




# SHAPE: A Sample-Adaptive Hierarchical Prediction Network for Medication Recommendation

Sicen Liu , Xiaolong Wang , Jingcheng Du , Yongshuai Hou , Xianbing Zhao , Hui Xu, Hui Wang, Yang Xiang , *Member, IEEE*, and Buzhou Tang , *Member, IEEE*

**Abstract**—Effectively medication recommendation with complex multimorbidity conditions is a critical yet challenging task in healthcare. Most existing works predicted medications based on longitudinal records, which assumed the encoding format of intra-visit medical events are serialized and information transmitted patterns of learning longitudinal sequence data are stable. However, the following conditions may have been ignored: 1) A more compact encoder for intra-relationship in the intra-visit medical event is urgent; 2) Strategies for learning accurate representations of the variable longitudinal sequences of patients are different. In this article, we proposed a novel Sample-adaptive Hierarchical medicAtion Prediction nETwork, termed SHAPE, to tackle the above challenges in the medication recommendation task. Specifically, we design a compact intra-visit set encoder to encode the relationship in the medical event for obtaining visit-level representation and then develop an inter-visit longitudinal encoder to learn the patient-level longitudinal representation efficiently. To endow the model with the capability of modeling the variable visit length, we introduce a soft curriculum learning method to

assign the difficulty of each sample automatically by the visit length. Extensive experiments on a benchmark dataset verify the superiority of our model compared with several state-of-the-art baselines.

**Index Terms**—Medication recommendation, curriculum learning, set encoder, electronic health record (EHR) data-mining.

## I. INTRODUCTION

RECENTLY, massive health data have offered the opportunity to assist clinical decision-making through deep learning [1], [2], [3], [4], [5], [6]. Effective and safe medication combination recommendation for patients who suffer from multiple diseases is an essential task in healthcare [7], [8], [9]. There are a lot of research interests in medication recommendation task [10], [11], [12], [13], [14], [14], [15], [16], [17], [18]. The intuitive goal of medication recommendation is to predict medication sequences for a particular patient based on complex health conditions. Existing strategies of medication recommendation can be categorized into two types: 1) *Instance-based methods*, which recommend medication sequences only based on the current hospital visit (e.g., diagnosis, procedure) [19], [20], [21]. The instance-based setting will ignore the temporal dependencies on the patient's health records. To overcome this issue, 2) *Longitudinal-based methods* were proposed to leverage the longitudinal patient records to predict personalized medication. Most longitudinal methods pursue enhanced representations of patient health status based on the historical health records (e.g., diagnosis, procedure) and use this patient representation to conduct medication recommendations [22], [23], [24], [25], [26], [27], [28], [29].

Despite the significance and value of the methods in the longitudinal methods, they still suffer from two critical limitations: 1) One problem with existing longitudinal works is that they neglect the compact intra-relationships between medical events within each visit. In other words, they ignore the relationship between the same type of medical codes during a visit. 2) Existing longitudinal models are static. Namely, all samples go through the same fixed computation flow. This may be powerless on the shorter records, which lack historical information.

On the one hand, existing longitudinal methods use the historical code sequences (e.g., medication, diagnosis) within each

Manuscript received 15 October 2022; revised 31 May 2023 and 7 August 2023; accepted 21 September 2023. Date of publication 28 September 2023; date of current version 6 December 2023. This work was supported in part by the National Natural Science Foundations of China under Grants 62276075, 62276082, U1813215, 61876052, and 62106115, in part by the Science and Technology Planning Project of Shenzhen Municipality under Grant JCYJ20190806112210067, in part by the National Key R&D Program of China under Grant 2021ZD0113402, in part by the National Natural Science Foundation of Guangdong, China under Grant 2019A1515011158, in part by the Major Key Project of PCL under Grant PCL2021A06, and in part by the Strategic Emerging Industry Development Special Fund of Shenzhen under Grant 20200821174109001. (*Corresponding authors: Yang Xiang; Buzhou Tang.*)

Sicen Liu, Xianbing Zhao, and Buzhou Tang are with the Department of Computer Science, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: liusicen@stu.hit.edu.cn; zhaoxianbing\_hitsz@163.com; tangbuzhou@gmail.com).

Xiaolong Wang is with the Department of Computer Science, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China (e-mail: wangxl@insun.hit.edu.cn).

Jingcheng Du is with the Melax Tech, Inc., Houston, TX 77030 USA, and also with The University of Texas Health Science Center, Houston, TX 77030 USA (e-mail: jingcheng.du@melaxtech.com).

Yongshuai Hou and Yang Xiang are with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: houyush@pcl.ac.cn; xiangy@pcl.ac.cn).

Hui Xu and Hui Wang are with the Gennlife (Beijing) Technology Company, Ltd., Beijing 10089, China (e-mail: xuhui@gennlife.com; wanghui@gennlife.com).

Digital Object Identifier 10.1109/JBHI.2023.3320139

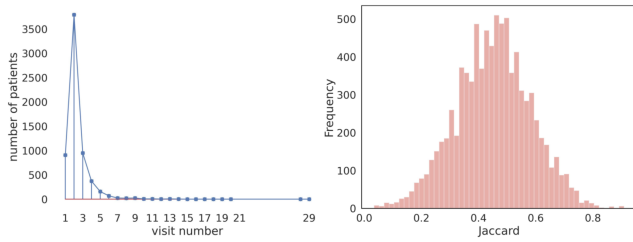


Fig. 1. Histogram of visit counts of MIMIC-III dataset (left) and the histogram of Jaccard between current medications and historical medications (right).

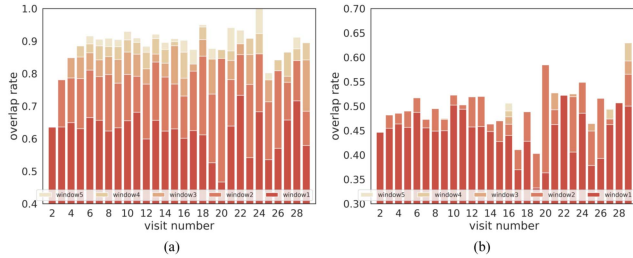


Fig. 2. Statistics of (a) Medication overlap rate and (b) Jaccard coefficients in various visits with different window sizes.

visit to present the complex patient’s health condition, where medical events are adopted independently and sparsely represented methods to obtain equal contributions representation in the current record. Most of them use multi-hot embedding methods to encode the structured data sequences. However, the impact of medical events varies for each patient, especially for patients with multimorbidity. For instance, during a visit, the health condition differs a lot between a patient diagnosed with both *Chronic systolic heart failure* and *Septic shock* and a patient diagnosed with both *Septic shock* and *Acute respiratory failure*. Previous methods ignore the compact intra-relationship of these medical events and the variable importance of each code for the patient.

On the other hand, such longitudinal patterns rely on historical health information and are powerless to the short visit that lacks historical records. As shown in Fig. 1, we conduct the statistic on the MIMIC-III [30] dataset. We can see that most visit lengths are short than thrice. For each visit, we calculate the Jaccard between current medications and past medications. We can see that a large portion of prescribed medicines are similar to those recommended before, which means the results of medication recommendations rely on historical medication records. Additionally, we conduct fine-grained statistics of the MIMIC-III dataset, as shown in Fig. 2. We calculate the proportion of medications that have appeared in history and the Jaccard with various visit windows. We can see that in the more extended visits, a large portion of drug sequences have been recommended before. However, the prevalence of short visit records in real-world clinical scenarios often lacks crucial historical medication information that could be referenced for treatment decisions. This phenomenon illustrates that a more robust strategy that could model the accurate representation of the variable longitudinal sequences is urgent.

To overcome these challenges, we proposed a novel Sample-adaptive Hierarchical medication Prediction nEtwork, named **SHAPE**, to learn a more accurate representation of patients. In SHAPE, we develop a hierarchical patient representation framework. Concretely, we first tailor an intra-visit set encoder to learn the visit-level representation and then design an inter-visit longitudinal encoder for learning the patient-level longitudinal representation. By performing the intra-visit set encoder and inter-visit longitudinal encoder, collaborative information latent in longitudinal historical interactions is explicitly hierarchical encoded. To enhance the ability to represent various lengths of visit records, we adopt a soft curriculum learning method to help our SHAPE model learn these data patterns by assigning the difficulty weight to each sample. The experiments on a public dataset demonstrate the effectiveness of our proposed model.

The main contributions of this work are three-fold:

- We present a hierarchical encoder mechanism towards medication recommendation, which could dig for a more accurate representation from the various records of the patient. In particular, we first design an intra-visit set encoder to encode the medical events and obtain visit-level representation, and then develop an inter-visit longitudinal encoder for learning the patient-level longitudinal information.
- We design an adaptive curriculum learning module for variable patient visit records, especially for the short ones, which aims at an adaptive learning strategy over time and the length of patient records to improve the effectiveness of medication recommendations.
- Extensive experimental results on the public benchmark dataset validate the effectiveness and superiority of our proposed method.

## II. RELATED WORK

### A. Medication Recommendation

Existing medication recommendation algorithms can be categorized into instance-based methods and longitudinal approaches. Instance-based algorithms extract patient information only from current visits. For example, LEAP [20] extracts patient representation from the current visit record and decomposes the medication recommendation into a sequential decision-making process. Longitudinal-based methods are designed to leverage temporal dependencies within the patient’s historical information. For example, RETAIN [22] uses two-level attention, which models the longitudinal information based on recurrent neural networks (RNN). GAMENet [24] uses augmented memory neural networks to fuse the drug-drug interactions and store the historical drug record to model the patient representation. MICRON [25] pays attention to the changes in patient health records and uses residual-based network inference to update the sequential representation. COGNet [27] conditionally generates the medication combinations either copied from the historical drug records or directly generate new drugs. These existing efforts, however, still suffer from the following limitations. Existing work ignores that the intra-visit medical events may pay variable effects on differing the health status of the patient. Most of them

use multi-hot embedding to encode the medical events in the current visit and ignore the difference of each medical event in intra-visit records. In this article, we proposed a hierarchical architecture to learn the comprehensive patient representation. We use an intra-visit set encoder to learn a more accurate representation of intra-visit medical events and develop an inter-visit longitudinal encoder to learn longitudinal information about the patient.

## B. Curriculum Learning

The conventional curriculum learning methods formalized the organized learning process of humans and animals, which illustrates gradually more complex ones [31]. Alex et al. derived two distinct indicators (i.e., rate of increase in prediction accuracy and rate of increase in network complexity) of the learning process as the reward signal to maximize learning efficiency automatically [32]. Guy et al. introduce sorted samples with different scoring functions to assign the learning difficulty of each instance [33]. Recently, curriculum learning has been applied to different medical tasks. Basu et al. propose a curriculum inspired by human visual acuity, which reduces the texture biases for gallbladder cancer detection [34]. Guo et al. demonstrate the application of curriculum learning for drug molecular design [35]. Gu et al. utilized curriculum learning to improve the training efficiency of molecular graph learning [36]. According to Figs. 1 and 2, we found that the short and new visits samples account for most of the entire dataset. The conventional longitudinal methods are hard to fit this pattern because lacking a flexible ability to model the scenarios where the patients do not have enough historical medication records and diagnosis information about their health condition. In this article, we propose a sample-adapting curriculum learning algorithm to assign the difficulty of each instance automatically.

## III. PROBLEM FORMULATION

### A. Electrical Health Records (EHR)

Patient electronic health record (EHR) data contains comprehensive medical information about the patient. Formally, EHR for patient  $j$  can be represented as a sequence  $X_j = (x_j^1, x_j^2, \dots, x_j^T)$ , where  $T$  is the corresponding totally visits number for patient  $j$ . For the single visit  $x_j^t$  of patient  $j$  at the  $t$ -th visit, where  $t \in \{1, 2, \dots, T\}$ , we ignore the index  $j$  of the patient to simplify notation. Then, the visit record is represented as  $x^t = (D^t, P^t, M^t)$ , where  $D^t \subseteq \{d_1, d_2, \dots, d_{|D|}\}$  denotes the set of diagnoses appeared in  $t$ -th visit,  $P^t \subseteq \{p_1, p_2, \dots, p_{|P|}\}$  denotes the set of procedures and  $M^t \subseteq \{m_1, m_2, \dots, m_{|\mathcal{M}|}\}$  denotes the set of medications appeared in  $t$ -th visit.  $|D|$ ,  $|P|$  and  $|\mathcal{M}|$  indicate the cardinality of corresponding element sets.

### B. DDI Graph

The medications may interact with other medications when prescribed, while the adverse drug-drug interactions (DDIs) graph records this interaction of adverse drug events. The DDI graph can be denoted as  $\mathcal{G}_d = \{\mathcal{V}, \mathcal{E}_d\}$ , where node set  $\mathcal{V}$

$\{m_1, m_2, \dots, m_{|\mathcal{M}|}\}$  represent the set of medications. The  $\mathcal{E}_d$  is the edge set of known DDIs between a pair of drugs. Adjacency matrix  $A_d \in \mathbb{R}^{|\mathcal{M}| \times |\mathcal{M}|}$  are defined to the construction of the graphs. When the  $A_d[i, j] = 1$  means the  $i$ -th medication and  $j$ -th one could interact with each other.

### C. Medication Recommendation Problem

Given a patient EHR sequence  $[x^1, x^2, \dots, x^t]$  and the DDI graph  $\mathcal{G}_d$ . For the multi-visit records patient, which includes the current diagnosis, procedure codes  $[D^t, P^t]$  and the historical records  $[x^1, x^2, \dots, x^{t-1}]$ . Note that, for the record of new visit patients, there are only current diagnosis and procedure codes  $[D^1, P^1]$ . The goal is to train a model to effectively recommend multiple medications by generating multi-label output  $\hat{y}_t \subseteq \{m_1, m_2, \dots, m_{|\mathcal{M}|}\}$  for this patient.

## IV. THE SHAPE FRAMEWORK

In this section, we present the technical details of the proposed **SHAPE** framework. As illustrated in Fig. 3, our model includes three components: (1) an **intra-visit set encoder** that learns the visit-level representation of the patient from the EHR data. (2) an **inter-visit longitudinal encoder** that takes the visit-level representation as input to learn the longitudinal information of the patient. (3) an **adaptive curriculum learning module** that cooperates with the prediction phase in the training stage to dynamically assign the difficulty weight of each instance by the patient visit length to improve the effectiveness of medication recommendations. Finally, the drug output is obtained from the sigmoid output representation.

### A. Patient Representation

The patient representation aims to learn a dense vector to represent a comprehensive patient's status. The physicians recommend medications based on the current diagnosis and procedure information during a clinical visit. Furthermore, the clinician also references the history of diagnosis, procedure, and medication records when the patient has historical visit records. Since the SHAPE is proposed for the generic patient, we use the three codes as the model input in the following, and the medication codes are always behind the other two medical events. Note that, for the patient who only once visited with diagnosis and procedure record, we apply a padding embedding as the medication input.

1) **Code-Level Embedding**: For predict the medication of multi-visit, we use the  $[D^t, P^t, M^{t-1}]$  as the current input, where  $M^{t-1}$  is the previous medication record. We design three correspond embedding table  $E_d \in \mathbb{R}^{|D| \times dim}$ ,  $E_p \in \mathbb{R}^{|P| \times dim}$  and  $E_m \in \mathbb{R}^{|\mathcal{M}| \times dim}$ , where the  $dim$  is the dimension of the embedding space. For the  $t$ -th visit, the set of medical events  $d^{(t)} \in D^t$ ,  $p^{(t)} \in P^t$ , and  $m^{(t-1)} \in M^{t-1}$  was transfer to the embedding space.

$$d_e^{(t)} = d^{(t)} E_d \quad (1)$$

$$p_e^{(t)} = p^{(t)} E_p \quad (2)$$

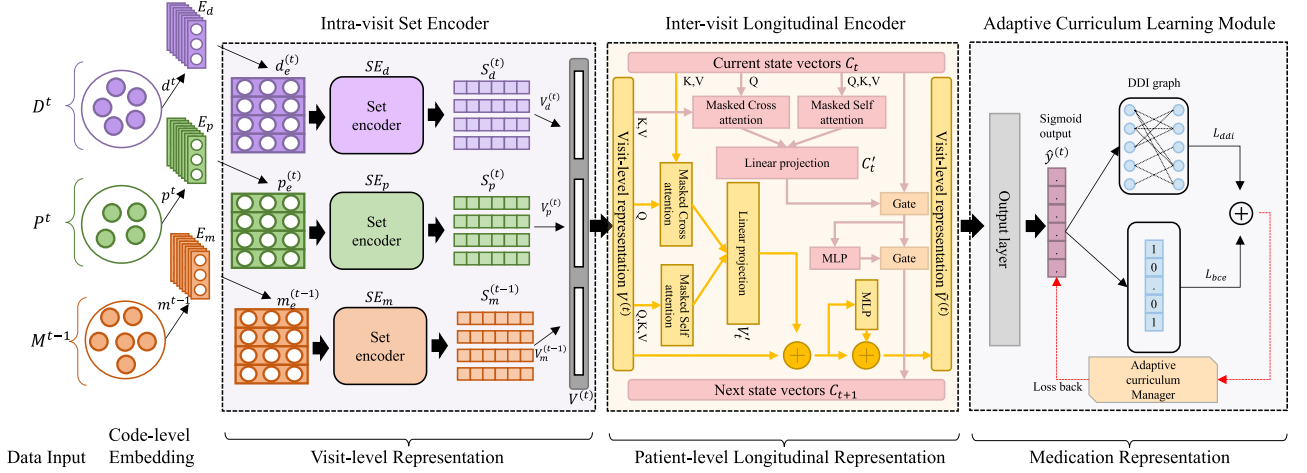


Fig. 3. Framework of our proposed SHAPE. There are three components: (1) Intra-visit Set Encoder captures the intra-relationship of the code-level medical events and summarizes it to the current visit-level representation. (2) Inter-visit Longitudinal Encoder to model the longitudinal information of the patient. (3) An Adaptive curriculum learning module automatically assigns each sample's difficulty according to the patient's visit length.

$$m_e^{(t-1)} = m^{(t-1)} E_m \quad (3)$$

2) *Intra-Visit Set Encoder*: Unlike the previous works [25], [26], which use the code embedding representation of the medical events as the patient representation. We employ the code-level embedding as the input of the set encoder to learn the code-level relationship and then integrate the code-level information into the visit-level representation. Inspired by the Set-Transformer [37], we employ inducing point methods to compress medical code representations into a more compact space for modeling the impact of intra-visit medical events and introduce the Intra-visit Set Encoder. The set encoder contained two *Induced Set Attention Block* (ISAB). In the ISAB, in addition to the set  $X \in \mathbb{R}^{m \times d}$ , a new trainable parameter vector  $I \in \mathbb{R}^{n \times d}$ , called inducing points, is introduced to model pairwise interactions among the elements in the input set. The ISAB has the two major sub-layers: *Multi-Head Attention* (MHA) and *row-wise FeedForward layer* (rFF), the functions are defined as:

$$MHA(Q, K, V) = [head_1, head_2, \dots, head_h] \quad (4)$$

$$head_i = Att(QW_i^Q, KW_i^K, VW_i^V) \quad (5)$$

$$Att(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{s}}\right)V \quad (6)$$

$$rFF(X) = Relu(XW_{rFF} + b_{rFF}) \quad (7)$$

where  $Q \in \mathbb{R}^{n_q \times d}$ ,  $K \in \mathbb{R}^{n_k \times d}$ ,  $V \in \mathbb{R}^{n_v \times d}$  are the inputs of attention  $Att(\cdot)$ ,  $W_i^Q \in \mathbb{R}^{d \times d_q}$ ,  $W_i^K \in \mathbb{R}^{d \times d_k}$ ,  $W_i^V \in \mathbb{R}^{d \times d_v}$ , and  $d_q = d_k = d_v = d/h$ .  $W_{rFF} \in \mathbb{R}^{d \times d}$  and  $b_{rFF} \in \mathbb{R}^d$  are learnable parameters. The  $[\cdot]$  means the concatenate operation. The ISAB is defined as:

$$ISAB(X) = LN(H + rFF(H)) \quad (8)$$

$$H = LN(X + MHA(X, Y, Y)) \quad (9)$$

$$Y = LN(Z + rFF(Z)) \quad (10)$$

$$Z = LN(I + MHA(I, X, X)) \quad (11)$$

where  $LN$  is layer normalization operation. The set-encoder is defined as:

$$SE_*(X) = ISAB(ISAB(X)) \quad (12)$$

where  $* \in \{d, p, m\}$ .

Given the code-level embedding representation, the output of the diagnosis set encoder is formulated as follows:

$$S_d^{(t)} = SE_d(d_e^{(t)}) \quad (13)$$

Similar to the diagnosis set encoder, the output of the procedure set encoder and medication set encoder are formulated as  $S_p^{(t)} = SE_p(p_e^{(t)})$ ,  $S_m^{(t-1)} = SE_m(m_e^{(t-1)})$ . After obtaining the code-level set representation of the three medical events, we combine them to visit-level representation  $V^{(t)}$  as the health status of the patient in the current visit. The visit-level representation is defined as:

$$V^{(t)} = [V_d^{(t)}, V_p^{(t)}, V_m^{(t-1)}] \quad (14)$$

where the  $V_d^{(t)}, V_p^{(t)}, V_m^{(t-1)}$  is the summation of code-level representation, and  $[\cdot]$  is the concatenate operation.

3) *Inter-Visit Longitudinal Encoder*: Previous works usually employ Recurrent Neural Networks (RNN) to model the dynamic patient history for learning longitudinal representations of patients. As the success of the attention mechanism in sequence task [38], [39], [40], it will be helpful to combine the attention mechanism and RNN pattern. Inspired by the Block-Recurrent Transformer (BRT) [41], which applies a transformer layer in a recurrent fashion along the sequence input. Differing from the basic BRT, we have followed the GPT [40], added the masked vector to prevent information leaks while modeling patient longitudinal visit records, and named Recurrent Attention Block (RAB). The RAB mainly includes the update stream between the hidden state vector and the visit-level representation. The

hidden state vector carries the patient temporal information, and the visit-level representation updates the information based on the historical state representation. For the state vector, the update function is formulated as follows:

$$C_{t+1} = g_2(MLP(g_1(C'_t, C_t)), g_1(C'_t, C_t)) \quad (15)$$

$$g_*(X, Y) = X \odot f + z \odot i \quad (16)$$

$$f = \sigma(W_f Y + b_f + 1) \quad (17)$$

$$i = \sigma(W_i Y + b_i - 1) \quad (18)$$

$$z = \tanh(W_z Y + b_z) \quad (19)$$

where  $MLP$  is multi-layer perceptron,  $\odot$  is the Hadamard product,  $W_f \in \mathbb{R}^{n_f \times d_f}$ ,  $W_i \in \mathbb{R}^{n_i \times d_i}$ ,  $W_z \in \mathbb{R}^{n_z \times d_z}$  are trainable weight matrices, and  $b_f \in \mathbb{R}^{d_f}$ ,  $b_i \in \mathbb{R}^{d_i}$ ,  $b_z \in \mathbb{R}^{d_z}$  are trainable bias vectors. The  $g_* \in \{g_1, g_2\}$  is the gate mechanism.  $C'_t$  is the combination of masked self-attention on the current hidden state  $C_t$  and the masked cross-attention with the visit-level representation  $V^{(t)}$ ,

$$C'_t = W'_c([Att(C_t, C_t, C_t), Att(C_t, V^{(t)}, V^{(t)}))] + b'_c \quad (20)$$

where  $W'_c \in \mathbb{R}^{n'_c \times d'_c}$  and  $b'_c \in \mathbb{R}^{d'_c}$  are learnable parameters.

The update stream of visit-level representation selects the longitudinal information from the hidden state and visit-level information from the current visit and is defined as:

$$\hat{V}^{(t)} = MLP(V^{(t)'} + V^{(t)}) + (V^{(t)'} + V^{(t)}) \quad (21)$$

where  $MLP$  is a multi-layer perceptron.  $V^{(t)'}$  is the concatenate of visit-level representation masked self-attention and masked cross-attention with the current hidden state, where a central feature is to delegate a considerable portion of the information update responsibility to the process for generating attention weights. The formulation is:

$$V^{(t)'} = W'_v([Att(V^{(t)}, V^{(t)}, V^{(t)}), Att(V^{(t)}, C_t, C_t))] + b'_v \quad (22)$$

where  $W'_v \in \mathbb{R}^{n'_v \times d'_v}$  and  $b'_v \in \mathbb{R}^{d'_v}$  are trainable parameters.

**4) Adaptive Curriculum Learning Module:** This module includes the prediction layer and the adaptive curriculum manager. After obtaining the updated patient-level representation  $\hat{V}^{(t)}$ , the final medication representation is generated through an output layer, which is defined as:

$$\hat{y}^{(t)} = \sigma(W_o \hat{V}^{(t)} + b_o) \quad (23)$$

where  $\sigma$  is sigmoid function, and  $W_o \in \mathbb{R}^{n'_o \times |\mathcal{M}|}$ ,  $b_o \in \mathbb{R}^{|\mathcal{M}|}$  are learnable parameters.

- **Supervised Multi-label Classification Loss:** The recommendation of medication combinations can be treated as a multi-label prediction task. We use the binary cross entropy loss  $l_{bce}$  as the multi-label task loss function, and  $l_{bce}$  is defined as:

$$\begin{aligned} \mathcal{L}_{bce} = & - \sum_t \sum_i^{|\mathcal{M}|} m_i^{(t)} \log(\hat{y}_i^{(t)}) \\ & + (1 - m_i^{(t)}) \log(1 - \hat{y}_i^{(t)}) \end{aligned} \quad (24)$$

where  $m_i^{(t)}$  and  $\hat{y}_i^{(t)}$  means the medical code at  $i$ -th coordinate at  $t$ -th visit.

- **Drug-Drug Interaction Loss:** The DDI loss is designed to control the DDI rate of generated medication combinations. Following the previous work [26], formally:

$$\mathcal{L}_{ddi} = - \sum_t \sum_{i,j}^{|\mathcal{M}|} (A_d \odot (\hat{y}^{(t)\top} \hat{y}^{(t)})) \quad (25)$$

where  $\odot$  is the Hadamard product.

- **Combined Loss Functions:** During the training, we noticed that the accuracy and the DDI rate often increase together, mainly due to the drug-drug interaction in real-world clinical scenarios. It is important to balance the multi-label classification loss and the DDI loss. Finally, we use a penalty weight  $\alpha$  over the DDI loss for training. The final loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{bce} + \alpha \mathcal{L}_{ddi} \quad (26)$$

where  $\alpha$  is a pre-defined hyperparameter. By presetting different  $\alpha$ , our SHAPE model could meet a different level of DDI requirements (the details of selecting the  $\alpha$  are shown in the DISCUSSION section).

- **Adaptive Curriculum Manager:** As shown in Fig. 2(a), although the medication combinations of most long visit records have been recommended before and are easy to predict, the short one lacking historical medication information is the most frequent situation in real-life clinical scenarios, which may be hard to predict accurately. To address this issue, we propose an adaptive curriculum manager that dynamically assigns complex coefficients to each patient and adopts the curriculum learning framework to train our SHAPE model. Specifically, we combine the visit length of the patient  $l_t$  into the training schema, where we modified the coefficient of learning rate with calculate  $\frac{I+l_t}{I_{max}}$  (i.e., (28)) to adjust the learning rate at the Adam [42] optimizer. The details of the adaptive curriculum manager are as follows:

$$\theta_t = \theta_{t-1} - \frac{\hat{\gamma} \mu_t}{\sqrt{\eta_t} + \epsilon} \quad (27)$$

$$\hat{\gamma} = \gamma \left(1 - \frac{I + l_t}{I_{max}}\right) \quad (28)$$

$$\mu_t = \frac{\beta_1 \mu_{t-1} + (1 - \beta_1) g_t}{1 - \beta_1} \quad (29)$$

$$\eta_t = \frac{\beta_2 \eta_{t-1} + (1 - \beta_2) g_t^2}{1 - \beta_2} \quad (30)$$

$$g_t = \nabla_{\theta} f_t(\theta_{t-1}) \quad (31)$$

where  $\epsilon$  is a constant added to the denominator to improve numerical stability,  $\gamma$  is the learning rate,  $I$  is the current training iteration number,  $l_t$  is the current visit length,  $I_{max}$  is the pre-defined maximum iteration number, and  $\mu_t, \eta_t$  is the parameter of the first moment and the second moment of Adam,  $\beta_1, \beta_2$  is the coefficient of the moment, the  $f(\theta)$  is the objective function, and  $\theta$  are parameters

TABLE I  
DATA STATISTICS

| Item                                  | Size                |
|---------------------------------------|---------------------|
| # of visits/ # of patients            | 14,995 / 6,350      |
| diag. / proc. / med. set size         | 1,958 / 1,430 / 131 |
| avg. / max. # of visits               | 2.37 / 29           |
| avg. / max. # of diagnoses per visit  | 10.51 / 128         |
| avg. / max. # of procedure per visit  | 3.84 / 50           |
| avg. / max. # of medication per visit | 11.44 / 65          |
| total # of DDI pairs                  | 448                 |

waiting to update,  $\nabla(\cdot)$  is the derivative operation. The adaptive curriculum manager is banded with the parameter update. Eq (28) is the critical step of the optimizer of the objective. We use the current iteration number and the current patient visit length to select the learning difficulty automatically.

## B. Inference

The SHAPE is trained end-to-end, and in the inference phase, the safe drug combination recommendation is generated from the sigmoid output  $\hat{y}^{(t)}$ , where we fix the threshold value as 0.5 to predict the label set. Then, the final predicted medication combinations correspond to the following:

$$\hat{Y}^{(t)} = \left\{ \hat{y}_i^{(t)} | \hat{y}_i^{(t)} > 0.5, 1 \leq i \leq |\mathcal{M}| \right\} \quad (32)$$

## V. EXPERIMENTS

In this section, we introduce the experiment details and conduct evaluation experiments to demonstrate the effectiveness of our SHAPE model.<sup>1</sup>

### A. Dataset

We use the EHR data from the Medical Information Mart for Intensive Care (MIMIC-III).<sup>2</sup> It contains 46,520 patients and 58,976 hospital admissions from 2001 to 2012. We conduct experiments on a benchmark released by COGNet [27], which is based on the MIMIC-III dataset for a fair comparison. Following the COGNet, we selected Top-40 severity DDI types from TWOSIDES [43], and we converted the drug code into ATC Third Level codes<sup>3</sup> to align with the DDI graph nodes. Finally, we followed the setting of COGNet and divided the dataset into training, validation, and test sets by the ratio of 4 : 1 : 1. The statistics of the post-processed data are reported in Table I.

### B. Metrics

We use three efficacy metrics: Jaccard, F1, and Precision-Recall Area Under Curve (PRAUC) combinations to evaluate the recommendation efficacy. Additionally, we also showed the DDI rate, and the number of predicted medications following the previous works [26], [27].

The Jaccard for the patient is calculated as below:

$$Jaccard = \frac{1}{T} \sum_{t=1}^T \frac{|M^t \cap \hat{Y}^{(t)}|}{|M^t \cup \hat{Y}^{(t)}|} \quad (33)$$

where the  $M^{(t)}$  is the ground-truth medication set sequence at  $t$ -th visit and the  $\hat{Y}^{(t)}$  is the predicted medication combinations.

The F1 of the patient is calculated as follows:

$$F1 = \frac{1}{T} \sum_{t=1}^T 2 \times \frac{P_t * R_t}{P_t + R_t} \quad (34)$$

$$P_i = \frac{|M^i \cap \hat{Y}^{(i)}|}{|\hat{Y}^{(i)}|} \quad (35)$$

$$R_i = \frac{|M^i \cap \hat{Y}^{(i)}|}{|M^i|} \quad (36)$$

The PRAUC is calculated with the ground truth code's predicted probability of each medication code.

$$PRAUC = \frac{1}{T} \sum_{t=1}^T \sum_{k=1}^{|\mathcal{M}|} P(k)_t (R(k)_t - R(k-1)_t) \quad (37)$$

where  $P(k)_t, R(k)_t$  are the precision and recall at the cut-off  $k$ -th threshold in the ordered retrieval list.

DDI rate aims to measure the interaction between the recommended medication combinations, which is calculated as follows:

$$DDI = \frac{1}{T} \sum_{t=1}^T \frac{\sum_{i=1}^{|\hat{Y}^{(t)}|} \sum_{j=i+1}^{|\hat{Y}^{(t)}|} \mathbf{1}\{A_d[\hat{Y}_i^{(t)}, \hat{Y}_j^{(t)}] = 1\}}{\sum_{i=1}^{|\hat{Y}^{(t)}|} \sum_{j=i+1}^{|\hat{Y}^{(t)}|} 1} \quad (38)$$

where  $A_d$  is the known knowledge of the DDI matrix.  $\hat{Y}_i^{(t)}$  denoted the  $i$ -th recommended medication and  $\mathbf{1}\{\cdot\}$  means to return 1 when the  $\{\cdot\}$  is true, otherwise, return 0.

### C. Baseline

We compare the SHAPE model with the following methods from different perspectives: *conventional machine learning method*, such as Logistic Regression(LR). *Instance-based methods*: LEAP [20], 4SDrug [21]. *Longitudinal-based methods*: RETAIN [22], DMNC [23], GAMNet [24], MICRON [25], SafeDrug [26], COGNet [27]. Specifically, LEAP [20] uses an attention mechanism to encode the diagnosis sequence step by step. 4SDrug [21] designs an attention-based method to augment the symptom representation and leverages the DDI graph to generate the current drug sequence. RETAIN [22] employs the attention gate mechanism to model the patient longitudinal information. DMNC [23] proposes a memory network to capture more interaction in the patient EHR record. GAMNet [24] combines the RNN and graph neural network to recommend medication combinations. MICRON [25] leverages a residual-based network to update the patient representation according to the new feature change. SafeDrug [26] utilizes drugs' molecule structures in the medication recommendation. COGNet [27] proposes a conditional generation model to copy or predict drugs according to the patient representation.

<sup>1</sup>[Online]. Available: <https://github.com/sherry6247/SHAPE>

<sup>2</sup>[Online]. Available: <https://mimic.physionet.org/>

<sup>3</sup>[Online]. Available: [https://www.whocc.no/atc/structure\\_and\\_principles/](https://www.whocc.no/atc/structure_and_principles/)

TABLE II  
PERFORMANCE COMPARISON ON THE MIMIC-III DATASET

| Model          | Jaccard                | F1                     | PRAUC                  | DDI                    | Avg.# of Drugs   |
|----------------|------------------------|------------------------|------------------------|------------------------|------------------|
| LR             | 0.4865 ± 0.0021        | 0.6434 ± 0.0019        | 0.7509 ± 0.0018        | 0.0829 ± 0.0009        | 16.1773 ± 0.0942 |
| LEAP(2017)     | 0.4521 ± 0.0024        | 0.6138 ± 0.0026        | 0.6548 ± 0.0033        | 0.0731 ± 0.0008        | 18.7138 ± 0.0666 |
| 4SDrug(2022)   | 0.4646 ± 0.0012        | 0.6263 ± 0.0012        | 0.7604 ± 0.0016        | <b>0.0540 ± 0.0004</b> | 14.6389 ± 0.0710 |
| RETAIN(2016)   | 0.4887 ± 0.0028        | 0.6481 ± 0.0027        | 0.7556 ± 0.0033        | 0.0835 ± 0.0020        | 20.4051 ± 0.2832 |
| DMNC(2018)     | 0.4864 ± 0.0025        | 0.6529 ± 0.0030        | 0.7580 ± 0.0039        | 0.0842 ± 0.0011        | 20.0000 ± 0.0000 |
| GAMNet(2019)   | 0.5067 ± 0.0025        | 0.6626 ± 0.0025        | 0.7631 ± 0.0030        | 0.0864 ± 0.0006        | 27.2145 ± 0.1141 |
| MICRON(2021)   | 0.5100 ± 0.0033        | 0.6654 ± 0.0031        | 0.7687 ± 0.0026        | 0.0641 ± 0.0007        | 17.9267 ± 0.2172 |
| SafeDrug(2021) | 0.5213 ± 0.0030        | 0.6768 ± 0.0027        | 0.7647 ± 0.0025        | 0.0589 ± 0.0005        | 19.9178 ± 0.1604 |
| COGNet(2022)   | 0.5336 ± 0.0011        | 0.6869 ± 0.0010        | 0.7739 ± 0.0009        | 0.0852 ± 0.0005        | 28.0903 ± 0.0950 |
| <b>SHAPE</b>   | <b>0.5513 ± 0.0009</b> | <b>0.7017 ± 0.0008</b> | <b>0.7906 ± 0.0009</b> | 0.0677 ± 0.0003        | 20.9949 ± 0.1189 |

The best results are highlighted in bold.

#### D. Parameter Setting

Here, we list the implementation details of SHAPE. We set the hidden dimension as 128 and use the Adam optimizer [42] with an initial learning rate  $1 \times 10^{-3}$  for 50 epochs. We fixed the random seed as 2023 to ensure the reproducibility of the model. Our model is implemented by Pytorch 1.7.1 based on Python 3.8.13 and training on two GeForce RTX 3090 GPUs, and an early-stopping mechanism was utilized. For a fair comparison, in the testing stage, we follow the previous work CONGNet [27], which random sample 80% data from test data for a round of evaluation. We repeat this process 10 times and calculate the mean and standard deviation as the final result we reported.

#### E. Result Analysis

As shown in Table II, our proposed model SHAPE outperforms all baselines with the higher Jaccard, F1, and AUPRC and increased by nearly 2% compared to the previous best model. The conventional LR and the Instance-based methods are poor as they only consider the patient’s health condition at the current visit. The performance of RETAIN and DMNC are comparable because both utilize the RNN architecture to capture the longitudinal information. The GAMENet introduces an additional DDI graph and fused it with the EHR co-occurrence graph, resulting in further performance improvement. SafeDrug leverages the drugs’ molecule structures to improve the performance of medication recommendations. Unlike most longitudinal algorithms, which focus on the historical record, the MICRON proposes using the residual network to capture changes in medications. The COGNet proposes the copy or prediction mechanism to generate the medication sequence since the statistics show that most medication codes have been recommended in historical EHR records. However, it fails to consider the short visit, which may not be enough historical reference, especially for the newly and secondly admission patients.

Compared with the baseline methods, our SHAPE model achieves state-of-the-art performance. On the one hand, it designs an intra-visit set encoder to collect the most informative medical events of each patient automatically. On the other hand, we develop an inter-visit longitudinal encoder to capture the longitudinal pattern, which inherits the merit of RNN and the attention mechanism. Besides, our adaptive curriculum manager assigns the difficulty of each sample based on the visit length accordingly. Hence, our SHAPE performance is better than the other methods.

We also noticed in Table II that the 4SDrug achieves the lowest and most charming DDI rate of predicted medication combinations. However, when considering the results shown in Fig. 4, the 4SDrug method achieves the lowest DDI rate, which is likely due to the lower count of predicted medication codes compared to other methods. Our observations indicate that the DDI rate tends to increase with the number of predicted medications. This lower DDI rate phenomenon also appears in the MICRON model since there are few predicted medications.

Furthermore, we noticed that the MIMIC-III dataset has an average DDI rate of 0.0875 itself, which means there is a large number of DDI phenomena in real-world practice. Based on this fact, our SHAPE also achieves a lower DDI rate and higher accuracy of medication recommendations, indicating the effectiveness of our proposed method.

To further validate that our SHAPE model can better model the short visit and even the new visit problem and recommend medication effectively, we investigate the performance of various visits with different models. As shown in the right picture of Fig. 1, there are severe long-tail phenomena in the MIMIC-III dataset, and most patients have less than five times admission records. We take patients’ first five visit records in the test set for visualization. We compared SHAPE with the COGNet and 4SDrug since (1) the COGNet achieves the best performance of the existing methods, and (2) the 4SDrug method uses the set-orient method to learn the code-level representation and uses the DDI loss to control the output predicted. As shown in Fig. 4, our SHAPE model is superior to the COGNet on the three metrics (i.e., Jaccard, F1, and PRAUC). Especially, our SHAPE achieves higher performance in the short visit length and shows an increasing trend. These results may directly show the power of SHAPE to solve the problem shown in Fig. 1, in which the short visit records are the critical samples. The higher accuracy of these samples is helpful for most situations in real-world clinical practices. On the contrary, the 4SDrug is always under the COGNet and SHAPE. The reason may be that the 4SDrug is an instance-based method that ignores temporal longitudinal information.

## VI. DISCUSSION

Upon analyzing the results in Table II, we can conclude that our proposed model SHAPE achieved the best performance compared to the LR and *Instance-base* and *Longitudinal-base* methods. The success of SHAPE is ascribed to the three

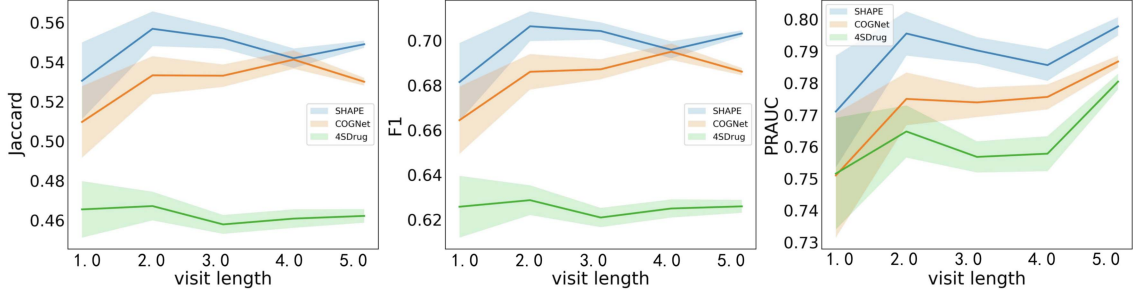


Fig. 4. Performance of different visit lengths with the various models.

TABLE III  
ABLATION STUDY FOR DIFFERENT SHAPE MODULES ON MIMIC-III DATASET

| Model                       | ISE | ILE | ACLM | DDI loss            | Jaccard                | F1                     | PRAUC                  | DDI rate               |
|-----------------------------|-----|-----|------|---------------------|------------------------|------------------------|------------------------|------------------------|
| SHAPE                       | ✓   | ✓   | ✓    | ✓                   | <b>0.5513 ± 0.0009</b> | <b>0.7017 ± 0.0008</b> | <b>0.7906 ± 0.0009</b> | 0.0677 ± 0.0003        |
| SHAPE <sub>w/SA</sub>       |     | ✓   | ✓    | ✓                   | 0.5404 ± 0.0008        | 0.6922 ± 0.0013        | 0.7845 ± 0.0011        | 0.0681 ± 0.0007        |
| SHAPE <sub>w/oISE</sub>     |     | ✓   | ✓    | ✓                   | 0.5280 ± 0.0011        | 0.6828 ± 0.0010        | 0.7739 ± 0.0009        | 0.0716 ± 0.0005        |
| SHAPE <sub>w/oILE</sub>     | ✓   |     | ✓    | ✓                   | 0.5243 ± 0.0016        | 0.6793 ± 0.0014        | 0.7718 ± 0.0018        | 0.0699 ± 0.0003        |
| SHAPE <sub>w/oACLM</sub>    | ✓   | ✓   |      | ✓                   | 0.5314 ± 0.0020        | 0.6856 ± 0.0018        | 0.7768 ± 0.0020        | 0.0660 ± 0.0004        |
| SHAPE <sub>w/oDDIloss</sub> | ✓   | ✓   | ✓    |                     | 0.5483 ± 0.0016        | 0.6989 ± 0.0014        | 0.7880 ± 0.0012        | 0.0857 ± 0.0005        |
| SHAPE                       | ✓   | ✓   | ✓    | $\alpha_{ddi}=0.1$  | 0.5411 ± 0.0014        | 0.6934 ± 0.0014        | 0.7848 ± 0.0012        | 0.0559 ± 0.0003        |
| SHAPE                       | ✓   | ✓   | ✓    | $\alpha_{ddi}=0.09$ | 0.5421 ± 0.0012        | 0.6945 ± 0.0011        | 0.7843 ± 0.0012        | 0.0571 ± 0.0004        |
| SHAPE                       | ✓   | ✓   | ✓    | $\alpha_{ddi}=0.08$ | 0.5451 ± 0.0017        | 0.6968 ± 0.0015        | 0.7850 ± 0.0015        | 0.0601 ± 0.0003        |
| SHAPE                       | ✓   | ✓   | ✓    | $\alpha_{ddi}=0.05$ | <b>0.5513 ± 0.0009</b> | <b>0.7017 ± 0.0008</b> | <b>0.7906 ± 0.0009</b> | <b>0.0677 ± 0.0003</b> |
| SHAPE                       | ✓   | ✓   | ✓    | $\alpha_{ddi}=0.02$ | 0.5507 ± 0.0012        | 0.7015 ± 0.0010        | 0.7919 ± 0.0010        | 0.0787 ± 0.0004        |
| SHAPE                       | ✓   | ✓   | ✓    | $\alpha_{ddi}=0.01$ | 0.5496 ± 0.0017        | 0.7009 ± 0.0014        | 0.7893 ± 0.0019        | 0.0844 ± 0.0005        |

The best results are highlighted in bold.

modules we proposed (i.e., the Intra-visit Set Encoder (ISE), the Inter-visit Longitudinal Encoder (ILE), and the Adaptive Curriculum Learning Module (ACLM)), and it achieved a lower DDI rate with our proposed combined loss function. To verify the effectiveness of each module we proposed, we designed the ablation experiments, SHAPE<sub>w/oISE</sub>: which remove the intra-visit set encoder and summarize the code-level to visit-level representation directly. SHAPE<sub>w/oILE</sub>: which uses the recurrent neural network to replace the inter-visit longitudinal encoder for learning the longitudinal information. SHAPE<sub>w/oACLM</sub>: which means removing the step of (28) and using the basic Adam optimizer to optimize the SHAPE. SHAPE<sub>w/oDDIloss</sub>: which only uses the multi-label classification loss function as the objective to train the model. We also compared the self-attention (SA) to investigate the effectiveness of our proposed compact intra-visit set encoder, SHAPE<sub>w/SA</sub>: which replaces the set encode as self-attention.

Table III shows the results for the different variants of SHAPE. As expected, when randomly removing the three modules we proposed. The performance brought a significant deterioration to the complete SHAPE model. The results of the DDI rate of SHAPE<sub>w/oDDIloss</sub> illustrate the effectiveness of the combination loss function. Overall, the SHAPE outperforms all variant models, which means each component is integral to SHAPE. Compared with the SHAPE, the SHAPE<sub>w/SA</sub> drops performance on total metrics, demonstrating that a more compacted encoder is more suitable to model the complex medical event code sequence.

Moreover, the performance drop of SHAPE<sub>w/oACLM</sub> can be observed in Table III, indicating that it is important to consider

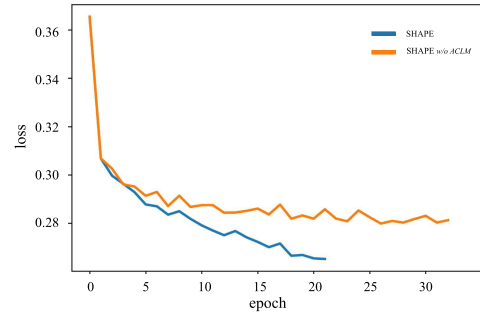


Fig. 5. Loss comparison on SHAPE and SHAPE<sub>w/oACLM</sub> regarding different numbers of train epochs.

the visit length as the guidance to assign the complex coefficient in the model of each patient. To explore the impact of the ACLM module, we conducted experiments to visualize the loss trajectory between SHAPE and SHAPE<sub>w/oACLM</sub>. As shown in Fig. 5, it can be seen that compared to SHAPE<sub>w/oACLM</sub>, SHAPE has a significant decrease in loss and converges quickly. This demonstrates the vital importance of the ACLM module, as it can automatically assign difficulty coefficients to each sample and learn more suitable parameters for various visit records.

Furthermore, to achieve a satisfactory trade-off for the DDI rate phenomenon in the medication combinations generated by SHAPE, we explore the hyperparameter  $\alpha$  in (26). The details are also shown in the second half of Table III, according to the results of Table III, we can conclude that: (1) the DDI rate of predicted medication combinations is gradually increasing with the decline of  $\alpha_{ddi}$ . (2) before the  $\alpha > 0.05$ , the performance of



TABLE IV  
EXAMPLE RECOMMENDED MEDICATIONS FOR A GIVEN PATIENT HEALTH CONDITION ON MIMIC-III

| Case 1        | Code  | hit | missed | error |
|---------------|---|-----|--------|-------|
| Diagnosis:    | 03849,5770,5849,5761,1623,5859,496,99592,57450,7904,28800,V1582,40390,4439,2720,28522,1991                          |     |        |       |
| Procedure:    | 5185,5188,5187,5293   |     |        |       |
| Ground truth: | D06A,J01C,N05A,A04A,J01M,A12A,A07A,N02A,A12B,A01A,B05C,N02B,B01A,A12C,A02B,A06A (16 codes)                          |     |        |       |
| Model         | Predicted codes   |     |        |       |
| 4SDrug        | N02B,A01A,A02B,A06A,B05C,A12A,A12C,A07A,C03C,A12B,N02A,J01M,B01A,J01D,A04A,R03A,R01A                                | 13  | 3      | 4     |
| COGNet        | N02B,A01A,A02B,A06A,B05C,A12A,A12C,C01C,A07A,C07A,A12B,N02A,J01M,B01A,R03A,R01A,J01C                                | 13  | 3      | 4     |
| SHAPE         | N02B,A01A,A02B,A06A,B05C,A12A,A12C,A07A,C07A,C03C,A12B,N02A,J01M,B01A,A04A,J01C                                     | 14  | 2      | 2     |
| Case 2        | Code  | hit | missed | error |
| Diagnosis:    | 03842,78552,V427,99592,4019,25000,V1007   |     |        |       |
| Procedure:    | 3893,03311  |     |        |       |
| Ground truth: | J05A,D11A,D01A,J01E,R05C,A07E,C01C,N05B,J01D,N06A,A12A,A07A,N02A,A12B,A01A,B05C,N02B,B01A,A12C,A02B,A06A (21 codes) |     |        |       |
| Model         | Predicted codes   |     |        |       |
| 4SDrug        | N02B,A01A,A02B,A06A,B05C,A12C,C01C,A07A,C07A,C03C,N02A,B01A,J01D,D11A,A07E,J05A,J01E,L04A                           | 15  | 6      | 3     |
| COGNet        | N02B,A01A,A02B,A06A,B05C,A12A,A12C,C01C,A07A,C07A,A12B,C02D,N06A,B01A,D01A,N05A,D11A,A04A,A07E,J05A,J01E,J01C,C02C  | 17  | 4      | 6     |
| SHAPE         | N02B,A01A,A02B,A06A,B05C,A12A,A12C,C01C,A07A,A12B,N02A,N06A,B01A,D01A,J01D,D11A,A07E,J05A,J01E,J01C,L04A            | 19  | 2      | 2     |

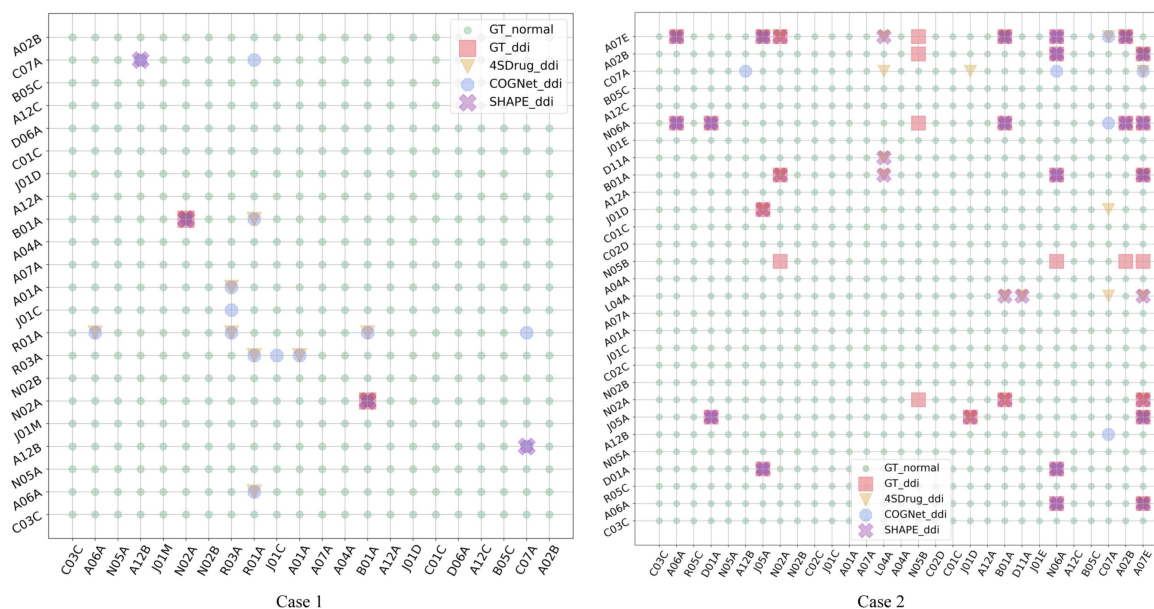


Fig. 6. Visualization DDI of the case study. *Case 1* is a new admission patient. *Case 2* is a secondary admission patient. In a chessboard, the red square corresponds to the DDI in the ground truth; the green point corresponds there are not appear DDI in the ground truth; the blue circle corresponds to the DDI in the predicted medications with 4SDrug, the inverted yellow triangle corresponds to the DDI predicted medications with COGNet, the purple cross corresponds to the DDI in the predicted medications with SHAPE. Best viewed in color.

other metrics is suppressed, which indicates the DDI rate and the accuracy performance of the predicted medication combination almost linearly decreases with the penalty weight. However, when the  $\alpha < 0.05$ , the performance of SHAPE fluctuated. Combined with the previously mentioned that the MIMI-III dataset has a 0.0875 DDI rate itself, which means not the lowest DDI rate is the superior optimal selection of clinical practice.

To intuitively demonstrate the advantages of SHAPE over the two baseline models, we analyze some examples to show the predicted results. We choose the short or new visit patients to demonstrate the model effect on harder predicted cases. Due to space constraints, we use the International Classification of

Disease (ICD) code to represent the diagnosis and procedure information and the ATC code to represent the medications. As shown in Table IV, *case 1* is a new admission patient, the doctor prescribed ground truth medication based on the diagnosis and procedure information of the patient's current visit. *Case 2* is a secondary admission patient, and we list the second record in *Case 2*. In *Case 2*, the physician combines the current health condition and the patient's historical record to prescribe medication. Overall, the SHAPE performed the best with 14 correct and 19 correct medications in two cases and achieved the lowest miss or error in the two cases. Furthermore, we noticed that in new visit *Case 1*, the instance-based method 4SDrug also achieves comparable performance with COGNet,

probably because of the instance-based approach against the single visit problem.

As shown in Fig. 6, we visualize the DDI status in two cases of each model, where the symmetric matrix shows the drug-drug relationship of the combination of medications. The point of  $GT_{normal}$  means there is no DDI in ground truth medication combinations, and  $GT_{ddi}$  means there probably is DDI in the ground truth medication combinations. The empty rows and columns mean these codes do not appear in the ground truth medications. We noticed in *Case 1* our SHAPE only generates two pairs of medication which maybe suffers the drug-drug interaction, on the contrary, the 4SDrug and COGNet generate five pairs (i.e., [A01 A, R03A], [A06 A, R01A], [N02 A, B01A], [B01 A, N02A], [B01 A, R01A]) and eight pairs (i.e., [A01 A, R03A], [A06 A, R01A], [C07 A, A12B], [C07 A, R01A], [A12B, C07A], [N02 A, B01A], [B01 A, N02A], [B01 A, R01A]). In the DDI of *Case 2*, we find that the DDI phenomenon in real-life scenarios exceeds ten pairs of medications. Our SHAPE simultaneously hits most situations similar to the ground truth medication prescribed by doctors, which hints that SHAPE can provide a safer way to recommend medication combinations.

There are also several limitations of the current study. Firstly, we only used diagnosis and procedure information for the side information to infer the medication and ignored others, such as vital signs and laboratory test records. Secondly, we only evaluate the SHAPE model on a public dataset, which also limits the generalizability of the model.

## VII. CONCLUSION

In this article, we proposed a sample adaptive hierarchical medication prediction network, named SHAPE, to better learn the accurate representation of the patient. Concretely, we first present an intra-visit set encoder to capture medical events relationship from the code-level perspective, which is usually ignored in most current works. Then, we developed an inter-visit longitudinal encoder to learn the visit-level longitudinal representation, which inherits the merits between attention and the RNN. Additionally, we designed an adaptive curriculum learning module that references patients' personalities to automatically assign each patient's difficulty for improving the performance of medication recommendations. Experiment results on the public benchmark dataset demonstrate that SHAPE outperforms existing medication recommendation algorithms by a large margin. We also investigate the performance of short visits and new visit samples, which shows that the SHAPE can effectively figure out the medication recommendation with the short admission of patients. Further ablation study results also suggest the effectiveness of each module of our proposed SHAPE.

## REFERENCES

- [1] A. Rajkumar et al., "Scalable and accurate deep learning with electronic health records," *NPJ Digit. Med.*, vol. 1, no. 1, pp. 1–10, 2018.
- [2] A. Esteva et al., "A guide to deep learning in healthcare," *Nature Med.*, vol. 25, no. 1, pp. 24–29, 2019.
- [3] F. Ma et al., "A general framework for diagnosis prediction via incorporating medical code descriptions," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2018, pp. 1070–1075.

- [4] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," *J. Amer. Med. Inform. Assoc.*, vol. 24, no. 2, pp. 361–370, 2017.
- [5] S. Liu et al., "A hybrid method of recurrent neural network and graph neural network for next-period prescription prediction," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 12, pp. 2849–2856, 2020.
- [6] S. Liu, X. Wang, Y. Xiang, H. Xu, H. Wang, and B. Tang, "Multi-channel fusion LSTM for medical event prediction using EHRs," *J. Biomed. Inform.*, vol. 127, 2022, Art. no. 104011.
- [7] M. Shani, Y. Schonmann, D. Comaneshter, and A. Lustman, "The relationship between patient medication adherence and following preventive medicine recommendation," *J. Amer. Board Fam. Med.*, vol. 34, no. 6, pp. 1157–1162, 2021.
- [8] P. Symeonidis, S. Chairistanidis, and M. Zanker, "Recommending what drug to prescribe next for accurate and explainable medical decisions," in *Proc. IEEE 34th Int. Symp. Comput.-Based Med. Syst.*, 2021, pp. 213–218.
- [9] Y. An, L. Zhang, M. You, X. Tian, B. Jin, and X. Wei, "MeSIN: Multi-level selective and interactive network for medication recommendation," *Knowl.-Based Syst.*, vol. 233, 2021, Art. no. 107534.
- [10] M. Wang, J. Chen, and S. Lin, "Medication recommendation based on a knowledge-enhanced pre-training model," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, 2021, pp. 290–294.
- [11] J. Shang, T. Ma, C. Xiao, and J. Sun, "Pre-training of graph augmented transformers for medication recommendation," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 5953–5959.
- [12] S. Zhang, J. Li, H. Zhou, Q. Zhu, S. Zhang, and D. Wang, "MERITS: Medication recommendation for chronic disease with irregular time-series," in *Proc. IEEE Int. Conf. Data Mining*, 2021, pp. 1481–1486.
- [13] N. Mahmoud and H. Elbeh, "IRS-T2D: Individualize recommendation system for type2 diabetes medication based on ontology and SWRL," in *Proc. 10th Int. Conf. Inform. Syst.*, 2016, pp. 203–209.
- [14] W. Zhao, X. Jiang, K. Wang, X. Sun, G. Hu, and G. Xie, "Construction of guideline-based decision tree for medication recommendation," *Stud. Health Technol. Inform.*, pp. 1–11, 2020.
- [15] Y. Wang, W. Chen, D. Pi, and L. Yue, "Adversarially regularized medication recommendation model with multi-hop memory network," *Knowl. Inf. Syst.*, vol. 63, no. 1, pp. 125–142, 2021.
- [16] Y. Wang, W. Chen, D. Pi, L. Yue, S. Wang, and M. Xu, "Self-supervised adversarial distribution regularization for medication recommendation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 3134–3140.
- [17] Y. Wang, W. Chen, D. Pi, L. Yue, M. Xu, and X. Li, "Multi-hop reading on memory neural network with selective coverage for medication recommendation," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 2020–2029.
- [18] Y. Si et al., "Deep representation learning of patient data from electronic health records (EHR): A systematic review," *J. Biomed. Inform.*, vol. 115, 2021, Art. no. 103671.
- [19] F. Gong, M. Wang, H. Wang, S. Wang, and M. Liu, "SMR: Medical knowledge graph embedding for safe medicine recommendation," *Big Data Res.*, vol. 23, 2021, Art. no. 100174.
- [20] Y. Zhang, R. Chen, J. Tang, W. F. Stewart, and J. Sun, "LEAP: Learning to prescribe effective and safe treatment combinations for multimorbidity," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2017, pp. 1315–1324.
- [21] Y. Tan et al., "4SDrug: Symptom-based set-to-set small and safe drug recommendation," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2022, pp. 3970–3980.
- [22] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 3512–3520.
- [23] H. Le, T. Tran, and S. Venkatesh, "Dual memory neural computer for asynchronous two-view sequential learning," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 1637–1645.
- [24] J. Shang, C. Xiao, T. Ma, H. Li, and J. Sun, "GAMENet: Graph augmented memory networks for recommending medication combination," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 1126–1133.
- [25] C. Yang, C. Xiao, L. Glass, and J. Sun, "Change matters: Medication change prediction with recurrent residual networks," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 3728–3734.
- [26] C. Yang, C. Xiao, F. Ma, L. Glass, and J. Sun, "SafeDrug: Dual molecular graph encoders for safe drug recommendations," 2021, *arXiv:2105.02711*.
- [27] R. Wu, Z. Qiu, J. Jiang, G. Qi, and X. Wu, "Conditional generation net for medication recommendation," in *Proc. ACM Web Conf.*, 2022, pp. 935–945.

- [28] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 10, pp. 1419–1428, 2018.
- [29] Y. Ren, Y. Shi, K. Zhang, X. Wang, Z. Chen, and H. Li, "A drug recommendation model based on message propagation and DDI gating mechanism," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 7, pp. 3478–3485, Jul. 2022.
- [30] A. E. Johnson et al., "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, pp. 1–9, 2016.
- [31] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.
- [32] A. Graves, M. G. Bellemare, J. Menick, R. Munos, and K. Kavukcuoglu, "Automated curriculum learning for neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1311–1320.
- [33] G. Hacohen and D. Weinshall, "On the power of curriculum learning in training deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2535–2544.
- [34] S. Basu, M. Gupta, P. Rana, P. Gupta, and C. Arora, "Surpassing the human accuracy: Detecting gallbladder cancer from USG images with curriculum learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20886–20896.
- [35] J. Guo et al., "Improving de novo molecular design with curriculum learning," *Nature Mach. Intell.*, vol. 4, no. 6, pp. 555–563, 2022.
- [36] Y. Gu, S. Zheng, Z. Xu, Q. Yin, L. Li, and J. Li, "An efficient curriculum learning-based strategy for molecular graph learning," *Brief. Bioinf.*, vol. 23, no. 3, 2022, Art. no. bbac099.
- [37] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3744–3753.
- [38] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
- [39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
- [40] A. Radford et al., "Improving language understanding by generative pre-training," 2018.
- [41] D. Hutchins, I. Schlag, Y. Wu, E. Dyer, and B. Neyshabur, "Block-recurrent transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 33248–33261.
- [42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
- [43] N. P. Tatonetti, P. P. Ye, R. Daneshjou, and R. B. Altman, "Data-driven prediction of drug effects and interactions," *Sci. Transl. Med.*, vol. 4, no. 125, 2012.