Patient-Visit-Spanned Hypergraph Learning for EHR-based Diagnosis Prediction

Ye Yuan[1], Haiyan Wang[1], Lun Hu[2] and Xin Luo[1,*]
[1]College of Computer and Information Science, Southwest University
[2]Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences
{yuanyekl, luoxin}@swu.edu.cn, why040680@email.swu.edu.cn, hulun@ms.xjb.ac.cn

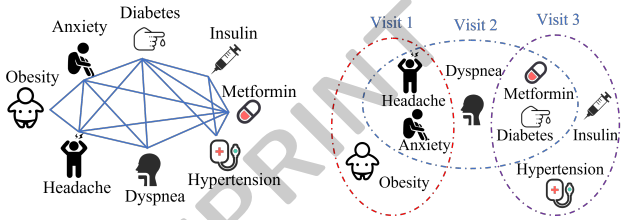
Abstract

Hypergraphs effectively model complex interactions in structured Electronic Health Records (EHR). Consequently, Hypergraph Neural Networks (HGNNs) are commonly applied to EHR-based diagnosis prediction. However, existing HGNNs struggle to capture patient-visit long-range dependencies when processing EHR-derived hypergraphs. To tackle this issue, we propose a Patient-Visit-Spanned HyperGraph Learning (PVHGL) framework specifically designed for diagnosis prediction. Concretely, PVHGL initially constructs a unified patient-visit hypergraph that integrates visit records from all patients, enabling the capture of shared healthcare patterns across the patient population. Subsequently, it incorporates a Transformer architecture enhanced by two structure-encoding matrices to facilitate one-step message propagation, which preserves both local and global hypergraph structural properties effectively. Additionally, the framework integrates a medical code co-occurrence matrix to explicitly guide the learning process by highlighting critical medical code interactions. Comprehensive experiments on three real-world datasets show that the proposed PVHGL significantly outperforms state-of-the-art baselines in diagnosis prediction.

1 Introduction

The widespread adoption of Electronic Health Records (EHR) has fueled deep learning-based predictive modeling. By uncovering meaningful clinical patterns, these models enhance understanding of patients' health states and support key tasks such as early diagnosis [Che and Liu, 2017], medication recommendation [Kim *et al.*, 2025], and risk stratification [Liu *et al.*, 2022], ultimately empowering proactive interventions and improving patient outcomes [Wu *et al.*, 2025].

According to existing studies, sequence-based models [Luo *et al.*, 2020] and graph neural networks (GNN)-based models [Poulain and Beheshti, 2024] are two mainstream approaches for processing EHR. In particular, sequence-based approaches typically represent clinical visits as temporal sequences to learn dependencies across clinical events [Chen *et al.*, 2025]. However, they often overlook complex interactions among heterogeneous medical codes, including diagnoses, medications, and procedures. In contrast, GNN-based models represent EHR as graphs, where medical codes co-occurring within a clinical visit serve as nodes and edges represent their co-occurrence relationships [Chan *et al.*, 2024], as illustrated in Figure 1(a). Although these models can capture interactions between medical codes, they essentially focus on pairwise relationships, which hinders sufficient exploration of the high-order code interactions.



(a) Vanilla graph of EHR (b) Hypergraph of EHR

Figure 1: An illustration of EHR for a patient with three clinical visits. (a) A vanilla graph connects pairs of medical codes that co-occur within a visit. (b) A hypergraph represents each visit as a hyperedge that links all medical codes recorded during that visit.

Hypergraphs [Feng *et al.*, 2019] have been proposed as an extension of vanilla graphs, where each hyperedge connects multiple nodes, thereby overcoming the limitations of pairwise modeling. As illustrated in Figure 1(b), this structure is well suited for modeling EHR, since it captures the intricate connections among medical codes and clinical visits. Accordingly, Hypergraph Neural Networks (HGNNs) [Cai *et al.*, 2022] have been applied to process hypergraphs. However, existing HGNNs often *rely on a localized two-step message propagation mechanism, thereby failing to fully utilize the patient-visit long-range dependencies within EHR-derived hypergraphs*. Specifically, such a defect can be described from the following two perspectives:

- Two-step (node–hyperedge–node) message propagation mechanism fails to explicitly model interactions between visits across clinically similar patients. Even when two visits involve a high overlap of shared codes,

*Corresponding author.

their interactions can only be implicitly inferred via the shared nodes. Such interaction effects often necessitate multiple alternating update iterations to be fully manifested. *Consequently, the long-range dependencies among patients are neglected*.

- Localized propagation scheme relies solely on information from adjacent hyperedges, which overlooks long-range dependencies spanning multiple hyperedges. Hence, it limits the model's ability to capture implicit interactions among medical codes across distant visits in EHR-derived hypergraphs. For instance, a patient is initially diagnosed with obesity, but the resulting diabetes is identified only after several visits. *Consequently, the long-range dependencies among visits are neglected.*

In light of the aforementioned challenges, we propose **PVHGL**, a novel **P**atient-**V**isit-Spanned **H**yper**G**raph **L**earning framework for precise EHR-based diagnosis prediction. PVHGL is designed to fully utilize the patient-visit long-range dependencies within EHR-derived hypergraphs. Specifically, it first constructs a unified hypergraph containing all clinical visits of EHR data, enabling the model to capture both local and global structural properties of the hypergraph. Additionally, the framework integrates a Transformer architecture augmented with two structure-encoding matrices, enabling one-step message propagation that effectively preserves both local and global structural properties of the hypergraph. Furthermore, a medical code co-occurrence matrix is incorporated to explicitly guide the learning process by accentuating the critical interactions among medical codes. The main contributions are summarized as follows:

- **Transformer-based one-step message propagation.** A Transformer architecture enhanced by two structure-encoding matrices for medical codes and clinical visits is designed, which preserves the local and global structural properties via efficient one-step message propagation. It overcomes the localized two-step propagation limitation of conventional HGNNs, enabling modeling of patient-visit long-range dependencies directly.
- **Co-occurrence-guided attention recalibration.** A medical code co-occurrence matrix is incorporated to explicitly recalibrate attention weights, which introduces the prior clinical knowledge to emphasize high-order interactions that reflect clinically meaningful dependencies.
- **Extensive experiments on clinical EHR datasets.** Extensive experiments are conducted on three real-world datasets, and the results demonstrate that PVHGL consistently outperforms state-of-the-art baselines in EHR-based diagnosis prediction[1].

2 Related Work

2.1 Sequence-based Diagnosis Predictor

Early studies primarily utilize sequence models to process EHR for diagnosis prediction. RETAIN [Choi *et al.*, 2016] employs two Recurrent Neural Networks (RNNs) to perform interpretable prediction. Dipole [Ma *et al.*, 2017] incorporates a bi-directional RNN with multiple attention mechanisms to capture past and future visit information. ConCare [Ma *et al.*, 2020] introduces a personalized clinical feature embedding framework that leverages multi-head self-attention to capture dynamic and static healthcare contexts. HiTANet [Luo *et al.*, 2020] utilizes a time-aware encoder to learn visit-level temporal information. However, these approaches are limited in modeling the intricate dependencies among diverse medical codes [Wu *et al.*, 2021].

2.2 GNN-based Diagnosis Predictor

GNNs have demonstrated their effectiveness in modeling EHR. For example, GRAM [Choi *et al.*, 2017] and G-BERT [Shang *et al.*, 2019] construct medical ontology graphs based on domain knowledge to enhance representations of medical codes. InceptionGCN [Kazi *et al.*, 2019] improves diagnosis prediction accuracy by incorporating geometric inception modules that effectively capture structural variations in convolution operations. To uncover hidden disease relationships, Chet [Lu *et al.*, 2022] explores horizontal connections within the medical ontology graph. DDHGNN [Wen *et al.*, 2022] constructs a bipartite graph composed of patients and diseases, and employs a hierarchical aggregation method that accounts for the graph's spatial-temporal properties. However, these methods inherently impede a thorough exploration of high-order interactions since only pairwise relationships between medical codes are utilized [Wu *et al.*, 2023].

2.3 HGNN-based Diagnosis Predictor

Recently, several studies have incorporated HGNNs for diagnosis prediction owing to their ability to capture high-order interactions through hyperedges. For instance, EHR2HG [Sun *et al.*, 2022] constructs two types of hypergraphs to jointly learn patient and disease embeddings, which are aggregated to generate diagnosis predictions. HCL [Cai *et al.*, 2022] constructs a hypergraph alongside two vanilla graphs and then applies contrastive learning to generate patient embeddings. MCHN [Zhang *et al.*, 2024a] constructs multiple hypergraphs from different medical code views, and designs code-level and visit-level attention mechanisms to aggregate clinical information. TACCO [Zhang *et al.*, 2024b] leverages hypergraph modeling to uncover meaningful groupings of medical concepts and visits by jointly analyzing their interactions. However, they fail to fully harness the patient-visit long-range dependencies in EHR-based hypergraphs.

3 Preliminaries

3.1 Electronic Health Records

EHR consists of longitudinal visit records collected from a cohort of patients. Formally, let $\mathcal{R} = \{r_u \mid u \in \mathcal{U}\}$ denote the complete set of visit records, where $\mathcal{U}$ denotes the set of patients and $|\mathcal{U}|$ represents the total number of individuals. Each record $r_u = (v_1^u, v_2^u, \ldots, v_{T-1}^u)$ comprises a sequence of clinical visits, with $(T-1)$ representing the number of observed clinical visits. Here, v_j^i denotes the j-th visit of the i-th

[1]Code and Supplementary Material are available at https://github.com/Forwhyy/PVHGL

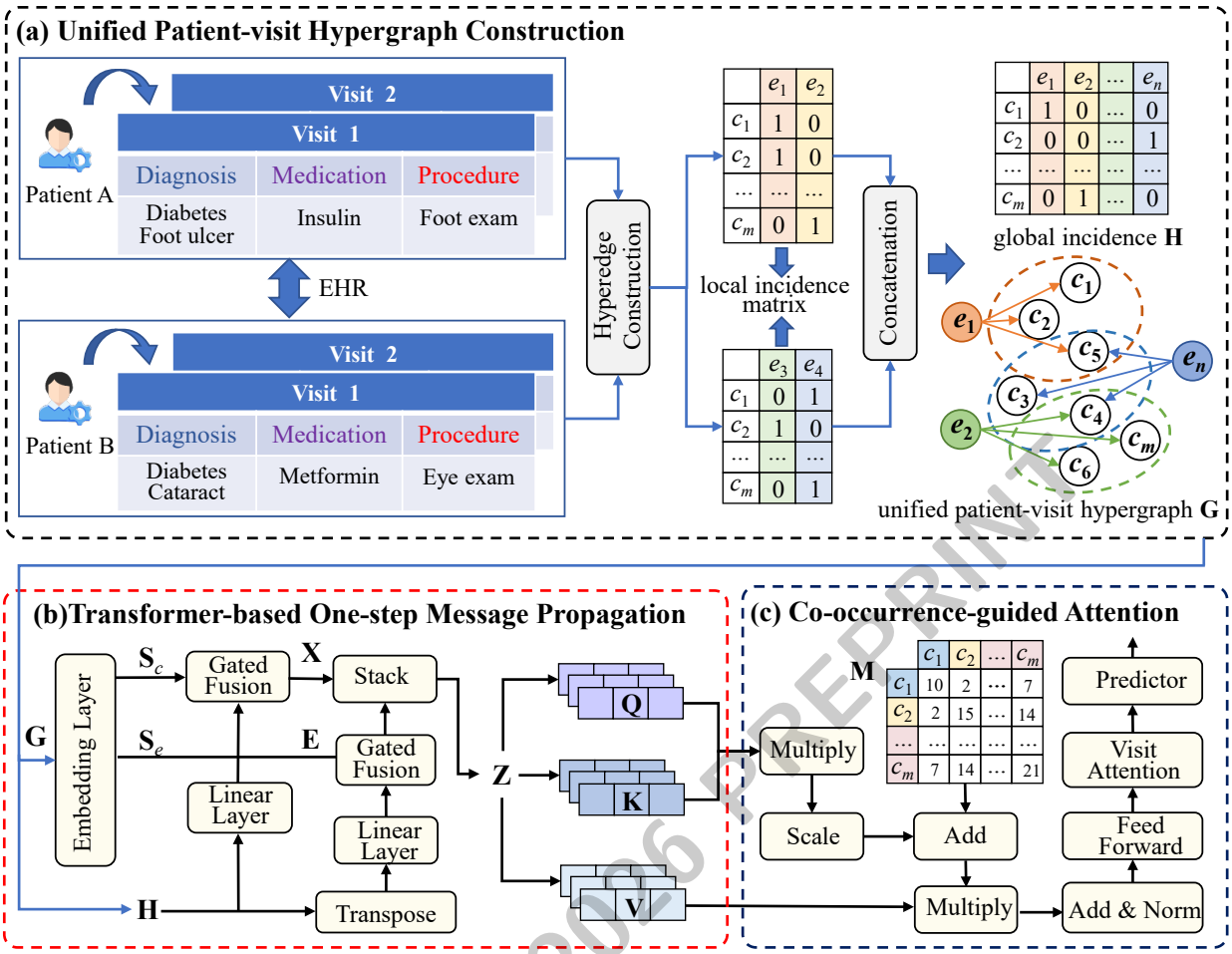


Figure 2: An overview of our proposed PVHGL framework.

patient. Each clinical visit v is characterized by a set of medical codes, including diagnoses, procedures, and medications, which are derived from standardized medical classification systems. To maintain clarity, we hereafter refer to clinical visit(s) as 'visit(s)' and medical code(s) as 'code(s).'

3.2 Task Definition

Let q denote the number of distinct diagnoses. Each visit at time step t is represented by a binary vector $y_t \in \{0,1\}^q$, where $y_{t,i} = 1$ indicates the presence of the i-th diagnosis. Given a patient with $(T-1)$ historical visits, the diagnosis prediction aims to estimate the diagnosis label probabilities for the T-th visit, denoted as $\hat{y}_T \in [0,1]^q$.

$$\hat{y}_T = \text{Prediction}(v_1, v_2, \ldots, v_{T-1}) \tag{1}$$

3.3 Hypergraph Neural Networks

Conventional HGNNs typically adopt a two-step message propagation mechanism to learn node embeddings. The first step updates each hyperedge embedding by aggregating messages from its connected nodes:

$$x_e^{(l)} = f_{\text{node}\to\text{edge}}\left(\left\{x_v^{(l-1)} \mid v \in \mathcal{V}(e)\right\}\right) \tag{2}$$

where $x_e^{(l)}$ and $x_v^{(l-1)}$ denote the embeddings of hyperedge e at the l-th layer and node v at the $(l-1)$-th layer, respectively. $\mathcal{V}(e)$ represents the set of nodes connected to hyperedge e. In the second step, each node embedding is updated by aggregating messages from its associated hyperedges:

$$x_v^{(l)} = f_{\text{edge}\to\text{node}}\left(\left\{x_e^{(l)} \mid e \in \mathcal{E}(v)\right\}\right) \tag{3}$$

Here, $\mathcal{E}(v)$ denotes the set of hyperedges of node v. The functions $f_{\text{node}\to\text{edge}}$ and $f_{\text{edge}\to\text{node}}$ serve as aggregation operators that enable message passing between nodes and hyperedges.

4 Methodology

This section presents the proposed PVHGL framework in detail. As illustrated in Figure 2, PVHGL consists of three key components: (1) unified patient-visit hypergraph construction, (2) Transformer-based one-step message propagation, and (3) co-occurrence-guided attention recalibration.

4.1 Unified Patient-visit Hypergraph Construction

A distinctive property of EHR is the inherent many-to-many structure. Specifically, each visit typically involves multiple codes, and each code often appears across multiple visits.

This characteristic induces complex high-order interactions that vanilla graphs fail to capture. To address it, we construct a unified patient-visit hypergraph that integrates visits from different patients, as depicted in Figure 2(a). Concretely, each unique code is modeled as a node, and each visit is represented as a hyperedge that connects all co-occurring codes. By concatenating hyperedges across the entire patient cohort, the unified hypergraph removes patient-specific boundaries and facilitates the joint modeling of high-order code interactions and population-wide healthcare patterns.

Formally, the unified hypergraph is represented by $G = (\mathcal{V}, \mathcal{E}, H)$. Here, $\mathcal{V} = \{c_1, c_2, \ldots, c_m\}$ denotes the set of nodes, each representing a unique code, and m is the total number of codes. The hyperedge set $\mathcal{E} = \{e_1, e_2, \ldots, e_n\}$ corresponds to visits, where n is the total number of visits. The topology of the unified hypergraph is denoted by the incidence matrix $H \in \mathbb{R}^{m \times n}$. Each entry h_{ij} is defined as:

$$h_{ij} = \begin{cases} 1 & \text{if node } c_i \in \text{ hyperedge } e_j, \\ 0 & \text{if node } c_i \notin \text{ hyperedge } e_j \end{cases} \tag{4}$$

By concatenating the local incidence matrix specific to each patient, we obtain the global incidence matrix of the unified patient-visit hypergraph.

4.2 Transformer-based One-step Message Propagation

Conventional HGNNs first aggregate node information via hyperedges and then propagate it back to nodes to obtain the node embeddings. Moreover, only adjacent hyperedges are processed during the learning process. Such a localized two-step message propagation mechanism fails to capture the implicit patient-visit long-range dependencies. Hence, we design a Transformer architecture enhanced by two structure-encoding matrices to implement one-step message propagation, which effectively preserves both local and global hypergraph structural properties.

Local hypergraph structure learning. This module captures local structural context from both code and visit perspectives. First, we embed codes and visits into a d-dimensional latent space, denoted as $Z_c \in \mathbb{R}^{m \times d}$ and $Z_e \in \mathbb{R}^{n \times d}$, respectively. Note that visit embeddings are initialized by aggregating connected code embeddings to preserve code-visit interactions and mitigate random initialization noise.

As introduced above, the structure of the hypergraph is encoded by an incidence matrix H. Hence, it naturally facilitates the extraction of local structure information. Let S_c and S_e be the local structure-encodings of codes and visits, respectively. These encodings are computed as follows:

$$\begin{aligned} S_c &= HW_c \in \mathbb{R}^{m \times d} \\ S_e &= H^\top W_e \in \mathbb{R}^{n \times d} \end{aligned} \tag{5}$$

where $W_c \in \mathbb{R}^{n \times d}$ and $W_e \in \mathbb{R}^{m \times d}$ are learnable weight matrices.

To mitigate the loss of crucial information during the local structure encoding process, we incorporate a gating mechanism that dynamically modulates the contributions of structural encodings. Specifically, we compute the scalar gates $g_{c_i} \in [0, 1]$ and $g_{e_j} \in [0, 1]$, which serve as attention weights to regulate the importance of structural information for each code i and visit j:

$$\begin{aligned} g_{c_i} &= \text{sigmoid}\left(\sum_{j=1}^{n} h_{i,j}\, \beta_j^{(e)}\right) \\ g_{e_j} &= \text{sigmoid}\left(\sum_{i=1}^{m} h_{i,j}\, \beta_i^{(v)}\right) \end{aligned} \tag{6}$$

where $\beta_j^{(e)}$ and $\beta_i^{(v)}$ are learnable scalar weights corresponding to the j-th visit and the i-th code, respectively. The final local structure-aware embeddings are obtained as follows:

$$\begin{aligned} X &= g_c \odot S_c + (1 - g_c) \odot Z_c \\ E &= g_e \odot S_e + (1 - g_e) \odot Z_e \end{aligned} \tag{7}$$

where $\odot$ denotes element-wise multiplication applied row-wise. The matrices X and E represent the local structure-aware embeddings for codes and visits, respectively. Intuitively, when a code exhibits strong correlation with specific visits, its embeddings are primarily shaped by structural information. Conversely, if the associations are weak, the model relies more on the raw input embedding. A similar interpretation applies to the gating on visits.

Global hypergraph structure learning. The codes often exhibit complex interactions that span multiple visits. Moreover, clinicians frequently leverage the experiences of similar visits in making clinical decisions. Inspired by these observations, we adopt a Transformer architecture to effectively model both the long-range dependencies among codes and the inter-visit relationships inherent in EHR.

To preserve the local structure of the hypergraph during representation learning, we incorporate the structure-aware embeddings obtained in Eq. (7) as inputs to the Transformer:

$$Z^{(0)} = \begin{bmatrix} X \\ E \end{bmatrix} \tag{8}$$

where $Z^{(0)}$ denotes the initial embedding matrix. We employ a self-attention mechanism to identify the most informative interactions between codes and visits, enabling the model to aggregate global structural information. It is formulated as:

$$\text{Self-Att}(Z^{(l)}) = A^{(l)} Z^{(l)} W_V^{(l)} \tag{9}$$

where $A^{(l)}$ denotes the weight matrix computed as:

$$A^{(l)} = \text{softmax}\left(\frac{Z^{(l)} W_Q^{(l)} (Z^{(l)} W_K^{(l)})^\top}{\sqrt{d}}\right) \tag{10}$$

where $W_Q^{(l)}$, $W_K^{(l)}$, and $W_V^{(l)}$ are learnable parameter matrices corresponding to queries, keys, and values in the l-th layer, respectively. The dimensionality parameter d determines the scaling factor of the dot products.

In our context, the explicit positional information is unavailable in the dataset. Hence, we omit positional encoding. Finally, the embeddings at the $(l+1)$-th layer are obtained through a residual connection:

$$Z^{(l+1)} = \text{LayerNorm}\left(Z^{(l)} + \text{Self-Att}(Z^{(l)})\right) \tag{11}$$

Compared with conventional HGNNs employing a localized two-step propagation scheme, our approach enables the joint learning of code and visit embeddings via a one-step propagation process, which eliminates the need for adjacent hyperedges. It directly computes pairwise attention across all nodes and extracts the local and global structural properties of the hypergraph.

4.3 Co-occurrence-guided Attention Recalibration

Recent studies have shown that a key challenge in existing Transformer-based models is over-globalization [Xing *et al.*, 2024]. This issue arises because the attention mechanism tends to assign excessive weights to distant nodes, overlooking informative neighbors that are critical to downstream tasks. In clinical practice, it is widely acknowledged that certain codes frequently co-occur. Such high-frequency co-occurrence patterns typically indicate strong clinical associations. Motivated by this, we introduce a medical code co-occurrence matrix to guide attention toward clinically meaningful code pairs.

Specifically, the co-occurrence matrix $M \in \mathbb{R}^{m \times m}$ is built by counting the frequency of code pairs that co-occur within the same visit. Note that M is only built on training data. For each visit in which codes i and j appear together, the entries m_{ij} and m_{ji} are incremented by 1. To prevent the self-attention of each code from being overly suppressed after normalization, the diagonal entry m_{ii} is set to the maximum value in the i-th row of M. Then M is normalized to align with the scale of attention scores:

$$\hat{M}_{ij} = \frac{M_{ij}}{\sum_{k=1}^{m} M_{ik}} \tag{12}$$

After obtaining the code-to-code attention weights from Eq.(10), we recalibrate them by adding corresponding values from the co-occurrence matrix, which emphasizes attention on code pairs with strong clinical relevance:

$$\hat{A}_{ij}^{(l)} = \delta_{i,j<m} \cdot A_{ij}^{(l)} + \alpha \cdot \hat{M}_{ij} \tag{13}$$

where α is the control coefficient, and δ is an indicator function that equals 1 if both indices i and j are less than m, and 0 otherwise. The recalibrated attention weights are normalized using the softmax function:

$$\tilde{A}_{ij}^{(l)} = \frac{\exp\left(\hat{A}_{ij}^{(l)}\right)}{\sum_{k=1}^{m+n} \exp\left(\hat{A}_{ik}^{(l)}\right)} \tag{14}$$

Finally, the embedding matrix $Z^{(l)}$ is updated using the recalibrated attention weights $\tilde{A}^{(l)}$ following the update procedure described in Eq. (9).

4.4 Diagnosis Prediction

To effectively leverage historical visit information, we extract the sequence of visit embeddings $[z_1^u, z_2^u, \ldots, z_{T-1}^u]$ for patient r_u from the final layer. Subsequently, a location-aware attention mechanism is applied to aggregate these embeddings into a patient embedding:

$$a_u = \text{softmax}([z_1^u, z_2^u, \ldots, z_{T-1}^u]W_a) \tag{15}$$

$$p_u = a_u[z_1^u, z_2^u, \ldots, z_{T-1}^u]^\top \tag{16}$$

where W_a is a learnable context vector and a_u denotes the attention weights over visits. The final prediction is generated based on the aggregated patient embedding p_u:

$$\hat{y}_T = \text{sigmoid}(\text{MLP}(p_u)) \tag{17}$$

where $\hat{y}_T$ represents the predicted probabilities for the q possible diagnoses. Let y_T denote the true diagnosis labels. The model is trained by minimizing binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \left(\mathbf{y}_i^\top \log \hat{\mathbf{y}}_i + (\mathbf{1} - \mathbf{y}_i)^\top \log(\mathbf{1} - \hat{\mathbf{y}}_i) \right) \tag{18}$$

5 Experiments

5.1 Experimental Settings

Datasets. We evaluate the methods on three real-world datasets, including MIMIC-III [Johnson *et al.*, 2016], MIMIC-IV [Johnson *et al.*, 2023], and eICU [Pollard *et al.*, 2018]. We retain only patients with at least two visits. Table 1 reports the dataset statistics, and the datasets are partitioned into training/validation/test sets with 70%-10%-20%.

Dataset	MIMIC-III	MIMIC-IV	eICU
# visits	14086	29303	15574
# patients	5442	10000	7177
# medicines	300	281	78
# diagnoses	1956	1134	868
# procedures	1392	899	1000
avg # diagnoses per visit	13.70	13.42	4.37

Table 1: Statistics of processed datasets.

Baselines. We divide baselines into three paradigms: (i) sequence-based models, i.e., RETAIN [Choi *et al.*, 2016], Transformer [Vaswani *et al.*, 2017], StageNet [Gao *et al.*, 2020], and HiTANet [Luo *et al.*, 2020]; (ii) GNN-based models, i.e., SafeDrug [Yang *et al.*, 2021], CGL [Lu *et al.*, 2021], Chet [Lu *et al.*, 2022], Sherbet [Lu *et al.*, 2023], and ADRL [Cheng *et al.*, 2025]; and (iii) HGNN-based models, i.e., AllSet (Transformer) [Chien *et al.*, 2022], HypEHR [Xu *et al.*, 2023], MCHN [Xing *et al.*, 2024], TACCO [Zhang *et al.*, 2024b], and Shy [Yu *et al.*, 2025].

Implementation Details. To ensure a fair comparison, we tune all competing methods via grid search over their hyperparameters and evaluate them on the same dataset splits. To account for stochasticity in training and initialization, we repeat each experiment five times with different random seeds and report the mean and standard deviation across runs.

Evaluation Metrics. Weighted F_1 score (w-F_1) and top-k recall (R@k) are selected as evaluation metrics. We set $k \in \{10, 20\}$ since most visits have around 13 diagnoses.

5.2 Performance Comparison

As shown in Table 2, PVHGL consistently outperforms all baselines. For instance, on MIMIC-III, PVHGL yields a 24.93% improvement in w-F_1 over the best sequence-based

Models	MIMIC-III			MIMIC-IV			eICU		
	w-F_1	R@10	R@20	w-F_1	R@10	R@20	w-F_1	R@10	R@20
RETAIN	19.23±0.21	24.12±0.16	33.72±0.25	26.20±0.34	30.13±0.38	41.06±0.37	46.88±0.19	67.51±0.27	76.11±0.33
Transformer	18.81±0.36	23.90±0.25	33.24±0.31	25.59±0.33	29.90±0.36	40.56±0.61	47.39±0.25	68.20±0.21	75.29±0.28
StageNet	18.32±0.11	22.62±0.39	32.29±0.44	22.76±0.26	27.80±0.43	38.60±0.34	47.60±0.23	67.40±0.18	74.63±0.29
HiTANet	20.69±0.14	24.18±0.22	33.76±0.15	27.56±0.23	31.53±0.19	41.30±0.26	48.16±0.22	68.32±0.16	76.52±0.23
CGL	22.21±0.09	24.71±0.07	34.39±0.11	29.82±0.06	32.76±0.06	42.75±0.15	50.67±0.12	71.99±0.15	78.23±0.17
SafeDrug	22.92±0.14	25.02±0.22	34.04±0.21	30.31±0.12	33.78±0.23	43.61±0.25	51.82±0.15	71.28±0.26	79.51±0.14
Chet	22.07±0.18	24.92±0.15	34.56±0.14	29.83±0.13	34.09±0.12	43.95±0.09	51.31±0.09	71.94±0.11	78.89±0.07
Sherbet	23.28±0.12	25.06±0.20	35.15±0.23	30.44±0.31	34.47±0.22	44.65±0.24	52.01±0.13	72.10±0.14	79.17±0.13
ADRL	23.14±0.21	25.33±0.17	35.30±0.23	31.19±0.16	34.10±0.23	44.25±0.15	50.05±0.19	71.16±0.15	79.34±0.22
AllSet	23.95±0.24	26.22±0.18	35.99±0.13	31.51±0.22	34.23±0.13	44.95±0.14	53.76±0.22	73.58±0.19	80.35±0.22
HypEHR	23.67±0.10	25.45±0.21	35.53±0.11	32.05±0.35	34.50±0.26	45.26±0.21	52.33±0.22	72.25±0.26	79.63±0.31
MCHN	23.54±0.08	25.37±0.11	35.51±0.09	30.55±0.06	33.98±0.07	44.83±0.07	53.12±0.10	72.33±0.11	80.33±0.15
TACCO	24.12±0.17	26.04±0.11	36.11±0.18	31.86±0.30	35.12±0.27	45.75±0.34	52.77±0.25	72.67±0.26	80.65±0.28
Shy	23.30±0.12	25.44±0.09	35.53±0.06	30.33±0.11	34.23±0.09	44.70±0.10	52.11±0.13	72.25±0.23	79.84±0.26
PVHGL	**25.85**±0.20	**27.52**±0.18	**37.79**±0.25	**34.74**±0.24	**37.43**±0.23	**47.35**±0.12	**55.18**±0.26	**74.33**±0.15	**81.35**±0.23

Table 2: Performance comparison with mean value and standard deviation. The best results are marked in **bold**.

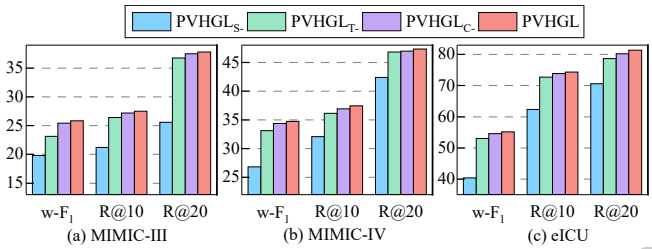


Figure 3: Ablation results on three datasets.

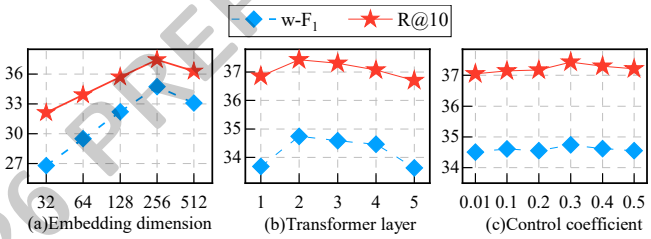


Figure 4: Hyperparameter testing on MIMIC-IV.

model, i.e., HiTANet. This significant gain is largely due to the limitations of sequence methods, which linearize visits and thus struggle to capture the structural dependencies among codes. We further observe that HGNN-based methods generally outperform graph-based counterparts, suggesting the importance of modeling high-order code dependencies. The HGNN-based models, except the proposed PVHGL, are restricted to message passing along the code→visit and visit→code directions. Consequently, message passing remains inherently two-stage, which makes it difficult to capture long-range code interactions and prevents explicit learning across similar visits. In contrast, PVHGL first employs two structure-encoding matrices to preserve the hypergraph's high-order semantics. It then couples code and visit embeddings to jointly model all possible code–visit interactions, enabling one-step message propagation. This design facilitates both long-range dependency modeling and explicit cross-visit knowledge sharing, ultimately leading to the best performance.

5.3 Ablation Study

To assess the contribution of each component in PVHGL, we construct three ablated variants. Specifically, PVHGL$_{S-}$ excludes structure encoding, PVHGL$_{T-}$ removes the Transformer, and PVHGL$_{C-}$ discards the co-occurrence matrix. As shown in Figure 3, ablating any component results in performance degradation, indicating that each module contributes positively. The most significant drop is observed for PVHGL$_{S-}$, highlighting the necessity of modeling high-order code interactions. The substantial decline observed in PVHGL$_{T-}$ verifies the effectiveness of the Transformer in capturing patient-visit long-range dependencies. Meanwhile, PVHGL$_{C-}$ yields a milder degradation, validating the contribution of high-frequency co-occurrence patterns in capturing clinically meaningful associations.

5.4 Hyperparameter Analysis

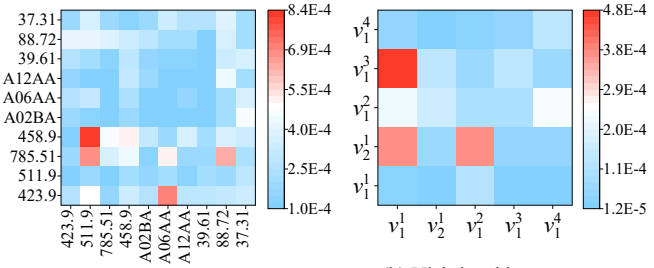
We investigate the sensitivity of three hyperparameters on MIMIC-IV: d, l, and α. As shown in Figure 4, increasing d and l improves performance initially, but further enlarging either leads to degradation, likely due to overfitting and over-propagation that impair generalization. Accordingly, α peaks at intermediate values by balancing co-occurrence information with structural signals. Deviating from this range either overfits to co-occurrence patterns or suppresses them, leading to poorer performance.

5.5 Visualization of Attention Weights

To verify whether PVHGL captures meaningful interactions both among codes and across visits, we visualize the code-

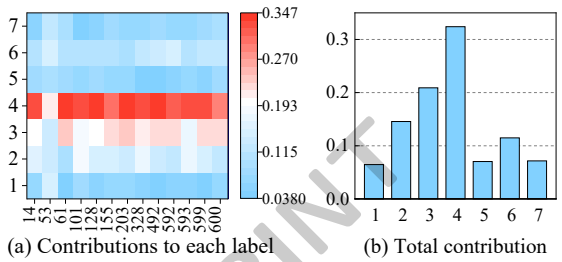
Visit	Diagnosis	Medication	Procedure
v_1^1	272.4; 410.71; 414.01; 274.9; 725.0	A12BA; A06AB; B01AB; B01AC; B05CX; B01AA; C07AB; C01DA; M04AA; N02AA	39.50; 37.22
v_2^1	272.4; 427.31; 274.9	B01AB; B01AA; C01BD; C01CA; C03CA	37.34; 99.25
v_1^3	410.71; 414.01; 403.91; 599.0	A06AB; B01AB; B01AC; C07AB; N02BE	37.61; 36.05

Table 3: Medical codes associated with selected clinical visits in MIMIC-IV.



(a) Code-leve hotmap (b) Visit-level hotmap

Figure 5: Attention heatmaps on MIMIC-IV.



(a) Contributions to each label (b) Total contribution

Figure 6: Visualization of visit-level contributions to diagnosis prediction: a case from MIMIC-IV.

level and visit-level attention heatmaps in Figure 5.

Code-level Attention Weights. We randomly sample ten codes, i.e., four diagnoses, three medications, and three procedures. Horizontal and vertical axes are annotated with the corresponding codes. As shown in Figure 5(a), hypotension (458.9) and pleural effusion (511.9) exhibit the highest attention score. Both are common complications of heart failure, suggesting that PVHGL effectively captures clinically meaningful co-occurrence patterns. In addition, we observe a high correlation between cardiogenic shock (785.51) and pleural effusion (511.9), which likely reflects a pathophysiological dependency. Specifically, cardiogenic shock results in significantly reduced cardiac output and systemic hypoperfusion. This hemodynamic compromise can lead to fluid overload, which in turn may cause pleural effusion. It demonstrates that PVHGL transcends simple co-occurrence patterns and successfully captures visit-spanned long-range dependencies.

Visit-level Attention Weights. We randomly sample five visits from four patients. As shown in Figure 5(b), v_j^i denotes the j-th visit of the i-th patient. Notably, strong correlations are observed among v_1^1, v_2^1, and v_1^3. To better understand the underlying factors, we present detailed visit information in Table 3. Specifically, v_1^3 and v_1^1 share two diagnoses (410.71 and 414.01) and three medications, while v_2^1 shares three medical codes with v_1^1. This demonstrates the presence of similar disease and treatment strategies both within and across patients. By tackling semantically related visits across different patients, PVHGL effectively captures patient-spanned long-range dependencies.

5.6 Interpretability of Prediction

PVHGL possesses satisfactory interpretability. We randomly sample a patient with six visits. The left side of Figure 6 presents the attention weights between each visit (y-axis) and the predicted diagnoses (x-axis). Brighter regions in the heatmap indicate higher attention weights, implying that specific visits play a more influential role in predicting certain diagnoses. Figure 6(b) shows the normalized contribution of each visit to the final prediction. Notably, the fourth visit yields the highest contribution since it consistently receives high attention scores across multiple diagnosis codes. This observation suggests that PVHGL can effectively identify and prioritize clinically informative visits within a patient's longitudinal history. Such interpretability provides valuable insights into the decision-making process and aligns with clinical intuition. Hence, it enables clinicians to make more evidence-based and personalized clinical decisions promptly, such as optimizing treatment plans, identifying high-risk subgroups, and prioritizing intervention strategies.

6 Conclusion

In this work, we propose PVHGL, a novel hypergraph learning framework that addresses the limitation of capturing patient-visit long-range dependencies for EHR-based diagnosis prediction. To achieve this, we first construct a unified patient-visit hypergraph that integrates population-wide visits to model shared healthcare patterns. Further, we propose a structure-enhanced Transformer equipped with dual encoding matrices to facilitate one-step message propagation while preserving local and global structural properties efficiently. Finally, we introduce the co-occurrence-guided attention recalibration mechanism to incorporate clinical priors, emphasizing medically significant interactions. Extensive experimental evaluations demonstrate the effectiveness of PVHGL. In the future, we plan to utilize sparse attention and gradient checkpointing to reduce the quadratic computational complexity of Transformer, thereby improving the scalability of PVHGL. Moreover, we plan to collaborate with clinical doctors to validate the effectiveness of PVHGL.

Acknowledgments

This research is supported in part by the New Chongqing Youth Innovation Talent Projectand under Grant CSTB2025YITP-QCRCX0054, and in part bythe National Natural Science Foundation of China under grant 62372385.

References

[Cai *et al.*, 2022] Derun Cai, Chenxi Sun, Moxian Song, Baofeng Zhang, Shenda Hong, and Hongyan Li. Hypergraph contrastive learning for electronic health records. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 127–135, 2022.

[Chan *et al.*, 2024] Tsai Hor Chan, Guosheng Yin, Kyongtae Bae, and Lequan Yu. Multi-task heterogeneous graph learning on electronic health records. *Neural Networks*, 180:106644, 2024.

[Che and Liu, 2017] Zhengping Che and Yan Liu. Deep learning solutions to computational phenotyping in health care. In *Proceedings of the 2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1100–1109. IEEE, 2017.

[Chen *et al.*, 2025] Jiayuan Chen, Changchang Yin, Yuanlong Wang, and Ping Zhang. Predictive modeling with temporal graphical representation on electronic health records. In *Proceedings of the 2025 International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5763–5771, 2025.

[Cheng *et al.*, 2025] Hengliang Cheng, Shibo Li, Tao Shen, and Weihua Li. Enhancing diagnosis prediction with adaptive disease representation learning. *Artificial Intelligence in Medicine*, 163:103098, 2025.

[Chien *et al.*, 2022] Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. You are allset: A multiset function framework for hypergraph neural networks. In *International Conference on Learning Representations*, 2022.

[Choi *et al.*, 2016] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512, 2016.

[Choi *et al.*, 2017] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. Gram: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 787–795, 2017.

[Feng *et al.*, 2019] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3558–3565, 2019.

[Gao *et al.*, 2020] Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M Glass, and Jimeng Sun. Stagenet: Stage-aware neural networks for health risk prediction. In *Proceedings of The Web Conference 2020 (WWW)*, pages 530–540. ACM, 2020.

[Johnson *et al.*, 2016] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[Johnson *et al.*, 2023] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H Lehman, Leo Anthony Celi, and Roger G Mark. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 2023.

[Kazi *et al.*, 2019] Anees Kazi, Shayan Shekarforoush, S. Arvind Krishna, Hendrik Burwinkel, Gerome Vivar, Karsten Kortüm, Seyed-Ahmad Ahmadi, Shadi Albarqouni, and Nassir Navab. Inceptiongcn: Receptive field aware graph convolutional network for disease prediction. In *International Conference on Information Processing in Medical Imaging (IPMI)*, pages 73–85. Springer, Cham, 2019.

[Kim *et al.*, 2025] Taeri Kim, Jiho Heo, Hyunjoon Kim, and Sang-Wook Kim. Hi-dr: Exploiting health status-aware attention and an ehr graph+ for effective medication recommendation. In *AAAI Conference on Artificial Intelligence*, 2025.

[Liu *et al.*, 2022] Yuxi Liu, Shaowen Qin, Zhenhao Zhang, and Wei Shao. Compound density networks for risk prediction using electronic health records. In *Proceedings of the 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1078–1085. IEEE, 2022.

[Lu *et al.*, 2021] Chang Lu, Chandan K. Reddy, Prithwish Chakraborty, Samantha Kleinberg, and Yue Ning. Collaborative graph learning with auxiliary text for temporal event prediction in healthcare. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3529–3535, 2021.

[Lu *et al.*, 2022] Chang Lu, Tian Han, and Yue Ning. Context-aware health event prediction via transition functions on dynamic disease graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4567–4574, 2022.

[Lu *et al.*, 2023] Chang Lu, Chandan K. Reddy, and Yue Ning. Self-supervised graph learning with hyperbolic embedding for temporal health event prediction. *IEEE Transactions on Cybernetics*, 53(4):2124–2136, 2023.

[Luo *et al.*, 2020] Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. 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*, 2020.

[Ma *et al.*, 2017] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diag-

nosis 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*, 2017.

[Ma *et al.*, 2020] Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiangtao Wang, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 833–840, 2020.

[Pollard *et al.*, 2018] Tom Pollard, Alistair Johnson, Jesse Raffa, Leo A. Celi, Omar Badawi, and Roger Mark. eICU Collaborative Research Database v2.0. PhysioNet, 2018.

[Poulain and Beheshti, 2024] Raphael Poulain and Rahmatollah Beheshti. Graph transformers on ehrs: Better representation improves downstream performance. In *The Twelfth International Conference on Learning Representations*, 2024.

[Shang *et al.*, 2019] Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of graph augmented transformers for medication recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5953–5959, 2019.

[Sun *et al.*, 2022] Ziyou Sun, Xiaohui Yang, Zhiquan Feng, Tao Xu, Xue Fan, and Jinglan Tian. EHR2HG: Modeling of EHRs data based on hypergraphs for disease prediction. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1730–1733. IEEE, 2022.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[Wen *et al.*, 2022] Qianlong Wen, Zhongyu Ouyang, Jianfei Zhang, Yiyue Qian, Yanfang Ye, and Chuxu Zhang. Disentangled dynamic heterogeneous graph learning for opioid overdose prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2009–2019, 2022.

[Wu *et al.*, 2021] Tong Wu, Yunlong Wang, Yue Wang, Emily Zhao, and Yilian Yuan. Leveraging graph-based hierarchical medical entity embedding for healthcare applications. *Scientific Reports*, 11(1), 2021.

[Wu *et al.*, 2023] Jialun Wu, Kai He, Rui Mao, Chen Li, and Erik Cambria. MEGACare: Knowledge-guided multiview hypergraph predictive framework for healthcare. *Information Fusion*, 100:101939, 2023.

[Wu *et al.*, 2025] Jialun Wu, Kai He, Rui Mao, Xuequn Shang, and Erik Cambria. Harnessing the potential of multimodal ehr data: A comprehensive survey of clinical predictive modeling for intelligent healthcare. *Information Fusion*, page 103283, 2025.

[Xing *et al.*, 2024] Yujie Xing, Xiao Wang, Yibo Li, Hai Huang, and Chuan Shi. Less is more: On the over-globalizing problem in graph transformers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[Xu *et al.*, 2023] Ran Xu, Mohammed K Ali, Joyce C Ho, and Carl Yang. Hypergraph transformers for ehr-based clinical predictions. In *AMIA Summits on Translational Science Proceedings 2023*, page 582, 2023.

[Yang *et al.*, 2021] Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. In *International Joint Conference on Artificial Intelligence*, 2021.

[Yu *et al.*, 2025] Leisheng Yu, Yanxiao Cai, Minxing Zhang, and Xia Hu. Self-explaining hypergraph neural networks for diagnosis prediction. In *ACM Conference on Health, Inference, and Learning*, 2025.

[Zhang *et al.*, 2024a] Xin Zhang, Xueping Peng, Weimin Chen, Weiyu Zhang, Xiaoqiang Ren, and Wenpeng Luv. Multi-channel hypergraph network for sequential diagnosis prediction in healthcare. In *Proceedings of the 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 2937–2942, 2024.

[Zhang *et al.*, 2024b] Ziyang Zhang, Hejie Cui, Ran Xu, Yuzhang Xie, Joyce C. Ho, and Carl Yang. Tacco: Task-guided co-clustering of clinical concepts and patient visits for disease subtyping based on ehr data. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6324–6334, 2024.