FFBDNet: Feature Fusion and Bipartite Decision Networks for Recommending Medication Combination

Zisen Wang[1,2] , Ying Liang[1], and Zhengjun Liu[1,2]

[1] Research Center for Ubiquitous Computing Systems, Institute of Computing Technology, Chinese Academy of Science, Beijing, China
liangy@ict.ac.cn
[2] School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China
{wzs_0120, 17863102086}@163.com

Abstract. Recommending medication combinations for patients is an essential part of artificial intelligence in the healthcare field. Existing approaches improve the effect of recommendations by considering how to make full use of patients' electronic health records or by introducing additional external knowledge, but there is still room for improving the fusion of heterogeneous and diverse knowledge and the effect between accuracy and drug-drug interaction (DDI) rate. To fill this gap, we propose the Feature Fusion and Bipartite Decision Networks (FFBDNet) to leverage external knowledge and improve accuracy and DDI rate. FFBDNet is equipped with a patient feature encoder which extract useful information from current and historical visits of patient to supplement the patient's health status, a medication feature encoder which can easily fuse the heterogeneous and diverse external knowledge of medications as feature, and a bipartite decision module to give medication recommendation results. FFBDNet also has a greedy loss function to improve accuracy and DDI rate. We demonstrate the effectiveness of FFBDNet by comparing with several state-of-the-art methods on a benchmark dataset. FFBDNet outperformed all baselines in all effective measures, reduced relatively the DDI rate by 97.65% from existing EHR data, and also is shown to improve 1.02% on Jaccard similarity.

Keywords: Medication combination prediction, External knowledge, Drug-Drug interaction, Data mining, Attention.

1 Introduction

Today, abundant health data, such as longitudinal electronic health records (EHR) and massive medical data available on the web enable researchers and doctors to build better predictive models for clinical decision making [1, 2]. Among other things, recommending effective and safe medication combinations is an important task, in particular to help patients with complex medical conditions [3, 4], and the primary objective is to personalize a safe combination of medications for a particular patient based

on the patient's electronic health records. In recent years, more and more researchers try to use neural network to model the recommendation process, so as to assist doctors make better and more efficient clinical decisions when facing a large number of patients. There are basically two types of these approaches: 1) Sequential decision-making models that look at recommending medication combinations to patients as a multi-step decision-making task, see [5-8].However, most decision-making tasks require a predetermined order or an appropriate reward function, which is difficult to define and will eventually affect the effect of the recommendation. 2) Multi-label classification models such as [4, 9-11] that view the medication combination recommendation as a multi-label classification task, so as to avoid the rationality of the order of the medication recommendation in the model prediction. However, they still suffer from the following limitations.

Fuse of External Knowledge. External knowledge refers to the medical data other than EHR, such as age and gender of patients, conflict relationship and molecular structure of medications, and in the medication combination recommendation, it usually refers to the external knowledge of medications. Existing works [4, 12] improve the effect of recommendation by introducing additional external knowledge of medications, but they have poor scalability for new external knowledge. New external knowledge can usually introduce new information for recommendation tasks, and better fusion of external knowledge can better support the model.

Effect between Accuracy and DDI rate. In medication combination recommendation, it is very important to avoid unnecessary drug-drug interaction as much as possible, so as to ensure the safety of recommendation results. Some existing works [8, 13] improve the accuracy and DDI rate for recommendation by explicitly or implicitly introducing DDI knowledge into training, such as implicitly adjusting DDI rate through reward function, or directly designing DDI loss to reduce DDI rate. However, there is still room for improvement in the effect between accuracy and DDI rate. Especially for the DDI rate, as the essential factor to measure the safety of medication combination recommendation, the DDI rate of the existing works is still at a high level.

To address these, we propose a Feature Fusion and Bipartite Decision Networks for medication combination recommendation, named FFBDNet, to fuse the external medical knowledge and to improve recommend effect. We believe that different external knowledge can introduce new information to assist recommendation. Our FFBDNet has the following contributions.

We propose a feature fusion module to fuse heterogeneous and diverse knowledge. The attention mechanism is used to extract the previous medical visit information related to the patient's current visit. A variety of non Euclidean space features of medications are encoded by graph convolution network. By concatenating new external knowledge in the feature coding stage, it can easily realize the fusion of external features.

We propose a bipartite decision module to make a joint decision for medication recommendation. It consists of two doctor models: direct doctor and recombination doctor. The direct doctor directly uses the patient's representation for recommendation, and the recombination doctor recombines the medications based on the similarity

between the patient and the drugs. Finally, the recommendation results of the two doctor models are fused to complete the joint decision-making.

We design a greedy loss to reduce the DDI rate of medication combination recommendation results. The greedy mask is used to filter high conflict medications in greedy loss, and experiments show that, compared with several state-of-the-art methods on real EHR data, greedy loss can avoid almost all DDI in the medication combination, while still maintaining a good recommendation accuracy.

2 Related Works

2.1 Medication Recommendation

The existing medication combination recommendation methods can be basically divided into two types: sequential decision-making and multi-label classification. Sequential decision-making models decompose one recommendation process into multi-step medication decision-making, see [5-8, 14]. For example, LEAP[6] uses recurrent neural network (RNN) to model the decision-making process, and uses content-based attention mechanism to capture label instance mapping to predict medication at each step. COMPNet[8] transforms the medication combination recommendation task into a disordered Markov decision process (MDP) problem, and designs a deep Q-learning (DQL) mechanism to learn the correlation and adverse interactions between medications. Multi-label classification models realize medication combination recommendation by predicting multiple labels for patients at one time, see [4, 10-13, 15]. Among them, GAMENet[4] customizes a memory storage module for external knowledge and extract external features from EHR graph and DDI graph by graph convolution network, so as to improve the effect of multi-label classification for medication recommendation. SafeDrug[12] specially designs an encoder to capture drug molecular knowledge, which is composed of global message passing neural network (MPNN) and local bipartite learning module, explicitly models the medication conflict process, and realizes medication recommendation to patients. Despite their initial success, there is still room for improvement in the effect between accuracy and DDI rate, as well as the poor fusion of additional external knowledge caused by structural customization.

In view of the success of the existing works through the use of external knowledge, in this paper, we design a feature fusion module that is easy to fuse the external features for the medication combination recommendation task, and design a greedy loss to optimize the effect between accuracy and DDI rate.

2.2 Medication Representation

The medical data related to medication is often non Euclidean space structure, which is often modeled by graph convolution neural network (GCN) in the existing works. Initializing each node in non Euclidean space data, GCN uses neighbor iterative aggregation to update nodes, and finally obtains the informative latent feature representations of each node [16-20]. At the beginning, it achieved good results in social networks. And with the development, it has been successfully applied in the field of

medicine in recent years. For example, Ma et al. [21] use GCN to encode each node in the medical graph to obtain an interpretable embedded representation of the medication. Zitnik et al. [22] construct a two-layer multimodal medication interaction graph, and use GCN to capture the conflict relationship between medications. The representations of medication molecules are commonly modeled by molecular descriptors [23] and medication fingerprint [24], and David et al. [25] use GCN to capture the deep semantic features of medication fingerprint. Huang et al. [26] use medication pairs to capture medical features, and directly model medication molecule graph based GCN [27].

In this paper, we will use GCN to encode a variety of non Euclidean space medical data of medications, so as to capture and utilize the medication feature of different knowledge sources.

3 Problem Formulation

Electrical Health Records (EHR). In longitudinal EHR data, each patient *n* can be represented as a sequence of multivariate observations: $R^{(n)} = [r_1^{(n)}, r_2^{(n)}, \cdots, r_{T^{(n)}}^{(n)}]$ where $n \in \{1,2,\cdots,N\}$, N is the total number of patients; $T^{(n)}$ is the number of visits of the n-th patient. To reduce clutter, the algorithms will be described for a single patient and drop the superscript (n) whenever it is unambiguous. Each history record $r_t = [c_t^d, c_t^p, c_t^m](t < T)$ of a patient for t-th visit is concatenation of corresponding diagnoses codes c_t^d, procedure codes c_t^p and medications codes c_t^m. And current record $r_T = [c_T^d, c_T^p]$ of a patient is concatenation of corresponding diagnoses codes c_T^d, procedure codes c_T^p. For simplicity, c_t^* is used to indicate the unified definition for different type of medical codes. $c_t^* \in \{0,1\}^{|C^*|}$ is a multi-hot vector, where C^* is the medical code set and $|C^*|$ is size of set C^*.

External Knowledge of Medication. In this paper, there are three kinds of external knowledge of medication: EHR graph, DDI graph and molecule graph. EHR graph contains the co-occurrence knowledge of medications, and can be denoted as $G^E = \{V^E, E^E\}$, where $V^E = C^m$ is the node set of all medications and E^E is the edge set of known combination medication in EHR database. DDI graph contains the conflict knowledge between medications, and can be denoted as $G^D = \{V^D, E^D\}$, where $V^D = C^m$ is the node set of all medications and E^D is the edge set of known DDIs between a pair of medications. molecule graph A contains the molecular composition knowledge of medications, which is similar to the root word in natural language processing, and can be denoted as $G^{m_i} = \{V^{m_i}, E^{m_i}\}$, where V^{m_i} is the node set of all molecular units of medication $m_i \in C^m$ and E^{m_i} is the edge set of known molecular structure of M_i. For simplicity, G^* is used to indicate the unified definition for different type of medical knowledge graphs, and adjacency matrix $A^* \in \mathbb{R}^{|V^*|\times|V^*|}$ is defined to clarify the construction of edge E^*.

Medication Combination Recommendation. Given medical codes of the current visit at time T (excluding medication codes) c_T^d, c_T^p, patient history $[r_1, r_2, \cdots, r_{T-1}]$ and external knowledge graph G^E, G^D, G^{m_i}, we want to recommend

multiple medications by predicting multi-label output, while the predicted results are as close to the ground truth as possible and the DDI rate is as low as possible.

4 The FFBDNet

As illustrated in Fig. 1, FFBDNet includes the following components: a patient feature encoder, a medication feature encoder, and a bipartite decision module. Next, we will first introduce these modules and then provide details of training and inference of FFBDNet.

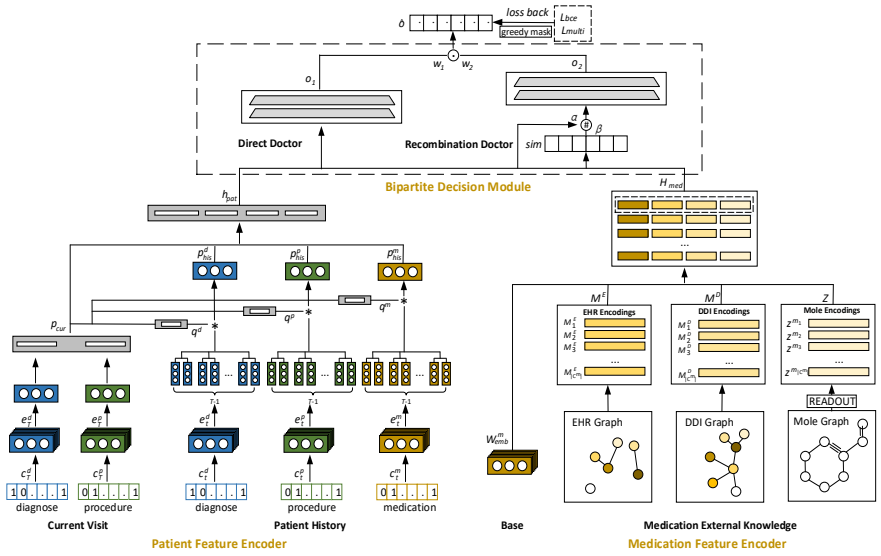

Fig. 1. The FFBDNet: We first encode current visit and patient history by attention mechanism to generate the patient health representation h_{pat} in Eq. (1-5). Then, we encode and concatenate the basic and external knowledge of medications to generate the medication representation H_{med} in Eq. (6-9). Direct doctor model is used to make medication recommendation o_1 based on the patient's representation directly in Eq. (10), and recombination doctor model recombines medications based on the similarity between patient and each medication to generate recommendation result o_2 in Eq. (11-12). Finally, we make a joint decision $\hat{o}$ based on the results of the bipartite doctor model in Eq. (13).

4.1 Patient Feature Encoder

From EHR data, patient health can be encoded by their current visit, which includes diagnosis and procedure information, and patient history, which includes diagnosis, procedure and medication information. Firstly, through EHR embedding, the sparse EHR data is mapped to the dense vector space. Then, current visit encoder is used to encode the patient's current health status. And by taking the patient's current visit code as a query, patient history encoder is used to capture the historical visit information from EHR based on the attention mechanism. Finally, by fusing the patient visit and

history code, the patient representation is generated to represent the final medical feature of the patient.

EHR Embedding. As mentioned before, a visit r_t consists of $[c_t^d, c_t^p, c_t^m]$ where each of c_t^* is a multi-hot vector at the t-th visit. The multi-hot vector c_t^* is binary encoded showing the existence of each medical codes recorded at the t-th visit. Like [4] used a linear embedding of the input vector, we derive EHR embeddings for c_t^d, c_t^p, c_t^m separately at the t-th visit as follows.

$$e_t^* = c_t^* W_{emb}^* \tag{1}$$

where $W_{emb}^* \in \mathbb{R}^{|C^*| \times dim}$ is the embedding matrix to learn. Thus the t-th visit r_t is transformed to $\widehat{r_t} = [e_t^d, e_t^p, e_t^m]$.

Current Visit Encoding. Then, concatenate the diagnosis and procedure of the patient at time *T* to encode the current visit of the patient as follows:

$$p_{cur} = NN_{cur}(e_T^d \# e_T^p) \tag{2}$$

where $NN_{cur}(\cdot): \mathbb{R}^{2dim} \to \mathbb{R}^{2dim}$ is a feed-forward neural network and # is the concatenation operation. The patient's current health status is encoded by the current diagnosis and procedure, so as to provide necessary information support for medication recommendation.

Patient History Encoding. We believe that the patient history can supplement the current health status, but not all history will help the current recommendation. Therefore, we use the attention mechanism to extract the current helpful information from patient history (including diagnosis, procedure and medication) to reduce the noise caused by unnecessary historical data. We derive history encodings for e_t^d, e_t^p, e_t^m separately as follows

$$q^* = NN_{qry}^*(p_{cur}) \tag{3}$$

$$p_{his}^* = \sum_{t=1}^{T-1} NN_{val}^*(e_t^*) \text{Softmax}(NN_{key}^*(e_t^*) q^*) \tag{4}$$

where $NN_{qry}^*(\cdot): \mathbb{R}^{2dim} \to \mathbb{R}^{dim}$ is the feed-forward neural network of query transform, $NN_{key}^*(\cdot): \mathbb{R}^{dim} \to \mathbb{R}^{dim}$ is the feed-forward neural network of key transform and $NN_{val}^*(\cdot): \mathbb{R}^{dim} \to \mathbb{R}^{dim}$ is the feed-forward neural network of value transform.

Patient Representation. The final patient representation is generated by concatenating the current and historical information of the patient. We follow a common and effective approach to first concatenate two vectors as a double-long vector, and then apply a feed-forward neural network as follow,

$$h_{pat} = NN_{pat}(p_{cur} \# p_{his}^d \# p_{his}^p \# p_{his}^m) \tag{5}$$

where $NN_{pat}(\cdot): \mathbb{R}^{5dim} \to \mathbb{R}^{dim}$ is a feed-forward neural network and # is the concatenation operation. For the fusion of external knowledge, the existing work usually introduces external knowledge by customizing a feature encoder for specific external knowledge, which leads to poor scalability of new external knowledge. And for our method, it is convenient to expand new useful information sources, such as the patient's age, gender and others that may be helpful to the description of the patient's

health, by using the attention mechanism and concatenate operation. Finally, the effect of recommendation will be improved easily by introducing the new and effective external knowledge.

4.2 Medication Feature encoder

In order to make use of the attributes and dependence of medications to further improve the recommendation effect, we additionally use EHR graph, DDI graph and molecule graph to encode medications and generate the feature representations. Firstly, for base encoding, the medication embedding matrix in the EHR embedding is used to represent the basic information of medication in recommendation. Then, through external knowledge encoding, the non Euclidean space external knowledge of the medication is coded based on the graph convolution network and a readout pooling function. Finally, by fusing the medication base and external code, the medication information table is generated to represent the final medical feature of all medications.

Base Encoding. In order to represent the basic information of medications in the recommendation process, W_{emb}^{m} is directly used to represent the basic attribute matrix of medications, which is the same as in Eq. (1), and each row vector in the matrix represents one medication.

External Knowledge Encoding. As mentioned before, the external knowledge of medication includes EHR graph, DDI graph and molecule graph, which is represented by A^E, A^D and A^{m_i}. Firstly, each $A^* \in \mathbb{R}^{|V^*|\times|V^*|}$ is preprocessed respectively as follows:

$$\hat{A}^* = \hat{D}^{*-\frac{1}{2}}(I + A^*)\hat{D}^{*-\frac{1}{2}} \tag{6}$$

where $\hat{D}^*$ is the diagonal matrix of A^* and I is identity matrix. Then we apply GCN on each $\hat{A}^*$ to learn improved embeddings respectively,

$$M^* = \hat{A}^*\sigma(\hat{A}^*W_{g1}^*)W_{g2}^* \tag{7}$$

where σ is a nonlinear activation function and $W_{g1}^* \in \mathbb{R}^{|V^*|\times dim}$, $W_{g2}^* \in \mathbb{R}^{dim\times dim}$ are the graph convolution matrix to learn. And the model depth can be deepened by increasing the number of convolution matrix layers. Then, each node in the external knowledge graph is encoded into M^*, where each row vector of $M^{\mathrm{E}} \in \mathbb{R}^{|C^m|\times dim}$ and $M^{\mathrm{D}} \in \mathbb{R}^{|C^m|\times dim}$ represents one medication, and each matrix represents one medication for $M^{m_i} \in \mathbb{R}^{|V^{m_i}|\times dim}$. In order to get the molecule representation of medications, referring to [12], M^{m_i} is pooled by a readout function to obtain the representation of the molecule knowledge of the medication, which calculates the average of all molecule nodes as follows:

$$z^{m_i} = \mathrm{READOUT}(\{M_j^{m_i} | j = 1, \dots, |V^{m_i}|\}) \tag{8}$$

where z^{m_i} is the molecule representation of the medication m_i, $M_j^{m_i}$ is the row vector of M^{m_i} and $|V^{m_i}|$ is the total number of the constructed molecule of the medication m_i. Then, the z^{m_i} of all medications are stacked to obtain the molecule matrix $Z = [z^{m_1}, z^{m_2}, \dots, z^{m_{|C^m|}}]^{\mathrm{T}}$ of medications.

Medication Information Table. Finally, we concatenate the different encodings of medications as the medication information table,

$$H_{med} = NN_{med}(W_{emb}^{m} \# M^{\mathrm{E}} \# M^{\mathrm{D}} \# Z) \tag{9}$$

where each row vector of $H_{med} \in \mathbb{R}^{|C^m| \times dim}$ is the representation of one medication, $NN_{med}(\cdot): \mathbb{R}^{4dim} \rightarrow \mathbb{R}^{dim}$ is a feed-forward neural network to learn and # is the concatenation operation. For the fusion of external knowledge, similar to the patient representation, it is easy to realize the fusion by adding new external features during vector concatenating.

4.3 Bipartite Decision Module

We use two doctor models to recommend medication combinations. Different doctor models use different encoding features to support the flexible fusion of external knowledge. Firstly, the direct doctor model only considers the patient representation to directly recommend the medication combination. And the recombination doctor model calculates the similarity between patient and each medication based on the patient representation and medication information table, and then recombines the medications based on the similarity calculation results to realize recommendation. Finally, we combine the recommendation results of the two doctor models to make a joint decision and complete the final recommendation for the patient.

Direct Doctor. For this doctor model, we directly use the patient representation for recommendation, and it can work when the feature of medications is missing. We use double-layer feed-forward neural network to project the patient representation and generate the probability of each medication in the recommended combination,

$$o_1 = NN_{o_1}(h_{pat}) \tag{10}$$

where $o_1 \in \mathbb{R}^{|C^m|}$ is directly retrieved using patient representation and $NN_{o_1}(\cdot): \mathbb{R}^{dim} \rightarrow \mathbb{R}^{|C^m|}$ is a feed-forward neural network to learn. When implemented, $NN_{o_1}(\cdot)$ is a two-layer network and its hidden layer is activated by relu.

Recombination Doctor. Recombination doctor calculate the similarity between patient and each medication, recombine medications based on the similarities and patient's representation, and generate the patient's medication combination result. We first use the patient representation h_{pat} and the medication information table H_{med} to calculate the similarity between the patient and each medication,

$$sim = \mathrm{cosine}(H_{med}, h_{pat}) \tag{11}$$

where $sim \in \mathbb{R}^{|C^m|}$ is the similarity of all medications and $\mathrm{cosine}(\cdot)$ is the function of cosine similarity. Then, input the similarity results into a double-layer feed-forward neural network to calculate the recombination, and input the patient representation together to adjust and guide the recombination process, and generate the recommendation results of the recombination doctor,

$$o_2 = NN_{o_2}(\alpha sim \# \beta h_{pat}) \tag{12}$$

where $o_2 \in \mathbb{R}^{|C^m|}$ is the result of recombination based on similarity and $NN_{o_2}(\cdot): \mathbb{R}^{|C^m|+\dim} \rightarrow \mathbb{R}^{|C^m|}$ is a feed-forward neural network to learn. $\alpha, \beta \in \mathbb{R}^1$ are trainable fusion weights, which are used to adjust the effect of similarity and patient representation on doctor model decision-making.

Joint Decision-making. Finally, the attention mechanism is used to adjust the decision weight of the two doctor models to realize joint decision-making,

$$\hat{o} = \text{sigmoid}(w_1 \odot o_1 + w_2 \odot o_2) \tag{13}$$

where $w_1, w_2 \in \mathbb{R}^{|C^m|}$ are trainable weight vectors, which integrate and adjust the importance of two doctors' decisions on different medications.

4.4 Model Training and Inference

In the training phase, the FFBDNet is trained end-to-end. We need to find the optimal parameters to realize medication combination recommendation. In order to improve the accuracy and DDI rate, we propose greedy loss to adjust the process of model training. And in the inference phase, we set a threshold δ, and determine the final medication combination to be recommended by picking those medications whose model prediction probability is greater than δ.

Multi-label Prediction Loss. We view the medication combination recommendation as a multi-label classification task. Therefore, we use two common multi-label classification loss functions as the objective function of our model, namely the binary cross entropy loss L_{bce} and the multi-label margin loss L_{multi}. L_{bce} makes the prediction result of the model closer to the growth truth, and L_{multi} makes the predicted probability of ground truth labels has at least 1 margin larger than others. Thus, threshold value is easier to be fixed when predicting.

$$L_{bce} = \sum_i^{|C^m|} y_i \log(\hat{o}_i) + (1 - y_i)\log(1 - \hat{o}_i) \tag{14}$$

$$L_{multi} = \sum_i^{|C^m|} \sum_{j \in Y} \frac{\max(0, 1 - (\hat{o}_j - \hat{o}_i))}{|Y|} \tag{15}$$

where y is the ground truth of the medication combination and Y is the index set of ground truth label.

Greedy Loss. We achieve greedy loss by multiplying L_{bce} and L_{multi} by greedy mask, which is used to shield high conflict medications,

$$\hat{L}_{bce} = \sum_i^{|C^m|} mask_i y_i \log(\hat{o}_i) + (1 - mask_i y_i)\log(1 - \hat{o}_i) \tag{16}$$

$$\hat{L}_{multi} = \sum_i^{|C^m|} \sum_{j \in Y} \frac{\max(0, mask_j(1 - (\hat{o}_j - \hat{o}_i)))}{|Y|} \tag{17}$$

$$L_{greedy} = \lambda_1 \hat{L}_{bce} + \lambda_2 \hat{L}_{multi} \tag{18}$$

where $\lambda_1, \lambda_2 > 0$ are the mixture weights and $mask_i$ is the greedy mask of the i-th medication of the patient. The essence of greedy loss is to explicitly reduce the co-occurrence frequency of conflict medications, so that the model can reduce the impact of conflict medications on parameters in the back-propagation process when learning

statistical knowledge. In detail, the greedy mask can be obtained by algorithm 1, in which the balance between accuracy and DDI rate can be adjusted by setting different greedy scale.

Inference. In inference phase, we apply a threshold δ = 0.5 on the output in Eq. (13) to predict medication combination.

$$\hat{Y} = \{i | \hat{o}_i > \delta, 1 \leq i \leq |C^m|\} \tag{19}$$

where $\hat{o}_i$ is the probability of each medication predicted by the model. Before the final inference, based on the loss function of Eq. (18), the model will be calibrated through the back-propagation algorithm to make the predictive scores as close as possible to the probabilities of medications occurrence in the actual scene. The effect of calibration will be affected by the data difference between training samples and actual scene, but it can be alleviated by limiting the number of training iterations or other methods to prevent over fitting. And then, we choose all medications with $\hat{o}_i$ greater than δ as the recommendation result.

Algorithm 1: Greedy mask generation algorithm

Input : Training ground truth $\{y_i, i \in [1, \dots, |C^m|]\}$, DDI adjancy A^D, greedy scale S

Output: greedy mask $\{mask^{(i)}, i \in [1, \dots, |C^m|]\}$

initialize $mask_k^{(i)} = 1 \quad \forall k = 1 \dots K$
initialize $MSet = set()$
for $i = 1 \dots |C^m|$ **do**
  if $y_i = 1$ **do**
    $add(MSet, i)$
  end if
end for
while $True$
  initialize $M = dict()$
  initialize $fine = True$
  for $\forall pair(n, m)\ in\ MSet$ **do**
    if $A_d[n, m] = 1$ **do**
      $M[n] += 1$
      $M[m] += 1$
    end if
  end for
  if $\max(M) > S$ **do**
    $fine = False$
  end if
  if $fine$ **do**
    break
  else
    $i = \text{argmax}(M)$
    $delete(MSet, i)$
  end if
end while

5 Experiment

We compare FFBDNet with the patient's actual EHR data, take the medication combination actually accepted by the patient as the ground truth, and take the output by FFBDNet as the prediction, and measure the accuracy of recommendations by comparing the differences between the ground truth and prediction. We also calculate the DDI rate in the prediction of FFBDNet by using the real medication confliction. In addition, we evaluate FFBDNet by comparing against other baselines on recommendation accuracy and DDI rate. FFBDNet is implemented in PyTorch [28] and trained with 8GB memory and Nvidia 2060 GPU.

Dataset. The experiments are carried out on MIMIC-III [29]. We follow the procedure similar to [12] to process the medical codes in the experiments. The NDC drug code in MIMIC-III is mapped to third level ATC code as prediction label. The statistics of the postprocessed data is reported in Table 1.

Baselines. We compare our model with the following baseline and state-of-the-art algorithms.

- **Logistic Regression (LR)**, multi-label classification model, is a logistic regression with L2 regularization. Binary relevance technique [30] is used to handle multi-label output.

Table 1. Statistics of the Data.

# patients	6,350
# clinical events	15,016
# diagnosis	1,958
# procedure	1,426
# medication	145
avg # of visits	2.36
avg # of diagnosis	10.51
avg # of procedure	3.84
avg # of medication	8.80
# medication in DDI knowledge base	123
# DDI types in knowledge base	40

- **RETain[**14], sequential decision-making model, can integrate recent visits through reverse time attention, and provide sequential prediction of medication combination.
- **Leap[**6], sequential decision-making model, decomposes medication recommendation into a continuous decision-making process, models the decision-making process with a cyclic decoder, and automatically determines the appropriate amount of medications.
- **GAMENet[**4], multi-label classification model, integrates the drug-drug interactions knowledge by a memory module, and models longitudinal patient records as the query. By using query vector to extract the information in the memory module of medications, medication combination recommendation is carried out.
- **CompNet[**8], sequential decision-making model, views the medication combination recommendation as an order-free Markov Decision Process (MDP) prob-

lem and designs a Deep Q Learning (DQL) mechanism to learn correlative and adverse interactions between medicines.

- **AMANet[**10], multi-label classification model, integrate both attention and memory to realize asynchronous multi-view learning, and focus on the dual-view sequences. The sequence is saved as the patient's historical memory, and the medication combination is recommended by querying the memory.
- **SafeDrug**[12], multi-label classification model, uses the medications' molecular structure and models DDIs to make safe medication recommendation as much as possible. Finally, the model combines and decodes the medication information for medication combination recommendation.

Metrics. We use five efficacy metrics: DDI rate, Jaccard Similarity Score (Jaccard), Average F1 (F1), Precision Recall AUC (PRAUC), and # of medications to evaluate the recommendation efficacy.

To measure the prediction accuracy, we use Jaccard, F1, PRAUC and # of medications to calculate the gap between the ground truth and the model prediction to describe the treatment efficacy of recommendation [10,12,13]. Jaccard is defined as the size of the intersection divided by the size of the union of ground truth and predicted medication set,

$$\text{Jaccard} = \frac{|Y \cap \hat{Y}|}{|Y \cup \hat{Y}|} \tag{20}$$

where Y is the index set of ground truth label and $\hat{Y}$ is the index set of model predicted label. Precision (P) ,Recall (R), and F1 are defined as:

$$\text{P} = \frac{|Y \cap \hat{Y}|}{|Y|}, \text{R} = \frac{|Y \cap \hat{Y}|}{|\hat{Y}|} \tag{21}$$

$$F1 = \frac{2\text{PR}}{\text{P+R}} \tag{22}$$

To measure medication safety, we use DDI Rate and relative DDI Rate (△ DDI Rate %),

$$\text{DDI Rate} = \frac{\sum_{i,j} A^m[i,j]}{\sum_{i,j} 1} \tag{23}$$

$$\Delta\text{DDI Rate}\% = \frac{\text{DDI Rate} - \text{DDI Rate (EHR)}}{\text{DDI Rate (EHR)}} \tag{24}$$

where A^m is the adjacency matrix of DDI graph and DDI Rate (EHR) is the DDI rate of the ground truth in EHR. And We randomly divide the dataset into training, validation, and test with ratio 4:1:1 and report the performance from the test set.

Knowledge Source Support. Table 2 lists the support of the baseline methods for different knowledge sources. For these methods that use external knowledge, they customize the feature encoder for specific external knowledge to capture the effective information, which limits the scalability of other external knowledge. For our method, we can support the integration of all different external knowledge of patients and medications, so that we can easily improve the amount of model information by introducing external knowledge, so as to improve the effect of recommendation.

Table 2. Knowledge Source Support of Baselines.

Methods	Knowledge Source Support
LR	EHR
RETAIN	EHR
Leap	EHR
GAMENet	EHR, DDI graph, EHR graph
CompNet	EHR, DDI graph
AMANet	EHR
SafeDrug	EHR, molecule graph

Performance Comparison. Table 3 compares the performance of different approaches on accuracy and DDI rate. Compared with the baselines, FFBDNet can introduce more information into the final decision-making process through the fusion of multiple external knowledge, so as to improve the discrimination ability of the model. Results show that FFBDNet has the highest score with respect to Jaccard, PR-AUC and F1. For FFBDNet(greedy), by using the greedy mask, the co-occurrence frequency of high conflict medications can be reduced. And results show that it can not only avoid almost all DDI while reaching the lowest DDI rate, but also still maintain the accuracy at a high level compared with the SafeDrug that emphasizes security.

As for the baseline, sequential decision-making models such as Leap, Retain and CompNet yield poor results. Similar to the conclusion of previous work [12], multi-label prediction model (GAMENet, AMANet, SafeDrug) might be more straightforward and effective in the medication recommendation task. The accuracy of AMANet can reach a high level, but it does not consider the problem of DDI. Both GAMENet and SafeDrug consider DDI in the process of model training. Although SafeDrug can get low DDI rate, it has low accuracy compared with our greedy method.

Table 3. Performance Comparison on MIMIC-III (ground truth DDI rate is 0.0808).

Methods	DDI Rate	ΔDDI	Jaccard	PRAUC	F1	# of Med.	# of parameters
LR	0.0724 ±0.0009	-10.40% ±1.11%	0.4543 ±0.0021	0.7550 ±0.0018	0.6142 ±0.0019	14.23 ±0.09	-
RETAIN	0.0810 ±0.0025	+0.25% ±3.07%	0.4882 ±0.0020	0.7529 ±0.0014	0.6487 ±0.0018	15.83 ±0.31	291,034
Leap	0.0693 ±0.0010	-14.23% ±1.67%	0.4442 ±0.0025	0.6452 ±0.0030	0.6071 ±0.0024	18.83 ±0.17	439,196
GAMENet	0.0798 ±0.0011	-1.24% ±1.32%	0.5146 ±0.0024	0.7657 ±0.0015	0.6694 ±0.0021	19.77 ±0.34	455,002
CompNet	0.0761 ±0.0008	-5.82% ±1.01%	0.4933 ±0.0019	0.7573 ±0.0020	0.6587 ±0.0017	19.33 ±0.21	961,412
AMANet	0.0879 ±0.0023	+8.79% ±2.82%	0.5195 ±0.0021	0.7772 ±0.0027	0.6739 ±0.0020	20.13 ±0.25	1,799,575
SafeDrug	0.0267 ±0.0009	-66.95% ±0.16%	0.4030 ±0.0025	0.6991 ±0.0024	0.5582 ±0.0020	25.56 ±0.11	406,170
FFBDNet(greedy)	**0.0019** ±0.0002	**-97.65%** ±0.28%	0.4361 ±0.0014	0.7061 ±0.0021	0.5978 ±0.0015	14.31 ±0.12	227,750
FFBDNet	0.0717 ±0.0016	-11.26% ±2.01%	**0.5292** ±0.0020	**0.7777** ±0.0010	**0.6833** ±0.0017	19.69 ±0.30	227,750

Multi Feature Ablation Study. We control the introduction of different knowledge to observe the effect of increasing information sources on the model results. It can be observed in Table 4 that some external knowledge bring new information to the model, so as to improve the final effect. FFBDNet can integrate the medication feature into the recommendation by using the recombination doctor model in the bipartite decision module, and it finally achieves the best results when all the information is used. Thus, in medication combination recommendation task, the effect of introducing new information sources by fusing heterogeneous and diverse external knowledge is verified

Greedy Ablation Study We evaluate greedy loss and show that accuracy and DDI rate can be controlled by greedy scale. The ground truth DDI rate in MIMIC-III is 0.0808. Table 5 shows the results of different greedy scales. It can be found that the larger the greedy scale, the greater the accuracy of the model and the greater the DDI rate. When the greedy scale is infinite, the accuracy of the model is the highest. The greedy loss provides a way for doctors to control the tradeoff between accuracy and DDI rate in recommendation.

Table 4. Multi Feature Ablation Study.

Patient	Medication	DDI Rate	ΔDDI	Jaccard	PRAUC	F1	# of Med.
current	-	0.0641 ±0.0009	-20.67% ±1.11%	0.5039 ±0.0018	0.7593 ±0.0019	0.6611 ±0.0016	18.95 ±0.24
current, history	-	0.0771 ±0.0012	-4.58% ±1.48%	0.5173 ±0.0015	0.7661 ±0.0018	0.6732 ±0.0014	20.36 ±0.16
current, history	base	0.0735 ±0.0014	-9.03% ±1.72%	0.5204 ±0.0013	0.7712 ±0.0007	0.6751 ±0.0012	19.90 ±0.17
current, history	base, EHR	0.0739 ±0.0006	-8.54% ±0.73%	0.5239 ±0.0019	0.7754 ±0.0013	0.6790 ±0.0017	19.65 ±0.16
current, history	base, EHR, DDI	0.0726 ±0.0016	-10.15% ±1.98%	0.5241 ±0.0017	0.7761 ±0.0015	0.6816 ±0.0016	19.33 ±0.21
current, history	Base, EHR, DDI, molecule	**0.0717** ±0.0016	**-11.26%** ±2.01%	**0.5292** ±0.0020	**0.7777** ±0.0010	**0.6833** ±0.0017	19.69 ±0.30

Table 5. Greedy Ablation Study.

Greedy Scale	DDI Rate	Jaccard	PRAUC	F1	# of Med.
1	0.0019±0.0002	0.4361±0.0014	0.7061±0.0021	0.5978±0.0015	14.31±0.12
2	0.0105±0.0005	0.4748±0.0015	0.7287±0.0014	0.6356±0.0015	16.25±0.18
3	0.0208±0.0004	0.4957±0.0018	0.7421±0.0014	0.6544±0.0016	17.64±0.19
4	0.0277±0.0008	0.5032±0.0021	0.7528±0.0017	0.6608±0.0019	18.09±0.24
5	0.0349±0.0005	0.5072±0.0012	0.7615±0.0013	0.6646±0.0011	18.43±0.20
6	0.0410±0.0006	0.5145±0.0026	0.7694±0.0021	0.6709±0.0023	18.57±0.23
+∞	0.0717±0.0016	0.5292±0.0020	0.7777±0.0010	0.6833±0.0017	19.69±0.30

6 Conclusion

In this paper, we propose FFBDNet for medication combination recommendation, which is equipped with a patient feature encoder, a medication feature encoder and a

bipartite decision module. Based on the attention mechanism and the concatenating operation, the feature encoders can easily fuse external knowledge to increase the model information source. With using the encoder results of patient and medications, the bipartite decision module make a joint decision to realize medication combination recommendation through two doctor models. And we design a greedy loss, which uses the greedy mask to filter high conflict medications, to reduce the DDI rate. We evaluated FFBDNet using benchmark data. The experimental results show that FFBDNet outperforms the state-of-the-art methods. Besides, using greedy loss to participate in the model training, FFBDNet can avoid almost all DDI, while still maintaining a good recommendation accuracy. In the future, we will study how to efficiently extract and fuse the multi-feature of medications to further improve the accuracy of representation while ensuring the scalability of external knowledge. Code related to this paper is available at https://github.com/wangzssdwh/FFDBNet.

References

1. Edward, C., Mohammad, T.B., Andy, S., Walter F.S., Sun, J.M.: Doctor ai: Predicting clinical events via recurrent neural networks. In: Machine Learning for Healthcare Conference, pp. 301-318 (2016)
2. Xiao, C., Choi, E., Sun, J.M.,: Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. Journal of the American Medical Informatics Association 25(10), 1419-1428 (2018)
3. Shang, J.Y., Ma, T.F., Xiao, C., Sun, J.M.: Pre-training of graph augmented transformers for medication recommendation. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, pp. 1907-1913 (2018)
4. Shang, J.Y., Xiao, C., Ma, T.F., Li, H.Y., Sun, J.M.: Gamenet: Graph augmented memory networks for recommending medication combination. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence, pp. 1126-1133 (2019)
5. Li, C., Wang, B.Y., Pavlu, V., Aslam, J.A.: Conditional bernoulli mixtures for multi-label classification. In: Proceedings of the 33rd International Conference on Machine Learning, pp. 2482–2491 (2016)
6. Zhang, Y., Chen, R., Tang, J., Stewart, W. F., Sun, J.: 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, pp. 1315 – 1324 (2017)
7. Le, H., Tran, T., Venkatesh, S.: Dual memory neural computer for asynchronous two-view sequential learning. In: Proceedings of the 24rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1637 – 1645 (2018)
8. Wang, S., Ren, P., Chen, Z., Ren, Z., Ma, J., Rijke, M.: Order-free Medicine Combination Prediction with Graph Convolutional Reinforcement Learning. In: Proceedings of the 28th Conference on Information and Knowledge Management, pp. 1623 – 1632 (2019)
9. Jacek, M.B., Thomas A.L.: Predicting medications from diagnostic codes with recurrent neural networks. In: Proceedings of the International Conference on the 5th Learning Representations, pp. 100-119 (2017)
10. He, Y., Wang, C., Li, N., Zeng, Z.: Attention and memory-augmented networks for dual-view sequential learning. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 125 – 134 (2020).

11. Bhoi, S., Lee, M.L., Hsu, W., Fang, H.S., Tan, N.C., 2021. Personalizing medication recommendation with a graph-based approach. ACM Transactions on Information Systems 40(3), 55-79 (2021)
12. Yang, C., Xiao, C., Ma, F., Glass, L., Sun, J.: SafeDrug: Dual molecular graph encoders for recommending effective and safe drug combinations. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence, pp. 3735-3741 (2021)
13. Wang, Y., Chen, W., Pi, D., Yue, L., Wang, S., Xu, M.: Self-supervised adversarial distribution regularization for medication recommendation. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, pp. 3134 – 3140 (2021)
14. Choi, E., Bahadori, M.T., Sun, J., Kulas, J., Schuetz, A., Stewart, W.: Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In Advances in Neural Information Processing Systems, pp. 3504-3512 (2016)
15. An, Y., Zhang. L., Yang, H.: Prediction of treatment medicines with dual adaptive sequential networks. IEEE Transactions on Knowledge and Data Engineering, 1-14 (2021)
16. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (2017)
17. Hamilton, W. L., Ying, R., Leskovec, J.: Inductive representation learning on large graphs. In: Proceedings of the 31st Conference on Neural Information Processing Systems. 1024–1034 (2017)
18. Petar, V., Guillem, C., Arantxa, C., Adriana, R., Pietro L., Yoshua, B.: Graph attention networks. In: International Conference on Learning Representations (2018)
19. Chen, J., Ma, T., Xiao, C.: FastGCN: Fast learning with graph convolutional networks via importance sampling. In: International Conference on Learning Representations (2018)
20. Xue, H., Yang, L., Rajan, V.: Multiplex bipartite network embedding using dual hypergraph convolutional networks. In: Proceedings of the 30th Web Conference, pp. 1649-1660 (2021)
21. Ma, T., Xiao, C., Zhou, J., Wang, F.: Drug similarity integration through attentive multiview graph autoencoders. CoRR abs/1804.10850. (2018)
22. Zitnik, M., Agrawal, M., Leskovec, J.: Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics 34(13), 457 – 466 (2018)
23. Mauri, A., Consonni, V., Pavan, M.: Dragon software: An easy approach to molecular descriptor calculations. Communications in MCC 56(2), 237-248 (2006)
24. Rogers, D., Hahn, M.: Extended-connectivity fingerprints. Journal of chemical information and modeling 50(5), 742-754 (2010)
25. David, K.D., Dougal M., Jorge I., Rafael B., Timothy H. Alan A., and Ryan P.A.: Convolutional networks on graphs for learning molecular fifingerprints. In: Proceedings of the 29th Conference on Neural Information Processing Systems (2015)
26. Huang, K.X., Fu, T.F., Xiao, C., Glass, L., Sun, J.M.,: Deeppurpose: a deep learning based drug repurposing toolkit. Bioinformatics (2020)
27. Huang, K.X., Xiao, C., Hoang, T., Glass, L., Sun, J.M.: Caster: Predicting drug interactions with chemical substructure representation. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, pp. 702-709 (2020)
28. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
29. Johnson, A.E., Pollard, T.J., Shen, L., Li, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data 3(1), 1-9 (2016)
30. Luaces, O., D´ıez, J., Barranquero, J., delCoz, J.J., Bahamonde, A.: Binary relevance efficacy for multilabel classification. Progress in Artificial Intelligence 1(4), 303–313 (2012)