Time-aware Heterogeneous Graph Transformer with Adaptive Attention Merging for Health Event Prediction

Shibo Li[a], Hengliang Cheng[a], Runze Li[b], Weihua Li[a,*]

[a]*School of Information Science and Engineering, Yunnan University, Kunming, 650500, China*
[b]*College of Letters and Science, University of California, Los Angeles, 90095, United States*

Abstract

The widespread application of Electronic Health Records (EHR) data in the medical field has led to early successes in disease risk prediction using deep learning methods. These methods typically require extensive data for training due to their large parameter sets. However, existing works do not exploit the full potential of EHR data. A significant challenge arises from the infrequent occurrence of many medical codes within EHR data, limiting their clinical applicability. Current research often lacks in critical areas: 1) incorporating disease domain knowledge; 2) heterogeneously learning disease representations with rich meanings; 3) capturing the temporal dynamics of disease progression. To overcome these limitations, we introduce a novel heterogeneous graph learning model designed to assimilate disease domain knowledge and elucidate the intricate relationships between drugs and diseases. This model innovatively incorporates temporal data into visit-level embeddings and leverages a time-aware transformer alongside an adaptive attention mechanism to produce patient representations. When evaluated on two healthcare datasets, our approach demonstrated notable enhancements in both prediction accuracy and interpretability over existing methodologies, signifying a substantial advancement towards per-

*Corresponding author.
Email addresses: lishibo@stu.ynu.edu.cn (Shibo Li), chenghl@mail.ynu.edu.cn (Hengliang Cheng), runzeli@g.ucla.edu (Runze Li), liweihua@ynu.edu.cn (Weihua Li)

sonalized and proactive healthcare management.

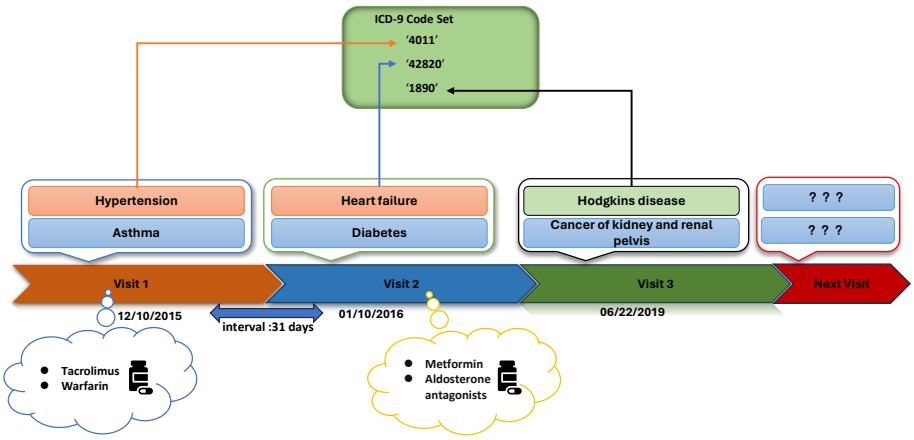
Keywords: Health event prediction, Heterogeneous graph learning, Medical knowledge graph, Time-aware transformer, Adaptive attention merging

1. Introduction

Electronic Health Records (EHR) encapsulate a wealth of patient visit information within medical institutions, encompassing diverse clinical data such as diagnoses, admission times, medical histories, and prescribed drugs. The adoption of EHR across numerous healthcare models has facilitated significant advances in disease prediction through deep learning models like Recurrent Neural Networks (RNNs)(Bai et al., 2018; Yin et al., 2019a; Ma et al., 2017) and Convolutional Neural Networks (CNNs)(Nguyen et al., 2017). Utilizing EHR enhances not only the accuracy of disease prediction but also broadens its application to various health prediction tasks, including mortality rates, hospital stay durations, risk assessment, and medication recommendations. Through deep learning, intricate relationships between patient data and diseases can be deciphered from the voluminous EHR data, aiding physicians in evaluating patient health and tailoring care. Despite these advancements, challenges persist in effectively leveraging diagnostic features for learning:

1. **Comprehensively assimilate knowledge derived from the medical domain**. The GRAM(Choi et al., 2017), an ontology-based model, leverages the hierarchical structure of medical ontologies to represent various medical diseases effectively. Building on GRAM, the KAME(Ma et al., 2018b) method enhances the disease prediction performance by utilizing high-level knowledge. However, these methods primarily focus on the hierarchical relationships between diseases and their ancestors, neglecting the horizontal (co-occurrence) relationships among different diseases. In the realm of health event prediction, offering a reliable explanation for the implicit relationships between diseases continues to be a significant challenge.

Figure 1 An example of a patient's visit record sequences



2. **Heterogeneously learning disease representations with rich meanings**. In clinical diagnostics, patients diagnosed with diverse diseases may receive identical or similar drugs. Current works(Lu et al., 2021) identify concealed connections between diseases by analyzing patient-disease interactions. Nevertheless, the scarcity of clinical data within EHRs hampers the ability to derive significant disease representations based solely on disease co-occurrence. It is suggested that a complex relationship exists, suggesting diseases treated with the same drugs may demonstrate hidden correlations or similarities, thereby increasing the likelihood of patients receiving diagnoses for related diseases sequentially. This could be due to common biological targets or biological pathways linking these diseases, suggesting underlying shared biological mechanisms. Figure 1 shows the patient's visit record sequences. Through analyzing clinical records and drug usage, we aim to unravel the intricate associations between diseases, thereby enhancing our comprehension of disease mechanisms and improving health event prediction accuracy.
3. **Modeling temporal information on disease progression**. Patient admission times are meticulously documented in EHRs. However, many

existing models() inadequately leverage temporal information, leading to suboptimal outcomes and an inability to track the dynamic progression of diseases. The incorporation of temporal data enables the capture of the evolving nature of diseases over time. For example, Figure 1 displays the chronological sequence of a patient's visits alongside the disease's dynamic shifts. Integrating this temporal aspect allows models to intricately discern the critical phases and pathways of disease development. Such integration fosters a thorough comprehension of diseases' dynamic behavior, thereby enhancing the precision of predictive models and the efficacy of clinical decisions.

To overcome these limitations, we propose an innovative model THAM, a novel **T**ime-aware **H**eterogeneous graph Transformer with **A**daptive attention **M**erging for health event prediction, which amalgamates hierarchical disease representation with insights from medical domain knowledge, the implicit connections between diseases and drugs, and temporal data from patient visits. Initially, we apply medical domain knowledge to structure disease representations hierarchically. Subsequently, a heterogeneous graph neural network is employed to derive meaningful disease insights by exploiting both the observed co-occurrences of diseases during patient visits and the interactions between disease manifestations and drug use. Furthermore, we have designed two stages: a preliminary evaluation stage and a comprehensive evaluation stage. During the preliminary evaluation phase, we introduce a Time-aware Transformer featuring a local-based attention mechanism designed to ascertain the preliminary attention weights for each patient visit. This method incorporates time information into the visit vectors via specific non-linear functions, thereby overcoming the constraints associated with a monotonically decaying time function. In the subsequent comprehensive evaluation phase, we posit that a patient's most recent visit record comprehensively reflects their disease progression. Consequently, we designate the embedding vector of the latest visit as the "comprehensive vector." This vector serves as the query vector, while the time interval data

are converted into key vectors using particular non-linear functions, facilitating the generation of comprehensive attention weights for each visit through the dot-product attention mechanism. Finally, the Adaptive attention merging mechanism is employed to acquire representations for patients by incorporating both types of attention. The main contributions of this work are summarized as follows:

- We harness the extensive knowledge within the medical domain to capture the hierarchical correlations among diseases. Furthermore, we suggest acquiring disease representations endowed with rich meanings via drug-disease heterogeneous co-occurrence graphs and disease ontology co-occurrence graphs.

- We have designed two different stages: the preliminary evaluation stage and the comprehensive evaluation stage. In the preliminary stage, time data are integrated into the representation of medical visits. The comprehensive evaluation stage amalgamates information from individual visits with overall visit data to analyze disease progression. Additionally, it learns the relationship between the comprehensive visit vector and temporal data, proficiently capturing the dynamics of disease evolution over time.

- We conducted experiments on two real-world public datasets to evaluate the performance of the proposed model. The results indicate that THAM outperforms state-of-the-art models in terms of prediction accuracy, confirming the validity of the proposed model.

2. Related Work

The widespread adoption of deep learning techniques in recent years has spurred their application in predictive analyses utilizing EHRs. These deep learning approaches have achieved demonstrably superior predictive accuracy compared to traditional machine learning models.

2.1. Models leveraging external knowledge

GRAM(Choi et al., 2017), KAME(Ma et al., 2018b), DMKAP(Li et al., 2023) and some models(Ma et al., 2019; Ye et al., 2021b) enhance the quality of medical representation learning by utilizing the hierarchical information of nodes in the medical ontology knowledge graph, their aim is to leverage the static attention mechanism built on the knowledge DAG. PRIME(Ma et al., 2018a) proposed a log-linear model that automatically learns the importance of different disease knowledge. Furthermore, to address the problem of data incompleteness in the medical field, some papers(Zhang et al., 2019; Yin et al., 2019b; Li et al., 2020a) combine the KnowLife knowledge graph with clinical expertise to compensate for this deficiency. CGL(Lu et al., 2021) constructs a patient-disease observation graph and a disease ontology graph using clinical observation information and medical knowledge. It learns representations using collaborative graph methods while incorporating unstructured text data. G-BERT(Shang et al., 2019) is a model that combines GNN(Scarselli et al., 2008) and BERT(Devlin et al., 2018). It fully utilizes ICD-9-CM[1](Slee, 1978) hierarchical information and introduces the language model pre-training paradigm into the healthcare domain. GNDP(Li et al., 2020b) learns the spatial and temporal patterns from patients' sequential graph, in which the domain knowledge is naturally infused. MedPath(Ye et al., 2021a) extracts personalized knowledge graphs (PKGs) from large-scale online medical knowledge graphs and learns PKG embeddings using GNNs. Chet(Lu et al., 2022) constructs a global disease co-occurrence graph with multiple node attributes based on the patient's medical histories, simulating disease transition processes. However, these works only consider limited relationships between disease knowledge and lack consideration for the temporal information of disease progression.

[1]International Classification of Diseases, Ninth Revision, Clinical Modification

2.2. Models capturing temporal relationships

This line of research focuses on acquiring the temporal characteristics and dependencies within the context of patient visit sequences. Electronic Health Records (EHRs) are not only sequential but also temporal. Each visit in the EHR data is accompanied by a timestamp, as the progression of diseases is inherently connected to time. T-LSTM(Baytas et al., 2017) effectively handles irregular time intervals in longitudinal medical records using a time decay strategy, thereby capturing the underlying structure in these irregular time series. DoctorAI(Choi et al., 2016a) employs a Recurrent Neural Network (RNN) to forecast patient diagnoses in subsequent visits and the time interval between their current and upcoming appointments. RETAIN(Choi et al., 2016b) introduces a reverse time attention model based on RNNs, leveraging two RNNs to learn the weights of visits and medical codes within visits. Dipole(Ma et al., 2017) models longitudinal EHR data using bidirectional RNN and applies three attention mechanisms. Additionally, Concare(Ma et al., 2020) improves multi-head self-attention by considering the time intervals between consecutive visits. Timeline(Bai et al., 2018) develops a timeline model to capture the time intervals between visits, enhancing prediction accuracy. Concare and Timeline both acknowledge the attenuation of relevant patient information if there is a time gap between consecutive visits. However, these works consider the correlation between diseases and time but ignore the cross-sectional and longitudinal hierarchical relationships among diseases.

3. Methodology

3.1. Problem Formulation

Electronic Health Records (EHRs) comprise numerous short-term or long-term visit records for patients. Let $C = \{c_1, c_2, \ldots, c_{|C|}\}$ denote the set of medical codes in the EHR dataset, where $|C|$ represents the total number of medical codes in the dataset. Similarly, let $D = \{d_1, d_2, \ldots, d_{|D|}\}$ denote the

set of all drugs used by patients in the EHR dataset, where $|D|$ represents the total number of drugs in the dataset.

EHR dataset. Let $P = \{p^u \mid u \in U\}$, where U is the set of patients in P, and $p^u = \{V_1^u, V_2^u, \ldots, V_T^u\}$ represents all visit records for patient u. Each visit $V_i^u = \{C_i^u, D_i^u\}$, where C_i^u and D_i^u are subsets of C and D. Let r_t represent the temporal information corresponding to the visit V_t. Then, $\Delta^u = \{\Delta_1, \Delta_2, \ldots, \Delta_T\}$, where $\Delta_t = r_t - r_{t-1}$.

Disease Prediction Task. The core objective of this task is to predict the occurrence of diseases in the $(T+1)$-th visit based on the previous T visit records for a given patient u. This can be represented by a binary vector $\hat{y} \in \{0, 1\}$, where $\hat{y}_i = 1$ indicates that disease c_i is predicted in $C_{(T+1)}^u$.

Heart Failure Prediction Task. [2]The core objective of this task is to predict a binary value $\hat{y} \in \{0, 1\}$ based on the previous T visit records for a given patient u. $\hat{y}_i = 1$ indicates that patient u is predicted to be diagnosed with heart failure in the $(T + 1)$-th visit.

For convenience, we will remove the superscript u from p^u, V_i^u, C_i^u, D_i^u and Δ^u in the rest of this paper.

3.2. Overview of the proposed model

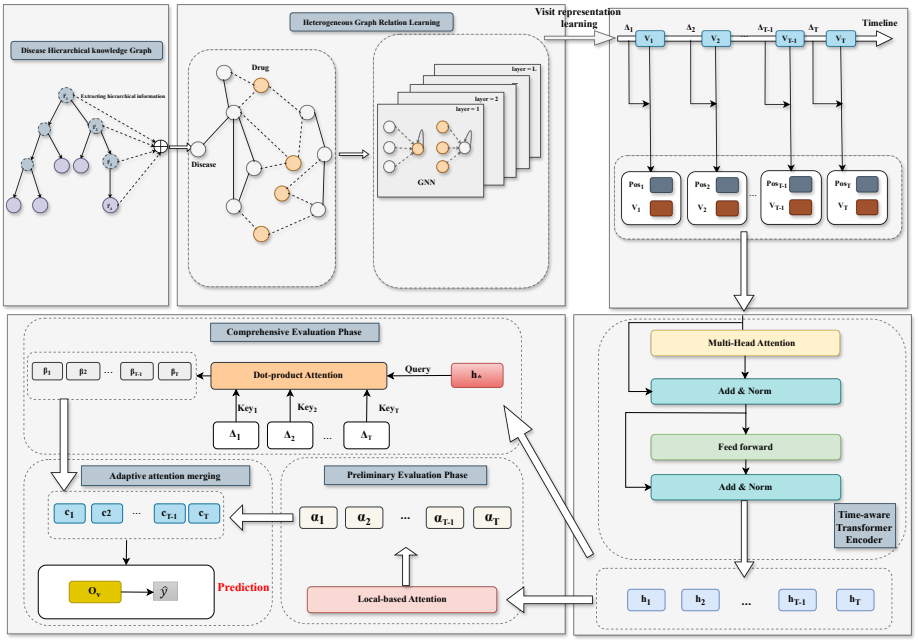
The model we proposed aims to fully utilize patient visit records in EHR data to predict the future health events of patients. In this section, we will elaborate on the seven main components of the model, and the schematic diagram of the model is shown in Figure 2.

3.2.1. Hierarchical Representation for Medical Codes

In the medical domain, contemporary disease classification systems such as ICD-9-CM and ICD-10(Organization, 2004) are employed to systematically categorize disease concepts at various levels using medical coding, thus establishing

[2]The codes of heart failure start with 428 in ICD-9-CM

Figure 2 The model structure of the proposed model.



a hierarchical structure akin to a tree. In this structure, each node is linked to a single parent node, with leaf nodes often denoting specific diseases and their ancestor nodes representing broader disease concepts. For instance, Hepatitis is classified as a specific disease, whereas Viral infection serves as its broader category. Typically, medical codes assigned during patient visits correspond to specific diseases (leaf nodes). Nonetheless, we contend that the representation of a specific disease should encompass both the disease itself and its broader category since diseases sharing common ancestors may exhibit similarities. Consequently, we recursively generate virtual child nodes for each non-leaf node and fill them into the virtual leaf nodes.

We posit that the hierarchical structure comprises H layers, with each layer h hosting m_h nodes. An embedding matrix is established for every layer within this hierarchy. Consequently, the embedding matrix pertinent to layer h is expressed as $L_h \in \mathbb{R}^{m_h \times m_c}$, where m_c denotes the embedding size. We select corresponding embedding vectors for disease c_i based on its position and that

of its ancestors at various levels in the tree. We then construct the hierarchical representation $L_i \in \mathbb{R}^{Hm_c}$ of c_i by concatenating the embedding vectors from each level: $L_i = l_1^i \oplus l_2^i \oplus \ldots \oplus l_H^i$, where $\oplus$ represents concatenation. This process culminates in the generation of a comprehensive embedding matrix for all diseases, represented as $L \in \mathbb{R}^{|C| \times Hm_c}$.

3.2.2. Graph Definition

In healthcare, it is common for patients to be diagnosed with a combination of certain diseases, such as chronic obstructive pulmonary disease (COPD) and heart failure, likely due to shared risk factors. We hypothesize that diseases diagnosed during the same visit, as well as distinct diseases treated with the same drug, may exhibit similarities. This hypothesis is grounded in the assumption that different diseases might share common biological targets; drugs that interact with these targets can modulate or influence physiological processes, resulting in therapeutic effects, which in turn suggest underlying similarities among the diseases. To explore these potential connections, we posit the following assumptions to reveal hidden relationships between diseases:

- **Disease similarity derived from medical concepts**. If two diseases belong to the same abstract disease concept, there may be some medical similarity between them.
- **Disease similarity derived from drug usage**. When two diseases are treated with identical or comparable drugs, it suggests the possibility of a similarity between the diseases. This similarity arises from the potential sharing of common biological targets, further leading to the inference that these diseases may also possess similar risk factors.

Based on the above assumptions, we constructed a drug-disease heterogeneous co-occurrence graph and a disease ontology co-occurrence graph, denoted as $M = \{M_{DC}, M_{CC}\}$.

M_{DC} is a heterogeneous drug-disease co-occurrence graph derived from EHR data, with nodes representing drugs and medical codes. We utilize a matrix

$B_{DC} \in \mathbb{R}^{|D|\times|C|}$ to represent the graph M_{DC}. Whenever a patient is diagnosed with disease c_j and concurrently uses drug d_i during a visit, we insert an edge $\overrightarrow{(d_i, c_j)}$ into the graph M_{DC} and let $B_{DC}[i][j] = B_{DC}[i][j]+1$. Subsequently, we normalize the B_{DC}. M_{CC} is a disease ontology co-occurrence graph also derived from EHR data, with nodes symbolizing medical codes. If two distinct diseases c_i and c_j are simultaneously diagnosed in a patient's visit record, we add two edges $\overrightarrow{(c_i, c_j)}$ and $\overleftarrow{(c_i, c_j)}$ into the graph M_{CC}. However, we conjecture that the mutual influence between two diseases is not symmetrical. For instance, while patients with asthma might frequently develop sinusitis, the reverse is less common. Thus, to mitigate computational complexity and disregard low-frequency co-occurrences, we introduce a threshold λ. Only nodes that meet the definition of formula 1 will be considered.

$$K_i = \{c_j \mid \frac{e_{ij}}{\sum_{j=1}^{|C|} e_{ij}} \geq \lambda\} \tag{1}$$

The e_{ij} represents the co-occurrence frequency of c_i and c_j. Then, we define the adjacency matrix $A_{CC} \in \mathbb{R}^{|C|\times|C|}$ to store the edge weights of the graph M_{CC}:

$$A_{CC}[i][j] = \begin{cases} 0 & \text{if } i = j \text{ or } c_j \notin K_i, \\ \frac{e_{ij}}{\sum_{c_j \in K_i} e_{ij}} & \text{otherwise.} \end{cases} \tag{2}$$

A_{CC} is an asymmetric matrix and quantifies the extent of mutual influence between two diseases. We contend that constructing this matrix enhances interpretability.

3.2.3. Heterogeneous Graph Relation Learning

We have designed a graph neural network (GNN) learning method that leverages a heterogeneous co-occurrence graph and a disease ontology co-occurrence graph to derive meaningful representations of diseases. Initially,each drug is assigned an embedding vector, $N \in \mathbb{R}^{|D|\times m_d}$ is the embedding matrix of all drugs with the size of m_d. We set $H_D^{(0)} = N$, $H_C^{(0)} = L$, and $H_C^{(l)} \in \mathbb{R}^{|C|\times m_c^{(l)}}$, $H_D^{(l)} \in \mathbb{R}^{|D|\times m_d^{(l)}}$ representing the hidden features of medical codes and drugs at the l-th layer.

- **Aggregation:** We incorporate two different co-occurrence matrices, A_{CC} and B_{DC}, as contextual information into node embeddings, and map the medical code features $H_C^{(l)}$ to the drug dimension, serving as the aggregation operation in the GNN:

$$M_D^{(l)} = H_D^{(l)} + B_{DC} H_C^{(l)} W_{CD}^{(l)} \in \mathbb{R}^{|D| \times m_d^{(l)}} \tag{3}$$

Here $W_{CD}^{(l)} \in \mathbb{R}^{m_c^{(l)} \times m_d^{(l)}}$ is a trainable parameter utilized for mapping medical code embeddings to the dimension of drug embeddings. We perform mapping of $H_D^{(l)}$ to the dimension of medical codes and subsequently aggregate the two co-occurrence matrixs as contextual information into the embeddings of the corresponding nodes:

$$M_C^{(l)} = H_C^{(l)} + B_{DC}^T H_D^{(l)} W_{DC}^{(l)} + A_{CC} H_C^{(l)} \in \mathbb{R}^{|C| \times m_c^{(l)}} \tag{4}$$

$W_{DC}^{(l)} \in \mathbb{R}^{m_d^{(l)} \times m_c^{(l)}}$ is also a trainable parameter used for mapping drug embeddings to the dimension of medical code embeddings.

- **Update:** We use the following formula 5 to update the hidden representation of medical codes and drugs as the update operation in GNN. Assuming a total of L layers, the update formula for each layer is as follows:

$$H_{\{D,C\}}^{(l+1)} = \sigma(\text{BatchNorm}(M_{\{D,C\}}^{(l)} W_{\{D,C\}}^{(l)})) \tag{5}$$

$W_{\{D,C\}}^{(l)}$ maps $M_{\{D,C\}}^{(l)}$ to the $(l+1)$-layer, and then we use the BatchNorm to normalize the hidden representation. σ represents the non-linear activation function. Here, we use LeakyReLU(Xu et al., 2015), which helps alleviate the gradient vanishing problem. Additionally, the small slope introduced by LeakyReLU increases the non-linearity of the model, which is particularly important for GNNs as they need to capture complex graph structure information. By enhancing non-linearity, it helps GNNs learn complex relationships between nodes more effectively.

3.2.4. Representation of Visits

We posit that the representation of a visit should be obtained by averaging the embeddings corresponding to the diseases diagnosed during that visit. Therefore, the initial embedding vector o_t of visit t should be represented as:

$$o_t = \frac{1}{|C_t|} \sum_{c_j \in C_t} H_C^j \in \mathbb{R}^{m_c^{(L)}} \tag{6}$$

Although RNN-based models(Choi et al., 2016a,b; Ma et al., 2017) consider the role of temporal information and operate under the premise that disease information decays at a stable rate, this assumption may not always apply. Particularly for some chronic diseases, the progression can be markedly slow, often leading to intervals exceeding a year between patient visits. For patients with such conditions, if the diagnostic codes from two sequential visits are similar, this might suggest that the disease has not intensified. In these instances, the attenuation of time-sensitive information should be less severe, rather than unduly diminishing the significance of the data. Hence, we introduce a function designed to integrate temporal data into the visit vector, thereby establishing the final visit vector v_t:

$$f_t = W_f \left(1 - \tanh\left(\left(W_e \frac{\Delta_t}{180} + b_e\right)^2\right)\right) + b_f \in \mathbb{R}^{m_c^{(L)}} \tag{7}$$

$$v_t = o_t + f_t \in \mathbb{R}^{m_c^{(L)}} \tag{8}$$

Here $W_e \in \mathbb{R}^a$, $b_e \in \mathbb{R}^a$, $W_f \in \mathbb{R}^{m_c^{(L)} \times a}$, and $b_f \in \mathbb{R}^{m_c^{(L)}}$. In the patient's visit sequence, if the interval between the occurrences of one disease and another disease is shorter, formula 7 is easier to be activated. To simplify the representation, we will use m instead of $m_c^{(L)}$ in the rest of the paper.

3.2.5. Preliminary Evaluation Phase

For each patient's visit record, we can obtain an input matrix $V = [v_1, v_2, \ldots, v_T]$. We generate a corresponding positional encoding for all visits in order, The generated positional encodings will be added to the medical visit vector v_t to obtain

a new representation of the visit v'_t:

$$Pos_{(t,2i)} = \sin\left(\frac{t}{10000^{2i/m}}\right) \in \mathbb{R}^m \tag{9}$$

$$Pos_{(t,2i+1)} = \cos\left(\frac{t}{10000^{2i/m}}\right) \in \mathbb{R}^m \tag{10}$$

$$v'_t = v_t + Pos_t \tag{11}$$

Where m represents the dimension size of the visit embedding, i is the detention of the position embedding Pos. The Time-aware Transformer Encoder (denoted as TTE) is employed to capture the long-term dependency between each visit:

$$[h_1, h_2, \ldots, h_T] = TTE([v'_1, v'_2, \ldots, v'_T]) \in \mathbb{R}^{T\times z} \tag{12}$$

Where $h_t \in \mathbb{R}^m$ represents the hidden representation of each visit, we use local-based attention(Luong et al., 2015) to calculate the preliminary attention weight α for each visit. This operation simulates the behavior of doctors during diagnosis, as they highly focus on visit history related to the target disease.

$$\alpha = Softmax([h_1, h_2, \ldots, h_T]w_\alpha) \in \mathbb{R}^T \tag{13}$$

Here $w_\alpha \in \mathbb{R}^b$ is a context vector for local-based attention and α is the attention weight of visit.

3.2.6. Comprehensive Evaluation Phase

In medical practice, doctors typically assess disease progression and forecast outcomes by synthesizing data from individual visits with overarching diagnostic information. We maintain that a patient's latest visit record encompasses comprehensive details of their disease trajectory(Choi et al., 2016b; Luo et al., 2020). Consequently, we set $h_* = h_T$ and designate h_* as the comprehensive visit vector. Initially, h_* is converted into a query vector Q:

$$Q = LeakyReLU(W_Q h_* + b_Q) \in \mathbb{R}^q \tag{14}$$

Where $W_Q \in \mathbb{R}^{q \times m}, b_Q \in \mathbb{R}^q$ are both trainable parameters. LeakyReLU allows negative values to have a small positive output, increasing the robustness of the results. When analyzing comprehensive diagnostic information, doctors need to combine the time information of disease onset to obtain the most important time points for the patient's condition. To simulate this process, we embed each temporal interval information Δ_t into the same space as the query vector, treating it as the key vector:

$$K_t = LeakyReLU\left(W_k\left(1 - \tanh\left(\left(W_t\frac{\Delta_t}{180} + b_t\right)^2\right)\right) + b_k\right) \in \mathbb{R}^q \quad (15)$$

Here $W_t \in \mathbb{R}^a, b_t \in \mathbb{R}^a, W_k \in \mathbb{R}^{q \times a}, b_k \in \mathbb{R}^q$ are trainable parameters. We employ scaled dot-product attention(Vaswani et al., 2017) to learn the correlation between the comprehensive visit vector and the temporal information. This enables us to derive the comprehensive attention weight β:

$$\beta = Softmax\left(\frac{QK^T}{\sqrt{q}}\right) \in \mathbb{R}^T \quad (16)$$

3.2.7. Adaptive attention merging

We have derived two distinct attention weights: the preliminary attention weight α and the comprehensive attention weight β. The preliminary evaluation phase serves as an initial assessment of each visit's significance and its temporal association, whereas the comprehensive evaluation phase offers a retrospective analysis of temporal information's relevance. The amalgamation of these two weights yields more robust attention weights. Thus, we introduce an adaptive attention merging mechanism, the comprehensive visit vector is mapped into a two-dimensional space and normalized through a Softmax layer:

$$\delta = Softmax(W_x h_* + b_x) \in \mathbb{R}^2 \quad (17)$$

Where $W_x \in \mathbb{R}^{2 \times q}, b_x \in \mathbb{R}^2$ are trainable parameters.We concatenate the preliminary attention weight α with the comprehensive attention weight β to obtain robust attention weight γ:

$$\gamma = \alpha \oplus \beta \in \mathbb{R}^{T \times 2} \quad (18)$$

Subsequently, we generate the overall attention weights η:

$$\eta = \gamma \odot \delta \in \mathbb{R}^{T \times 2} \tag{19}$$

Where $\odot$ denotes the element-wise multiplication, which utilizes broadcasting mechanisms.Finally, we normalize the overall attention weight and obtain the overall attention score η'_t for each visit, as shown below:

$$\eta'_t = \frac{\eta_t}{\sum_{i=1}^{T} \eta_i} \tag{20}$$

3.2.8. Prediction and Inference

After obtaining the overall attention weight for each visit, we can obtain the patient's output through attention pooling:

$$O = \sum_{t=1}^{T} \eta'_t \, h_t \in \mathbb{R}^m \tag{21}$$

We use a multi-layer perceptron with a sigmoid activation function on the model's output O to compute the predicted probability $\hat{y}$. In the Diagnosis prediction task, we predict the diseases the patient will have at the $T+1$ visit, it is a multi-label classification. In the Heart failure prediction task, we predict whether the patient will be diagnosed with heart failure at the $T+1$ visit, it is a binary classification. Therefore, the loss function of this model is binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{|N|} \sum_{i=1}^{|N|} \left(y_i^T \log(\hat{y}_i) + (1 - y_i)^T \log(1 - \hat{y}_i) \right) \tag{22}$$

y is the true label of medical codes or heart failure, $|N|$ is the number of samples. During the inference stage, we set the model to eval mode and obtain the medical code embeddings H_C after GNN learning, and combine them with the patient's visit representation and time information. Given a new patient for inference, we continue to execute and make predictions from Eq.(8). Algorithm 1 describes the overall training process of the proposed THAM.

Algorithm 1 Training Procedure of THAM

Input: Training set T_t, and validation set T_v

Output: Trained model parameter

Randomly initialize the parameter ω of THAM and drug embedding matrix N
Obtain the hierarchical embedding matrix L of all diseases based on the medical knowledge graph
Construct heterogeneous drug-disease co-occurrence matrix B_{DC} and disease ontology co-occurrence matrix A_{CC} from T_t
Set $H_D^{(0)} = N$, $H_C^{(0)} = L$
for $epoch = 1$ to $EPOCH$ **do**
    Randomly shuffle the order of samples in training set T_t.
    for $(p, \Delta, y) \in T_t$ **do**
        for $l = 0$ to $L - 1$ **do**
            $M_{\{D,C\}}^{(l)} = Aggregation(H_{\{D,C\}}^{(l)}, A_{CC}, B_{DC})$
            $H_{\{D,C\}}^{(l+1)} = Update(M_{\{D,C\}}^{(l)})$
        end for
        Calculate the preliminary visit embeddings o using Eq.(6)
        Calculate the final visit embeddings v using Eq.(7)-(8)
        Calculate the new visit embedding v' using Eq.(9)-(11)
        Utilizing transformer TTF, encode v' to derive h according to Eq.(12)
        Calculate the preliminary attention weight α using Eq.(13)
        Calculate the comprehensive attention weight β using Eq.(14)-(16)
        Calculate the overall attention score η' for each visit using Eq.(17)-(20)
        $O = \sum_{t=1}^{T} \eta'_t h_t$
        $\hat{y} = MlpWithSigmoid(O)$
        Calculate the prediction loss $\mathcal{L}$ using Eq.(22)
        Update model parameters ω according to the gradient of $\mathcal{L}$
    end for
    Calculate the average validation loss $\mathcal{L}_v$ using validation set T_v
    if $\mathcal{L}_v < \mathcal{L}_v^{min}$ **then**
        $\omega_{best} = \omega$
        $\mathcal{L}_v^{min} = \mathcal{L}_v$
    end if
end for

Table 1: Statistics of MIMIC-III and MIMIC-IV datasets

Dataset	**MIMIC-III**	**MIMIC-IV**
# patients	7,493	10,000
Max. # visit	42	93
Avg. # visit	2.66	3.79
# codes	4,880	5985
Max. # codes per visit	39	39
Avg. # codes per visit	13.06	13.51
# drugs	3202	3070
Max. # drugs per visit	164	193
Avg. # drugs per visit	37.36	25.38

4. Experiments

4.1. Experimental Setup

4.1.1. Dataset

To evaluate our proposed model, we focused on two extensively recognized datasets in the realm of critical care research: MIMIC-III(Johnson et al., 2016) and MIMIC-IV(Johnson et al., 2023). Table 1 displays the comprehensive details pertaining to the MIMIC-III and MIMIC-IV datasets. Both datasets emanate from the extensive de-identified clinical data collected at the Beth Israel Deaconess Medical Center in Boston, Massachusetts, encompassing detailed records from patients admitted to the Intensive Care Units (ICUs). MIMIC-III covers data from over 40,000 ICU admissions between 2001 and 2012, incorporating a vast spectrum of information including patient demographics, vital signs, laboratory test results, diagnoses, and diagnostic codes. MIMIC-IV extends this dataset, covering approximately 60,000 ICU admissions from 2008 to 2019, thus providing an updated and expanded database that reflects more recent clinical practices and patient demographics.

To ensure a comprehensive analysis, we selected patients from MIMIC-IV

who were admitted between 2013 and 2019, avoiding temporal overlap with the MIMIC-III dataset and ensuring the distinctiveness of the patient cohorts under investigation. And we included patients who had multiple visits (# of visits ≥ 2) in order to eliminate cases where there were no visit records available as labels.We adopted a randomized approach to divide both datasets into training, validation, and testing segments. This partitioning facilitates a balanced assessment of the model's predictive accuracy and generalizability. Specifically, for MIMIC-III, the data was divided into 6000 training, 500 validation, and 993 testing samples. For MIMIC-IV, the distribution comprised 8000 training, 1000 validation, and 1000 testing samples. In the context of heart failure prediction, the label will be assigned as 1 if the patient is diagnosed with heart failure during their most recent visit.

This methodical preparation and segmentation of the datasets are critical for evaluating the model's capability to accurately predict outcomes and events based on the rich clinical data available. By treating the last visit of a patient as the label and all preceding visits as features, we aim to harness the longitudinal data structure inherent in these databases, thereby enhancing the model's ability to forecast critical care outcomes with higher precision and reliability. Through this analytical framework, our research endeavors to contribute significantly to the advancement of predictive modeling in critical care, ultimately aiming to improve patient outcomes through data-driven insights and interventions.

4.1.2. Baselines

To evaluate the performance of our proposed model, it is necessary to compare it with various state-of-the-art models in the fields of electronic health record analysis and disease prediction. We selected the following methods as baselines:

- **RNN/CNN/Attention-based model**: Dipole(Ma et al., 2017), RETAIN(Choi et al., 2016b), Deepr(Nguyen et al., 2017) and Timeline(Bai et al., 2018).

- **Graph-based model**: GRAM(Choi et al., 2017), KAME(Ma et al., 2018b), G-BERT(Shang et al., 2019), CGL(Lu et al., 2021), Chet(Lu et al., 2022) and BioDynGraph(Li et al., 2024).

4.1.3. Parameter Settings

We use the Xavier method to randomly initialize the embeddings for diseases and drugs. Sinusoidal Position Embeddings are used to generate position embeddings.

- **In the disease prediction task**. On the MIMIC-III dataset, the embedding sizes for m_c, m_d are 48 and 64. The layer number L of GNN is 2. The hidden dimensions $m_c^{(1)}, m_c^{(2)}$ and $m_d^{(1)}$ are 64, 192 and 64, $a = 64, q = 64, b = 32, \lambda = 0.01$. For the hyper-parameters of Time-aware Transformer Encoder, we set the multi-head number as 4, the number of encoder layer is 1, and the size of middle feed-forward network as 1024. On the MIMIC-IV dataset, both m_c and m_d are set to 64, and $m_c^{(2)}$ is set to 256. The remaining parameters are consistent with those on the MIMIC-III dataset. We set the number of epochs to 200, with an initial learning rate of 1e-1. The learning rate is decayed to 1e-2, 1e-3, and 1e-4 at epochs 10, 100, and 200 respectively.

- **In the heart failure prediction task**. On the MIMIC-III dataset, we set $m_c = 7$ and $m_d = 16$, $m_c^{(1)}$, $m_c^{(2)}$, and $m_d^{(1)}$ set to 10, 28, and 16 respectively. $a = 16, q = 16, b = 32$. We set the number of encoder layers to 1. On the MIMIC-IV dataset, m_c and m_d are set to 5 and 16, and $m_c^{(1)}$, $m_c^{(2)}$, and $m_d^{(1)}$ set to 10, 20 and 16 respectively. Other parameters remain the same as in the MIMIC-III dataset. We set the number of epochs to 100, with an initial learning rate of 1e-2. The learning rate is decayed to 1e-3, 1e-4, and 1e-5 at epochs 2, 3, and 20 respectively.

We use the Adam(Kingma & Ba, 2014) as the optimizer. The model is implemented using Python 3.10.13 and PyTorch 1.12.0 with CUDA 11.5, running

on a machine with an Intel E5-2697 CPU, 251GB memory, and GeForce RTX 3090 GPU.

4.1.4. Experiment Evaluation

We use weighted F_1 score (w-F_1) and recall at k (R@k) as performance evaluation metrics for disease prediction. w-F_1 is the weighted sum of F_1 scores for all disease codes, with a higher w-F_1 indicating higher accuracy in disease prediction. R@k represents the coverage of correctly predicted diseases among the top-k predictions, with a higher R@k indicating higher coverage. As for heart failure prediction, the evaluation metrics are AUC and F_1 score. AUC measures the area under the Receiver Operating Characteristic (ROC) curve, and its magnitude is positively correlated with the ability to distinguish between positive and negative cases. The F_1 score is the harmonic mean of precision and recall, aiming to provide a balanced performance measure considering both precision and recall. A higher F_1 score indicates better overall performance in terms of false positive and false negative rates.

4.2. Experiment Result

4.2.1. Diagnosis prediction and Heart Failure prediction results

In this section, we evaluated the performance of the THAM in comparison to existing baselines using two public datasets. The models were independently trained five times with distinct parameter initializations, with outcomes reported as mean(standard deviation). Table 2 showcases the evaluation metrics: w-F_1 (%) and R@k (%), where k is set at [10,20]. Since the average diagnosis number in a visit is around 13. THAM surpassed other models, which can be chiefly attributed to its comprehensive exploitation of EHR data. By uncovering hidden drug-disease associations and leveraging temporal visit information, THAM can trace the trajectory of disease progression. In contrast, CGL's limited approach focuses solely on patient-disease interactions, yielding less nuanced insights into diseases. THAM also surpasses Chet, which learns disease combinations and transitions, demonstrating the superiority of THAM's

disease representation. We notice that G-BERT has a lower w-f1 score, which may be due to the removal of pre-training in the original model and its inability to handle simple sequences effectively. GRAM and KAME also achieve relatively lower scores, which may be attributed to their use of static graphs for disease representation learning, without capturing dynamic features of user activity. Additionally, our proposed THAM performs significantly better on the MIMIC-IV dataset compared to the MIMIC-III dataset, possibly because the MIMIC-IV dataset is larger, indicating that our proposed model benefits from more training data to fully demonstrate its effectiveness.

Table 3 presents the results of using AUC (%) and F_1 (%) for heart failure evaluation, showing that our proposed model performs better compared to other baseline models. Additionally, we noticed that the performance metrics of all models are better on the MIMIC-IV dataset than on MIMIC-III. We believe that the main reason for this improvement is the larger training set available in MIMIC-IV, as models based on deep learning require a sufficient amount of data to learn satisfactory parameters.

Table 2: Diagnosis prediction results on MIMIC-III and MIMIC-IV using w-F_1 (%) and R@k (%).

	MIMIC-III			**MIMIC-IV**		
Models	**w-F_1**	**R@10**	**R@20**	**w-F_1**	**R@10**	**R@20**
RETAIN	20.43 (0.30)	26.15 (0.20)	34.78 (0.22)	24.71 (0.24)	28.02 (0.47)	34.46 (0.13)
Dipole	19.35 (0.33)	24.98 (0.27)	34.02 (0.21)	23.69 (0.24)	27.39 (0.34)	35.48 (0.29)
Deepr	18.87 (0.21)	24.74 (0.25)	33.47 (0.17)	24.08 (0.17)	26.29 (0.25)	33.93 (0.21)
Timeline	20.46 (0.18)	25.75 (0.13)	34.83 (0.14)	25.26 (0.30)	29.00 (0.21)	37.13 (0.39)
GRAM	21.52 (0.10)	26.51 (0.09)	35.80 (0.09)	23.50 (0.11)	27.29 (0.27)	36.36 (0.30)
KAME	21.10 (0.13)	24.97 (0.18)	33.99 (0.24)	21.88 (0.17)	25.10 (0.22)	34.85 (0.15)
G-BERT	19.88 (0.19)	25.86 (0.12)	35.31 (0.13)	24.49 (0.20)	27.16 (0.06)	35.86 (0.19)
BioDynGrap	25.21 (0.14)	28.15 (0.15)	38.10 (0.12)	27.09 (0.18)	30.13 (0.21)	38.65 (0.18)
CGL	21.92 (0.12)	27.13 (0.30)	36.49 (0.15)	25.41 (0.08)	28.52 (0.42)	37.15 (0.29)
Chet	22.63 (0.08)	28.64 (0.13)	37.87 (0.09)	26.35 (0.13)	30.28 (0.09)	38.69 (0.15)
THAM	**25.46 (0.07)**	**31.00 (0.16)**	**41.10 (0.14)**	**30.79 (0.22)**	**35.30 (0.16)**	**44.90 (0.20)**

Table 3: Heart failure prediction results on MIMIC-III and MIMIC-IV using AUC (%) and F_1 (%)

Models	**MIMIC-III**		**MIMIC-IV**	
	AUC	F_1	**AUC**	F_1
RETAIN	83.21 (0.26)	71.32 (0.17)	89.02 (0.26)	67.38 (0.21)
Dipole	82.08 (0.29)	70.35 (0.21)	88.69 (0.24)	66.22 (0.15)
Deepr	81.36 (0.13)	69.54 (0.08)	88.43 (0.18)	61.36 (0.12)
Timeline	82.34 (0.31)	71.03 (0.24)	87.53 (0.13)	66.07 (0.21)
GRAM	83.55 (0.19)	71.78 (0.14)	89.61 (0.12)	68.94 (0.19)
KAME	82.88 (0.12)	72.03 (0.07)	89.05 (0.15)	69.36 (0.22)
G-BERT	81.50 (0.24)	71.18 (0.12)	87.26 (0.12)	68.04 (0.17)
BioDynGraph	75.13 (0.12)	68.15 (0.17)	87.00 (0.08)	69.02 (0.11)
CGL	84.19 (0.16)	71.77 (0.10)	89.05 (0.15)	69.36 (0.22)
Chet	86.14 (0.14)	73.08 (0.09)	90.83 (0.09)	71.14 (0.15)
THAM	**87.13 (0.07)**	**74.82 (0.11)**	**93.57 (0.16)**	**76.49 (0.20)**

4.2.2. Ablation Study

In order to investigate the effectiveness of components of the model, we performed an ablation experiment. Specific components of the model were either removed or modified: THAM$_{a-}$ randomly initializing the disease embedding matrix, THAM$_{b-}$ without embedding time information, and THAM$_{c-}$ not using the adaptive attention merging mechanism. The ablation experiment was conducted on the MIMIC-IV dataset:

- **THAM**$_{a-}$: Instead of connecting embedding vectors at different levels, we randomly initialize the embedding matrix of diseases. This contrast is intended to emphasize the importance of hierarchical information in diseases.

- **THAM**$_{b-}$: We remove the embedded time vector in Eq.(8) and directly use the o_t from Eq.(6) as the final visit vector for subsequent predictions. This contrast aims to explore the importance of time information.

- **THAM**$_{c-}$: We cancel the comprehensive evaluation phase and use the preliminary attention weights obtained from Eq.(13) as the overall attention weights for subsequent predictions, without using the Adaptive attention merging mechanism. This approach aims to demonstrate that the most recent medical records contain all the information about the disease progression. It is essential to fully utilize the most recent medical records.

- **THAM**$_{d-}$: Building upon THAM$_{c-}$, we continue to remove the embedded time vector and retain the structure of Transformer to learn hidden states and utilize the local-based attention mechanism to learn patient representation.

Table 4: Diagnosis prediction and heart failure prediction for THAM variants on the MIMIC-IV dataset.

Models	**Diagnosis**			**Heart failure**	
	w-F_1	**R@10**	**R@20**	**AUC**	**F_1**
THAM$_{a-}$	28.68	33.34	42.13	91.30	74.51
THAM$_{b-}$	30.48	34.82	44.21	93.10	75.72
THAM$_{c-}$	29.58	34.18	43.13	92.42	75.24
THAM$_{d-}$	28.77	34.07	42.78	91.98	74.83
THAM	**30.79**	**35.30**	**44.90**	**93.57**	**76.49**

The results of the ablation experiments are presented in Table 4. We noticed that for THAM$_{a-}$, which utilizes a randomly initialized disease embedding matrix, all the metrics except w-F_1 show a significant decrease. This indicates the importance of obtaining meaningful disease representations by leveraging the hierarchical relationships among diseases, as it has a crucial impact on achieving good patient representations. In the case of THAM$_{b-}$, the decline in metrics is not substantial. Despite not incorporating time information in the visit representation, it still outperforms all ablation models. This can be credited to

its utilization of medical domain knowledge, including hierarchical embedding matrices. Furthermore, during the comprehensive evaluation phase, it learns the correlation between comprehensive visits and time information, further affirming the effectiveness of time information modeling. On the other hand, THAM$_{c-}$ retains the time information embedding but forsakes the comprehensive evaluation phase, resulting in comparatively inferior performance compared to THAM$_{b-}$. This validates that relying solely on preliminary representations leads to a significant loss of crucial information and impairs predictive performance. THAM$_{d-}$ discards both time information embedding and the comprehensive evaluation phase, it slightly outperforms THAM$_{a-}$ in all metrics. We postulate that meaningful disease representations have a more substantial impact on model performance compared to time information. These findings collectively constitute a comprehensive ablation study, accentuating the significance of each component within THAM. Upon analyzing the results in Table 4, it is evident that even with the removal of individual components, THAM's performance remains superior to the current baseline models. This further underscores the robustness and superiority of our proposed model, emphasizing the importance of leveraging medical domain knowledge, the relationship between drugs and diseases, as well as the significance of time information.

4.2.3. Prediction Analysis

- **Emerging diseases**. The term "Emerging diseases" refers to ailments identified in subsequent patient visits that were not present in earlier visits.
- **Occurred diseases**. The term "Occurred diseases" refers to diseases that have also appeared in early visits during subsequent patient visits.

Our objective is to leverage the ability to predict such emerging diseases as a measure of a model's capacity to learn diagnostic similarity between patients. While proficient prediction of previously diagnosed diseases is a baseline expectation, the ability to identify new, potential diagnoses based on similar patient data is equally crucial. In this context, patients treated with the same drug are

considered similar, and a diagnosis of an emerging disease in one patient might be predictive for the other. The R@k ($k = 20, 40$) is employed to analyze the performance of different models in predicting both previously diagnosed and emerging diseases, given the relatively small number of newly predicted diseases by each model. This metric reflects the proportion of accurately predicted occurred or emerging diseases against the total confirmed diagnoses. GRAM, CGL, and Chet were selected as comparison models due to their shared utilization of hierarchical (horizontal and vertical) disease relationships. This selection facilitates the assessment of the effectiveness of our proposed drug-disease ontology graph and disease ontology graph. As shown in Table 5, the test set results demonstrate that our proposed model THAM, achieves better performance in predicting both emerging and occurred diseases compared to existing baseline models. These findings substantiate the efficacy of our proposed heterogeneous graph and disease ontology graph learning approach in leveraging patient similarity patterns to predict potential future diagnoses.

Table 5: R@k of predicting occurred/emerging diseases on MIMIC-III.

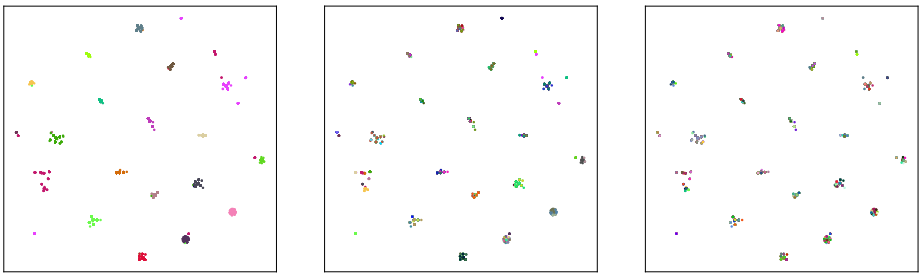
Models	**Occurred diseases**		**Emerging diseases**	
	R@20	R@40	R@20	R@40
GRAM	21.05	23.11	15.32	22.50
CGL	21.79	25.13	16.33	23.58
Chet	19.93	22.70	16.80	24.25
THAM	**22.48**	**25.45**	**17.01**	**24.50**

4.3. Interpretability analysis

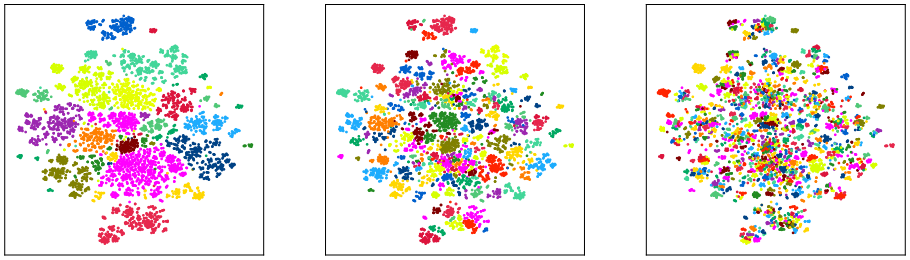
In this section, we discuss the representations of diseases and drugs trained by the model. The diseases in the ICD-9-CM standard are classified into different categories. To demonstrate our model's disease classification ability and illustrate the similarity among diseases, we utilize t-SNE(Van der Maaten & Hinton, 2008) to visualize the embedding vectors of 4,880 diseases and 3,202 drugs

from the MIMIC-III dataset. Additionally, we compare the disease embedding vectors produced by several baseline models that incorporate the hierarchical relationship of diseases. In Figure 3, the different colors represent the various categories of diseases classified by the ICD-9-CM standard. From Figure 3, it is evident that all models have successfully classified diseases into corresponding clusters according to real-world classification standards, this indicates that we have successfully learned excellent disease representations by leveraging the correlation between drugs and diseases. Compared to CGL, THAM has a more distinctive way of classifying diseases. As shown in Figure 4, we map the disease embedding vectors into a 3D space and the drug embedding vectors into a 2D space. It can be observed that THAM still possesses excellent disease classification capability. Therefore, we can infer that obtaining better disease representations through the utilization of drug and time information is crucial.

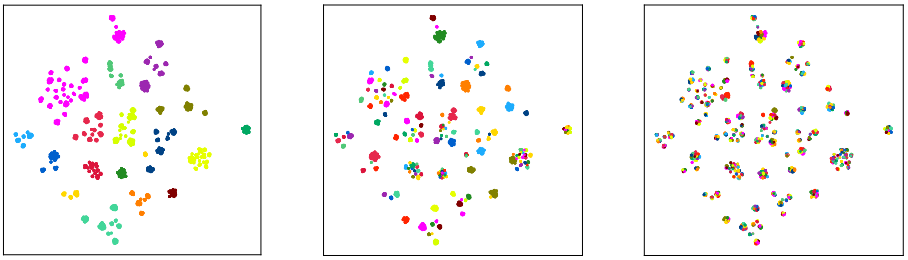
Figure 3 Code embeddings in three levels acquired by the GRAM, CGL, and THAM models. Each level represents different disease types, as indicated by the corresponding colors.



(a) GRAM level 1 **(b)** GRAM level 2 **(c)** GRAM level 3



(d) CGL level 1 **(e)** CGL level 2 **(f)** CGL level 3

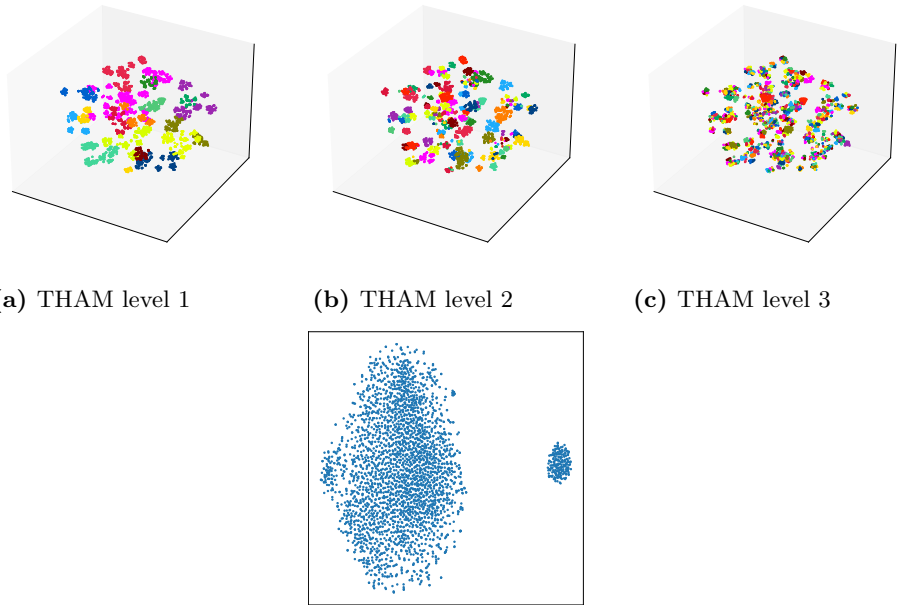


(g) THAM level 1 **(h)** THAM level 2 **(i)** THAM level 3

4.4. Parameter sensitivity analysis

We conducted a comprehensive sensitivity analysis on the model's hyperparameters to ascertain their impact on performance. This analysis was carried out on the MIMIC-III and MIMIC-IV datasets, using disease prediction metrics as indicators. Modifications included varying the dimension m of disease codes,

Figure 4 3D spaces of code embeddings acquired by model THAM and drug embeddings in 2D space.

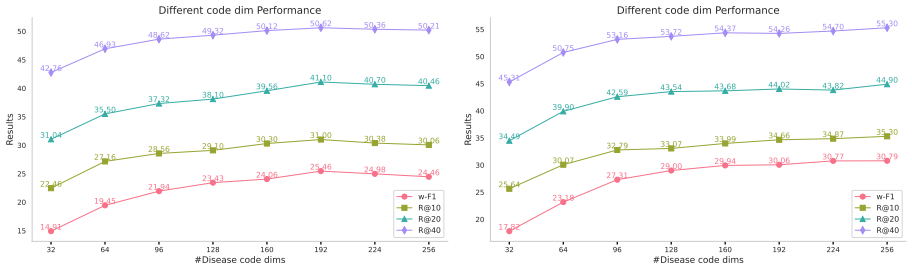


(a) THAM level 1 **(b)** THAM level 2 **(c)** THAM level 3

(d) Drug embeddings

initially set at 32 and incrementally increased by 32 up to a maximum of 256. In this evaluation of disease code dimensions, we set the number of layers for the encoder to 2. On the MIMIC-III dataset, the model exhibited its optimal performance when the disease code dimension was set to 192, with most indicators reaching their peak values. The scores were as follows: w-F_1 at 25.46%, R@10 at 31.00%, R@20 at 41.10%, and R@40 at 50.62%, surpassing other configurations. It is noteworthy that the model's performance improved gradually as the disease code dimension increased from 32 to 192, at which point all indicators reached their peak values. Beyond this dimension, all indicators showed slight declines. On the MIMIC-IV dataset, the model exhibited the best overall predictive performance with a disease code dimension of 256. This observation suggests that increasing the disease code dimension on both datasets can result in excellent performance. These findings imply that the proposed model neces-

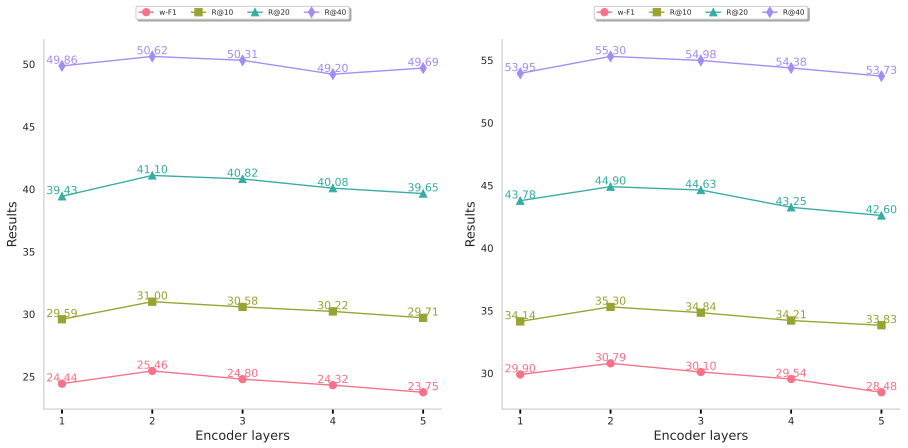
Figure 5 The Impact of code dimensions on Performance of MIMIC-III and MIMIC-IV.



(a) MIMIC-III **(b)** MIMIC-IV

sitates more parameters for effectively learning and representing complex data features. For a visualization of the model's performance across varying disease code dimensions, refer to Figure 5. In addition, we conducted an evaluation of the sensitivity of the model's encoder layers. Initially, we set the disease code dimension to the previously determined optimal value. The number of encoder layers was incrementally increased from 1 to 5. On the MIMIC-III dataset, the model achieved its best predictive performance with 2 encoder layers, yielding a w-F_1 at 25.46%, R@10 at 31.00%, R@20 at 41.10%, and R@40 at 50.62%. These scores outperformed other configurations, but further increases in the number of encoder layers resulted in slight declines in performance. Similarly, on the MIMIC-IV dataset, the model also peaked with 2 encoder layers, achieving a w-F_1 score of 30.79%, R@10 score of 35.30%, R@20 score of 44.90%, and R@40 score of 55.30%, followed by gradual declines. These findings indicate that an excessive number of encoder layers does not necessarily improve the predictive performance of the model. Thus, the results confirm that setting the number of encoder layers to 2 can achieve highly favorable performance on both datasets. The performance of the model with varying numbers of encoder layers can be observed in Figure 6. Setting the disease code dimension to 192 and the number of encoder layers to 2 has both showcased remarkable performance on both datasets. This underscores the model's robustness across hyperparameters.

Figure 6 The Impact of Encoder Layers on Performance of MIMIC-III and MIMIC-IV.



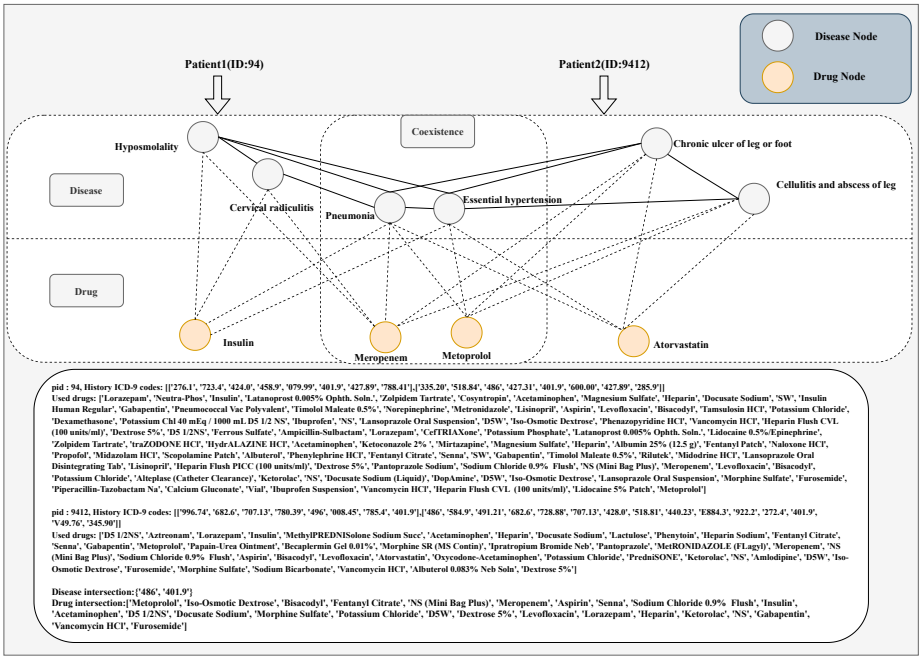
(a) MIMIC-III **(b)** MIMIC-IV

The analysis of parameter sensitivity yields valuable insights into the model's optimal performance.

4.5. Case Study

We randomly selected two patients, with IDs 92 and 9412, from the MIMIC-III dataset. By analyzing their historical admission records, we extracted a heterogeneous subgraph that offers insights into our proposed method for heterogeneous graph learning. In Figure 7, diseases are represented by grey circles, and drugs by orange nodes. The weights of the edges between diseases indicate their co-occurrence frequency, while the weights of the dashed edges connecting diseases and drugs also represent their co-occurrence frequency. Notably, both patients were treated with the same drug, such as Meropenem, and were diagnosed with pneumonia simultaneously. This suggests that these patients may have similar or related diseases in the future. The construction of a heterogeneous graph allows us to uncover hidden relationships between drugs and diseases. To enhance the interpretability of the model, paths and weights in the graph are converted to corresponding adjacency matrices. To maintain conci-

Figure 7 Heterogeneous subgraphs extracted from the visit records of Patient 1 and Patient 2.



sion, only a subset of the diagnosed diseases and drug records of the patients are displayed in the figure, while the complete historical admission records of the two patients are recorded below in Figure 7.

5. Conclusion and future work

This paper introduces THAM, a model that utilizes heterogeneous graph learning methods, models time information and Adaptive attention merging mechanism. It aims to learn meaningful representations of diseases and drugs, enabling the prediction of future health events and exploration of disease progression over time. We demonstrate the superiority of THAM over baseline methods using the widely-used MIMIC-III and MIMIC-IV datasets. THAM effectively leverages visit data from electronic health records (EHR) and showcases its efficacy in predicting health events, thereby enhancing personalized

and prospective healthcare management. In ablation study, we analyze the contributions of hierarchical disease information, time information, and adaptive attention merging mechanisms. In conclusion, THAM presents a novel strategy that significantly improves the accuracy of health event prediction.

In the future, our plans include exploring the expansion of the model's capabilities to integrate a broader range of medical ontologies and electronic medical record systems. This expansion will further enhance its applicability and accuracy. Additionally, we will investigate the potential of THAM to adapt to real-time data inputs, supporting dynamic and continuous patient monitoring systems. Furthermore, this model can be extended to incorporate multimodal data such as imaging and genomic information. This extension will greatly enrich the predictive capabilities of the model and provide a more comprehensive view of patient health.

CRediT authorship contribution statement

Shibo Li: investigation, software, writing—original draft preparation, writing—review and editing, visualization, formal analysis, data curation. **Hengliang Cheng:** writing—review and editing, software, visualization. **Runze Li:** writing—review and editing. **Weihua Li:** Conceptualization, methodology, investigation, supervision, project administration, funding acquision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The link to the dataset used in the paper is as follows: (1) `https://mimic.mit.edu/docs/iii/` (2) `https://mimic.mit.edu/docs/iv/`

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant 32060151, and the Yunnan Provincial Foundation for Leaders of Disciplines in Science and Technology, China under Grant 202305AC160014, and the Innovation Research Foundation for Graduate Students of Yunnan University under Grant ZC-23234341.

References

Bai, T., Zhang, S., Egleston, B. L., & Vucetic, S. (2018). Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 43–51).

Baytas, I. M., Xiao, C., Zhang, X., Wang, F., Jain, A. K., & Zhou, J. (2017). Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 65–74).

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016a). Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference* (pp. 301–318). PMLR.

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

Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., & Stewart, W. (2016b). Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, *29*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, .

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B. et al. (2023). Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, *10*, 1.

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. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, *3*, 1–9.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, .

Li, Q., You, T., Chen, J., Zhang, Y., & Du, C. (2024). Biodyngrap: Biomedical event prediction via interpretable learning framework for heterogeneous dynamic graphs. *Expert Systems with Applications*, *244*, 122964.

Li, R., Yin, C., Yang, S., Qian, B., & Zhang, P. (2020a). Marrying medical domain knowledge with deep learning on electronic health records: a deep visual analytics approach. *Journal of medical Internet research*, *22*, e20645.

Li, W., Li, H., Yang, B., Zhou, L., Yang, X., Zhang, M., & Wang, B. (2023). Knowledge-aware representation learning for diagnosis prediction. *Expert Systems*, *40*, e13175.

Li, Y., Qian, B., Zhang, X., & Liu, H. (2020b). Knowledge guided diagnosis prediction via graph spatial-temporal network. In *Proceedings of the 2020 SIAM International Conference on Data Mining* (pp. 19–27). SIAM.

Lu, C., Han, T., & Ning, Y. (2022). Context-aware health event prediction via transition functions on dynamic disease graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 4567–4574). volume 36.

Lu, C., Reddy, C. K., Chakraborty, P., Kleinberg, S., & Ning, Y. (2021). Collaborative graph learning with auxiliary text for temporal event prediction in healthcare. *arXiv preprint arXiv:2105.07542*, .

Luo, J., Ye, M., Xiao, C., & Ma, F. (2020). Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 647–656).

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, .

Ma, F., Chitta, R., Zhou, J., You, Q., Sun, T., & Gao, J. (2017). Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1903–1911).

Ma, F., Gao, J., Suo, Q., You, Q., Zhou, J., & Zhang, A. (2018a). Risk prediction on electronic health records with prior medical knowledge. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1910–1919).

Ma, F., Wang, Y., Xiao, H., Yuan, Y., Chitta, R., Zhou, J., & Gao, J. (2019). Incorporating medical code descriptions for diagnosis prediction in healthcare. *BMC medical informatics and decision making*, *19*, 1–13.

Ma, F., You, Q., Xiao, H., Chitta, R., Zhou, J., & Gao, J. (2018b). Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 743–752).

Ma, L., Zhang, C., Wang, Y., Ruan, W., Wang, J., Tang, W., Ma, X., Gao, X., & Gao, J. (2020). Concare: Personalized clinical feature embedding via

capturing the healthcare context. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 833–840). volume 34.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*.

Nguyen, P., Tran, T., Wickramasinghe, N., & Venkatesh, S. (2017). Deepr: A convolutional net for medical records. *IEEE Journal of Biomedical and Health Informatics*, *21*, 22–30. doi:10.1109/JBHI.2016.2633963.

Organization, W. H. (2004). *International Statistical Classification of Diseases and related health problems: Alphabetical index* volume 3. World Health Organization.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE transactions on neural networks*, *20*, 61–80.

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

Slee, V. N. (1978). The international classification of diseases: ninth revision (icd-9).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, .

Ye, M., Cui, S., Wang, Y., Luo, J., Xiao, C., & Ma, F. (2021a). Medpath: Augmenting health risk prediction via medical knowledge paths. In *Proceedings of the Web Conference 2021* (pp. 1397–1409).

Ye, M., Cui, S., Wang, Y., Luo, J., Xiao, C., & Ma, F. (2021b). Medretriever: Target-driven interpretable health risk prediction via retrieving unstructured medical text. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (pp. 2414–2423).

Yin, C., Zhao, R., Qian, B., Lv, X., & Zhang, P. (2019a). Domain knowledge guided deep learning with electronic health records. In *2019 IEEE International Conference on Data Mining (ICDM)* (pp. 738–747). doi:10.1109/ICDM.2019.00084.

Yin, C., Zhao, R., Qian, B., Lv, X., & Zhang, P. (2019b). Domain knowledge guided deep learning with electronic health records. In *2019 IEEE International Conference on Data Mining (ICDM)* (pp. 738–747). IEEE.

Zhang, X., Qian, B., Li, Y., Yin, C., Wang, X., & Zheng, Q. (2019). Knowrisk: an interpretable knowledge-guided model for disease risk prediction. In *2019 IEEE International Conference on Data Mining (ICDM)* (pp. 1492–1497). IEEE.