accurate model with low DDI rate meanwhile.

Weight Annealing for DDI Loss (WA). In training, the accuracy and DDI rate often increase together [50]. However, what we expect is an accurate model with low DDI rate in the meantime. Hence, finding the balance between and safety and accuracy is the key. Similar to [50, 67], we denote the expected DDI rate as ϕ (recall its definition in the beginning of Sec. 1), which is a hyper-parameter requiring fine-tuning to strike a balance between accuracy and safety. If the DDI rate $\rho \geq \phi$, we have to optimize DDI loss and can adjust the value of α dynamically according to the gap between ρ and ϕ. Otherwise, we only consider the multi-label prediction loss. Similar to [53], we also want to stabilize and facilitate training stage. Motivated by this, we propose a weight annealing scheme. Specifically, we adjust α by the following strategy:

$$\alpha = \begin{cases} 1 & \rho < \phi \\ \min\{1, \exp(\tau(1-\frac{\rho}{\phi}))\} & \rho \geq \phi \end{cases}, \quad (14)$$

where ρ is the current DDI rate and τ is a hyper-parameter.

Notice that $(1-\exp(\tau(1-\frac{\rho}{\phi})))$ is a **concave** function of ρ instead of the linear one in [67], which we call linear adjusting in this paper.

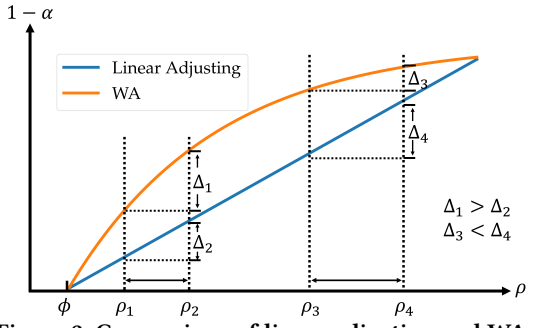

Figure 2: Comparison of linear adjusting and WA.

As shown in Fig. 2, when the gap between ρ and ϕ is still large (e.g. on interval $[\rho_3, \rho_4]$), the value of α will decrease slowly to keep the weight of DDI loss still at a high level. When the gap is small (e.g. on interval $[\rho_1, \rho_2]$), α will decline sharply and the model turns its focus on optimizing multi-label prediction loss.

5 EXPERIMENTS

In this section, we conduct extensive experiments to make a comprehensive evaluation of our proposed MoleRec.

5.1 Setup Protocol

Due to that limited EHR data are publicly available, in line with [67], we adopt bootstrapping sampling in the evaluation stage, which is better suited for limited samples [10, 46]. Ten times of bootstrapping sampling is conducted with mean and standard deviation reported. All experiments are conducted on a Linux machine with 48 CPU cores, 128GB memory and a 10.8GB NVIDIA RTX2080Ti GPU.

Dataset. We use the EHR data from **MIMIC-III** [22]. In line with [50, 67], we split train-validation-test by $\frac{2}{3}$-$\frac{1}{6}$-$\frac{1}{6}$. It contains 14,995 visits from 6,350 patients. It includes 131 drugs and the average number of used drug is 19.19.

Evaluation Metrics. We use four popular metrics in medication recommendation, Drug-Drug-Interaction Rate (DDI), Jaccard Similarity Score (Jaccard), F1-score and Precision Recall AUC (PRAUC). Among them, DDI is a measurement for safety that computes the rate of the combination prediction that contains two or multiple drugs whose relation in the DDI matrix is positive. while the others are for accuracy as commonly used in general recommendation literature. We give the detailed definitions of evaluation metrics adopted in our paper as follows:

Drug-Drug-Interaction Rate (DDI). For a certain patient x, the corresponding DDI is defined as:

$$\text{DDI} = \frac{\sum_{i=1}^{N_x} \sum_{k,l \in \{j:\hat{\mathbf{o}}_j^{(t)}=1\}} \mathbf{1}\{\mathbf{D}_{kl}=1\}}{\sum_{i=1}^{N_x} \sum_{k,l \in \{j:\hat{\mathbf{o}}_j^{(t)}=1\}} 1}, \quad (15)$$

where N_x represents the total number of visits for patient x, $\mathbf{o}^{(t)}$ denotes the multi-label predictions at the t-th visit, $\mathbf{o}_j^{(t)}$ denotes the j-th entry of $\mathbf{o}^{(t)}$, $\mathbf{D}$ is the prior DDI relation matrix and $\mathbf{1}$ is an indicator function which returns 1 when $\mathbf{D}_{kl}=1$, otherwise 0.

Jaccard Similarity Score (Jaccard). For a certain patient x at the t-th visit, the definition of Jaccard is as follows:

$$\text{Jaccard}^{(t)} = \frac{|\{i:\hat{\mathbf{o}}_i^{(t)}=1\} \cap \{i:\mathbf{o}_i^{(t)}=1\}|}{|\{i:\hat{\mathbf{o}}_i^{(t)}=1\} \cup \{i:\mathbf{o}_i^{(t)}=1\}|}, \quad (16)$$

Table 1: Performance on MIMIC-III in terms of DDI rate, Jaccard, F1-score and PRAUC. The best and the runner-up results are highlighted in bold and underline respectively under t-tests, at the level of 95% confidence level.

Method	DDI ↓	Jaccard ↑	F1-score ↑	PRAUC ↑	Avg.# of Drugs
Logistic Regression	0.0816 ± 0.0007	0.4924 ± 0.0030	0.6509 ± 0.0027	0.7589 ± 0.0026	16.7474 ± 0.1131
ECC [48]	0.0817 ± 0.0007	0.4856 ± 0.0031	0.6438 ± 0.0028	0.7590 ± 0.0026	16.2578 ± 0.0992
RETAIN [8]	0.0871 ± 0.0013	0.4866 ± 0.0034	0.6471 ± 0.0032	0.7593 ± 0.0035	18.5941 ± 0.2186
LEAP [69]	0.0760 ± 0.0008	0.4540 ± 0.0027	0.6158 ± 0.0025	0.6598 ± 0.0026	18.6739 ± 0.0661
DMNC [30]	0.0801 ± 0.0011	0.4550 ± 0.0031	0.6160 ± 0.0031	0.6757 ± 0.0029	20.0000 ± 0.0000
GAMENet [50]	0.0859 ± 0.0005	0.5037 ± 0.0015	0.6601 ± 0.0014	0.7673 ± 0.0024	27.2603 ± 0.1929
SafeDrug [67]	0.0773 ± 0.0006	0.5126 ± 0.0028	0.6691 ± 0.0023	0.7655 ± 0.0022	20.8940 ± 0.1086
4SDrug [51]	**0.0703 ± 0.0011**	0.4800 ± 0.0027	0.6404 ± 0.0024	0.7611 ± 0.0026	16.1684 ± 0.1280
MoleRec	0.0724 ± 0.0008	**0.5305 ± 0.0033**	**0.6843 ± 0.0029**	**0.7736 ± 0.0027**	21.0893 ± 0.1788

where $\hat{\mathbf{o}}^{(t)}$ and $\mathbf{o}^{(t)}$ denote the multi-label predictions and ground-truth recommendation, respectively. Note that $*_i$ represents the i-th entry of $*$. Then, we take the average over all the patient's visits to obtain the final Jaccard Similarity Score for patient x,

$$\text{Jaccard} = \frac{1}{N_x}\sum_{i=1}^{N_x}\text{Jaccard}^{(i)}, \tag{17}$$

where N_x represents the total number of visits for patient x.

F1-score. We first provide the definitions of **Precision** and **Recall** for a patient x at the t-th visit,

$$\text{Precision}^{(t)} = \frac{|\{i:\hat{\mathbf{o}}_i^{(t)}=1\}\cap\{i:\mathbf{o}_i^{(t)}=1\}|}{|\{i:\hat{\mathbf{o}}_i^{(t)}=1\}|}, \tag{18}$$

$$\text{Recall}^{(t)} = \frac{|\{i:\hat{\mathbf{o}}_i^{(t)}=1\}\cap\{i:\mathbf{o}_i^{(t)}=1\}|}{|\{i:\mathbf{o}_i^{(t)}=1\}|}. \tag{19}$$

The F1-score is the harmonic mean of Precision and Recall,

$$\text{F1}^{(t)} = \frac{2}{\frac{1}{\text{Precision}^{(t)}}+\frac{1}{\text{Recall}^{(t)}}}. \tag{20}$$

Then, we average over all visits and obtain F1 score for patient x,

$$\text{F1} = \frac{1}{N_x}\sum_{i=1}^{N_x}\text{F1}^{(i)}, \tag{21}$$

where N_x represents the total number of visits for patient x.

Precision Recall AUC (PRAUC). Note that we treat medication combination recommendation as an information retrieval problem. For the patient x at the t-th visit, PRAUC is defined as follows:

$$\text{PRAUC}^{(t)} = \sum_{k=1}^{|\mathcal{M}|}\text{Precision}(k)^{(t)}\Delta\text{Recall}(k)^{(t)}, \tag{22}$$

$$\Delta\text{Recall}(k)^{(t)} = \text{Recall}(k)^{(t)} - \text{Recall}(k-1)^{(t)}, \tag{23}$$

where k is the rank in the sequence of the retrieved drugs, $|\mathcal{M}|$ denotes the number of drugs. $\text{Precision}(k)^{(t)}$ represents represents the precision at cut-off k in the ordered retrieval list and $\Delta\text{Recall}(k)^{(t)}$ denotes the change of recall from drug $k-1$ to k. We also average over all visits and then obtain the PRAUC value for patient x,

$$\text{PRAUC} = \frac{1}{N_x}\sum_{i=1}^{N_x}\text{PRAUC}^{(i)}, \tag{24}$$

where N_x represents the total number of visits for patient x.

Compared Methods. We compare our method with the representative state-of-the-art baselines as follows:

Logistic Regression (LR) is an instance-based classifier with L_2 regularization. We adopt One-vs-the-Rest (OvR) multi-class strategy for the multi-label prediction setting, i.e., fitting one classifier per class. We implement LR by *scikit-learn* package and adopt One-vs-the-Rest (OvR) multi-class strategy for the multi-label prediction setting, i.e., fitting one classifier per class. LBFGS is chosen as the optimizer.

Ensemble Classifier Chain (ECC) [48] is a multi-label model that arranges LR classifiers into a chain. Each classifiers gets the predictions of the preceding classifiers in the chain as features. We implement a 10-member ensemble of ClassifierChains by *scikit-learn* package. Each ClassifierChain adopts LBFGS as optimizer.

RETAIN [8] is based on a two-level neural attention model that detects influential past visits and significant clinical variables within those visits. We implement two 64-dim GRUs as the two-level RNN. The dropout rate for the output embedding is set to 0.5. We adopt Adam as the optimizer and learning rate ranges from $\{1e-5, 5e-4, 1e-4\}$.

LEAP [69] formulates medication recommendation to sequential decision making, with recurrent decoder to model label dependencies and content-based attention to capture label instance mapping. We also search learning rate from $\{1e-5, 5e-4, 1e-4\}$. Due to the fact that LEAP is a sequence-based models, we set the max drug combination size to 20.

DMNC [30] is multi-view sequential learning method via memory augmented neural network based on differentiable neural computer (DNC). We use 64-dim embedding tables to represent medical codes. Using *dnc* package, we implement a 16-cell DNC as encoder. Decoder is a 64-dim GRU. Same as LEAP, we also set the max drug combination size to 20.

GAMENet [50] is based on memory networks with memory bank enhanced by integrated drug usage, DDI graphs and dynamic memory with patient history. We use hyperparameters provided by authors yield the best results. We adopt 64-dim embedding tables and 64-dim GRU as RNN.

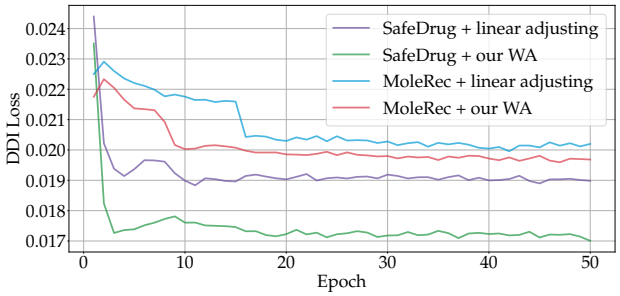

Figure 3: Comparison of combinations of backbone models (ours and SafeDrug) and loss adjusting strategies.

SafeDrug [67] is a recent work that extracts and encodes molecule structure information to augment the medication recommendation task. We look for the optimal learning rate over $\{1e-5, 5e-4, 1e-4\}$, and other parameters that yield the best results in the original paper are used in our experiments.

4SDrug [51] is designed for recommending small drug sets to ensure less drug-drug interactions. The learning rate is searched within $\{1e-2, 5e-3, \ldots, 1e-5\}$.

5.2 Implementation Details.

For embedding tables $\mathbf{E}_p$, $\mathbf{E}_d$ and $\mathbf{E}_s$, we set the embedding size to 64. For RNN_d and RNN_p, we use gate recurrent unit (GRU) with 64 hidden units. The dropout rate of two GRU cells' inputs is set as 0.7. As metioned before, we choose a 4-layer GIN [65] with hidden embedding size of 64. For each molecule graph, the 9-dimensional initial node features contains atomic number and chirality, as well as other additional atom features, while the 3-dimensional edge features contains bond type, bond stereochemistry and whether the bond is conjugated [17]. The feed-forward network FF_1 and FF_3 are implemented as one biased linear layer and the feed-forward network FF_2 is implemented as two biased linear layers. We choose the hyper-parameters according to the performance on the validation set, where threshold $\delta = 0.5$, weight $\beta = 0.95$, $\tau = 2.5$ and the target DDI rate $\phi = 0.06$. The model is built on *Pytorch 1.9.0* and the model parameters are trained with Adam optimizer.

5.3 Quantitative Study

This study evaluates the accuracy and safety of our method and the baselines as summarized previously. Each baseline is configured by using default settings in the original paper or our fine-tuned parameters leading to best performance.

Table 1 summarizes the prediction performance of all methods. For these traditional classifiers that are not DL-based, LR and ECC, even though they recommend rather less drugs, there still exists a lot of drug-drug-interactions in the final medication combination set. Those methods treating medication recommendation task as sequence generation task, ECC, LEAP and DMNC, show poor performance among all methods. Hence, formulating medication recommendation into multi-label prediction may be more reasonable. RETAIN does not consider DDI relations, resulting in a high DDI rate. As for GAMENet, since it is highly dependent on historical combinations while the DDI rate may be high in ground-truth history, it also yields poor DDI rate. 4SDrug aims to reduce the size of recommended drug sets and we can see that 4SDrug does recommends the least drugs while sacrificing the accuracy. Due to encoding molecule structural information, SafeDrug nearly beats all other baselines. But it still makes predictions at a molecular entirety-level and neglects the effect of molecule substructures and the interactions among them. With modeling interactions among substructures and dependencies between patient's health condition and substructures, our method that is based on molecular substructure-aware representations can provide more accurate medication combination and outperforms all baselines with a notable margin. Additionally, equipped with the proposed weight annealing strategy, MoleRec gains further improvement on both accuracy and safety and still balance them well.

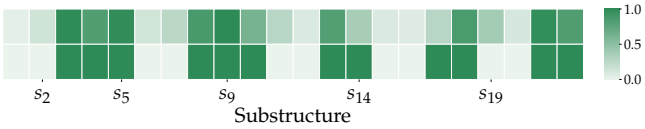

Figure 4: Top: relevancy prediction between patient's health and 22 substructures s_i. Bottom: binary value decided if s_i is a component of ground-truth drugs. There is a coherence between prediction and ground truth for this patient example.

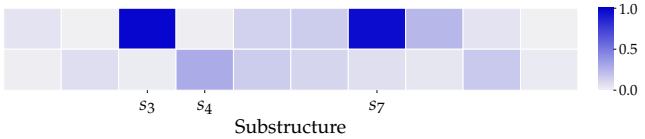

Figure 5: Top: relevancy prediction between patient's health condition and 10 substructures s_i which make up an integrated drug molecule. Bottom: attention score of the corresponding substructures learned by MoleRec. There is a deviation between the prediction and attention score for this drug example.

5.4 Qualitative Study

Analysis of Our Adjusting Strategy for Loss. To verify the effectiveness of our proposed WA in stabilize and facilitate the training procedure, as demonstrated in Fig. 3, we compare the combinations of different backbone models (ours and SafeDrug) and different loss adjusting strategies. From the DDI loss evolving tendencies on both SafeDrug and MoleRec with different strategies, we can see that equipped with our weight annealing strategy, DDI loss goes down at a faster rate. Our strategy is superior is mainly due to adaptively re-weighting DDI loss at a dynamic rate according to the gap between the current DDI rate and the target.

Case Study. We also qualitatively evaluate the correlations between patient's health condition and substructures learned by MoleRec. Fig. 4 shows the learned relevancy between a patient's health condition and some substructures. For the substructures composing ground-truth drugs e.g., s_5 and s_9, MoleRec learns that they are closely related to patient's health condition compared with s_2 and s_{19}. But s_{14} shows low relevancy, and Fig. 5 could explain this exception next. As for Fig. 5, it compares each substructure's relevancy with a patient's health condition and attention score to the whole drug. We can see that though substructure s_4 has the highest attention score, its relevancy is much lower than s_3 and s_7. The reason may lies in that the drug molecule has many structures s_4 but s_4 is

Table 2: Ablation study on MIMIC-III in terms of DDI rate, Jaccard, F1-score and PRAUC. The best and the runner-up results are highlighted in bold and underline respectively under t-tests, at the level of 95% confidence level. We also present the results of SafeDrug, and it shows the best overall performance among baselines.

Method	DDI ↓	Jaccard ↑	F1-score ↑	PRAUC ↑	Avg.# of Drugs
SafeDrug [67]	0.0773 ± 0.0006	0.5126 ± 0.0028	0.6691 ± 0.0023	0.7655 ± 0.0022	20.8940 ± 0.1086
MoleRec (our full version)	0.0724 ± 0.0008	0.5305 ± 0.0033	0.6843 ± 0.0029	0.7736 ± 0.0027	21.0893 ± 0.1788
– temporal dependencies within visits	0.0777 ± 0.0006	0.5259 ± 0.0026	0.6807 ± 0.0023	0.7665 ± 0.0025	21.5592 ± 0.1406
– substructure interaction module (SIM)	0.0756 ± 0.0006	0.5250 ± 0.0028	0.6797 ± 0.0024	0.7698 ± 0.0025	21.6722 ± 0.1634
– substructure relevancy module (SRM)	0.0812 ± 0.0006	0.4873 ± 0.0028	0.6455 ± 0.0026	0.7317 ± 0.0027	21.5606 ± 0.1743
– multi-label prediction loss	**0.0646 ± 0.0007**	0.4619 ± 0.0042	0.6216 ± 0.0039	0.7510 ± 0.0026	14.1569 ± 0.1664
BRICS → RECAP	0.0733 ± 0.0008	0.5298 ± 0.0029	0.6838 ± 0.0026	0.7733 ± 0.0024	20.9460 ± 0.1522
embedding table → GNN	0.0716 ± 0.0007	**0.5312 ± 0.0039**	**0.6850 ± 0.0033**	**0.7751 ± 0.0025**	20.5375 ± 0.1583
WA → linear adjusting	0.0749 ± 0.0008	0.5258 ± 0.0036	0.6804 ± 0.0031	0.7707 ± 0.0028	20.5263 ± 0.1142

not the critical factor to drug's function. Hence, we need to amplify the influence of s_3 and s_7 that truly determine drug's function. Two examples shows MoleRec can discriminate those substructures that really contribute to healing patients.

5.5 Ablation Study

We investigate the effects of different components to the final performance and summarize the results in Table 2.

Temporal dependencies within visits. To verify the superiority of longitudinal methods, we treat each visit as a single training sample instead of using a RNN to model their temporal dependencies. Compared to our longitudinal model, it shows inferior performance.

Substructure interaction module. We remove the substructure interaction module (SIM) from our model and conduct extensive experiments. Table 2 shows that the performance degrades especially in terms of safety, which empirically indicates that modeling the interactions among substructures could improve the recommendation safety to a great extent.

Substructure relevancy module. We also discard the proposed substructure relevancy module (SRM) to evaluate its impact. Because of the absence of SRM, the substructure-aware representation $\ddot{\mathbf{r}}_k^{(i)}$ for a given drug molecule k is recalculated by $\ddot{\mathbf{r}}_k^{(i)} = \sum_{s \in \mathcal{S}_k} \mathbf{a}_{k,s} \mathbf{r}_s$, where $\mathbf{r}_s$ is the substructure representation. Also, to make use of patient's information, we use the similar method as SafeDrug to recalculated the appearance probability of each drug in the prescription, $\hat{\mathbf{m}}^{(i)} = \sigma(\mathbf{R}_m^{(i)} \mathrm{FF}_4(\mathbf{e}^{(i)}))$, where all drugs' substructure-aware representations make up new $\mathbf{R}_m^{(i)}$ and $\mathrm{FF}_4 : \mathbb{R}^{2h} \to \mathbb{R}^h$ is a feed forward network. Compared to our proposed SRM, this method shows poorer performance. This suggests that SRM could effectively model the relevancy between patients and molecule substructures, and this substructure-grained drug recommendation is superior.

Multi-label margin loss. We experiment with leaving out the multi-label margin loss $\mathcal{L}_{multi}$. Then, our final objective is reduced to the weighted sum of only $\mathcal{L}_{bce}$ and $\mathcal{L}_{ddi}$. Without $\mathcal{L}_{multi}$, we can find that the our model recommends much less drugs. Hence, less DDI will occur and less ground-truth drugs are recommended in final prescriptions. Results suggest that $\mathcal{L}_{multi}$ could effectively make the predicted probability of ground truth labels has at least 1 margin larger than others.

Molecule segmentation strategy. For all experiments in Table 1, we adopt BRICS method to decompose molecules into substructures. To evaluate the sensitivity of our model to molecule segmentation strategy, we adopt another method called *retrosynthetic combinatorial analysis procedure* (RECAP) [34], which is also available as an API in RDKit. According to results in Table 2, RECAP and BRICS show competitive performance on our model and both outperform the baselines by notable margins. Hence, our model is robust to molecule segmentation strategy.

Substructure representation learning. We simply define a learnable embedding table $\mathbf{E}_s$, of which each row corresponds to a representation for a particular substructure. Similar to learning entirety-level molecular representations, we can also use a Graph Neural Network to learn substructure representations, i.e. utilizing structural information to enhance the learned representations. Hence, we substitute the original embedding table with GIN and report the results in Table 2. We find that it brings slight improvement over safety and accuracy. But it should be mentioned that the model using GIN has approximately 11.14% parameters[4] more than that using embedding table. Therefore, we stick to using embedding table because it is simple yet efficient.

Re-weighting DDI loss. In this paper, we propose a weight annealing strategy for DDI loss. We compare it with the linear adjusting proposed in SafeDrug [67]. We equipped our model with the linear adjusting and make comparisons with our proposed WA. Based on the results in Table 2, WA could help the model not only obtain safer and more accurate recommendation, but also find a better equilibrium point between safety and accuracy. We attribute this to the ability of WA to adaptively re-weight DDI loss.

6 CONCLUSION

We have proposed a novel molecular substructure-aware attentive method for medication combination recommendation, entitled MoleRec. It models the interactions among molecular substructures and the dependencies between patient's health condition and substructures. Moreover, we design an adjusting strategy to re-weight DDI loss in the training phase, which helps model find an better balance between recommendation accuracy and safety. Extensive experiments on MIMIC-III have demonstrated the effectiveness of our approach in terms of both accuracy and safety.

[4]Our model using embedding table has 456, 112 parameters while the one using 4-layer GIN has 506, 932 parameters

REFERENCES

[1] Eric Anderson, Gilman D Veith, and David Weininger. 1987. *SMILES, a line notation and computerized interpreter for chemical structures.* US Environmental Protection Agency, Environmental Research Laboratory.

[2] UM Ashwinkumar and KR Anandakumar. 2012. A web-based patient support system using artificial intelligence to improve health monitoring and quality of life. In *2012 Second International Conference on Advanced Computing & Communication Technologies*. IEEE, 101–105.

[3] Konstantinos Bougiatiotis, Fotis Aisopos, Anastasios Nentidis, Anastasia Krithara, and Georgios Paliouras. 2020. Drug-Drug Interaction Prediction on a Biomedical Literature Knowledge Graph. In *International Conference on Artificial Intelligence in Medicine*. Springer, 122–132.

[4] Chao Chen, Haoyu Geng, Nianzu Yang, Junchi Yan, Daiyue Xue, Jianping Yu, and Xiaokang Yang. 2021. Learning Self-Modulating Attention in Continuous Time Space with Applications to Sequential Recommendation. In *International Conference on Machine Learning*. PMLR, 1606–1616.

[5] Shicheng Cheng, Liang Zhang, Bo Jin, Qiang Zhang, and Xinjiang Lu. 2021. Drug target prediction using graph representation learning via substructures contrast. (2021).

[6] Wen-Hao Chiang, Li Shen, Lang Li, and Xia Ning. 2020. Drug-drug interaction prediction based on co-medication patterns and graph matching. *International Journal of Computational Biology and Drug Design* 13, 1 (2020), 36–57.

[7] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*.

[8] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in Neural Information Processing Systems* 29 (2016).

[9] Martin R Cowie, Juuso I Blomster, Lesley H Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, et al. 2017. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology* 106, 1 (2017), 1–9.

[10] Saha Dauji and Kapilesh Bhargava. 2016. Estimation of concrete characteristic strength from limited data by bootstrap. *Journal of Asian Concrete Federation* 2, 1 (2016), 81–94.

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

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Alexey Dosovitskiy and Josip Djolonga. 2019. You only train once: Loss-conditional training of deep networks. In *International Conference on Learning Representations*.

[14] R Scott Evans. 2016. Electronic health records: then, now, and in the future. *Yearbook of medical informatics* 25, S 01 (2016), S48–S61.

[15] William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems*. 1024–1034.

[16] Keith B Hoffman, Mo Dimbil, Colin B Erdman, Nicholas P Tatonetti, and Brian M Overstreet. 2014. The Weber effect and the United States Food and Drug Administration's Adverse Event Reporting System (FAERS): analysis of sixty-two drugs approved from 2006 to 2010. *Drug safety* 37, 4 (2014), 283–294.

[17] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: datasets for machine learning on graphs. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 22118–22133.

[18] Sabrina Jaeger, Simone Fulle, and Samo Turk. 2018. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of Chemical Information and Modeling* (2018).

[19] Shunyu Jiang, Fuli Feng, Weijian Chen, Xiang Li, and Xiangnan He. 2021. Structure-enhanced meta-learning for few-shot graph classification. *AI Open* 2 (2021), 160–167.

[20] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2018. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*. PMLR, 2323–2332.

[21] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2020. Multi-objective molecule generation using interpretable substructures. In *International Conference on Machine Learning*. PMLR, 4849–4859.

[22] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* (2016).

[23] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining*.

[24] Yu-Ting Kao, Shu-Fen Wang, Meng-Hsiu Wu, Shwu-Huey Her, Yi-Hsuan Yang, Chung-Hsien Lee, Hsiao-Feng Lee, An-Rong Lee, Li-Chien Chang, and Li-Heng Pao. 2022. A substructure-based screening approach to uncover N-nitrosamines in drug substances. *Journal of Food & Drug Analysis* 30, 1 (2022).

[25] B Karlik and E Oztoprak. 2007. Web-based telemedical consultation and diagnosis model by multiple artificial neural networks. *Ukrainian Journal of Telemedicine and Medical Telematics* 5, 2 (2007), 156–160.

[26] Justin Klekota and Frederick P Roth. 2008. Chemical substructures that enrich for biological activity. *Bioinformatics* 24, 21 (2008), 2518–2525.

[27] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran S. Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *International Conference on Machine Learning*. 5637–5664.

[28] Xiangzhe Kong, Zhixing Tan, and Yang Liu. 2021. Graphpiece: Efficiently generating high-quality molecular graph with substructures. *arXiv preprint arXiv:2106.15098* (2021).

[29] Greg Landrum et al. 2006. RDKit: Open-source cheminformatics. (2006).

[30] Hung Le, Truyen Tran, and Svetha Venkatesh. 2018. Dual memory neural computer for asynchronous two-view sequential learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1637–1645.

[31] Cecilia S Lee and Aaron Y Lee. 2020. Clinical applications of continual learning machine learning. *The Lancet Digital Health* 2, 6 (2020), e279–e281.

[32] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. 2019. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*. PMLR, 3744–3753.

[33] AA Leman and Boris Weisfeiler. 1968. A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Technicheskaya Informatsiya* 2, 9 (1968), 12–16.

[34] Lewell, XQ, Judd, DB, Watson, SP, Hann, and MM. 1998. RECAP - Retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J CHEM INFORM COMPUT SCI* (1998).

[35] Xuan Lin, Zhe Quan, Zhi-Jie Wang, Tengfei Ma, and Xiangxiang Zeng. 2020. KGNN: Knowledge Graph Neural Network for Drug-Drug Interaction Prediction.. In *International Joint Conference on Artificial Intelligence*, Vol. 380. 2739–2745.

[36] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. 2015. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677* (2015).

[37] Chen Ma, Peng Kang, and Xue Liu. 2019. Hierarchical gating networks for sequential recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 825–833.

[38] Lukasz Maziarka, Tomasz Danel, Slawomir Mucha, Krzysztof Rataj, Jacek Tabor, and Stanislaw Jastrzebski. 2020. Molecule Attention Transformer. *CoRR* abs/2002.08264 (2020).

[39] Siqi Miao, Mia Liu, and Pan Li. 2022. Interpretable and Generalizable Graph Learning via Stochastic Attention Mechanism. In *International Conference on Machine Learning*. PMLR, 15524–15543.

[40] Harry L Morgan. 1965. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation* (1965).

[41] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain Generalization via Invariant Feature Representation. In *International Conference on Machine Learning (ICML)*. 10–18.

[42] Arnold K Nyamabo, Hui Yu, Zun Liu, and Jian-Yu Shi. 2021. Drug–drug interaction prediction with learnable size-adaptive molecular substructures. *Briefings in Bioinformatics* (2021).

[43] Arnold K Nyamabo, Hui Yu, and Jian-Yu Shi. 2021. SSI–DDI: substructure–substructure interactions for drug–drug interaction prediction. *Briefings in Bioinformatics* (2021).

[44] Ziad Obermeyer and Ezekiel J Emanuel. 2016. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine* 375, 13 (2016), 1216.

[45] Chuleeporn Phanus-umporn, Watshara Shoombuatong, Veda Prachayasittikul, Nuttapat Anuwongcharoen, and Chanin Nantasenamat. 2018. Privileged substructures for anti-sickling activity via cheminformatic analysis. *RSC advances* (2018).

[46] Victor Picheny, Nam Ho Kim, and Raphael T Haftka. 2010. Application of bootstrap method in conservative estimation of reliability with limited samples. *Structural and Multidisciplinary Optimization* 41, 2 (2010), 205–217.

[47] Hsiao-Hsien Rau, Chien-Yeh Hsu, Yu-An Lin, Suleman Atique, Anis Fuad, Li-Ming Wei, and Ming-Huei Hsu. 2016. Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network. *Computer methods and programs in biomedicine* 125 (2016), 58–65.

[48] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine Learning* (2011).
[49] Farah Shamout, Tingting Zhu, and David A Clifton. 2020. Machine learning for clinical outcome prediction. *IEEE reviews in Biomedical Engineering* 14 (2020), 116–126.
[50] Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. 2019. Gamenet: Graph augmented memory networks for recommending medication combination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1126–1133.
[51] 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*. 3970–3980.
[52] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 565–573.
[53] Chenyang Tao, Shuyang Dai, Liqun Chen, Ke Bai, Junya Chen, Chang Liu, Ruiyi Zhang, Georgiy Bobashev, and Lawrence Carin Duke. 2019. Variational annealing of GANs: A Langevin perspective. In *International Conference on Machine Learning*. PMLR, 6176–6185.
[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).
[55] Meng Wang, Mengyue Liu, Jun Liu, Sen Wang, Guodong Long, and Buyue Qian. 2017. Safe medicine recommendation via medical knowledge graph embedding. *ArXiv e-prints* (2017).
[56] Pengyang Wang, Yanjie Fu, Yuanchun Zhou, Kunpeng Liu, Xiaolin Li, and Kien A Hua. 2020. Exploiting Mutual Information for Substructure-aware Graph Representation Learning.. In *International Joint Conference on Artificial Intelligence*. 3415–3421.
[57] Shuang Wang, Zhen Li, Shugang Zhang, Mingjian Jiang, Xiaofeng Wang, and Zhiqiang Wei. 2020. Molecular property prediction based on a multichannel substructure graph. *IEEE Access* 8 (2020), 18601–18614.
[58] Shuang Wang, Tao Song, Shugang Zhang, Mingjian Jiang, Zhiqiang Wei, and Zhen Li. 2022. Molecular substructure tree generative model for de novo drug design. *Briefings in bioinformatics* 23, 2 (2022), bbab592.
[59] Yanling Wang, Jing Zhang, Shasha Guo, Hongzhi Yin, Cuiping Li, and Hong Chen. 2021. Decoupling representation learning and classification for gnn-based anomaly detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1239–1248.
[60] Qitian Wu, Yirui Gao, Xiaofeng Gao, Paul Weng, and Guihai Chen. 2019. Dual Sequential Prediction Models Linking Sequential Recommendation and Information Dissemination. In *International Conference on Knowledge Discovery & Data Mining*. ACM, 447–457.
[61] Qitian Wu, Chenxiao Yang, Wentao Zhao, Yixuan He, David Wipf, and Junchi Yan. 2023. DIFFormer: Scalable (Graph) Transformers Induced by Energy Constrained Diffusion. In *The Eleventh International Conference on Learning Representations*.
[62] Qitian Wu, Hengrui Zhang, Xiaofeng Gao, Junchi Yan, and Hongyuan Zha. 2021. Towards open-world recommendation: An inductive model-based collaborative filtering approach. In *International Conference on Machine Learning*. 11329–11339.
[63] Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. 2022. Handling Distribution Shifts on Graphs: An Invariance Perspective. In *International Conference on Learning Representations*.
[64] Qitian Wu, Wentao Zhao, Zenan Li, David Wipf, and Junchi Yan. 2022. NodeFormer: A Scalable Graph Structure Learning Transformer for Node Classification. In *Advances in Neural Information Processing Systems*.
[65] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How Powerful are Graph Neural Networks?. In *International Conference on Learning Representations*.
[66] Chenxiao Yang, Qitian Wu, Qingsong Wen, Zhiqiang Zhou, Liang Sun, and Junchi Yan. 2022. Towards out-of-distribution sequential event prediction: A causal treatment. In *Advances in Neural Information Processing Systems*.
[67] Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. 2021. SafeDrug: Dual Molecular Graph Encoders for Recommending Effective and Safe Drug Combinations.. In *International Joint Conference on Artificial Intelligence*. 3735–3741.
[68] Nianzu Yang, Kaipeng Zeng, Qitian Wu, Xiaosong Jia, and Junchi Yan. 2022. Learning substructure invariance for out-of-distribution molecular representations. In *Advances in Neural Information Processing Systems*.
[69] Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart, and Jimeng Sun. 2017. LEAP: learning to prescribe effective and safe treatment combinations for multimorbidity. In *proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1315–1324.
[70] Xiaohan Zhao, Bo Zong, Ziyu Guan, Kai Zhang, and Wei Zhao. 2018. Substructure assembling network for graph classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[71] Jiajing Zhu, Yongguo Liu, Chuanbiao Wen, and Xindong Wu. 2020. DGDFS: Dependence Guided Discriminative Feature Selection for Predicting Adverse Drug-Drug Interaction. *IEEE Transactions on Knowledge and Data Engineering* (2020).

APPENDIX

A NOTATIONS

Table 3 lists the notations used to facilitate reading.

Table 3: Notations for facilitating reading. Note that $*^{(i)}$ corresponds to the i-th visit of a patient.

Notation	Description
$\mathbf{V}$	patient electronic health record sequence
$\mathbf{v}^{(i)}$	record of visit
$\mathbf{v}_d^{(i)}$	multi-hot diagnosis vector
$\mathbf{v}_p^{(i)}$	multi-hot procedure vector
$\mathbf{v}_m^{(i)}$	multi-hot medication vector
N_x	total number of visits for patient x
$\mathbf{D}$	DDI relation matrix
$\mathcal{D}, \mathcal{P}, \mathcal{M}$	diagnosis, procedure, medication code sets
$\mathcal{S}$	substructure code set
h	dimension of vector
$\mathbf{h}_d^{(i)}, \mathbf{h}_p^{(i)}$	hidden diagnosis and procedure states
$\mathbf{E}_d, \mathbf{E}_p$	embedding tables for diagnosis and procedure
$\mathbf{e}^{(i)}$	patient representation
$\mathbf{E}_m$	embedding table for medication
$\mathbf{E}_s$	initial embedding table for substructures
$\mathbf{E}_s^*$	embedding table for substructures after SIM
$\dot{\mathbf{r}}_k$	entirety-level representation of drug k
$\mathbf{r}_s$	representation of substructure s
$\mathbf{a}_{k,s}$	attention weight of substructure s to drug k
$\mathbf{c}^{(i)}$	relevancy between substructures and disease
$\tilde{\mathbf{r}}_s^{(i)}$	scaled representation of substructure s
$\ddot{\mathbf{r}}_k^{(i)}$	substructure-aware representation of drug k
$\mathbf{R}_m^{(i)}$	all drugs' substructure-aware representations
$\hat{\mathbf{m}}^{(i)}$	predicted probabilities
$\hat{\mathbf{o}}^{(i)}$	multi-label predictions
$\mathbf{o}^{(i)}$	ground-truth recommendation

B DATASET DESCRIPTION

MIMIC-III[5] is a large and freely-available database containing deidentified health-related data, supporting a range of analytic studies, e.g. spanning epidemiology and clinical decision-rule improvement. However, researchers are required to complete a recognized course in protecting human research participants that includes Health Insurance Portability and Accountability Act (HIPAA) requirements. Then, access will be granted.

C DETAILS OF GIN

Graph Isomorphism Network (GIN) [65] is a simple architecture, which generalizes the Weisfeiler-Lehman test [33]. Therefore, GIN can achieve maximum discriminative power among GNNs. Recently, it has been adopted by a rich literature of works related to

[5] https://physionet.org/content/mimiciii/1.4/

molecules [19, 39, 59] as their backbone. Recalling that we have introduced the general mechanisms of modern GNNs in Sec. 4.2, in particular, GIN updates the node representations as follows:

$$\begin{aligned}\mathbf{a}^{(k)}(u) &= \mathrm{AGG}^{(k)}\left(\left\{\mathbf{h}_n^{(k-1)}(v), \mathbf{h}_e(u,v) | v \in \mathcal{N}(u)\right\}\right),\\ \mathbf{h}_n^{(k)}(u) &= \mathrm{MLP}^{(k)}\left(\mathbf{a}^{(k)}(u) + \left(1+\epsilon^{(k)}\right)\cdot \mathbf{h}_n^{(k-1)}(u)\right),\end{aligned} \tag{25}$$

where ϵ could be set as a learnable parameter or a fixed scalar, and MLP denotes the composition of functions. In principle, in our proposed method, GIN can be replaced by other molecular encoding methods.

D CURRENT LIMITATIONS & OUTLOOK

We discuss the limitations of the present work and shed more lights on the potential future directions to pave the way for more follow-up advances in combinatorial drug recommendation.

Inductive Learning for New Drugs. Our model assumes that drugs and patients appearing in the testing set are all exposed to the model in the training stage, i.e., either the drug space or the patient space is shared by training and testing data. Such a setting is often called transductive learning in the literature. However, real-world recommendation system needs to interact with an open world where both new unseen users and items could appear in the future [62]. In the context of medication recommendation, there could be new drugs introduced to the system or new patients asking for medical care. And, the ideal model should be equipped with the ability to handle these new entities (that are unseen during training) without any re-training or fine-tuning on the new data. Such a problem setting is called inductive learning [15, 62] which refers to that the testing set contains new entities that are unseen by the model during training.

Distribution Shifts & Out-of-Distribution Generalization. Another promising direction is to consider the distribution shifts between training and testing data, which can be pervasive in practice and could largely impact the model performance as suggested by recent evidence [27]. There could be two types of distribution shifts in the drug recommendation.

- One category of distribution shifts lies in different contexts (like medical records in different hospitals or recorded at different seasons, etc.) behind the sequential data. For example, the model is trained with data collected in the summer and is expected to perform well on testing data collected in the winter. Due to different temperature and epidemic in different seasons, the training distributions can exhibit certain differences than the testing ones. A recent study [66] points out that traditional maximum likelihood estimation (used by most of existing models for sequential recommendation/prediction) would fail to generalize to new distributions due to the confounding bias of latent contexts. To alleviate this, it develops a causal intervention approach as an effective treatment for model training.
- Another category of distribution shifts could happen due to spurious features of drugs (e.g., molecule scaffolds or sizes). For example, the overall sizes of molecules are different between training and testing sets. These appearance characteristics should not affect the model's generalization ability since they have no causal relationship with the target disease. Despite this, these spurious features often correlate with the labels due to potential bias in observational data, which are called spurious correlation in the literature [41]. A recent work [68] formulates the out-of-distribution (OOD) generalization problem for molecular representation learning and proposes a domain-invariant learning approach that can address the spurious correlation and learn causal relation from the input molecule to its predicted properties. In the problem of this paper, we only focus on the correlation-based learning that targets the associations between the function of drugs and particular substructures, yet ignore their causal relation, which might affect the model's ability to handle new data from out-of-distributions.

As future works, one can naturally formulate the problem of drug recommendation (that is associated with both molecular representation learning and sequential prediction) under certain distribution shifts, and based on this explore effective methods.

More Powerful Graph Encoders. With the rapid development in graph machine learning community, there are quite a few recent works identifying more powerful encoder architectures for graph-structured data, i.e., graph Transformers [38, 61, 64], that show competitive or even superior performance than commonly used GNNs. As a direction orthogonal to our focus as well as above two points, one can extend the encoding framework in our work to some advanced version with Transformer-like architectures to achieve better expressivity and capacity for molecular representations.