

Self-supervised Hierarchical Representation for Medication Recommendation

Yuliang Liang¹, Yuting Liu¹, Yizhou Dang¹, Enneng Yang¹, Guibing Guo¹(✉), Wei Cai³, Jianzhe Zhao¹, and Xingwei Wang²

¹ Software College, Northeastern University, China
guogb@swc.neu.edu.cn

² School of Computer Science and Engineering, Northeastern University, China

³ Neusoft Research of Intelligent Healthcare Technology, Co. Ltd., Laboratory of Medical Information Engineering, China

Abstract. Medication recommender is to suggest appropriate medication combinations based on a patient’s health history, e.g., diagnoses and procedures. Existing works represent different diagnoses/procedures well separated by one-hot encodings. However, they ignore the latent hierarchical structures of these medical terms, undermining the generalization performance of the model. For example, “Respiratory Diseases”, “Chronic Respiratory Diseases” and “Chronic Bronchiti” have a hierarchical relationship, progressing from general to specific. To address this issue, we propose a novel hierarchical encoder named HIER to hierarchically represent diagnoses and procedures, which is based on standard medical codes and compatible with any existing methods. Specifically, the proposed method learns relation embedding with a self-supervised objective for incorporating the neighbor hierarchical structure. Additionally, we develop the position encoding to explicitly introduce global hierarchical position. Extensive experiments demonstrate significant and consistent improvements in recommendation accuracy across four baselines and two real-world clinical datasets. The code is available at <https://github.com/yuliang-liang/HEIR>.

Keywords: Self-supervised learning · Hierarchical representation · Medication recommendation.

1 Introduction

Medication recommendation systems understand patients’ health status from electronic health records (EHRs) and tailor a safe medication combination. In EHRs, each patient’s medical history is documented as a series of visits, typically involving diagnoses, procedures and medications. Existing works often represent diagnoses and procedures with one-hot encodings. The model then learns their embeddings in a low-dimensional space from patient data [22,17,23]. This embedding method organizes symbolic objects (e.g., words, entities and concepts) and has proven successful in numerous applications [3]. In healthcare, a universal language—namely, a standard coding system—is used to categorize diseases

and conditions hierarchically. This hierarchical structure enables constraints to be imposed on the embedding space, allowing the model to effectively learn commonalities within each disease category.

Recent studies in medication recommendation mainly focus on integrating drug-related knowledge into models, such as drug-drug interaction graphs [28,29], molecular structures of drugs [22,23], and medical knowledge graphs linking diseases with medications [26]. These methods improve the accuracy and safety of recommendations by enhancing medication representations. However, the representation of diagnoses and procedures, along with their hierarchical structures, remains underexplored.

In medication recommendations, patient heterogeneity introduces a wide range of diagnoses and procedures, making it difficult for the model to effectively learn their feature representations from scratch. For example, in the clinical dataset MIMIC-III [7], there are only 5,449 patients with at least two visits, yet exhibit 4,491 unique diagnoses and 1,412 unique procedures. On the other hand, the current models employ pooling operations (e.g., sum-pooling) to aggregate multiple diagnosis/procedure features as patient representation. Then, the model is optimized to match patients with medications, overlooking the structural properties (hierarchies) of diagnosis and procedure. We present the theoretical analysis in the following section. Due to the above issues, this paper leverages prior hierarchies to construct hierarchical representations rather than learning their complicated relationships from data.

In clinical practice, it is a well-recognized principle that patients with similar diagnoses are often prescribed similar medications. Motivated by this, we highlight the importance of the hierarchical structures of diagnoses and procedures and exploit them to obtain more accurate representations. Our approach is grounded in the use of standardized medical codes, in which diagnoses and procedures exhibit a tree-like structure (hierarchical structure) from general to specific, corresponding to their conceptions. To be more specific, each diagnosis and procedure is tagged by a unique code, i.e., ICD coding⁴ in EHRs. However, current methods map diagnoses and procedures directly to one-hot encodings, which unfortunately disregard their hierarchical structure.

To bridge the research gap, we propose the Hierarchical encoder named **HIER** for diagnosis and procedure representations guided by medical codes. We explicitly equip the model with hierarchical structures to enhance their representations and improve recommendation accuracy. Specifically, the proposed hierarchical encoder consists of two components: *Relation Embedding* and *Position Encoding*. The relation embedding aggregates their parental categories and is optimized with a self-supervised objective to capture neighbor hierarchical structure. The position encoding, on the other hand, is derived from segmented medical codes to introduce global hierarchical positions. Integrating these two components enables us to obtain hierarchical representations that are compati-

⁴ The International Classification of Diseases (ICD) is an international classification system used by healthcare providers to code and categorize diseases, conditions, etc.

ble with any medication recommendation model. The main contributions of this paper are as follows:

- We demonstrate, for the first time, that state-of-the-art medication recommendation models fail to adequately capture the structural relations of diagnoses and procedures.
- We introduce a novel hierarchical encoder, guided by medical codes, to integrate the hierarchical structure into the representation of medical terms (e.g., diagnoses and procedures). This hierarchical representation is universal and independent of the training process, compatible with any existing medication recommendation model.
- We conduct comprehensive experiments to validate the effectiveness of our method, demonstrating its significant and consistent performance improvements across four baselines and two real-world datasets.

2 Related Works

2.1 Medication Recommendation

The medication recommender provides the most suitable medications based on a patient’s health conditions. Based on the input of single or multiple visits, medication recommendation can be categorized into instance-based [27,5] and longitudinal methods [22,26]. Early studies are instance-based methods [27,16] and can only prescribe on single-visit patient data. LEAP [27] formulates the medication recommendation as a multi-instance, multi-label learning problem, mapping a set of instances, i.e., disease conditions to the associated medication sets. CompNet [16] formulates the task as an order-free Markov decision process, employing deep Q-Learning to capture correlations between medicines.

Recently, the longitudinal methods have attracted growing research interest [12,26,23]. Longitudinal-based methods make use of the temporal dependencies among multiple visits in patients’ medical history. RETAIN [2] is a two-level attention-based model that detects influential past visits and significant clinical variables within those visits. GAMENet [12] proposes a graph-augmented memory module to store historical drug information as references for further prediction. SafeDrug [22] extracts and encodes drug molecule structure to correlate with its functionalities and improve medication recommendation safety. MICRON [21] adopt a residual-based recurrent network and takes the changes in patient health records as input to update a hidden medication vector. MoleRec [23] leverage a molecular substructure-aware attentive method to model substructures’ interactions and relevancy to patient’s health condition. However, these methods do not target the representation of medical diagnoses and procedures, leaving much room for improvement.

2.2 The Representation of Medical Codes

ICD coding is a medical glossary established by the World Health Organization that provides specific coding rules and descriptions for diagnoses and procedures.

Recently, several works have exploited the hierarchy of ICD codes for health applications, mainly focusing on automated ICD coding [17,8,13] and disease prediction [1,11]. The hierarchical information of ICD codes reduces the semantic searching space of predictions. GRAM [1] employs a weighted average of the ancestors of a medical concept for diagnosis prediction but requires an expanded embedding matrix, which places a significant burden on training. [19] leverages graph convolutional network (GCN) to capture the hierarchical relationships among medical codes and their semantics. They employ a label-dependent attention mechanism, allowing the model to learn distinct clinical document representations for predicting diagnosis or procedure codes. HieNet [17] proposes a bidirectional hierarchy passage encoder to capture the codes' hierarchical representations and a progressive predicting method for automated ICD coding.

Notably, these end-to-end methods cannot be applied to medication recommendation systems. To bridge this gap, we aim to devise a simple yet effective hierarchical encoder to integrate their hierarchy into the models.

3 Preliminary

3.1 Electrical Health Records (EHR)

EHR data is a collection of patients' medical histories in the format of medical codes (e.g., diagnoses, procedures, and medications). Formally, the EHR for patient i can be represented as a sequence of visits: $\mathbf{X}_i = [\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \dots, \mathbf{x}_i^{(T_i)}]$ where T_i is the total number of visits for patient i . To minimize confusion, we will omit the patient index i as long as it is clear. Each visit of patient can be represented by $\mathbf{x}^{(t)} = [\mathbf{d}^{(t)}, \mathbf{p}^{(t)}, \mathbf{m}^{(t)}]$ containing the corresponding diagnosis codes $\mathbf{d}^{(t)} = \{d_1, d_2, \dots\} \in \mathbb{R}^{|\mathcal{D}|}$, procedure codes $\mathbf{p}^{(t)} = \{p_1, p_2, \dots\} \in \mathbb{R}^{|\mathcal{P}|}$ and medication codes $\mathbf{m}^{(t)} = \{m_1, m_2, \dots\} \in \mathbb{R}^{|\mathcal{M}|}$, where $\mathcal{D}, \mathcal{P}, \mathcal{M}$ are all possible code sets, and $|\cdot|$ denotes the cardinality.

3.2 Patient Representation

In medication recommendation, a patient is typically represented based on their diagnoses and procedures [12,22,26]. Current methods transform each diagnosis/procedure into a unique one-hot encoding, i.e., ID, subsequently projecting it into embeddings using two learnable embedding tables $\mathbf{E}_d \in \mathbb{R}^{|\mathcal{D}| \times \dim}$ and $\mathbf{E}_p \in \mathbb{R}^{|\mathcal{P}| \times \dim}$ for diagnoses and procedures, respectively. The embedding dimension is denoted by \dim . This process is formalized as follows:

$$\mathbf{d}_e^{(t)} = \mathbf{E}_d \cdot \mathbf{d}^{(t)}, \quad \mathbf{p}_e^{(t)} = \mathbf{E}_p \cdot \mathbf{p}^{(t)} \quad (1)$$

where $\mathbf{d}^{(t)}$ and $\mathbf{p}^{(t)}$ are the diagnosis and procedure sets (multi-hot vectors) for the visit t of a patient; $\mathbf{d}_e^{(t)}$ and $\mathbf{p}_e^{(t)}$ are the corresponding embedding sets with variable-size.

Generally, a patient has multiple diagnoses and multiple procedures in a visit. Therefore, their embeddings are aggregated using a pooling operator $\phi(\cdot)$, such as sum-pooling and average-pooling. The role of pooling is to merge a collection of diagnoses/procedure features into a single summary feature, while ensuring permutation invariance to input features. It can be formalized as follows:

$$\mathbf{e}_d^{(t)} = \phi(\mathbf{d}_e^{(t)}), \quad \mathbf{e}_p^{(t)} = \phi(\mathbf{p}_e^{(t)}) \quad (2)$$

where $\mathbf{e}_d^{(t)}, \mathbf{e}_p^{(t)} \in \mathbb{R}^{dim}$ are the aggregated diagnosis and procedure embeddings. They maintain a constant dimensionality across visit t .

For representing longitudinal patient data, two distinct Recurrent Neural Networks (RNNs) [4] are employed to separately model the dynamic histories of patient diagnoses and procedures in different visits.

$$\mathbf{h}_d^{(t)} = \text{RNN}_d(\mathbf{e}_d^{(t)}, \mathbf{h}_d^{(t-1)}), \quad \mathbf{h}_p^{(t)} = \text{RNN}_p(\mathbf{e}_p^{(t)}, \mathbf{h}_p^{(t-1)}) \quad (3)$$

where $\mathbf{h}_d^{(t)}, \mathbf{h}_p^{(t)} \in \mathbb{R}^{dim}$ are the hidden states of RNNs. The hidden states of $\mathbf{h}_d^{(0)}$ and $\mathbf{h}_p^{(0)}$ are initialize from zero.

Finally, we concatenate the two hidden states to generate the comprehensive patient representation $\mathbf{e}_{pt}^{(t)}$.

$$\mathbf{e}_{pt}^{(t)} = \text{Concate}(\mathbf{h}_d^{(t)}, \mathbf{h}_p^{(t)}) \quad (4)$$

As shown in Fig. 1(a), the patient representation is then processed through various model backbones to predict medication combinations (in this paper, we refer to the remaining modules of the model as the backbone). We do not show RNN in Fig. 1, because it is unnecessary for the instance-based medication recommendation methods.

3.3 International Classification of Diseases (ICD)

ICD provides a comprehensive list of codes for diseases, conditions, injuries, surgeries, and other medical procedures. We take extensively used ICD-9-CM⁵ as an example, which is an adaptation of ICD-9 specifically for use in the United States. Other versions of ICD codes follow a similar hierarchical structure. In ICD-9-CM, the three-digit codes are the heading for a category that can be further subdivided into more specific codes using fourth and/or fifth digits. We illustrate the detailed code structure for ‘‘Diabetes mellitus’’ as follows:

$$\underbrace{\boxed{2} \boxed{5} \boxed{0}}_{\text{Category}} . \underbrace{\boxed{0} \boxed{1}}_{\text{Etymology}} \quad (5)$$

– 250 *Diabetes mellitus*

⁵ <https://www.cms.gov/medicare/coding-billing/icd-10-codes/icd-9-cm-diagnosis-procedure-codes-abbreviated-and-full-code-titles>

- **250.0** *Diabetes mellitus without mention of complication*
 - * **250.01** *Diabetes mellitus without mention of complication, type I [juvenile type], not stated as uncontrolled*

where $\boxed{2}\boxed{5}\boxed{0}$ categorizes “Diabetes mellitus”. The fourth and fifth digits of $\boxed{0}\boxed{1}$ provide additional code notes for the etiology (manifestation) specific to the particular diabetic manifestation. Additionally, different code ranges are meaningful. For instance, 249-259 and 240-279 stand for “Diseases of other endocrine glands” and “Endocrine, nutritional and metabolic diseases, and immunity disorders”.

4 Proposed Method

In this section, we introduce a simple yet effective hierarchical encoder, **HIER**, to incorporate the hierarchical structures based on medical codes and construct the hierarchical representation. As illustrated in Fig. 1(b), in terms of design, our method can seamlessly integrate into the existing methods by replacing the traditional embedding layer. By doing so, we can effectively enhance feature representations, improving recommendation accuracy.

4.1 Hierarchical Encoder

This paper targets the medical code of diagnoses and procedures, but our method can also be used in other tree-structured terms. Without loss of generality, we refer to them as medical codes. As illustrated in Fig. 1(b), the proposed hierarchical encoder first aggregates the target medical code with higher-level codes (e.g., parent codes) to generate relation embedding. It can establish neighbor hierarchical relationships in the embedding space. Afterward, the position encoding introduces the global hierarchical position from the separated segments of codes. Finally, we obtain the hierarchical representation by merging the relation embedding with the position encoding.

Relation Embedding The relation embedding establishes relative hierarchical relationships in diagnosis and procedure embeddings. Firstly, we transform the medical coding system, such as ICD-9-CM, into two undirected tree-structured graphs for diagnoses and procedures, respectively. This structure describes the hierarchy (topology) among the medical codes, uniformly denoted by $T = (\mathcal{V}, \mathcal{E})$.

Then, we encode the target medical code v by aggregating its embedding with its parent codes denoted by $p(v)$ within the tree. For instance, the diagnosis code 250.01 can be derived through the sequence of higher-level codes (grandparent and parent codes) $250 \rightarrow 250.0 \rightarrow 250.01$. This aggregation from higher-level codes provides a relative hierarchical connection between the target code and its

Self-supervised Hierarchical Representation for Medication Recommendation

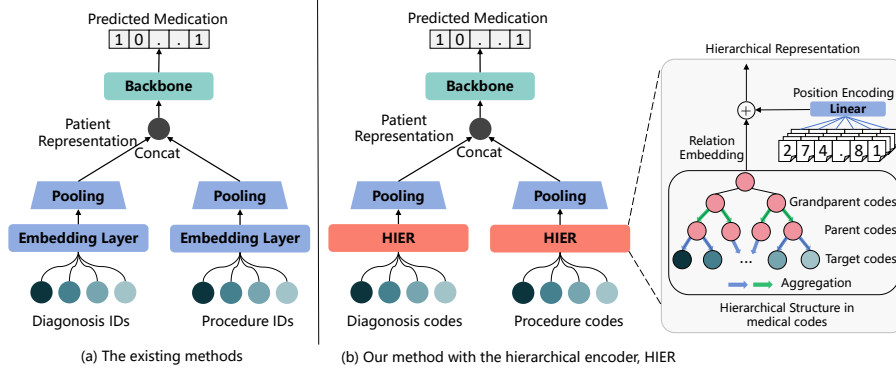


Fig. 1. An illustration of medication recommendation models, where (a) is the paradigm of existing methods; (b) is our method with the hierarchical encoder.

ancestors. We formalize the aggregation process as follows:

$$\mathbf{e}_v^{k+1} = \text{Aggregator}(\mathbf{e}_v^k, \mathbf{e}_{p(v)}^k) \quad (6)$$

$$= \sigma(\mathbf{W}^{k+1} \cdot \text{Concat}(\mathbf{e}_v^k, \mathbf{e}_{p(v)}^k)) \quad (7)$$

where $\mathbf{e}_v^{k+1} \in \mathbb{R}^{dim}$ is the aggregated embedding of v at layer $k+1$; $\mathbf{e}_v^k, \mathbf{e}_{p(v)}^k \in \mathbb{R}^{dim}$ are the embeddings of v and its parent $p(v)$ at layer k , respectively. σ is the rectified linear unit (ReLU) activation function. $\mathbf{W}^{k+1} \in \mathbb{R}^{2dim \times dim}$ is the weight matrix.

By aggregating K layers, we obtain the relation embedding for the medical code. We randomly initialize each node embedding at $k=0$ and optimize the weight matrix with a self-supervised objective (which will be introduced later) to preserve the hierarchical structure in the embedding space. Finally, the derived relation embedding can be used in various downstream tasks, such as medication recommendation. It can be formulated as follows:

$$\mathbf{r}_v \leftarrow \mathbf{e}_v^K, \forall v \in \mathcal{V} \quad (8)$$

where $\mathbf{r}_v \in \mathbb{R}^{dim}$ is the relation embedding for the medical code v , after K layers of aggregation.

Position Encoding The above relation embedding provides hierarchical information on the neighbor structure. To incorporate the global hierarchical feature, we devise a hierarchical position encoding for medical codes. Specifically, we first separate the medical code into a sequence of digits. Let $\{v_1, \dots, v_L\}$ represent the sequence of digits of the medical code v , where L is the number of digits. Then, each digit v_l is transformed to one-hot encoding. Finally, we concatenate

all one-hot encodings and obtain the multi-hot encoding.

$$\text{PE}_v = [\text{PE}_{(v_1)}, \text{PE}_{(v_2)}, \dots, \text{PE}_{(v_L)}] \quad (9)$$

where $\text{PE}_{(v_l)}$ is the one-hot encoding of the l -th digit in the medical code v ; PE_v is multi-hot position encoding. For the short code, zero padding is applied to the missing digits to maintain dimensional consistency. To adjust the position encoding to a specific dimension, we adopt the linear layer followed by a ReLU function σ . The ReLU activation function ensures that the projected position encoding remains positive, preventing symbol conflicts when facing the pooling operator. We denote the final position encoding as \mathbf{s}_v .

$$\mathbf{s}_v = \sigma(\text{Linear}(\text{PE}_v)) \quad (10)$$

By adding the relation embedding \mathbf{r}_v with the position encoding \mathbf{s}_v , we obtain the hierarchical representation of medical codes $\mathbf{z}_v \in \mathbb{R}^{dim}$.

$$\mathbf{z}_v \leftarrow \mathbf{r}_v + \mathbf{s}_v \quad (11)$$

4.2 Implementation

In order to seamlessly integrate our method into various SOTA medication recommendation backbones, we propose a simple and efficient hierarchical encoder plug-in to replace the traditional embedding layer, as shown in Fig. 1. The relation embedding is universal and generated independently from the model training, ensuring almost without additional computational and parameter overhead.

Specifically, we optimize the relation embedding in a self-supervised manner to preserve the hierarchical structure. We encourage the relation embedding \mathbf{r}_v of medical code v to be closer to the embedding of its parent code $\mathbf{r}_{p(v)}$ and farther away from the randomly selected one \mathbf{r}_{rand} . The self-supervision optimization objective is formulated as follows:

$$\mathcal{L} = -\frac{1}{N} \sum_{n=1}^N [\log(\mathbf{r}_v \cdot \mathbf{r}_{p(v)}) - \log(\mathbf{r}_v \cdot \mathbf{r}_{rand})] \quad (12)$$

where N is the number of samples in the batch; $\mathbf{r}_{p(v)}$ and \mathbf{r}_{rand} are the relation embedding of its parent and randomly selected code.

Next, we generate the position encoding \mathbf{s}_v as previously described, and add the relation embedding \mathbf{r}_v to obtain the hierarchical representation $\mathbf{z}_v \leftarrow \mathbf{r}_v + \mathbf{s}_v$. During training for specific tasks, we only adjust the parameters of derived relation embedding and the weight matrix of position encoding. The pseudocode of HIER is shown in Fig.2.

4.3 Discussion and Insights

In this subsection, we present some theoretical insights to show the limitations of existing methods. We analyze from two perspectives: *Alignment and Uniformity*, as well as *Feature Aggregation*. They can address two crucial questions:


```

# Replace torch.nn.Embedding with Relation_Embedding and Position_Embedding.
class Medication_Recommendation_Model(nn.Module):
    def __init__(self):
        ...
        # Original embeddings
        # procedure_embedding = Embedding(proc_num_embeddings, embedding_dim)
        # condition_embedding = Embedding(cond_num_embeddings, embedding_dim)

        # Relation-based embeddings
        procedure_relation_embedding = Relation_Embedding(proc_num_embeddings,
            embedding_dim)
        condition_relation_embedding = Relation_Embedding(cond_num_embeddings,
            embedding_dim)

        # Position-based embeddings
        procedure_position_embedding = Position_Embedding(proc_num_embeddings,
            embedding_dim)
        condition_position_embedding = Position_Embedding(cond_num_embeddings,
            embedding_dim)
        ...

    def forward(self, x):
        ...
        # Original features
        # procedure_features = procedure_embedding(procedures)
        # condition_features = condition_embedding(conditions)

        # Enhanced hierarchical features
        procedure_features = procedure_relation_embedding(procedures) + \
            procedure_position_embedding(procedures)
        condition_features = condition_relation_embedding(conditions) + \
            condition_position_embedding(conditions)
        ...

```

Fig. 2. HEIR Pseudocode in PyTorch implementation

What knowledge do existing models derive from the data? Inspired by [18], we analyze the Binary Cross-Entropy (BCE) loss, a widely used loss function in medication recommendation. We derive that the BCE loss favors the two properties of alignment and uniformity from Eq.(13) to Eq.(14). For the proof, please see the Appendix.

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\sigma(\hat{y}_i)) + (1 - y_i) \log(1 - \sigma(\hat{y}_i))] \quad (13)$$

$$= -\frac{1}{N} \sum_{i=1}^N \underbrace{[y_i f(u_i)^\top f(m_i)]}_{\text{Alignment}} - \underbrace{\log(1 + e^{f(u_i)^\top f(m_i)})}_{\text{Uniformity}} \quad (14)$$

where y_i and \hat{y}_i represent the label and prediction for sample i , respectively; σ is the activation function and N denotes the number of batch samples.

Assuming u is the patient and m is the medication, the neural network $f(\cdot)$ maps them into a low-dimensional embedding $f(u)$ and $f(m)$, respectively. If the matching score is determined through the dot product $f(u)^\top f(m)$, then we can have Eq. (14), where the optimization objective includes both Alignment and Uniformity terms. The first term encourages the alignment of paired patient u_i and medication m_i . If $y_i = 1$, it maximizes the inner product of $f(u_i)$ and $f(m_i)$. Consequently, the matched patient and medication features align closely. The second term preserves maximal information by the uniform distribution of patient and medication features on the unit hypersphere [18,15]. In summary, the existing models essentially optimize the objectives in terms of Alignment and Uniformity. That is, it matches the patient with possible medications while uniformly distributing the features of patients and medications.

Why do existing models fail to capture hierarchical relationships? In medication recommendation, the model aggregates multiple diagnosis/procedure features with the pooling operator ϕ (e.g., sum-pooling), which also makes neural networks permutation-invariant for the features. [24] shows that the pooling operator encodes the entire set (i.e., multiple features), and then the model backbone, in principle, recovers all the input information and performs relational reasoning from there. However, any such relational reasoning is not built into the model, and must be entirely learned, leading to an additional training burden [10,25,14]. Hence, the model focuses more on individual attributes of the inputs and less on encoding information about their interactions and relationships. In our case, the model struggles to adequately learn the relations of diagnoses and procedures, particularly their hierarchical relations. Therefore, in this paper, we explicitly incorporate the hierarchical structure of the medical codes rather than relying solely on learning from data.

5 Experiment

To verify the superiority and effectiveness of the proposed HIER method, we perform extensive experiments to answer the following research questions:

- **RQ1:** How does our method perform compared to various SOTA baselines on different datasets?
- **RQ2:** How do individual components affect the model performance?
- **RQ3:** How does our method improve convergence efficiency during training?
- **RQ4:** Do the traditional embedding method and our approach capture hierarchical relationships?

5.1 Experimental Setup

The models are trained on Ubuntu 22.04 with NVIDIA RTX 3090. The implementation is based on the Pyhealth framework [20]. We follow most of the hyper-parameter configures of the original baselines. We set the learning rate to

Table 1. The statistics of MIMIC-III and MIMIC-IV datasets.

Item	MIMIC-III	MIMIC-IV
# of visits / # of patients	14,141 / 5,449	147,406 / 43,939
diag. / proc. / med. space size	4,491 / 1,412 / 193	19,184 / 10,603 / 200
avg. / max # of visits	2.59 / 29	3.35 / 70
avg. / max # of diag. per visit	31.65 / 459	48.98 / 1,318
avg. / max # of proc. per visit	9.46 / 85	8.763 / 164
avg. / max # of med. per visit	29.74 / 105	23.83 / 111

$\{10^{-3}, 10^{-4}, 10^{-5}\}$, and set the epochs to 25 (sufficient to achieve optimal performance). For a fair comparison, we only modify the baselines by replacing the embedding layers with HIERs. Following the recent studies [22,23,26], we evaluate the model using Jaccard Similarity Score (Jaccard), F1 score, Precision-Recall AUC (PRAUC), and Drug-Drug Interaction rate (DDI). A lower DDI indicates better performance, while higher values are better for other metrics.

Datasets We perform the experiments on two extensively used datasets MIMIC-III [7] and MIMIC-IV [6]. The statistics of the pre-processed data are reported in Tab. 1. Following the pre-processing of [12,22], we filter out the patients with only one visit, as they do not reflect long-term health status and medication response. It is worth noting that we retain tail diagnoses and procedures for a more realistic clinical setting. We split the patients into training, validation, and test sets with a ratio of 8:1:1.

Baselines We consider the following popular and advanced baseline methods for medication recommendation. For a fair comparison, our method only modifies the baseline model’s embedding layers with HIER. Notably, there is no directly available hierarchical representation method for medication recommendation, as related works are end-to-end models for other medical applications.

- RETAIN [2] uses a two-level attention mechanism to detect influential past visits and significant clinical variables within those visits, e.g., key diagnoses.
- GAMENet [12] propose graph augmented memory networks to encode longitudinal patient data and drug-drug interaction graphs.
- SafeDrug [22] leverages drug molecule structures and DDI graph to recommend effective and safe medication combinations.
- MoleRec [23] propose a substructure-aware attentive method that models substructures’ interactions and relevancy to a patient’s health condition.

5.2 Performance Comparison (RQ1)

As shown in Tab. 2, we evaluate the performance of our hierarchical encoder HIER with four baseline methods (RETAIN, GAMENet, SafeDrug and MoleRec). We observe that our method consistently improves performance across

Table 2. Performance comparison with four baseline methods across two datasets.

Methods	MIMIC-III				MIMIC-IV			
	Jaccard(\uparrow)	F1-score(\uparrow)	PRAUC(\uparrow)	DDI(\downarrow)	Jaccard(\uparrow)	F1-score(\uparrow)	PRAUC(\uparrow)	DDI(\downarrow)
RETAIN [2]	0.4645	0.6224	0.7376	0.0680	0.4673	0.6197	0.7390	0.0629
	+Ours	0.4837	0.6405	0.7540	0.0660	0.4710	0.6236	0.7438
GAMENet [12]	0.4485	0.6075	0.7136	0.0679	0.4544	0.6073	0.7173	0.0652
	+Ours	0.4739	0.6307	0.7402	0.0663	0.4777	0.6305	0.7506
SafeDrug [22]	0.4541	0.6132	0.7174	0.0710	0.4519	0.6038	0.7121	0.0638
	+Ours	0.4745	0.6314	0.7405	0.0692	0.4704	0.6219	0.7420
MoleRec [23]	0.4572	0.6159	0.7157	0.0684	0.4302	0.5865	0.6962	0.0642
	+Ours	0.4627	0.6213	0.7252	0.0652	0.4324	0.5904	0.7000

all baselines in terms of Jaccard, F1 and PRAUC in both datasets. For example, our method improves GAMENet by 5.66%, 3.82% and 3.73% in Jaccard, F1, and PRAUC, respectively, on the MIMIC-III dataset. This is attributed to the fact that our method effectively improves the representations of diagnoses and procedures by incorporating their hierarchies. We also find that HIER reduces DDIs in certain cases, except for similar results compared to RETAIN and MoleRec on MIMIC-IV. While our method does not specifically target drug-drug interactions, the results suggest that improving the representations of diagnoses and procedures can help reduce DDIs.

It is noteworthy that the classic model, RETAIN, achieves the best performance except for DDIs than other baselines on both datasets. For example, in the MIMIC-III dataset, RETAIN achieves the best result of PRAUC 0.7376, while GAMENet, SafeDrug, and MoleRec achieve only 0.7136, 0.7174, and 0.7157, respectively. This is due to the realistic and challenging experimental setup that does not filter out tail diagnoses and procedures. RETAIN effectively predicts medications by identifying the importance of visits and the conditions of patients. The simple architecture of RETAIN is more effective than others. In addition, MoleRec’s performance collapses in MIMIC-IV. This may be because the model fails to correlate the fine-grained substructures of drug molecules with sparse and diverse diagnoses and procedures in this dataset. This phenomenon inspires us to explore more effective model architectures for future work.

5.3 Ablation Study (RQ2)

We conduct an ablation study for the relation embedding and the position encoding to evaluate their effectiveness. The experiments are conducted on the MIMIC-III dataset with the baseline of SafeDrug. We have two variants when only using relation embedding or position encoding denoted by ‘w/ E’ and ‘w/ P’, respectively. As shown in Table. 3, we observe that significant performance gains can still be achieved by using either relation embedding or position encoding. The position encoding can perform better than the relation embedding, thanks to its global hierarchical position. These results demonstrate the effectiveness of hierarchical information. Finally, our method obtained the best results on all metrics with the combination of the two components.

Self-supervised Hierarchical Representation for Medication Recommendation

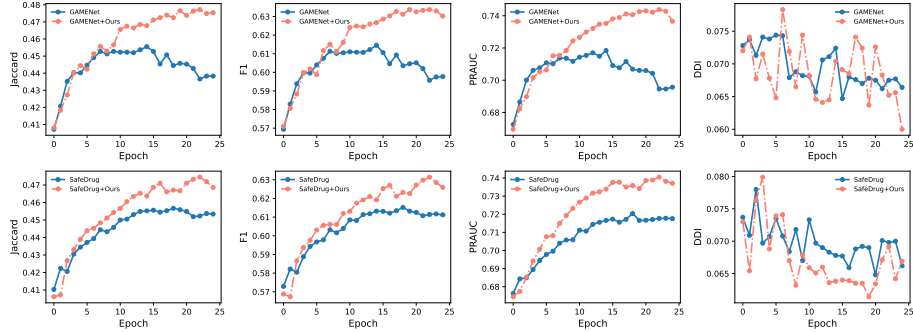


Fig. 3. Convergence efficiency curve in the test set during training stage: the first row for GAMENet and the second row for SafeDrug. Our method significantly improves the training convergence efficiency in both baselines.

Table 3. Ablation study on MIMIC-III dataset. The models with relation embedding or position encoding are denoted by ‘ $w/$ E’ and ‘ $w/$ P’, respectively.

Model	Jaccard(\uparrow)	F1-score(\uparrow)	PRAUC(\uparrow)	DDI(\downarrow)
w/o E&P	0.4541	0.6132	0.7174	0.0710
$w/$ E	0.4668	0.6245	0.7346	0.0698
$w/$ P	0.4717	0.6292	0.7368	0.0727
Ours	0.4745	0.6314	0.7405	0.0692

5.4 Convergence Efficiency (RQ3)

We also analyze the training convergence efficiency in comparison with two baselines, GAMENet and SafeDrug, as shown in Fig.3. We evaluate the performance on the test set during the training stage every epoch (maximum of 25). As training progresses, both GAMENet+Ours and SafeDrug+Ours can surpass the baseline models in terms of Jaccard, F1, and PRAUC. The baseline models (GAMENet and SafeDrug) overfit or stabilize around epoch 15. In contrast, the performance of our method continues to improve and has not yet reached its optimum, suggesting an enhanced generalization capability. The results illustrate that our method effectively accelerates training progress and improves the model’s generalization ability. We also find that our method slightly reduces the DDI, even though we do not control it.

5.5 Visualization and Analysis (RQ4)

The goal of our method is to incorporate the hierarchical structure into the representation of diagnoses and procedures via medical codes. To validate it, we employ the t-SNE [9] tool to visualize the traditional embeddings and relation

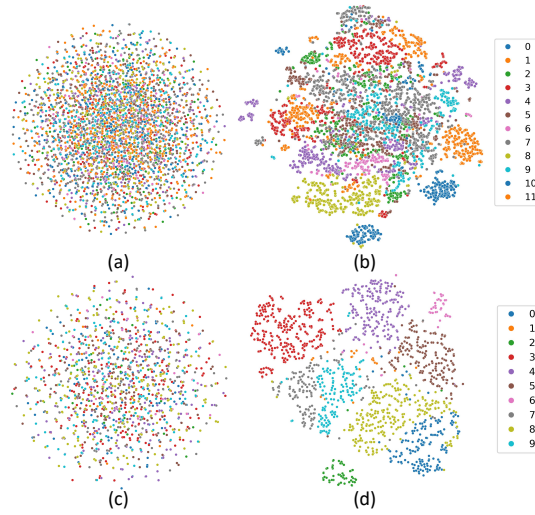


Fig. 4. The visualization of diagnoses and procedures in the 2-d space using t-SNE, where (a) and (c) for embeddings of diagnoses and procedures; (b) and (d) for our relation embeddings of diagnoses and procedures.

embeddings of diagnoses and procedures on a 2-dimensional space. As illustrated in Fig. 4, we visualize the diagnosis and procedure embeddings on the trained models SafeDrug and SafeDrug+Ours. The different colors of points represent the main medical category, corresponding to the first digit of medical codes (including supplementary classification "E000-E999" and "V01-V91").

As shown in Fig. 4(a)(c), the traditional embedding method exhibits a uniform distribution for both diagnoses and procedures. Their underlying relationships have not been captured. This is because the traditional embedding method fails to learn the hierarchical structure and treats diagnoses/procedures as separate and unrelated. In contrast, the results of our method in Fig. 4(b)(d) show distinct clusters of relation embeddings corresponding to their hierarchical structure of medical codes. The visualization results demonstrate the effectiveness and superiority of our method through the hierarchical structure of medical codes.

6 Conclusion and Future Work

In this paper, we proposed a simple yet efficient hierarchical encoder plug-in for medical codes in medication recommendation. It can effectively incorporate the hierarchical structure of diagnoses and procedures into their representation and improve recommendation accuracy. In future work, we intend to investigate the applicability of our hierarchical encoder across different healthcare applications. In addition, we hope to evaluate the model in diverse clinical settings to understand its generalizability and robustness.

Ethical Considerations and Clinical Impact In healthcare applications, it is essential to evaluate the ethical implications and potential clinical impact. While our method aims to enhance drug recommendations, it must be carefully assessed to ensure patient safety and clinical reliability. The model should complement, not replace, clinician judgment, providing additional insights without undermining accountability.

Acknowledgments. This work is partially supported by the National Natural Science Foundation of China under Grant (No. 62032013, 62102074), the Science and Technology projects in Liaoning Province (No. 2023JH3/10200005).

References

1. Choi, E., Bahadori, M.T., Song, L., Stewart, W.F., Sun, J.: Gram: graph-based attention model for healthcare representation learning. In: ACM SIGKDD international conference (2017)
2. 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 (NIPS) (2016)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Elman, J.L.: Finding structure in time. Cognitive science (1990)
5. Gong, F., Wang, M., Wang, H., Wang, S., Liu, M.: Smr: medical knowledge graph embedding for safe medicine recommendation. Big Data Research (2021)
6. Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., et al.: MIMIC-IV, a freely accessible electronic health record dataset. Scientific data (2023)
7. Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.W.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific data (2016)
8. Kaur, R., Ginige, J.A., Obst, O.: AI-based ICD coding and classification approaches using discharge summaries: A systematic literature review. Expert Systems with Applications (2023)
9. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of machine learning research (JMLR) (2008)
10. Murphy, R.L., Srinivasan, B., Rao, V., Ribeiro, B.: Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs. In: International Conference on Learning Representations (ICLR) (2019)
11. Placido, D., Yuan, B., Hjaltelin, J.X., Zheng, C., Haue, A.D., Chmura, P.J., Yuan, C., Kim, J., Umeton, R., Antell, G., et al.: A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. Nature medicine (2023)
12. Shang, J., Xiao, C., Ma, T., Li, H., Sun, J.: Gamenet: Graph augmented memory networks for recommending medication combination. In: the AAAI Conference on Artificial Intelligence (AAAI) (2019)
13. Teng, F., Zhang, Q., Zhou, X., Hu, J., Li, T.: Few-shot ICD coding with knowledge transfer and evidence representation. Expert Systems with Applications (2024)

14. Wagstaff, E., Fuchs, F.B., Engelcke, M., Osborne, M.A., Posner, I.: Universal approximation of functions on sets. *Journal of Machine Learning Research* (2022)
15. Wang, C., Yu, Y., Ma, W., Zhang, M., Chen, C., Liu, Y., Ma, S.: Towards representation alignment and uniformity in collaborative filtering. In: *ACM SIGKDD international conference*. pp. 1816–1825 (2022)
16. Wang, S., Ren, P., Chen, Z., Ren, Z., Ma, J., de Rijke, M.: Order-free medicine combination prediction with graph convolutional reinforcement learning. In: *ACM international conference on information and knowledge management (CIKM)*. pp. 1623–1632 (2019)
17. Wang, S., Tang, D., Zhang, L., Li, H., Han, D.: Hienet: Bidirectional hierarchy framework for automated icd coding. In: *International Conference on Database Systems for Advanced Applications (DASFAA)*. pp. 523–539 (2022)
18. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: *International conference on machine learning*. pp. 9929–9939 (2020)
19. Xie, X., Xiong, Y., Yu, P.S., Zhu, Y.: Ehr coding with multi-scale feature attention and structured knowledge graph propagation. In: *ACM international conference on information and knowledge management (CIKM)*. pp. 649–658 (2019)
20. Yang, C., Wu, Z., Jiang, P., Lin, Z., Gao, J., Danek, B., Sun, J.: PyHealth: A deep learning toolkit for healthcare predictive modeling. In: *ACM SIGKDD International Conference* (2023), <https://github.com/sunlabuiuc/PyHealth>
21. Yang, C., Xiao, C., Glass, L., Sun, J.: Change matters: Medication change prediction with recurrent residual networks. In: *International Joint Conference on Artificial Intelligence (IJCAI)* (2021)
22. Yang, C., Xiao, C., Ma, F., Glass, L., Sun, J.: Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. In: *International Joint Conference on Artificial Intelligence (IJCAI)* (2021)
23. Yang, N., Zeng, K., Wu, Q., Yan, J.: Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning. In: *the ACM Web Conference (WWW)* (2023)
24. Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R.R., Smola, A.J.: Deep sets. *Advances in neural information processing systems (NIPS)* (2017)
25. Zhang, Y., Hare, J., Prügel-Bennett, A.: FSPool: Learning set representations with featurewise sort pooling. In: *International Conference on Learning Representations (ICLR)* (2020)
26. Zhang, Y., Wu, X., Fang, Q., Qian, S., Xu, C.: Knowledge-enhanced attributed multi-task learning for medicine recommendation. *ACM Transactions on Information Systems (TOIS)* (2023)
27. Zhang, Y., Chen, R., Tang, J., Stewart, W.F., Sun, J.: Leap: learning to prescribe effective and safe treatment combinations for multimorbidity. In: *ACM SIGKDD international conference* (2017)
28. Zheng, Z., Wang, C., Xu, T., Shen, D., Qin, P., Huai, B., Liu, T., Chen, E.: Drug package recommendation via interaction-aware graph induction. In: *the ACM Web Conference (WWW)*. pp. 1284–1295 (2021)
29. Zheng, Z., Wang, C., Xu, T., Shen, D., Qin, P., Zhao, X., Huai, B., Wu, X., Chen, E.: Interaction-aware drug package recommendation via policy gradient. *ACM Transactions on Information Systems (TOIS)* (2023)