Context-Aware Safe Medication Recommendations with Molecular Graph and DDI Graph Embedding

Qianyu Chen[1], **Xin Li** *[1], **Kunnan Geng**[1], **Mingzhong Wang**[2]

[1] Beijing Institute of Technology
[2] University of the Sunshine Coast
{qychen,xinli,3120190993}@bit.edu.cn, mwang@usc.edu.au

Abstract

Molecular structures and Drug-Drug Interactions (DDI) are recognized as important knowledge to guide medication recommendation (MR) tasks, and medical concept embedding has been applied to boost their performance. Though promising performance has been achieved by leveraging Graph Neural Network (GNN) models to encode the molecular structures of medications or/and DDI, we observe that existing models are still defective: 1) to differentiate medications with similar molecules but different functionality; or/and 2) to properly capture the unintended reactions between drugs in the embedding space. To alleviate this limitation, we propose Carmen, a cautiously designed graph embedding-based MR framework. Carmen consists of four components, including patient representation learning, context information extraction, context-aware GNN, and DDI encoding. Carmen incorporates the visit history into the representation learning of molecular graphs to distinguish molecules with similar topology but dissimilar activity. Its DDI encoding module is specially devised for the non-transitive interaction DDI graphs. The experiments on real-world datasets demonstrate that Carmen achieves remarkable performance improvement over state-of-the-art models and can improve the safety of recommended drugs with proper DDI graph encoding.

Introduction

To benefit from the escalating growth of the volume of electronic health records (EHR), many deep learning models [Shang et al. 2019a,b; Choi et al. 2017, 2016b; Yang et al. 2021; Choi et al. 2016a] have been proposed to mine EHR efficiently. Specifically, a promising and essential application in healthcare is medication recommendation (MR) [Yang et al. 2021; Shang et al. 2019b,a], which aims at recommending medication combinations for patients according to their history EHR.

Medical concept representation, which represents and preserves the relationship between medical concepts in low-dimensional subspaces, has been adopted to aid deep learning models to improve the prediction accuracy and efficiency for MR tasks and has a substantial impact on the performance of MR models. With proper medical concept embedding, the recommendations can be achieved in the embedding spaces by measuring the "distance" between visits and medications. Due to the extreme complexity of EHR data and insufficient labeled data in MR tasks, learning appropriate representation for each medical concept is not trivial. Therefore, most existing approaches incorporate some existing medical "knowledge", among which molecular structures of medications are the most important, to enhance medical representation learning. The study of the molecular structures can be traced back to "molecular descriptors" [Mauri et al. 2006] and "molecular fingerprints" [Rogers and Hahn 2010; Duvenaud et al. 2015], which are shallow models with mathematical conversions and algorithms. Recent work started to focus on applying deep models to represent molecular structures. [Shin et al. 2019] developed a model which ingests SMILES strings and uses a self-attention mechanism to learn the drug structure. With graph neural networks (GNN), the topology structure of the molecules is applied to the models, such as [Yang et al. 2021], which proposed Dual Molecular Graph Encoders to learn molecular representations.

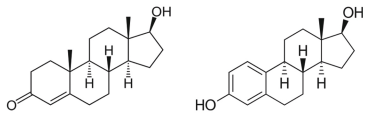

Figure 1: Testosterone (left) and Estradiol (right).

Molecular-based medication representation methods rely on graph representations of molecules, where atoms and bonds are represented by nodes and edges, respectively. However, converting molecules into graphs inevitably induces information loss as similar molecules might be converted to the same graph structure. In particular, the graph structures of pairs of stereoisomers are the same, even though they have different functionalities. Furthermore, as most existing methods simply adopt vanilla GNNs for molecular graph encoding, they encode medications with similar molecular graphs closer to the embedding space. Unfortunately, medications with similar molecular graphs or similar molecular 3D structures do not always indicate they have similar functionality. Fig. 1 illustrates that $Testosterone$ and $Estradiol$, which are two medications with completely different functionalities. With

*Corresponding author.

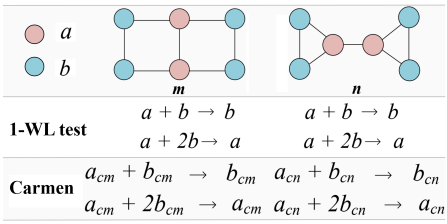

Figure 2: The molecule graphs for the medications m and n. The 1-WL test cannot distinguish them, but Carmen can. WL test updates one atom representation by taking an injective mapping on its neighborhood structure. Therefore, atom a/b will have the same representation in m and n as they have the same neighborhood. In comparison, the representation learned by Carmen considers the external context knowledge on how m and n are used with other medications.

vanilla GNNs, they would have almost identical representations, which might misguide the medication recommendation module to equal them and mistakenly recommend them to patients as another.

The crux of resolving the above issues lies in how to compensate for the information loss and empower GNNs to learn more distinguishable medication representations in case medications have similar molecular graphs but different functionalities. Motivated by the fact that medications with different functionality exhibit different co-occurring behaviors in EHR, it is important to properly utilize co-occurring medical concepts of each medication as its context information to favor the representation learning. Therefore, we proposed Carmen, a context-aware GNN module that enables GNN to inject the context information of each medication into its representations. For example, Fig. 2 demonstrates that Carmen is capable of distinguishing pairs of medications with similar molecular graphs but different use cases in the embedding space as Carmen considers their occurrence patterns in prescriptions while the 1-WL test (and vanilla GNNs) cannot.

In addition to the contextual information that can be learned from prescriptions, medication recommendations should also avoid drug combinations that have Drug-Drug Interactions (DDI) as they can lead to unintended reactions and side effects. DDI are usually provided/presented in a graph, where nodes represent drugs and edges represent interactions. Existing models either consider a DDI loss to regularize the objective function [Yang et al. 2021] or utilize a conventional message passing-based graph embedding to model the drug interactions [Shang et al. 2019b]. They directly or indirectly applied the "***transitivity***" feature, which plays an important role in learning social networks and knowledge graphs. "Transitivity" in a graph indicates that, if there is an edge between vertices v and u, and one between u and w, it is likely that v and w are also connected. However, we argue that the "transitivity" feature should not be applied to the DDI graph. For example, $digoxin$, which is in the $cardiac\ glycoside$ class of drugs, will cause gynecomastia and increase the risk of breast and uterus cancer while being administrated with $estrogens$. $digoxin$ is

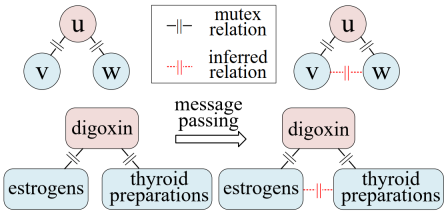

Figure 3: "Transitivity" via message passing results in inaccurate inference. The drugs of different colors are mutually exclusive. The link between the drugs declares DDI, which is a mutex relation. For the non-transitive DDI graph, the primitive message passing will mistakenly infer the existence of a mutex relation between v and w.

also suggested not to be taken with $thyroid\ preparations$, which may decrease the response to $digoxin$. But it is harmless to administrate $estrogens$ and $thyroid\ preparations$ together. Fig. 3 illustrates the example and its general form of "transitivity" inference, and we use "non-transitive" to describe the absence of transitivity. In comparison with existing models, Carmen applies reformative message dispelling to encode DDI with the "transitivity" feature off.

Our main contributions are summarized as follows:

- We recognized the issue of indistinguishable molecule graphs in MR tasks and developed a context-aware GNN that equips molecular graphs with the context information extracted from EHR, thus learning more distinguishable medication representations.
- We recognized the inappropriacy of encoding non-transitive DDI graphs with transitivity-favored message passing in existing work and proposed a non-transitive DDI encoding scheme. We theoretically demonstrated our method works better for embedding DDI graphs.
- The experiments proved that the use of context information improves the learned medication embeddings which in turn leads to more accurate recommendations. Meanwhile, with the help of proper DDI encoding, the recommendations eliminate unsafe drug combinations.

Related Work

Representation learning aims to convert the observed data into low-dimensional data informatively. Consequently, learning effective medical representations, primarily the representations of medical codes and patients' visit records [Choi et al. 2017; Shang et al. 2019b,a], has become an important topic in healthcare-related research. Based on the Skip-gram model, Med2vec [Choi et al. 2016a] learns representations by considering the co-occurrence information of the clinical concepts in EHR. To tackle the issue of the absence of long-term information in Med2Vec, [Choi et al. 2016b] proposed RETAIN to model the history of sequential dependencies. Leap [Zhang et al. 2017] generalizes treat recommendation to a sequential decision-making process with label dependency and label instance mapping considered. In comparison, besides the contextual information about medication co-occurrence, we also consider the repre-

sentation of molecular and DDI information in MR tasks.

Molecular information is pivotal for capturing the structure and property of drugs. CASTER [Huang et al. 2020] developed a deep auto-encoding module that takes SMILE strings as input to represent sub-structures of drugs. Nevertheless, SMILES owns a sequential structure and neglects the spatial information of the molecule. Alternatively, as GNNs demonstrated their effectiveness for processing topology structures, they have been widely applied to molecule representation. For example, GMPNN [Nyamabo et al. 2022] utilizes edges in molecular graphs as gates to control the flow of message passing, and CMPNN [Song et al. 2020] applies a communicative kernel to improve the molecular embedding by strengthening the message interactions. For MR tasks, SafeDrug [Yang et al. 2021] develops dual molecular graph encoders to embed global and local molecular structures. However, the insufficient encoding ability of vanilla GNNs limits the advantage of utilizing molecular knowledge [Beani et al. 2021]. And learning the effective embedding of molecules for MR remains a challenge.

Avoiding adverse DDI in drugs or/and predicting DDI in prescriptions via machine learning models have been widely studied, aiming to prevent adverse effects triggered by DDI in treatment and diagnosis. In existing work, leveraging the DDI information can be achieved by either developing a DDI loss function, such as SafeDrug [Yang et al. 2021], or regularizing the medication concept representation learning on the DDI graph, such as GAMENet [Shang et al. 2019b]. Moreover, SMR [Gong et al. 2021] utilizes a knowledge graph, where DDI appears as the relation connecting entities. However, as we argued before, the existing approaches simply utilize the conventional transitivity-favored message-passing scheme, neglecting the fact that the interactions in DDI graphs are non-transitive.

Basic Notation

Three sets of medical concepts are considered in the paper, including diagnosis, procedure, and medication, which are represented as $\mathcal{D} = \{\mathbb{d}_1, \mathbb{d}_2, \ldots, \mathbb{d}_{|\mathcal{D}|}\}$, $\mathcal{P} = \{\mathbb{p}_1, \mathbb{p}_2, \ldots, \mathbb{p}_{|\mathcal{P}|}\}$, and $\mathcal{M} = \{\mathbb{m}_1, \mathbb{m}_2, \ldots, \mathbb{m}_{|\mathcal{M}|}\}$, respectively. Formally, the EHR of each patient is a sequence of hospital visits, $< V^1, V^2, \ldots, V^{\mathcal{T}} >$, where V^t denotes the t^{th} hospital visit. For the t^{th} visit of i^{th} patient, V_i^t, can be presented as a triplet (d_i^t, p_i^t, m_i^t), where $d_i^t \in \{0,1\}^{|\mathcal{D}|}, p_i^t \in \{0,1\}^{|\mathcal{P}|}$, and $m_i^t \in \{0,1\}^{|\mathcal{M}|}$. For simplicity, in the following sections, we omit the patient subscript when there is no ambiguity. Table 1 shows the key notations in the paper.

Proposed Model

We propose a medication recommendation model with context-aware GNN (Carmen) to learn more distinguishable medication representations, thus making better medication recommendations. Fig. 4 illustrates the architecture of Carmen, which consists of four major components: patient representation learning, context information extraction, context-aware GNN, and DDI encoding.

The patient representation learning module processes diagnosis and procedure codes from the visit sequence and returns the patient representation. The context information extraction module distills context information for each medication from three co-occurrence matrices, yielding abstracted information. The context-aware GNN module skillfully injects the context information into message passing to enable GNN to distinguish medication with similar molecular structures. Meanwhile, the DDI encoding module properly represents the drugs in the non-transitive DDI graph, and the generated drug embeddings are combined with the representations from the context-aware GNN. Finally, we can make recommendations for each patient based on their representations and the learned medication embedding matrix.

Notation	Description
E_d	Diagnosis Embedding matrix
E_p	Procedure Embedding matrix
E_m	Medication Embedding matrix
h^t	Patient representation for the $t^{th} visit$
A_{md}	Medication-diagnosis co-occurrence matrix
A_{mp}	Medication-procedure co-occurrence matrix
A_{mm}	Medication-medication co-occurrence matrix
$\hat{y}$	Prediction of the current visit
y	Ground truth of the current visit

Table 1: Key Notations

Patient Representation Learning

To learn patient representations, given the historical diagnosis and procedure codes from the visit records $< V^1, V^2, \ldots, V^t >$, we start with encoding the diagnosis and procedure of each visit.

Visit Representation. Given a triplet of multi-hot vector (d^t, p^t, m^t) denoting the t^{th} visit, we convert d^t/p^t to a low-dimensional d_e^t/p_e^t by multiplying embedding matrix E_d/E_p with original multi-hot vector d^t/p^t. $E_d \in \mathbb{R}^{|\mathcal{D}| \times l}$ and $E_p \in \mathbb{R}^{|\mathcal{P}| \times l}$ denote the embedding matrix of diagnosis and procedure, respectively. The superscript l is the dimensionality of medical concept representation.

Patient Representation. Similar to the healthcare scenario where doctors always refer to patients' medical history to make a diagnosis, we utilize two GRUs to process the $< d_e^1, d_e^2, \ldots, d_e^t >$ sequence and the $< p_e^1, p_e^2, \ldots, p_e^t >$ sequence to capture longitudinal information of diagnosis view and procedure view:

$$d_h^t = GRU_d(d_e^t, d_h^{t-1}), \quad p_h^t = GRU_p(p_e^t, p_h^{t-1}), \qquad (1)$$

where $d_h^t, p_h^t \in \mathbb{R}^l$. To combine both diagnosis and procedure information, we use $h^t = W_h\,[d_h^t; p_h^t]$, where $[;]$ is the concatenate operator and W_h is a learnable weight matrix in $\mathbb{R}^{l \times 2l}$. Thus, we obtain the final patient representation $h^t \in \mathbb{R}^l$, which includes all the procedure and diagnosis information of the patient.

Context-Aware Medication Representation Learning

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph, $\mathcal{V}$ denotes the node set, $\mathcal{E}$ denotes the edge set, and $n = |\mathcal{V}|$ denotes the number

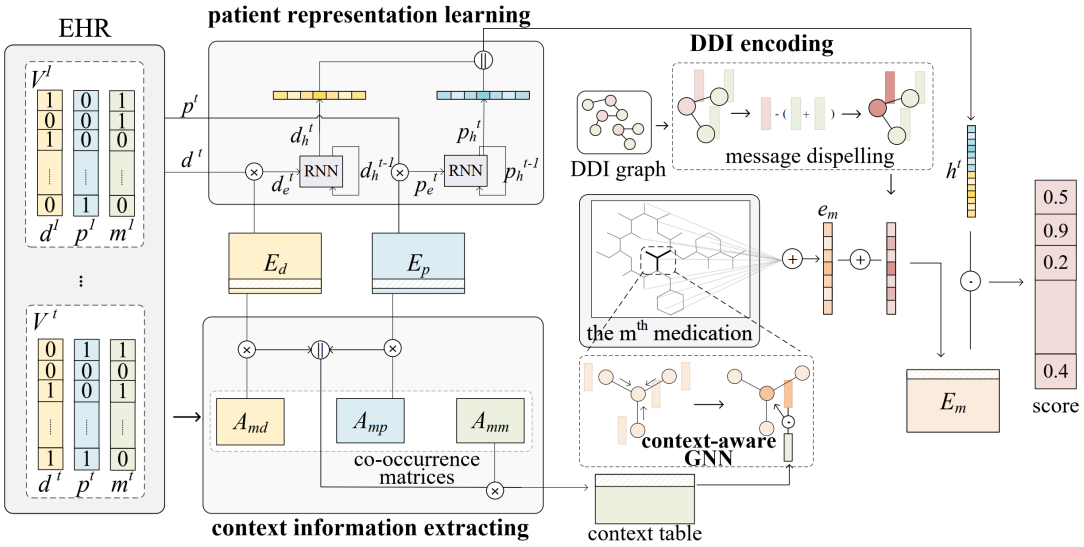

Figure 4: The framework of Carmen. Carmen has four main modules: 1) Patient representation learning module takes diagnosis d^t and procedure p^t as inputs to get the patient representation h^t; 2) Context information extracting module generates the context table C according to the three co-occurrence matrices A_{md}, A_{mp}, and A_{mm}; 3) Context-aware GNN injects the m^{th} medication's context C^m into its molecule representation; 4) DDI encoding module leverages message dispelling to separate adjacent drugs from the DDI graph. Based on patient representation h^t and medication representations E_m, the prediction score is computed for the patient.

of nodes in the graph. Given a node $v \in \mathcal{V}$, $\mathcal{N}(v) := \{u \mid \{u, v\} \in \mathcal{E}\}$ is the set of v's neighbors. In GNNs, each node v receives messages from its neighbors:

$$\begin{aligned} z^k_{\mathcal{N}(v)} &= \mathcal{AGG}(z^{k-1}_u, \forall u \in \mathcal{N}(v)), \\ z^k_v &= \mathcal{UPD}(z^{k-1}_v, z^k_{\mathcal{N}(v)}). \end{aligned} \tag{2}$$

Eq. (2) depicts the layer-wise operation of GNN, where $z^k_{\mathcal{N}(v)}$ denotes the output of the function $\mathcal{AGG}(\cdot)$ which gathers neighborhood information at GNN's k^{th} layer, and the function $\mathcal{UPD}(\cdot)$ updates node embedding with its corresponding aggregated neighborhood information. To learn the graph-level representations, a readout function $\mathcal{R}(\cdot) : \mathbb{R}^{n \times l} \longmapsto \mathbb{R}^l$ is then adopted to integrate the node representations learned by GNNs [Hassani and Khasahmadi 2020].

As discussed before, vanilla GNNs are inborn defective for medication representation due to the inadequate distinguishing power to medications with similar molecular structures. Therefore, we inject medication-specific information into the message-passing process of GNN.

Context Information Extraction. As different medications would exhibit different co-occurring behaviors in EHR datasets, we first construct three co-occurrence matrices from the training set: medication-diagnosis co-occurrence matrix $A_{md} \in \mathbb{R}^{|\mathcal{M}| \times |\mathcal{D}|}$, medication-procedure co-occurrence matrix $A_{mp} \in \mathbb{R}^{|\mathcal{M}| \times |\mathcal{P}|}$, and medication-medication co-occurrence matrix $A_{mm} \in \mathbb{R}^{|\mathcal{M}| \times |\mathcal{M}|}$. Rows of each matrix are normalized by L_1 $norm$.

Each medication is preliminarily represented by $C_d = A_{md}E_d$ and $C_p = A_{mp}E_p$, where $C_d, C_p \in \mathbb{R}^{|\mathcal{M}| \times l}$ can be considered as the preliminary medication context representations under the two views, which are then combined into the co-occurring information C_{dp} by $C_{dp} = [C_d; C_p]\, W_c$, where $W_c \in \mathbb{R}^{2l \times l}$ is also a learnable parameter matrix.

The combination information is captured by $C_{mm} = A_{mm}C_{dp}$, and is integrated with the co-occurring information C_{dp} by $C = C_{dp} + tanh\,(C_{dp}W_{s1}) \odot C_{mm}$, where $W_{s1} \in \mathbb{R}^{l \times l}$ taken by activation function $tanh(\cdot)$ is a feature attention layer, which aims to adaptively select valuable features in C_{mm} and filter out the trivial ones according to C_{dp}. $\odot$ denotes the element-wise product. Finally, we get the context information C of medications.

Context-aware GNN. The molecular graph of the m^{th} medication is represented as $\mathcal{G}_m = (\mathcal{V}_m, \mathcal{E}_m)$, where $\mathcal{V}_m$ and $\mathcal{E}_m$ denote the set of atoms and edges, respectively. Chemical bonds are modeled as edges. We aggregate the neighborhood information for each atom $v \in \mathcal{V}_m$ with

$$z^k_{\mathcal{N}(v)} = \sum_{\forall u \in \mathcal{N}(v)} \frac{W^k z^{k-1}_u}{\sqrt{a_u a_v}}, \tag{3}$$

where $W^k \in \mathbb{R}^{l \times l}$ is the weight matrix of k^{th} layer. a_u and a_v are the degrees of atom u and atom v, indicating the number of chemical bonds connecting them.

As existing methods are limited by the distinguishing power of vanilla GNNs, we design a novel aggregation form for atoms in the molecular graph, wherein each atom is encoded by its neighborhood information and the additional graph-level medication context information. The neighborhood information of each atom and the context embedding of medication C^m are aggregated as below:

$$\hat{z}^k_{\mathcal{N}(v)} = tanh\,(W_{s2}C^m) \odot z^k_{\mathcal{N}(v)}, \tag{4}$$

where $W_{s2} \in \mathbb{R}^{l \times l}$ with $tanh(\cdot)$ is another feature attention

layer, and $C^m \in \mathbb{R}^l$ is the m^{th} row of C for the m^{th} medication. Then, $\hat{z}^k_{\mathcal{N}(v)}$ together with a single self-connection representation z^{k-1}_v are integrated to infer the representation of atom v in k^{th} layer:

$$z^k_v = \epsilon z^{k-1}_v + \hat{z}^k_{\mathcal{N}(v)}, \tag{5}$$

where ϵ is a hyper-parameter for balancing the weight between the atom v and its neighbors. To summarize the atom representations into a graph-level medication representation, we leverage the readout function $\mathcal{R}(\cdot)$:

$$e_m = \mathcal{R}\left(\left\{z^K_v, \forall v \in \mathcal{V}\right\}\right), \tag{6}$$

where K denotes the layer number of GNN and e_m is the final representation of the m^{th} medication. Each medication is encoded by context-aware GNN in parallel and stored in medication embedding matrix $E_m \in \mathbb{R}^{|\mathcal{M}|\times l}$.

DDI Encoding

As we have highlighted, message passing-based encoding schemes, including conventional GCNs, cannot be simply applied to DDI graphs as they are non-transitive in nature. In this section, we propose an encoding scheme that favors non-transitive DDI graphs.

A DDI graph is an undirected graph $\mathcal{G}_{ddi} = (\mathcal{V}_{ddi}, \mathcal{E}_{ddi})$, where $\mathcal{V}_{ddi}$ and $\mathcal{E}_{ddi}$ denote its nodes and edges, respectively. Each node $v_{ddi} \in \mathcal{V}_{ddi}$ represents a medication and each edge $e_{ddi} \in \mathcal{E}_{ddi}$ indicates the presence of DDI between two medications. For DDI graphs, the conventional encoding process (Eq. (2)) is specified as:

$$\begin{aligned} z^k_{\mathcal{N}(v_{ddi})} &= \mathcal{AGG}_{ddi}(z^{k-1}_u, \forall u \in \mathcal{N}(v_{ddi})), \\ z^k_{v_{ddi}} &= \mathcal{UPD}_-(z^{k-1}_{v_{ddi}}, z^k_{\mathcal{N}(v_{ddi})}). \end{aligned} \tag{7}$$

We employ the attention mechanism [Veličković et al. 2017] for aggregation. As the connected drugs in a DDI graph in fact repel each other, their representations in a low-dimensional embedding space should be far apart. Consequently, the conventional $\mathcal{UPD}(\cdot)$ function (Eq. (5)) becomes invalid as it enforces neighbors to be close in the embedding space. Hence, we modify the update function as:

$$\mathcal{UPD}_-(z^{k-1}_{v_{ddi}}, z^k_{\mathcal{N}(v_{ddi})}) = \gamma z^{k-1}_{v_{ddi}} - z^k_{\mathcal{N}(v_{ddi})}, \tag{8}$$

where γ is the hyperparameter that adjusts the balance between drug v_{ddi} and its neighborhood. We refer to the process, including the aggregates and the dispelling updates, as the message dispelling, formally defined below.

Definition 1 (Message Dispelling)**.** *$G = (V, E)$ is an N-node undirected graph. A is the adjacency matrix of G and $D_i = \sum_j A_{ij}$. For nodes $V = \{v_1, \ldots, v_N\}$, X^t_i is the embedding of v_i after t times message dispelling, of which the basic form is $X^{t+1} = (I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}) X^t$.*

The following proposition proves that the embeddings of the drugs are estranged as the message dispelling proceeds, and the connected nodes are the first to be separated. Therefore, the message dispelling is practical for graphs with non-transitive structures, such as DDI graphs.

Proposition 1. *For a large enough $t = \tau$, $\|X^\tau_i - X^\tau_j\|_2 > \|X^0_i - X^0_j\|_2$.*

Proof. For a message dispelling

$$X^{t+1}_i = (I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}) X^t_i, \tag{9}$$

it can be rewritten as:

$$X^t_i = L X^{t-1}_i = L^t X^0_i, \tag{10}$$

where $L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ denotes the normalized graph Laplacian matrix. The eigendecomposition of L is $L = U \Lambda U^T$, where $U = ([u_1, ..., u_N])$, $u_i \in \mathbb{R}^N$ and $\Lambda = diag([\lambda_1, ..., \lambda_N])$. According to the properties of L, $\lambda_i \in [0, 2]$, and U is guaranteed to be an orthogonal matrix. Then Eq. (10) can be expanded as:

$$X^t_i = U \Lambda^t U^T X^0_i. \tag{11}$$

By setting $\hat{X} = U^T(X^0_i - X^0_j)$, we have

$$X^t_i - X^t_j = U \Lambda^t U^T (X^0_i - X^0_j) = U \Lambda^t \hat{X}. \tag{12}$$

We can induce that

$$\|X^t_i - X^t_j\|_2 = \sqrt{\sum_{j=1}^{N} (\sum_{i=1}^{N} \lambda^t_i u_{ij} \hat{x}_i)^2}, \tag{13}$$

where u_{ij} represents the j^{th} element of eigenvector u_i, and $\hat{x}_i$ denotes the i^{th} element of $\hat{X}$. With $\lambda_i \in [0, 2]$, when τ is large enough, we have

$$\sqrt{\sum_{j=1}^{N} (\sum_{i=1}^{N} \lambda^\tau_i u_{ij} \hat{x}_i)^2} > \sqrt{\sum_{j=1}^{N} (\sum_{i=1}^{N} u_{ij} \hat{x}_i)^2}. \tag{14}$$

Hence, we can conclude that $\|X^\tau_i - X^\tau_j\|_2 > \|X^0_i - X^0_j\|_2$. □

The embedding of the drugs learned by the DDI embedding module is then combined with e_m from Eq. (6) to facilitate the recommendation.

Prediction and Objectives

The matching score of each medication is the similarity between the patient representation h^t and the medication representation E_m as $\hat{y} = \sigma(LN(\tilde{E}_m \tilde{h}^t))$, where $\hat{y} \in \mathbb{R}^{|\mathcal{M}|}$, σ denotes the sigmoid activation function, and LN represents layer normalization operation. $\tilde{E}_m$ is row normalized E_m and $\tilde{h}^t$ is normalized h^t.

Objective. This paper formulates medication recommendation as a multi-class and multi-label classification task. First, we adopt binary cross-entropy (BCE) loss L_{bce} as part of the objective, and empirically utilize the multi-label hinge loss L_{margin}, aiming to keep a significant margin between the ground truth labels' scores and the others. Thus, the objective is the weighted sum of L_{bce} and L_{margin}:

$$\begin{aligned} L_{bce} &= -\sum_{i=1}^{|\mathcal{M}|} y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i), \\ L_{margin} &= -\sum_{i:y(i)=1} \sum_{j:y(j)=0} \frac{max\left(0, 1 - (\hat{y}_i - \hat{y}_j)\right)}{|\mathcal{M}|}, \\ L &= (1 - \alpha) L_{bce} + \alpha L_{margin}. \end{aligned} \tag{15}$$

Item	MIMIC-III	MIMIC-IV
# of patients	6350	9036
# of visits	15032	20616
Avg.# of visit per patient	2.37	2.28
Max.# of visit	29	28
# of unique diagnosis codes	1958	1892
# of unique procedure codes	1430	4939
# of unique medication codes	131	131

Table 2: Statistics of dataset

For the i^{th} medication, $\hat{y}_i$ is the prediction and y_i is the ground truth label. α is a predefined hyperparameter to control the proportion of two loss functions.

Experiments

We conducted extensive experiments for performance comparison between Carmen [1] and several state-of-the-art methods. To further verify the reliability and effectiveness of our model, two analytical studies are also provided.

Experimental Setting

Datasets. We evaluated our model on MIMIC-III [Johnson et al. 2016] and MIMC-IV [Johnson et al. 2018]. After the data preprocessing, we got 131 medications for recommendation when we set the ATC Third Level code as the target label. In more detail, each ATC Third Level code involves one or more medications and each medication corresponds to one ATC Third Level code.

Evaluation. We measured the performance with three common metrics, including Jaccard similarity, F1 score, and Precision-Recall AUC (PRAUC).

Baselines. To evaluate our work comprehensively, we compared **Carmen** with the state-of-the-art methods from different categories: 1) Shallow model baselines include Logistic Regression (**LR**) and Ensembles of Classifier Chains (**ECC**) [Read et al. 2011]; 2) Deep model baselines include **RETAIN** [Choi et al. 2016b], **Leap**[Zhang et al. 2017], **GAMENet**[Shang et al. 2019b], and **SafeDrug** [Yang et al. 2021]. We also introduce the variants of Carmen, including **Carmen w/o (context & ddi-enc)** which removes the context information injection and the DDI encoding [2] , **Carmen w/ ddi-loss** which replaces the DDI encoding with an additional DDI loss from SafeDrug, **Carmen w/ ddi-agg** which encodes the DDI graph by a conventional message passing, and **Carmen w/o ddi-enc** which removes the DDI encoding. Note that as we analyzed in Prop 1, message dispelling benefits from multiple layers while message passing will confront over-smoothing in this condition. For fairness, we set 2 layers for message passing and 9 for message dispelling.

[1]Code is available at https://github.com/bit1029public/Carmen.

[2]Carmen w/o (context & ddi-enc) simply utilizes the vanilla GNNs to encode molecules to represent medications.

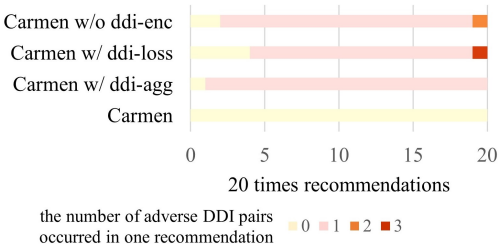

Figure 5: The number of occurrences of the adverse DDI among 20 times recommendations. Carmen w/o ddi-enc and Carmen w/ ddi-agg recommended one unsafe drug pair 17 times and 19 times (pink bars), respectively, and Carmen w/ ddi-loss recommended three unsafe drug pairs once (red bar). Carmen recommended none of unsafe drug pairs.

Performance Comparison

Carmen outperforms baselines. Table 3 lists the results of medication recommendations. Each model was executed five times with different seeds, and the mean and standard deviation of the results were presented. The results show that Carmen and its variants consistently outperform the baselines. The comparison between Carmen w/o ddi-enc and Carmen w/o (context & ddi-enc) demonstrates that the major performance gain is from the context information involved in the GNN forward process. Although both Carmen w/o (context & ddi-enc) and SafeDrug use the vanilla GNN to encode molecules, the design details of their graph encoders are significantly different, and SafeDrug has an extra local Bipartite encoder. Concretely, the degree of nodes has been proved to be discriminative information to encode graphs [Geerts, Mazowiecki, and Perez 2021] but is ignored in SafeDrug. This explains why Carmen w/o (context & ddi-enc) still achieves comparable performance.

DDI encoding guarantees safety. DDI knowledge can be included in two ways, DDI encoding and DDI loss. For DDI loss, it has trivial improvement for Carmen (Carmen w/ ddi-loss vs. Carmen w/o ddi-enc), notable upgrade for GAMENet, and almost no influence on SafeDrug. The reason is that the DDI knowledge is not only determined by the molecules but also implied in the visit records prescribed by physicians. GAMENet only utilizes the co-occurrence information so that extra DDI knowledge relieves the disadvantage of lacking molecular details (or other medication attributes). SafeDrug tends to make conservative recommendations to fit the DDI loss function, compromising accuracy as the DDI knowledge is not always consistent with visit records. The representation obtained from Carmen w/o ddi-enc is far more informative as it considers the contributions from both molecules and records, and it can be trained to dominate the output, thus loosening the negative constraint brought by DDI loss.

For DDI encoding, we observe that it even has some negative impact on accuracy as a trade-off for drug safety, which means drugs with DDI should not appear together. Given the fact that some EHR records have DDI in presence, instead of using the DDI rate [Shang et al. 2019b; Yang et al. 2021] as a metric, we evaluated the effectiveness brought by dif-

Method	MIMIC-III			MIMIC-IV		
	Jaccard (%)	**Prauc** (%)	**F1-score** (%)	**Jaccard** (%)	**Prauc** (%)	**F1-score** (%)
LR	$49.40_{\pm 00.14}$	$75.88_{\pm 00.19}$	$65.08_{\pm 00.11}$	$47.32_{\pm 00.18}$	$73.80_{\pm 00.16}$	$63.03_{\pm 00.17}$
ECC	$48.36_{\pm 00.14}$	$75.86_{\pm 00.18}$	$63.96_{\pm 00.10}$	$44.03_{\pm 00.10}$	$71.39_{\pm 00.06}$	$60.16_{\pm 00.09}$
RETAIN	$48.55_{\pm 00.15}$	$75.68_{\pm 00.12}$	$64.67_{\pm 00.12}$	$44.03_{\pm 00.10}$	$71.40_{\pm 00.02}$	$60.13_{\pm 00.08}$
Leap	$45.44_{\pm 00.18}$	$65.71_{\pm 00.30}$	$61.63_{\pm 00.17}$	$43.50_{\pm 00.23}$	$63.21_{\pm 00.40}$	$59.63_{\pm 00.22}$
GAMENet w/o ddi-loss	$50.93_{\pm 00.08}$	$76.24_{\pm 00.24}$	$66.56_{\pm 00.08}$	$47.50_{\pm 00.08}$	$73.96_{\pm 00.13}$	$63.33_{\pm 00.07}$
GAMENet	$51.59_{\pm 00.09}$	$76.84_{\pm 00.12}$	$67.15_{\pm 00.07}$	$47.84_{\pm 00.18}$	$73.95_{\pm 00.28}$	$63.63_{\pm 00.17}$
SafeDrug w/o ddi-loss	$50.42_{\pm 00.23}$	$76.11_{\pm 00.04}$	$66.20_{\pm 00.20}$	$48.52_{\pm 00.41}$	$74.09_{\pm 00.24}$	$64.31_{\pm 00.37}$
SafeDrug	$50.35_{\pm 00.15}$	$75.76_{\pm 00.18}$	$66.14_{\pm 00.39}$	$48.57_{\pm 00.62}$	$74.00_{\pm 00.31}$	$64.35_{\pm 00.54}$
Carmen w/o (context & ddi-enc)	$51.20_{\pm 00.14}$	$74.57_{\pm 00.07}$	$66.90_{\pm 00.12}$	$48.74_{\pm 00.27}$	$71.90_{\pm 00.18}$	$64.53_{\pm 00.23}$
Carmen w/o ddi-enc	$53.13_{\pm 00.11}$	$77.35_{\pm 00.23}$	$68.56_{\pm 00.08}$	$50.27_{\pm 00.09}$	$75.09_{\pm 00.09}$	$65.90_{\pm 00.08}$
Carmen w/ ddi-agg	$53.15_{\pm 00.09}$	$77.19_{\pm 00.14}$	$68.58_{\pm 00.08}$	$50.33_{\pm 00.11}$	$75.07_{\pm 00.10}$	$65.99_{\pm 00.11}$
Carmen w/ ddi-loss	$\mathbf{53.23_{\pm 00.18}}$	$\mathbf{77.36_{\pm 00.21}}$	$\mathbf{68.65_{\pm 00.16}}$	$\mathbf{50.49_{\pm 00.09}}$	$\mathbf{75.13_{\pm 00.13}}$	$\mathbf{66.15_{\pm 00.11}}$
Carmen	$52.67_{\pm 00.21}$	$76.52_{\pm 00.36}$	$68.12_{\pm 00.19}$	$50.06_{\pm 00.12}$	$74.62_{\pm 00.30}$	$65.69_{\pm 00.07}$

Table 3: Performance comparison on MIMIC-III and MIMIC-IV. Numbers in bold indicate the best performance.

ferent methods of applying DDI information. We tested each model 10 times on MIMIC-III&IV respectively, and counted the number of appearances of the "adverse DDI", which represents the DDI not appearing in the prescriptions of the test dataset. It can be observed from Fig. 5 that Carmen does not recommend any unsafe drug combinations. Conversely, Carmen w/ ddi-loss fails to reduce unsafe drug combinations. Likewise, Carmen w/ ddi-agg cannot guarantee not to recommend any adverse DDI. These results prove that our DDI encoding module can handle the inherent property of the non-transitive DDI graph and captures the relation between drugs to ensure the safety and reliability of drug use, but inevitably decreases the accuracy as some DDI are in the EHR test set. In contrast, the DDI loss function focuses more on numerical accuracy. It is deficient in capturing and leveraging the concrete DDI information between drugs, leading to the lack of inductive ability of the model.

Carmen improves the distinguishing power of GNN. To measure the impact of molecule similarity on making predictions, we introduced a "confusion index" η_i, which indicates how much confusion for the i^{th} medication is due to other medications with similarities to it:

$$\eta_i = \frac{\sum_{j!=i} n_j s_{ij}}{n_i + \sum_{j!=i} n_j s_{ij}}. \tag{16}$$

n_j and n_i denote the number of occurrences of the j^{th} and i^{th} medications in the training data. s_{ij} denotes the molecular similarity between i^{th} medication and j^{th} medication, which is defined by Dice similarity on their ECMP (Extended Connectivity Fingerprints) [Rogers and Hahn 2010]. When the molecular structure of medication is unique, the confusion index reaches its minimum value (0). Whereas, when the dataset is dominated by one single molecular structure, the confusion index approaches its maximum value (1). Therefore, the "confusion" provides a way to assess how challenging it is to predict a medicine properly.

We identified the medications that Carmen w/o ddi-enc always predicted better than the baseline models in all five rounds of experiments and obtained their respective confusion indexes. Then we plot the average improvement between two compared models on the Jaccard index (y-axis) with respect to the "confusion index" (x-axis) in Fig. 6. The x-axis starts from 0.95 and we calculated the average improvement within every 0.01 interval. It is evident that the majority of the medications that are better predicted have a high confusion index η, and the larger η becomes, the more significant the improvement of the Jaccard index is, indicating that the major gains in our model are from the medications with the larger η. This proves that our model is capable of differentiating the medications with similar molecules more effectively.

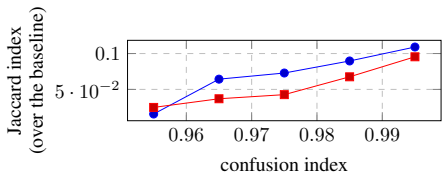

Figure 6: Model improvement regarding confusion index. Blue line represents Carmen w/o ddi-enc−Carmen w/o (context & ddi-enc), and red line represents Carmen w/o ddi-enc−SafeDrug.

Conclusion

This paper proposed a novel context-aware GNN (Carmen) for medication recommendations. Carmen extracts context information for each medication and injects it into GNN forward process, improving the distinguishing power of the vanilla GNNs. Notably, a DDI encoding module is developed to properly embed drugs, remedying the defect of the conventional message passing applied in the non-transitive DDI graph. The experimental results show that the proposed model remarkably outperforms state-of-the-art methods. We also verified that the major improvement is attributed to the context-aware GNN, and DDI encoding ensures the safety and reliability of the recommendation.

Ethics Statement

Although our model was tested on public anonymous data (MIMIC-III&IV) with ethical approval, its future applications may involve patient data collection and access, which should follow the principle of respect for autonomy. In practice, the prediction generated by the model is meant to assist physicians and is by no means replacing them.

Acknowledgments

This work has been partially supported by NSFC under Grant No. 62276024 and No. 92270125.

References

Beani, D.; Passaro, S.; Létourneau, V.; Hamilton, W.; Corso, G.; and Liò, P. 2021. Directional graph networks. In *International Conference on Machine Learning*, 748–758. PMLR.

Choi, E.; Bahadori, M. T.; Searles, E.; Coffey, C.; Thompson, M.; Bost, J.; Tejedor-Sojo, J.; and Sun, J. 2016a. Multi-layer representation learning for medical concepts. In *proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1495–1504.

Choi, E.; Bahadori, M. T.; Song, L.; Stewart, W. F.; and Sun, J. 2017. GRAM: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 787–795.

Choi, E.; Bahadori, M. T.; Sun, J.; Kulas, J.; Schuetz, A.; and Stewart, W. 2016b. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. *Advances in Neural Information Processing Systems*, 29: 3504–3512.

Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; and Adams, R. P. 2015. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Advances in Neural Information Processing Systems*, 28: 2224–2232.

Geerts, F.; Mazowiecki, F.; and Perez, G. 2021. Let's agree to degree: Comparing graph convolutional networks in the message-passing framework. In *International Conference on Machine Learning*, 3640–3649. PMLR.

Gong, F.; Wang, M.; Wang, H.; Wang, S.; and Liu, M. 2021. SMR: medical knowledge graph embedding for safe medicine recommendation. *Big Data Research*, 23: 100174.

Hassani, K.; and Khasahmadi, A. H. 2020. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*, 4116–4126. PMLR.

Huang, K.; Xiao, C.; Hoang, T.; Glass, L.; and Sun, J. 2020. Caster: Predicting drug interactions with chemical substructure representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 702–709.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Li-wei, H. L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3: 160035.

Johnson, A. E. W.; Stone, D. J.; Celi, L. A.; and Pollard, T. J. 2018. The MIMIC Code Repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1): 32–39.

Mauri, A.; Consonni, V.; Pavan, M.; and Todeschini, R. 2006. DRAGON software: an easy approach to molecular descriptor calculations. *Match Communications in Mathematical & in Computer Chemistry*, 56(2): 237–248.

Nyamabo, A. K.; Yu, H.; Liu, Z.; and Shi, J.-Y. 2022. Drug–drug interaction prediction with learnable size-adaptive molecular substructures. *Briefings in Bioinformatics*, 23(1): bbab441.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2011. Classifier chains for multi-label classification. *Machine learning*, 85(3): 333–359.

Rogers, D.; and Hahn, M. 2010. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.*, 50(5): 742–754.

Shang, J.; Ma, T.; Xiao, C.; and Sun, J. 2019a. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346*.

Shang, J.; Xiao, C.; Ma, T.; Li, H.; and Sun, J. 2019b. Gamenet: Graph augmented memory networks for recommending medication combination. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1126–1133.

Shin, B.; Park, S.; Kang, K.; and Ho, J. C. 2019. Self-attention based molecule representation for predicting drug-target interaction. In *Machine Learning for Healthcare Conference*, 230–248. PMLR.

Song, Y.; Zheng, S.; Niu, Z.; Fu, Z.-H.; Lu, Y.; and Yang, Y. 2020. Communicative Representation Learning on Attributed Molecular Graphs. In *IJCAI*, 2831–2838.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Yang, C.; Xiao, C.; Ma, F.; Glass, L.; and Sun, J. 2021. SafeDrug: Dual Molecular Graph Encoders for Safe Drug Recommendations. *CoRR*, abs/2105.02711.

Zhang, Y.; Chen, R.; Tang, J.; Stewart, W. F.; and Sun, J. 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.