

HypeMed: Enhancing Medication Recommendations with Hypergraph-Based Patient Relationships

XIANGXU ZHANG, Gaoling School of Artificial Intelligence, Renmin University of China, China

XIAO ZHOU^{*†‡}, Gaoling School of Artificial Intelligence, Renmin University of China, China

HONGTENG XU^{†‡}, Gaoling School of Artificial Intelligence, Renmin University of China, China

JIANXUN LIAN, Microsoft Research Asia, China

Medication recommendation aims to generate safe and effective medication sets from health records. However, accurately recommending medications hinges on inferring a patient’s latent clinical condition from sparse and noisy observations, which requires both (i) preserving the visit-level combinatorial semantics of co-occurring diagnoses/procedures and (ii) leveraging informative historical references through effective, visit-conditioned retrieval. Most existing methods fall short in one of these aspects: graph-based modeling often fragments higher-order intra-visit patterns into pairwise relations, while inter-visit augmentation methods commonly exhibit an imbalance between learning a globally stable representation space and performing dynamic retrieval within it. To address these limitations, this paper proposes HypeMed, a two-stage hypergraph-based framework unifying intra-visit coherence modeling and inter-visit augmentation. HypeMed consists of two components: MedRep for representation pre-training and SimMR for similarity-enhanced recommendation. In the first stage, MedRep encodes clinical visits as hyperedges via knowledge-aware contrastive pre-training, creating a globally consistent, retrieval-friendly embedding space. In the second stage, SimMR performs dynamic retrieval within this space, fusing retrieved references with the patient’s longitudinal data to refine medication prediction. Evaluation on real-world benchmarks shows that HypeMed outperforms state-of-the-art baselines in both recommendation precision and DDI reduction, simultaneously enhancing the effectiveness and safety of clinical decision support. The implementation is publicly available at <https://github.com/xansar/HypeMed>.

CCS Concepts: • **Information systems** → **Data mining**; • **Applied computing** → **Health informatics**.

Additional Key Words and Phrases: Medication Recommendation, Electronic Health Records, Hypergraph

ACM Reference Format:

Xiangxu Zhang, Xiao Zhou, Hongteng Xu, and Jianxun Lian. 2024. HypeMed: Enhancing Medication Recommendations with Hypergraph-Based Patient Relationships. *ACM Trans. Inf. Syst.* XX, X, Article XXX (March 2024), 28 pages. <https://doi.org/XXXXXXXX.XXXXXX>

1 Introduction

Healthcare remains a fundamental societal challenge [2, 3, 26], intensified by rapid urbanization [15, 17, 24, 27, 50, 53] and recurring public health crises [16, 57]. As a key application of Artificial Intelligence (AI) [14, 49, 51, 58] in healthcare [18, 35], medication recommendation systems [7, 37, 43, 47, 54] aim to generate safe and effective medication combinations from patients’ Electronic Health Records (EHRs) [8, 30]. EHRs provide a longitudinal trace of diagnoses,

*Corresponding author.

† Also with Beijing Key Laboratory of Research on Large Models and Intelligent Governance.

‡ Also with Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE.

Authors’ Contact Information: Xiangxu Zhang, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China, xansar@ruc.edu.cn; Xiao Zhou, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China, xiaozhou@ruc.edu.cn; Hongteng Xu, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China, hongtengxu313@gmail.com; Jianxun Lian, Microsoft Research Asia, Beijing, China, jianxun.lian@outlook.com.

2024. Manuscript submitted to ACM

Manuscript submitted to ACM

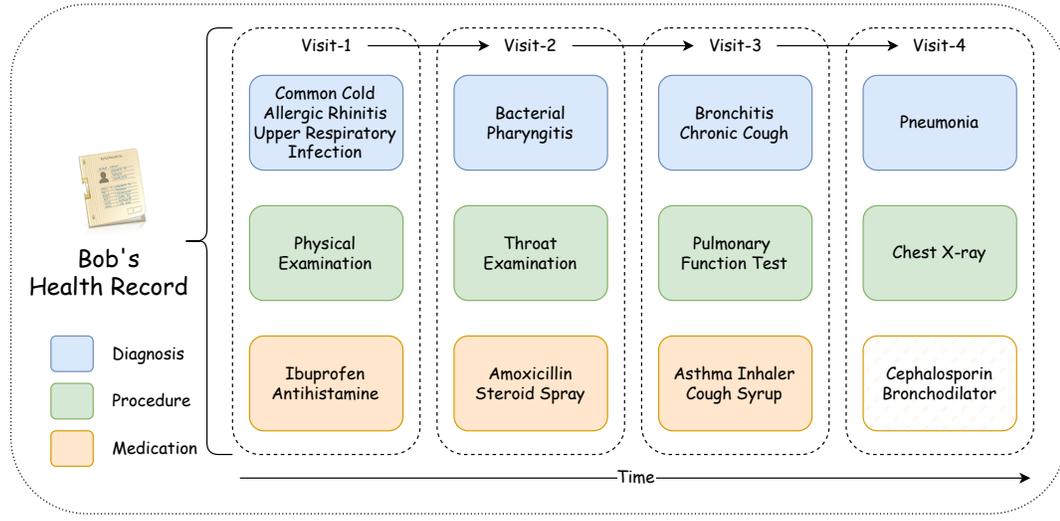


Fig. 1. Overview of Bob's EHR spanning four medical visits. Each visit includes diagnosis, procedure, and medication codes, illustrating the progression of related respiratory conditions over time. Note that the medication section in Visit-4 is left unfilled, indicating it as the target for medication recommendation.

procedures, and medications across visits (Fig. 1). While EHRs offer a wealth of information, the inherent complexity and high dimensionality of such data make it difficult to capture the complete clinical picture.

Despite steady progress [21, 40, 42, 43, 47, 48], the central difficulty remains unchanged: *medication recommendations hinge on accurately inferring a patient's latent clinical condition from sparse, noisy observations*. The observed medical codes provide only indirect, discrete evidence of an underlying physiological state, making condition reconstruction inherently ambiguous. To approximate the true patient status, a model must integrate evidence not only from the current observations but also from their *relationships* in EHRs. In this paper, we use *patient relationships* to refer to two complementary forms of clinical evidence: (i) *intra-visit* set-level co-occurrence among diagnoses/procedures/medications, which reflects coherent syndromic patterns, and (ii) *inter-visit* connections that link the current visit to relevant visits in the patient trajectory, including prior visits of the same patient and clinically similar visits from other patients. However, existing methods often fail to unify these two forms of evidence, yielding a fragmented and superficial estimate of the underlying condition.

From the intra-visit perspective, a dominant bottleneck is *semantic fragmentation*. Most approaches [28, 37, 43, 47] model co-occurrence with simple graphs, breaking inherently high-order clinical patterns into pairwise edges and thus diluting visit-level combinatorial semantics. For instance, Charcot's triad—fever, right-upper-quadrant pain, and jaundice—is individually non-specific, but their *co-occurrence* is highly suggestive of acute cholangitis and warrants urgent care [10, 34]. This motivates modeling a visit as a set-level interaction, rather than an accumulation of pairwise relations. **From the inter-visit perspective**, the key is to *retrieve the right historical references* that can effectively complement the current visit for decision making [4, 5, 38]. Existing methods often exhibit an imbalance between static representation learning and dynamic retrieval. Representation-centric approaches (e.g., DGCL [28], PROMISE [42]) mainly focus on learning a globally stable and static embedding space by aggregating historical/external signals,

yet they typically lack an explicit mechanism to perform visit-conditioned retrieval tailored to the current context. In contrast, retrieval-centric approaches (e.g., GAMENet [37], COGNet [43], VITA [21], DAPSNNet [40]) emphasize on-the-fly retrieval in the embedding space to augment the current input, but often rely on representations that are not explicitly optimized for retrieval, making the retrieved neighbors less semantically aligned with the current clinical context. This imbalance motivates a unified design that jointly shapes a retrieval-friendly representation space and performs context-aware retrieval within it.

To address these limitations, we propose HypeMed, a hypergraph-based framework that reconstructs latent clinical conditions by unifying intra-visit combinatorial interaction modeling with inter-visit reference augmentation, thereby capturing patient relationships both within and across visits. To mitigate *semantic fragmentation*, we model each clinical visit as a hyperedge, so that the representation is learned from the set-level co-occurrence of diagnoses and procedures rather than decomposed pairwise links. Based on this structure, we introduce **Medical Entity Relevance Representation (MedRep)**, a knowledge-aware hypergraph contrastive pre-training stage that encodes high-order interactions into a globally consistent embedding space. On top of the MedRep space, to alleviate the representation–retrieval imbalance, we further propose **Similar Visit Enhanced Medication Recommendation (SimMR)**, which performs visit-conditioned dynamic retrieval directly in this hyperedge-aware metric space. This coupled design makes the retrieved neighbors better aligned with the current visit context, enabling HypeMed to incorporate informative historical references and refine the estimated condition for medication recommendation. Extensive experiments on real-world benchmarks (MIMIC-III/IV and eICU) demonstrate that HypeMed consistently outperforms state-of-the-art baselines in recommendation accuracy while reducing drug–drug interaction (DDI) rates, thereby achieving a superior balance of clinical effectiveness and medication safety. Our main contributions are summarized as follows:

- We address semantic fragmentation in intra-visit modeling by representing each visit as a hyperedge and proposing a hypergraph-based contrastive pre-training module (MedRep) to encode knowledge-aware entity representations that capture combinatorial semantics.
- We alleviate the representation–retrieval imbalance in inter-visit augmentation by designing a similar visit enhanced medication recommendation module (SimMR) that performs visit-conditioned dynamic retrieval directly in the MedRep-optimized embedding space for context-aligned reference aggregation.
- Extensive experiments on public MIMIC-III/IV and eICU datasets demonstrate that HypeMed consistently improves recommendation accuracy while reducing the drug–drug interaction (DDI) rate, validating both effectiveness and safety.

2 Related Work

2.1 Medication Recommendation

Medication recommendation is an important subdomain of recommender systems [12, 13, 25, 29, 56]. Improving medication recommendation from EHRs hinges on two complementary capabilities: modeling *intra-visit* clinical context from the current observations, and leveraging *inter-visit* references from the patient trajectory and clinically similar visits from other patients.

Intra-visit reasoning with relational structures. A representative line of work leverages structured relational modeling—often via graph neural networks (GNNs) [36]—to capture concurrence among diagnoses, procedures, and medications, and to inject domain knowledge such as DDIs for safety-aware prediction [1, 6, 28, 37, 43, 47, 55]. While

effective, these methods predominantly operate on *pairwise* edges, which can fragment the visit-level context when the clinical meaning arises from the *joint* presence of multiple entities.

Inter-visit augmentation via similarity and retrieval. Early studies mainly adopted *instance-based* models that recommend medications from a single visit, without explicitly leveraging inter-visit context [11, 54]. Later, *longitudinal* models incorporated a patient’s historical visits to capture temporal dependencies and improve prediction [6, 7, 22, 28, 39, 43]. More recently, *external* augmentation extends the context beyond an individual patient by retrieving similar visits from the broader cohort, yet existing designs often emphasize only one side of the problem. PROMISE [42] mainly focuses on learning a globally stable, trajectory-level representation (e.g., via DTW-based aggregation), but provides limited visit-conditioned, on-the-fly retrieval for the current context. In contrast, DAPSNet [40] emphasizes dynamic retrieval at prediction time to augment the current visit, but relies on representations that are not explicitly optimized for retrieval, which may weaken the quality of retrieved references. However, these retrieval-based approaches often decouple similarity estimation from representation learning, which can induce a mismatch between the embedding space and the retrieval objective, ultimately limiting the utility of retrieved references.

Our approach. To fill these gaps, HypeMed unifies intra-visit coherence modeling with inter-visit reference augmentation in a single hypergraph framework. MedRep represents each visit as a hyperedge and performs knowledge-aware hypergraph contrastive pre-training to encode high-order interactions into a globally consistent, retrieval-friendly embedding space, mitigating semantic fragmentation. Building upon this space, SimMR performs visit-conditioned dynamic retrieval and integrates the retrieved references with the patient trajectory to refine latent condition estimation and improve medication recommendation.

2.2 Hypergraph Contrastive Learning in Recommendation

Hypergraph contrastive learning [23] has recently attracted increasing attention as an effective strategy to mitigate data sparsity and enhance representation robustness in recommender systems. For instance, in the context of session-based recommendation, DHCN [45] applies contrastive learning to align session representations derived from hypergraphs and their corresponding line graphs. MHCN [52] extends this idea to social recommendation by contrasting local and global node representations, while HCCF [44] integrates collaborative filtering with contrastive learning to alleviate over-smoothing and sparsity issues. Collectively, these methods demonstrate the potential of combining hypergraph structures with contrastive objectives to capture richer relational semantics. However, their applications remain largely confined to general recommendation scenarios. This paper adapts and extends hypergraph contrastive learning to medication recommendation, enabling set-level intra-visit modeling and inter-visit reference augmentation in a unified framework.

3 Preliminaries

This section introduces the notations and definitions used in this paper and outlines the process of constructing the hypergraph. The main symbols are summarized in Tab. 1.

3.1 Notations and Definitions

3.1.1 Definition of EHR Data. The fundamental unit of EHR data is the patient, and each patient’s EHR consists of multiple visits. Let $\mathcal{R} = \{\mathcal{S}^{(i)}\}_{i=1}^N$ represent the EHR data of N patients. Each patient i has a patient trajectory, represented as an ordered sequence of visits $\mathcal{S}^{(i)} = \langle \mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_{|\mathcal{S}^{(i)}|} \rangle$, where each visit $\mathcal{S}_t = \langle \mathbf{d}_t, \mathbf{p}_t, \mathbf{m}_t \rangle$ contains diagnosis, procedure, and medication multi-hot vectors, with $\mathbf{d}_t \in \{0, 1\}^{|\mathcal{D}|}$, $\mathbf{p}_t \in \{0, 1\}^{|\mathcal{P}|}$, and $\mathbf{m}_t \in \{0, 1\}^{|\mathcal{M}|}$. Given the

Table 1. Summary of main notations used in this paper.

| Notation | Definition |
|--|---|
| \mathcal{R} | EHR database containing all patients' visit sequences |
| $\mathcal{S}_t^{(i)}$ | The t -th visit of patient i (patient index i is omitted when clear) |
| $\mathcal{D}, \mathcal{P}, \mathcal{M}$ | Sets of all diagnosis, procedure, and medication codes |
| $\mathbf{d}, \mathbf{p}, \mathbf{m}$ | Multi-hot vectors for diagnoses, procedures, and medications |
| $\mathbf{v}^{\mathcal{D}}, \mathbf{v}^{\mathcal{P}},$ and $\mathbf{v}^{\mathcal{M}}$ | Visit representations derived from diagnoses, procedures, and medications |
| $\mathcal{H}_X = (\mathcal{N}_X, \mathcal{E}_X)$ | Hypergraph for domain $X \in \{\mathcal{D}, \mathcal{P}, \mathcal{M}\}$ |
| $\mathbf{Z}^X, \mathbf{U}^X$ | Node and hyperedge embeddings learned from \mathcal{H}_X |
| \mathbf{h}_t | Health status representation of visit t (derived from diagnoses and procedures) |
| $\mathbf{v}_{hist}, \mathbf{v}_{sim}$ | Representations from the historical-visit and similar-visit channels |
| \mathbf{v}_t | Final fused visit representation |
| \mathbf{y}_t | Predicted probability vector of medications for visit t |
| \mathbf{A} | Drug–drug interaction (DDI) adjacency matrix |

current diagnosis \mathbf{d}_t , procedure \mathbf{p}_t , and the EHR history $\langle \mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_{t-1} \rangle$, HypeMed aims to predict the medication set \mathbf{m}_t for the current visit. Patient indices (i) and domain subscripts are omitted when clear from context.

3.1.2 Definition of Hypergraph. A hypergraph is denoted as $\mathcal{H} = (\mathcal{N}, \mathcal{E})$, where \mathcal{N} and \mathcal{E} are the sets of nodes and hyperedges, respectively. Fig. 2 illustrates an example hypergraph with four nodes and two hyperedges, where $\mathcal{N} = \{n_i\}_{i=1}^4$ and $\mathcal{E} = \{e_i\}_{i=1}^2$. Each hyperedge can connect more than two nodes; for example, e_1 connects n_1, n_2 , and n_3 . $\mathcal{N}(n_i)$ denotes the set of hyperedges incident to node n_i , and $\mathcal{N}(e_i)$ represents the set of nodes contained in hyperedge e_i .

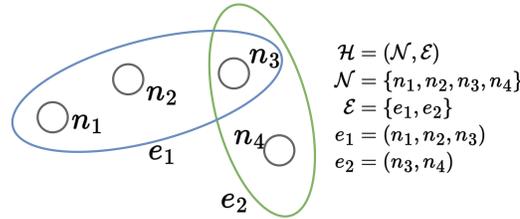


Fig. 2. An example of a hypergraph with four nodes and two hyperedges: e_1 connects nodes n_1, n_2 , and n_3 , while e_2 connects nodes n_3 and n_4 .

3.2 EHR Hypergraph Construction

When constructing the EHR hypergraph, we do not merge all medical entities into a single heterogeneous hypergraph. Instead, Alg. 1 outlines the process for building separate hypergraphs for each domain: diagnosis, procedure, and medication. This design is motivated by three factors: context specificity, complexity management, and scalability, as detailed below.

Algorithm 1 Hypergraph Construction for Diagnosis, Procedure, and Medication

```

1: function CONSTRUCTHYPERGRAPHS( $\{\mathcal{H}_X\}_{X \in \{\mathcal{D}, \mathcal{P}, \mathcal{M}\}}, \mathcal{R}$ )
2:    $N_{\mathcal{D}}, N_{\mathcal{P}}, N_{\mathcal{M}} \leftarrow \emptyset$  ▷ Initialize node sets
3:    $E_{\mathcal{D}}, E_{\mathcal{P}}, E_{\mathcal{M}} \leftarrow \emptyset$  ▷ Initialize hyperedge sets
4:   ADDENTITIESTOHYPERGRAPH( $\mathcal{D}, N_{\mathcal{D}}, E_{\mathcal{D}}, \mathcal{R}$ )
5:   ADDENTITIESTOHYPERGRAPH( $\mathcal{P}, N_{\mathcal{P}}, E_{\mathcal{P}}, \mathcal{R}$ )
6:   ADDENTITIESTOHYPERGRAPH( $\mathcal{M}, N_{\mathcal{M}}, E_{\mathcal{M}}, \mathcal{R}$ )
7:   return  $(N_{\mathcal{D}}, E_{\mathcal{D}}), (N_{\mathcal{P}}, E_{\mathcal{P}}), (N_{\mathcal{M}}, E_{\mathcal{M}})$ 
8: end function

9: procedure ADDENTITIESTOHYPERGRAPH( $\mathcal{X}, N_{\mathcal{X}}, E_{\mathcal{X}}, \mathcal{R}$ )
10:  for all entities  $x \in \mathcal{X}$  do
11:    Add  $x$  to  $N_{\mathcal{X}}$ 
12:  end for
13:  for all visits  $\mathcal{S}_t^{(i)} \in \mathcal{R}$  do
14:     $e^{\mathcal{X}} \leftarrow$  hyperedge for entities in  $\mathcal{S}_t^{(i)}$  of type  $\mathcal{X}$ 
15:    Add  $e^{\mathcal{X}}$  to  $E_{\mathcal{X}}$ 
16:  end for
17: end procedure

```

- **Context Specificity:** Diagnoses, procedures, and medications each carry distinct medical information and contextual meanings. By constructing separate hypergraphs, we can capture the unique contextual relationships inherent to each type of medical entity more precisely.
- **Complexity Management:** A single visit may contain as many as hundreds of medical entities. Incorporating all these entities into a single heterogeneous hyperedge could lead to excessively high complexity and introduce additional noise, making it more difficult to capture the relationships among nodes. Constructing separate hypergraphs reduces system complexity, making the model easier to train.
- **Flexibility and Scalability:** The structure of independently constructed hypergraphs offers greater flexibility for future model iterations. This design also facilitates extension to new entity types or updated datasets.

Taking the diagnosis hypergraph $\mathcal{H}_{\mathcal{D}} = (N_{\mathcal{D}}, E_{\mathcal{D}})$ as an example, it is constructed in two steps:

- **Node Set Construction:** Create nodes for all diagnostic entities in \mathcal{D} and add them to the hypergraph node set $N_{\mathcal{D}}$.
- **Hyperedge Set Construction:** Traverse all visits in the EHR training data. For each diagnostic vector $\mathbf{d}_t^{(i)}$ in visit $\mathcal{S}_t^{(i)}$, create a hyperedge e connecting all diagnostic entities contained in that visit, and add e to the hypergraph hyperedge set $E_{\mathcal{D}}$.

The construction procedures for $\mathcal{H}_{\mathcal{P}}$ and $\mathcal{H}_{\mathcal{M}}$ follow the same logic.

4 Framework

This section outlines the HypeMed framework, as illustrated in Fig. 3. The architecture primarily comprises the Medical Entity Relevance Representation Stage (MedRep) and the Similar Visit Enhanced Medication Recommendation Stage (SimMR). We also explored an end-to-end variant in early experiments, which resulted in an approximately 3% decrease in Jaccard score and lower training efficiency. Therefore, in this paper we adopt a two-stage training pipeline.

We adopt a two-stage training paradigm, consisting of pre-training followed by fine-tuning. First, we pre-train MedRep on the clinical hypergraph with a contrastive objective to learn medical *entity* (node) and *visit* (hyperedge) representations. At the start of SimMR, we instantiate two *trainable* embedding tables for entities and visit hyperedges, initialized with the pre-trained MedRep outputs. During fine-tuning, these embedding tables and all recommender parameters are optimized end-to-end under the recommendation loss. This preserves the semantic structure learned in pre-training while allowing task-specific adaptation at the embedding level, yielding stable optimization and improved downstream performance. The overall training pipeline is illustrated in Alg. 2.

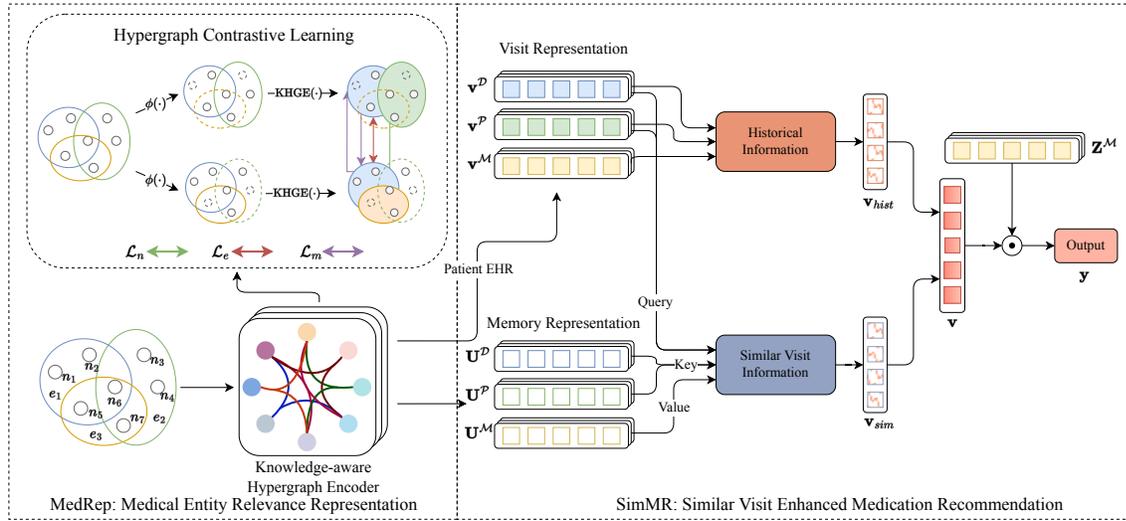


Fig. 3. Overall architecture of HypeMed. HypeMed comprises two stages: the Medical Entity Relevance Representation Stage (MedRep) and the Similar Visit Enhanced Medication Recommendation Stage (SimMR). MedRep focuses on encoding intra-visit set-level combinatorial semantics into a globally consistent, retrieval-friendly embedding space. SimMR integrates longitudinal history and visit-conditioned retrieved similar visits to refine latent condition estimation.

4.1 MedRep: Medical Entity Relevance Representation

In the MedRep stage, we obtain node and hyperedge representations through hypergraph-based contrastive learning. We first generate two augmented subgraphs of the original hypergraph and then apply a knowledge-aware hypergraph encoder (KHGE) to compute node and hyperedge embeddings for both views. Finally, we perform contrastive learning between these two views to enhance embedding quality. By modeling each visit as a hyperedge that jointly connects diagnoses, procedures, and medications, MedRep mitigates the intra-visit semantic fragmentation (Introduction) and yields context-aware entity and visit embeddings. Below we first introduce the encoder, followed by the augmentation and contrastive objectives.

4.1.1 Knowledge-aware Hypergraph Encoder. We design a knowledge-aware hypergraph encoder (Fig. 4) to capture contextual information and incorporate prior medical category knowledge. This encoder consists of two complementary modules: (1) a *Local Message Passing Network (LMPN)* that models neighborhood-level dependencies on the medical hypergraph, and (2) a *Knowledge-aware Global Attention Network (KGAN)* that incorporates structured medical knowledge

Algorithm 2 Two-Stage Training Pipeline for HypeMed

```

1: function TRAINMODEL( $\mathcal{H}, \mathcal{D}, \mathcal{P}, \mathcal{M}$ )
2:   Initialize KHGE parameters ▷ Stage 1: MedRep training
3:   Construct augmented subgraphs  $\mathcal{H}_1, \mathcal{H}_2$  by dropout on  $\mathcal{H}$ 
4:   for  $epoch = 1$  to  $e_{rep}$  do
5:      $(\mathbf{Z}_1, \mathbf{U}_1) \leftarrow \text{ENCODEHYPERGRAPH}(\mathcal{H}_1)$  ▷ Encode augmented hypergraph via the hypergraph encoder
6:      $(\mathbf{Z}_2, \mathbf{U}_2) \leftarrow \text{ENCODEHYPERGRAPH}(\mathcal{H}_2)$  ▷ Encode augmented hypergraph via the hypergraph encoder
7:     Optimize  $\mathbf{Z}_1, \mathbf{U}_1, \mathbf{Z}_2, \mathbf{U}_2$  using contrastive loss  $\mathcal{L}_{cl}$  ▷ Eq. (5)
8:   end for
9:   return  $\mathbf{Z}, \mathbf{U}$  ▷ Final entity and hyperedge embeddings

10: Initialize SimMR with  $\mathbf{Z}, \mathbf{U}$  ▷ Stage 2: SimMR training (embeddings initialized from MedRep and trainable)
11: for  $epoch = 1$  to  $e_{rec}$  do
12:   for all visits  $\mathcal{S}_t^{(i)}$  do
13:     Retrieve  $k$  similar visits for  $\mathbf{v}_{sim}$ 
14:      $\mathbf{v}_{hist} \leftarrow \text{COMPUTE HISTORICAL REPRESENTATION}(\mathcal{S}_t^{(i)})$  ▷ Aggregate past visits via temporal attention
15:      $\mathbf{v}_{sim} \leftarrow \text{COMPUTE SIMILAR VISIT REPRESENTATION}(\mathcal{S}_t^{(i)})$  ▷ Encode retrieved similar visits (Sec. 4.2.3)
16:     Fuse  $\mathbf{v}_{hist}$  and  $\mathbf{v}_{sim}$  to obtain  $\mathbf{v}_t$ 
17:     Optimize with total loss  $\mathcal{L}$ 
18:   end for
19: end for
20: return trained SimMR model
21: end function

```

(e.g., ICD/ATC hierarchies) to capture global correlations. The outputs of both components are fused via a two-layer feed-forward network with residual connections and layer normalization. The detailed formulations for each module are presented as follows.

(a) Local Message Passing Network (LMPN). Research on graph neural networks [36] and hypergraph neural networks [9] suggests that node representations derived from local message passing effectively model neighborhood relationships, making nodes connected by edges more similar. Thus, applying local message passing on a medical hypergraph enhances the similarity of medical entity representations within the same visit. The LMPN is implemented via a hypergraph attention layer followed by layer normalization and residual connections, as illustrated in Fig. 4. Local message passing involves two processes: hyperedge representation computation and node representation computation:

$$\begin{aligned}
\mathbf{u}_j^{(k+1)} &= \sum_{n_i \in \mathcal{N}(e_j)} \alpha_{ij}^{(k)} \mathbf{z}_i^{(k)}, \\
\mathbf{z}_i^{(k+1)} &= \sum_{e_j \in \mathcal{N}(n_i)} \alpha_{ij}^{(k)} \mathbf{u}_j^{(k)},
\end{aligned} \tag{1}$$

where the attention coefficient $\alpha_{ij}^{(k)}$ controls the message weight between node n_i and hyperedge e_j . Before aggregation, a learnable linear transformation is applied to refine both node and hyperedge embeddings.

(b) Knowledge-aware Global Attention Network (KGAN). To capture long-range dependencies and inject domain knowledge, we introduce the Knowledge-aware Global Attention Network (KGAN). Medical knowledge from ICD or

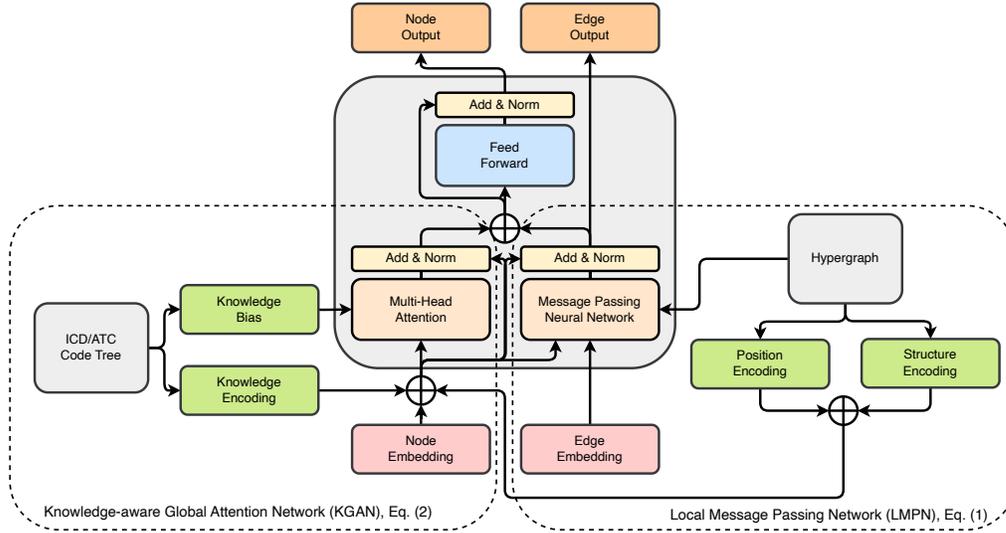


Fig. 4. The detailed architecture of the Knowledge-aware Hypergraph Encoder (KHGE). The encoder consists of a Local Message Passing Network (LMPN) and a Knowledge-aware Global Attention Network (KGAN). The LMPN focuses on message propagation over the medical hypergraph, while the KGAN integrates medical knowledge into a global self-attention mechanism. The two outputs are combined through a feed-forward fusion layer with residual connections.

ATC code hierarchies provides a structured prior indicating semantic closeness among entities. We first transform the hierarchical tree into a knowledge bias matrix Ω , where each element encodes the path distance between two entities and is mapped to a learnable bias parameter. This bias is then incorporated into the multi-head self-attention mechanism as:

$$\mathbf{Z}_g^{(k+1)} = \text{Softmax}\left(\frac{\mathbf{W}_q \mathbf{Z}^{(k)} (\mathbf{W}_k \mathbf{Z}^{(k)})^T + \Omega}{\sqrt{d}}\right) \mathbf{W}_v \mathbf{Z}^{(k)}. \quad (2)$$

Here, \mathbf{W}_q , \mathbf{W}_k , and \mathbf{W}_v are learnable projection matrices, and Ω provides a prior bias that allows the attention to favor medically related entities.

(c) Fusion and Multi-layer Aggregation. We integrate the node representations from LMPN and KGAN through a two-layer feed-forward network (FFN) with residual connections and layer normalization, where $\mathbf{Z}_\ell^{(k+1)}$ and $\mathbf{Z}_g^{(k+1)}$ denote the outputs of LMPN and KGAN, respectively. To mitigate over-smoothing, we perform a layer-wise average of intermediate representations:

$$\begin{aligned} \mathbf{Z}^{(k+1)} &= \text{FFN}\left(\mathbf{Z}_\ell^{(k+1)} + \mathbf{Z}_g^{(k+1)}\right), \\ \mathbf{Z} &= \frac{1}{L} \sum_{k=1}^L \mathbf{Z}^{(k)}, \quad \mathbf{U} = \frac{1}{L} \sum_{k=1}^L \mathbf{U}^{(k)}. \end{aligned} \quad (3)$$

This fusion mechanism effectively combines local relational learning with global knowledge reasoning.

4.1.2 Hypergraph Contrastive Learning. Existing research [23] indicates that hypergraph contrastive learning can enhance the representational capabilities of embeddings. Hypergraph contrastive learning brings the representations of similar nodes and hyperedges closer together while distancing those of dissimilar nodes and hyperedges. This aligns with our expectations for medical entity representation. Hence, we employ a hypergraph contrastive learning approach

based on nodes and hyperedges for the training phase of MedRep. Initially, we construct two augmented subgraphs based on the original hypergraph. Subsequently, we compute node and hyperedge representations based on these two subgraphs. Finally, contrastive learning is executed between these two sets of representations.

Augmented Subgraphs Construction. EHR data often contains a considerable amount of noise, which can negatively impact the learning of medical entity representations. To address this issue, constructing augmented subgraphs through random dropout of elements in the original visits has become crucial for mitigating the effects of noise [23, 41]. We independently perform two stochastic augmentations on the original hypergraph to obtain subgraphs \mathcal{H}_1 and \mathcal{H}_2 , by randomly dropping nodes, incidences, and features with predefined probabilities. Dropped elements are excluded when computing the contrastive objectives.

Contrastive Learning. After obtaining two augmented subgraphs \mathcal{H}_1 and \mathcal{H}_2 , we use the knowledge-aware hypergraph encoder to generate node and hyperedge embeddings for both views:

$$\begin{aligned} \mathbf{Z}_1, \mathbf{U}_1 &= \text{KHGE}(\mathcal{H}_1), \\ \mathbf{Z}_2, \mathbf{U}_2 &= \text{KHGE}(\mathcal{H}_2). \end{aligned} \quad (4)$$

To enhance the consistency of multi-level representations, we adopt a three-level contrastive learning strategy encompassing node-, hyperedge-, and membership-level objectives (as illustrated in the upper-left of Fig. 3). This design ensures that (1) nodes sharing similar contexts are mapped closely in the embedding space, (2) hyperedges containing related nodes exhibit similar representations, and (3) membership contrast bridges the alignment between node and hyperedge embeddings. The overall contrastive objective is formulated as:

$$\mathcal{L}_{cl} = \text{InfoNCE}(\mathbf{Z}_1, \mathbf{Z}_2) + \lambda_1 \text{InfoNCE}(\mathbf{U}_1, \mathbf{U}_2) + \lambda_2 \text{InfoNCE}(\mathbf{Z}_1, \mathbf{U}_2), \quad (5)$$

where λ_1 and λ_2 balance the contributions of hyperedge-level and membership-level contrastive terms. The InfoNCE loss [31] is defined as:

$$\text{InfoNCE}(\mathbf{U}, \mathbf{V}) = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\mathbf{u}_i \cdot \mathbf{v}_i / \tau)}{\sum_{j=1}^N \exp(\mathbf{u}_i \cdot \mathbf{v}_j / \tau)}, \quad (6)$$

where τ is the temperature parameter controlling the smoothness of the contrastive distribution.

During the MedRep stage, we optimize \mathcal{L}_{cl} to jointly learn node embeddings \mathbf{Z} (medical entities) and hyperedge embeddings \mathbf{U} (visits), thereby establishing a unified and semantically consistent representation space.

Upon completing MedRep, we export the learned embedding matrices for medical entities and visit embeddings (factorized into diagnosis, procedure, and medication subspaces). These matrices are then used in SimMR to *initialize* two *trainable* embedding tables (entities and visits), which are updated jointly with the recommender parameters. Specifically, the task-specific representations are obtained as:

$$\mathbf{Z}_{\mathcal{X}}, \mathbf{U}_{\mathcal{X}} = \text{KHGE}(\mathcal{H}_{\mathcal{X}}), \quad \mathcal{X} \in \{\mathcal{D}, \mathcal{P}, \mathcal{M}\}. \quad (7)$$

In this way, the pre-trained embeddings serve as a high-quality initialization for the recommendation task, enabling the model to leverage rich medical correlations captured during the contrastive learning phase.

4.2 SimMR: Similar Visit Enhanced Medication Recommendation

In the second stage, we train SimMR based on the representations obtained from MedRep, specifically optimizing for the medication recommendation task. In clinical practice, similar visits are often consulted as external evidence [4, 5, 38]. Accordingly, SimMR refines latent condition estimation for the current visit by augmenting it with two types of

inter-visit references: (i) information from the patient trajectory (temporal causality) and (ii) visit-conditioned retrieved similar visits (analogical evidence).

We first compute the representation of the current visit and then feed it into two channels to obtain complementary visit representations: a *historical* channel and a *similar-visit* channel. Finally, we fuse these two representations and apply a dot-product scorer to produce the medication probabilities. The pipeline is shown in the right panel of Fig. 3.

4.2.1 Visit Representation. A single visit may involve multiple domains of medical entities (diagnoses, procedures, medications) that exhibit semantic gaps. We therefore compute domain-specific visit representations for each visit.

We first mean-pool the entity embeddings in a given domain to form an initial visit representation (\mathbf{z}_t^X), then refine it via attention between it and the corresponding entity embeddings (like hyperedge computation in MedRep). This yields the domain-specific visit representation and facilitates subsequent similar-visit retrieval.

Concretely, for domain $X \in \{\mathcal{D}, \mathcal{P}, \mathcal{M}\}$, we compute

$$\mathbf{v}_t^X = \text{MHA}\left(\mathbf{z}_t^X, \mathbf{Z}^X, \mathbf{Z}^X\right), \quad (8)$$

where $\text{MHA}(\cdot)$ denotes multi-head attention and \mathbf{z}_t^X is the mean-pooled embedding of the entities in domain X for the t -th visit. We obtain $\mathbf{v}_t^{\mathcal{D}}$, $\mathbf{v}_t^{\mathcal{P}}$, and $\mathbf{v}_t^{\mathcal{M}}$ accordingly, and denote the medication label as $\mathbf{m}_t \in \{0, 1\}^{|\mathcal{M}|}$.

4.2.2 Historical Information Representation. The success of longitudinal medication recommendation methods [37, 43, 47] underscores the significance of historical information in the task of medication recommendation. This subsection describes how we compute visit representations using historical information.

Two types of information are particularly crucial in a patient trajectory. The first is the patient’s health status information, which includes diagnostic and procedure conditions. These visits directly reflect the patient’s health condition. The second type is medication information. Existing studies [43, 46] indicate that patients exhibit continuity in medication usage. Medications previously administered are likely to be used again in the future. Hence, modeling medication information can reflect the patient’s historical trends in medication usage.

However, considering all historical visits in the patient trajectory is impractical, as not all historical visits are relevant to the current one. Visits from the distant past may not provide helpful references for current medication. Therefore, we employ a window-based visit attention mechanism over the patient trajectory. The following formula can represent this process:

$$\mathbf{v}_{\text{hist}} = \text{MHA}(\mathbf{h}_t, \mathbf{V}_{t-k:t-1}, \mathbf{V}_{t-k:t-1}), \quad (9)$$

where \mathbf{h}_t is the current health-status embedding (from diagnoses and procedures), and $\mathbf{V}_{t-k:t-1} \in \mathbb{R}^{k \times d}$ stacks the k most recent visit embeddings. The resulting \mathbf{v}_{hist} captures recent diagnostic, procedural, and medication trends within the temporal window.

4.2.3 Similar Visit Information Representation. In the MedRep stage, we obtain embeddings for all training visits (hyperedges) and factorize them into domain-specific subspaces. Given \mathbf{h}_t , we retrieve the top- k similar visits in the health-status subspace to ensure consistency with the pre-trained geometry. This design is clinically motivated: we retrieve similar past visits as external evidence, and condition the retrieval on the current visit’s health-status embedding \mathbf{h}_t to ensure relevance. Let \mathbf{U}^H and \mathbf{U}^M denote the hyperedge embeddings in the health-status (diagnosis/procedure)

and medication subspaces, respectively. We aggregate the retrieved evidence via attention:

$$\mathbf{v}_{\text{sim}} = \text{MHA}\left(\mathbf{h}_t, \mathbf{U}_{\text{top-}k}^H, \mathbf{U}_{\text{top-}k}^M\right), \quad (10)$$

where $\mathbf{U}_{\text{top-}k}^H$ and $\mathbf{U}_{\text{top-}k}^M$ are from the retrieved k nearest visits in the corresponding subspaces.

To further improve retrieval alignment, we adopt an in-batch contrastive objective that aligns the current visit’s embeddings with those of the retrieved visits in the same representation spaces:

$$\mathcal{L}_{cl}^{\text{sim}} = \text{InfoNCE}\left(\mathbf{h}_t, \mathbf{U}^H\right) + \text{InfoNCE}\left(\mathbf{v}_t^M, \mathbf{U}^M\right), \quad (11)$$

where both terms use temperature-scaled inner-product similarity, consistent with the pre-trained geometry, and help mitigate representation–retrieval mismatch during fine-tuning by maintaining geometric consistency between the current visit and the retrieved visits.

4.2.4 Channels Fusion. We perform representation fusion after obtaining the visit representations from the two channels (historical information and similar visit information). We stack the two channel representations for the same visit and compute their weights using a two-layer MLP followed by a Softmax layer. This process is formulated as follows:

$$\mathbf{v}_t = \alpha_{\text{hist}} \mathbf{v}_{\text{hist}} + \alpha_{\text{sim}} \mathbf{v}_{\text{sim}}, \quad [\alpha_{\text{hist}}, \alpha_{\text{sim}}] = \text{Softmax}(\text{MLP}([\mathbf{v}_{\text{hist}} \parallel \mathbf{v}_{\text{sim}}])). \quad (12)$$

To capture information from different perspectives, we apply regularization constraints to them. We encourage the two representations to be decorrelated (approximately orthogonal). Concretely, we minimize their cosine similarity:

$$\mathcal{L}_{\text{orth}} = \left| \frac{\mathbf{v}_{\text{hist}} \cdot \mathbf{v}_{\text{sim}}}{\|\mathbf{v}_{\text{hist}}\| \|\mathbf{v}_{\text{sim}}\|} \right|. \quad (13)$$

4.2.5 Prediction and Objectives. We calculate the probability distribution for recommending medications by assessing the dot-product similarity between the representations of patient visits and medications. This can be expressed as follows:

$$\mathbf{y}_t = \sigma\left(\mathbf{v}_t^\top \mathbf{Z}^M\right), \quad (14)$$

where $\sigma(\cdot)$ is the logistic function applied element-wise. Given a threshold η , the predicted medication set is

$$\hat{m}_t = \{i \mid (\mathbf{y}_t)_i \geq \eta\}. \quad (15)$$

Here, $(\mathbf{y}_t)_i$ denotes the i^{th} element of \mathbf{y}_t . That is, \hat{m}_t is the set of indices whose predicted probabilities exceed η .

We adopt the binary cross-entropy loss, multi-label loss, and DDI loss used in SafeDrug [47], where the binary cross-entropy loss and multi-label loss are related to prediction accuracy, and the DDI loss pertains to the DDI rate. These loss functions are computed as follows:

- **Binary Cross Entropy Loss** is a typical loss function used in machine learning, particularly effective in multi-label classification settings. It calculates the loss by measuring the distance between the model’s predicted probabilities and the actual binary labels, penalizing predictions based on the divergence from the true labels. This makes it highly suitable for scenarios where each instance can belong to multiple classes independently.

$$\mathcal{L}_{bce} = - \sum_{i=1}^{|\mathcal{M}|} \left[\mathbf{m}_{t,i} \log y_{t,i} + (1 - \mathbf{m}_{t,i}) \log(1 - y_{t,i}) \right]. \quad (16)$$

- **Multi-label Loss** promotes model predictions where the presence of one label significantly exceeds the probability of an absent label by a specified margin, effectively learning both label occurrence and their relationships in

multi-label classification tasks.

$$\mathcal{L}_{multi} = \frac{1}{|\mathcal{M}|} \sum_{i,j:\mathbf{m}_t,i=1,\mathbf{m}_t,j=0} \max(0, 1 - (y_{t,i} - y_{t,j})). \quad (17)$$

- **DDI Loss** is designed to minimize the risk of adverse drug-drug interactions in predictive models. It penalizes predicted probabilities assigned to medication pairs with known DDIs, thus reducing the likelihood of recommending harmful medication combinations in clinical settings.

$$\mathcal{L}_{ddi} = \sum_{i,j} \mathbf{A}_{ij} y_{t,i} y_{t,j}. \quad (18)$$

Here, $|\mathcal{M}|$ is the number of medications, \mathbf{m}_t is the ground-truth multi-hot label, \mathbf{y}_t is the predicted probability vector, and $\mathbf{A} \in \mathbb{R}^{|\mathcal{M}| \times |\mathcal{M}|}$ is the DDI adjacency matrix. The total loss function can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{bce} + \lambda_{multi} \mathcal{L}_{multi} + \lambda_{ddi} \mathcal{L}_{ddi} + \lambda_{aux} (\mathcal{L}_{cl}^{sim} + \mathcal{L}_{orth}), \quad (19)$$

where λ_{multi} , λ_{ddi} , and λ_{aux} are balancing hyperparameters.

5 Experiments

To evaluate the performance of HypeMed against state-of-the-art (SOTA) models, we conduct extensive experiments on three widely used and publicly available benchmark datasets: MIMIC-III [20], MIMIC-IV [19], and eICU [32].

5.1 Experimental Settings

5.1.1 Datasets. MIMIC-III¹ and MIMIC-IV² are large-scale de-identified EHR databases from the Beth Israel Deaconess Medical Center, comprising data from over 40,000 ICU patients across diverse clinical conditions. **eICU**³ is a multi-center critical care database covering more than 200 hospitals across the United States and is used to assess model generalizability under cross-institutional settings.

Following prior work [6, 28, 43, 48], we adopt the data preprocessing pipeline from SafeDrug [47], which includes filtering low-frequency entities, removing patients with fewer than two visits, and converting codes into standardized ICD and ATC formats. For eICU, medications are recorded using GTC codes. Since procedure codes are unavailable, we use only diagnosis and medication information for this dataset. Moreover, because eICU lacks a DDI graph, we exclude DDI-related losses and metrics during both training and evaluation.

The statistics of the processed datasets are summarized in Tab. 2. Following common practice in contrastive representation learning (e.g., TriCL [23] and SGL [41]), we construct two independently augmented views of each hypergraph to form positive pairs for contrastive supervision. Empirically, this setup achieves an optimal balance between representation diversity and computational efficiency.

5.1.2 Evaluation. To comprehensively evaluate model performance, we employ five standard metrics: Jaccard Similarity \uparrow (%), F1-score \uparrow (%), and Precision-Recall AUC (PRAUC) \uparrow (%) for accuracy; DDI Rate \downarrow (%) for safety; and the average number of prescribed medications (# Med.) \downarrow for prescription compactness. The ground-truth DDI rates are 8.68% and 7.24% for the MIMIC-III and MIMIC-IV datasets, respectively.⁴ Given that the DDI ground truth is non-zero, our

¹<https://physionet.org/content/mimiciii/1.4/>

²<https://physionet.org/content/mimiciv/2.0/>

³<https://physionet.org/content/eicu-crd/2.0/>

⁴Due to discrepancies across prior reports [28, 39, 43], we recalculated the DDI ground truth following the standard procedure and provide reference code for reproducibility.

Table 2. Statistics of processed MIMIC-III/MIMIC-IV and eICU.

| Item | MIMIC-III | MIMIC-IV | eICU |
|---------------------------|-----------|----------|--------|
| # of patients | 6,350 | 9,036 | 7,855 |
| # of visits | 15,032 | 20,616 | 16,869 |
| avg. # of visits | 2.37 | 2.28 | 2.15 |
| # of unique diag. codes | 1,958 | 1,892 | 692 |
| # of unique proc. codes | 1,430 | 4,939 | - |
| # of unique med. codes | 131 | 131 | 46 |
| avg. # of diag. per visit | 10.51 | 13.62 | 4.32 |
| avg. # of proc. per visit | 3.84 | 3.55 | - |
| avg. # of med. per visit | 11.44 | 10.29 | 10.59 |

objective is not to eliminate DDIs entirely, but to ensure that the model’s recommended medication combinations exhibit a DDI rate below the empirical human benchmark. Symbols \uparrow and \downarrow respectively indicate *higher is better* and *lower is better*. To ensure a fair and consistent comparison, we adopt the conventional evaluation metrics used in the domain of medication recommendation [37, 43, 47]. The calculation methods for each metric are as follows:

- **Jaccard Similarity** between the predicted and ground-truth medication sets for an individual patient is computed as:

$$\text{Jaccard} = \frac{1}{T} \sum_{t=1}^T \frac{|m_t \cap \hat{m}_t|}{|m_t \cup \hat{m}_t|}, \quad (20)$$

where m_t denotes the ground-truth medication set and \hat{m}_t the predicted set at visit t .

- **F1-score** is defined as:

$$\text{F1} = \frac{1}{T} \sum_{t=1}^T \frac{2P_t R_t}{P_t + R_t}, \quad P_t = \frac{|m_t \cap \hat{m}_t|}{|\hat{m}_t|}, \quad R_t = \frac{|m_t \cap \hat{m}_t|}{|m_t|}. \quad (21)$$

Here, P_t and R_t denote the precision and recall for visit t , respectively, and T is the total number of visits.

- **PRAUC** represents the area under the precision–recall curve. Following prior work [43, 47], we compute it as the mean of per-visit average precision scores:

$$\text{PRAUC} = \frac{1}{T} \sum_{t=1}^T \text{AP}_t, \quad (22)$$

where AP_t is the average precision of visit t .

- **DDI Rate** measures the proportion of predicted medication pairs that have known drug–drug interactions:

$$\text{DDI} = \frac{1}{T} \sum_{t=1}^T \frac{\sum_{i=1}^{|\hat{m}_t|} \sum_{j=i+1}^{|\hat{m}_t|} \mathbb{1}\{\mathbf{A}_{\hat{m}_t^{(i)}, \hat{m}_t^{(j)}} = 1\}}{\sum_{i=1}^{|\hat{m}_t|} \sum_{j=i+1}^{|\hat{m}_t|} 1}, \quad (23)$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function and \mathbf{A} is the adjacency matrix of the DDI graph.

- **# Med.** quantifies the average number of medications prescribed per visit:

$$\#\text{Med.} = \frac{1}{T} \sum_{t=1}^T |\hat{m}_t|. \quad (24)$$

First, we calculate these metrics for each patient, and then we compute the average value across all patients to determine the final performance of the method.

Following prior work [6, 43, 47, 48], we evaluate all models using a bootstrapping protocol rather than cross-validation. Each model is trained on a fixed training set with hyperparameters tuned on a validation set. During evaluation, we perform ten rounds of bootstrapped testing by randomly sampling 80% of the test set with replacement, reporting the mean and standard deviation across rounds.

5.1.3 Baselines. To demonstrate the effectiveness of HypeMed, we compare it against representative medication recommendation methods, grouped into three categories: *machine-learning*, *instance-based*, and *longitudinal* approaches.

- **Machine Learning Methods.**

- **LR** is a logistic regression-based baseline that performs independent binary classification for each medication label.
- **ECC** [33] (Ensembled Classifier Chain) encodes diagnosis and procedure sets into multi-hot vectors and applies a chain of SVM classifiers to perform multi-label prediction. The chaining mechanism enables ECC to capture label dependencies and improve overall prediction accuracy.

- **Instance-based Methods.**

- **LEAP** [54] formulates medication recommendation as a sequential decision-making process. It employs a recurrent decoder with content-based attention to model label dependencies across visits.

- **Longitudinal Methods.**

- **RETAIN** [7] is an interpretable predictive framework that processes EHR data using a reverse-time attention mechanism, enhancing both prediction accuracy and clinical interpretability.
- **GAMENet** [37] integrates a DDI knowledge graph with a memory-augmented GCN to jointly model patient history and medication safety.
- **SafeDrug** [47] recommends effective and safe medication combinations by jointly modeling global and local molecular structures of medications. It also incorporates a controllable loss term to explicitly penalize DDIs.
- **COGNet** [43] introduces a “copy-or-predict” mechanism within an encoder-decoder framework. By leveraging historical medication usage, COGNet determines whether to replicate past prescriptions or predict new ones.
- **Carmen** [6] is a context-aware medication recommendation framework based on graph neural networks. It fuses molecular structure, contextual information, and DDI graph encoding to enhance both performance and safety.⁵
- **DGCL** [28] employs graph contrastive learning to jointly train medication knowledge graphs and EHR graphs, effectively controlling DDI levels and improving recommendation robustness.
- **VITA** [21] improves accuracy through relevant-visit selection and target-aware attention mechanisms, which better capture relationships between current and historical visits.
- **MoleRec** [48] utilizes molecular substructure-aware encoding and attention mechanisms to generate personalized and safer medication combinations.
- **DAPSNet** [40] employs dual attention mechanisms at both code and visit levels to construct comprehensive patient representations. It retrieves information from similar trajectories through a patient memory module and applies an information bottleneck to enhance robustness and safety.

⁵We were unable to successfully execute Carmen’s official implementation; therefore, we report the results provided in their paper [6], obtained under the same data preprocessing pipeline.

- **PROMISE** [42] introduces a pre-trained multimodal framework that integrates structured EHRs and clinical text via hypergraph and language encoders. By combining multimodal pre-training with controllable DDI optimization, it effectively balances accuracy and medication safety.

We adopted the hyperparameter configurations recommended in the original papers of all baseline models to ensure fair comparison. For baselines without publicly available implementations, we carefully re-implemented them according to the methodological details described in their papers.

Our method was implemented in Python 3.8.16 using PyTorch 1.13.1. Experiments were conducted on a workstation equipped with dual Intel Xeon Gold 5318Y CPUs, 251GB of RAM, and three NVIDIA A40 GPUs. The hyperparameter search space is summarized in Tab. 3.

Table 3. Hyperparameter search space.

| Hyperparameter | Range |
|---------------------------|--|
| Learning rate | $[5 \times 10^{-3}, 1 \times 10^{-4}]$ |
| Weight decay | $[1 \times 10^{-4}, 1 \times 10^{-5}]$ |
| Dropout ratio | $[0, 0.5]$ |
| Epochs | 50 |
| Batch size | 16 |
| Pretraining epochs | {150, 300, 500, 1500, 3000} |
| Pretraining learning rate | $\{1 \times 10^{-3}, 5 \times 10^{-4}\}$ |
| Pretraining weight decay | $\{1 \times 10^{-4}, 1 \times 10^{-5}\}$ |
| KHGE layers | {1, 2, 4} |
| Embedding dimension | 64 |
| λ_{multi} | $[1, 10^{-6}]$ |
| λ_{ddi} | $[1, 10^{-6}]$ |
| λ_{aux} | $[1, 10^{-6}]$ |

5.2 Performance Comparison

As shown in Tab. 4, Tab. 5, and Tab. 6, HypeMed achieves the best or highly competitive overall performance across the three benchmark datasets (**MIMIC-III**, **MIMIC-IV**, and **eICU**), particularly on Jaccard and F1, while maintaining favorable safety/compactness trade-offs, demonstrating strong generalization ability across diverse clinical settings.

Machine learning (ML) methods (LR and ECC) achieve relatively low performance, as they rely on simple multi-hot encodings and fail to capture temporal or relational dependencies among medical entities. **Instance-based methods** (e.g., LEAP) make predictions solely based on the current visit, ignoring both patient history and inter-visit references from similar visits, which further limits their accuracy. **Longitudinal methods** (e.g., RETAIN, GAMENet, SafeDrug, COGNet, DGCL, and MoleRec) perform substantially better by modeling temporal dependencies within patient histories. Among them, **SafeDrug** effectively reduces DDI rates by incorporating drug–drug interaction information, though at the cost of slightly lower accuracy. **COGNet** employs sequence generation with beam search, achieving modest accuracy gains but increasing the number of recommended medications and thus DDI risk. **VITA** retrieves relevant past visits via target-aware attention, yet its performance varies with dataset temporal structures. **Carmen**, **DGCL**, and **MoleRec** further integrate co-occurrence relationships, contrastive learning, and molecular structures, respectively, to enhance medication representations. However, these models mainly focus on intra-patient historical modeling and overlook inter-visit reference augmentation via visit-conditioned retrieval.

Table 4. Performance comparison on **MIMIC-III** in terms of Jaccard (%), F1 (%), PRAUC (%), DDI (%) and # Med.. Numbers in **bold** and underlined indicate the best and the second-best performance, respectively, according to t-tests at the 95% confidence level. The ground-truth rate of DDI in the MIMIC-III dataset is 8.68%. In the DDI column, values in **orange** indicate that the mean+std exceeds the ground truth (inferior), while values in **blue** indicate that the mean-std falls below it (superior).

| Model | Jaccard \uparrow | F1 \uparrow | PRAUC \uparrow | DDI \downarrow | # Med. \downarrow |
|----------|--------------------|------------------|------------------|------------------|---------------------|
| LR | 49.24 \pm 0.27 | 64.96 \pm 0.25 | 75.48 \pm 0.31 | 8.30 \pm 0.07 | 16.05 \pm 0.14 |
| ECC | 48.89 \pm 0.26 | 64.47 \pm 0.23 | 76.19 \pm 0.20 | 8.57 \pm 0.08 | 15.74 \pm 0.07 |
| LEAP | 45.50 \pm 0.18 | 61.68 \pm 0.17 | 65.46 \pm 0.42 | 7.82 \pm 0.08 | 18.57 \pm 0.08 |
| RETAIN | 48.60 \pm 0.34 | 64.66 \pm 0.34 | 75.98 \pm 0.43 | 8.83 \pm 0.10 | 18.92 \pm 0.13 |
| GAMENet | 50.18 \pm 0.20 | 65.84 \pm 0.19 | 76.48 \pm 0.23 | 8.88 \pm 0.09 | 27.39 \pm 0.19 |
| SafeDrug | 51.14 \pm 0.25 | 66.81 \pm 0.23 | 76.43 \pm 0.23 | 5.87 \pm 0.04 | 18.94 \pm 0.13 |
| DAPSNet | 51.68 \pm 0.30 | 67.22 \pm 0.26 | 76.64 \pm 0.40 | 5.96 \pm 0.05 | 21.14 \pm 0.15 |
| PROMISE | 52.33 \pm 0.25 | 67.91 \pm 0.21 | 80.58 \pm 0.17 | 9.33 \pm 0.15 | 15.27 \pm 0.13 |
| COGNet | 52.94 \pm 0.39 | 68.30 \pm 0.36 | 76.79 \pm 0.25 | 8.62 \pm 0.09 | 28.08 \pm 0.18 |
| Carmen | 52.67 \pm 0.21 | 68.12 \pm 0.19 | 76.52 \pm 0.36 | – | – |
| DGCL | 52.54 \pm 0.15 | 67.98 \pm 0.14 | 77.28 \pm 0.15 | 8.31 \pm 0.03 | 21.92 \pm 0.05 |
| VITA | 52.15 \pm 0.25 | 67.42 \pm 0.21 | 76.23 \pm 0.25 | 8.11 \pm 0.09 | 29.53 \pm 0.15 |
| MoleRec | 53.12 \pm 0.39 | 68.50 \pm 0.33 | 77.51 \pm 0.25 | 7.16 \pm 0.07 | 20.54 \pm 0.16 |
| HypeMed | 53.73 \pm 0.30 | 69.06 \pm 0.25 | 78.34 \pm 0.31 | 8.19 \pm 0.07 | 20.01 \pm 0.14 |

Table 5. Performance comparison on **MIMIC-IV**. The ground-truth DDI rate is 7.24%. Other notations are consistent with Tab. 4.

| Model | Jaccard \uparrow | F1 \uparrow | PRAUC \uparrow | DDI \downarrow | # Med. \downarrow |
|----------|--------------------|------------------|------------------|------------------|---------------------|
| LR | 47.65 \pm 0.20 | 63.31 \pm 0.20 | 74.08 \pm 0.22 | 7.07 \pm 0.09 | 13.85 \pm 0.09 |
| ECC | 45.82 \pm 0.16 | 61.39 \pm 0.15 | 73.58 \pm 0.14 | 7.11 \pm 0.14 | 13.08 \pm 0.14 |
| LEAP | 43.73 \pm 0.26 | 59.84 \pm 0.25 | 63.73 \pm 0.26 | 6.57 \pm 0.06 | 17.34 \pm 0.07 |
| RETAIN | 44.10 \pm 0.49 | 60.23 \pm 0.50 | 71.60 \pm 0.40 | 7.29 \pm 0.14 | 14.43 \pm 0.17 |
| GAMENet | 47.80 \pm 0.26 | 63.56 \pm 0.28 | 75.09 \pm 0.20 | 7.85 \pm 0.05 | 23.36 \pm 0.17 |
| SafeDrug | 48.47 \pm 0.17 | 64.24 \pm 0.17 | 74.27 \pm 0.22 | 6.11 \pm 0.11 | 18.15 \pm 0.11 |
| DAPSNet | 49.14 \pm 0.22 | 64.85 \pm 0.19 | 74.07 \pm 0.14 | 5.95 \pm 0.07 | 18.62 \pm 0.15 |
| PROMISE | 47.29 \pm 0.30 | 63.29 \pm 0.28 | 76.76 \pm 0.28 | 7.08 \pm 0.13 | 11.34 \pm 0.08 |
| COGNet | 50.38 \pm 0.27 | 65.87 \pm 0.23 | 74.99 \pm 0.15 | 7.79 \pm 0.15 | 25.43 \pm 0.19 |
| Carmen | 50.06 \pm 0.12 | 65.69 \pm 0.07 | 74.62 \pm 0.30 | – | – |
| DGCL | 49.98 \pm 0.13 | 65.62 \pm 0.12 | 75.52 \pm 0.14 | 7.50 \pm 0.04 | 18.41 \pm 0.07 |
| VITA | 50.31 \pm 0.29 | 65.75 \pm 0.26 | 74.80 \pm 0.20 | 7.45 \pm 0.11 | 26.37 \pm 0.18 |
| MoleRec | 50.64 \pm 0.23 | 66.23 \pm 0.21 | 75.14 \pm 0.32 | 6.78 \pm 0.07 | 17.74 \pm 0.11 |
| HypeMed | 51.05 \pm 0.28 | 66.50 \pm 0.25 | 76.23 \pm 0.28 | 6.65 \pm 0.11 | 16.53 \pm 0.16 |

Compared with **DAPSNet** and **PROMISE**, the performance improvements of HypeMed primarily stem from its *representation-consistent* design that couples representation learning and visit-conditioned inter-visit retrieval. Although DAPSNet and PROMISE also incorporate similarity information to improve recommendation accuracy, their retrieval mechanisms are *decoupled* from representation learning: DAPSNet relies on embedding-level similarity and answer-side matching, while PROMISE computes trajectory similarity via DTW followed by hierarchical attention. This separation can lead to a mismatch between the learned embedding space and the retrieval criterion. In contrast, HypeMed performs retrieval and aggregation within the same hypergraph representation space, enabling coherent use of inter-visit references to refine latent condition estimation and medication prediction. This representation–retrieval

Table 6. Performance comparison on **eICU**. There is no DDI information available in the eICU dataset. Other notations are consistent with Tab. 4.

| Model | Jaccard \uparrow | F1 \uparrow | PRAUC \uparrow | # Med. \downarrow |
|----------|--------------------|------------------|------------------|---------------------|
| LR | 39.96 \pm 0.22 | 55.33 \pm 0.21 | 73.24 \pm 0.36 | 7.55 \pm 0.07 |
| ECC | 33.41 \pm 0.26 | 47.66 \pm 0.29 | 72.66 \pm 0.28 | 5.62 \pm 0.05 |
| LEAP | 41.25 \pm 0.18 | 56.43 \pm 0.21 | 62.67 \pm 0.29 | 13.36 \pm 0.04 |
| RETAIN | 41.29 \pm 0.35 | 56.53 \pm 0.34 | 71.02 \pm 0.35 | 9.81 \pm 0.06 |
| GAMENet | 40.80 \pm 0.17 | 56.00 \pm 0.18 | 68.86 \pm 0.18 | 12.45 \pm 0.03 |
| SafeDrug | 41.19 \pm 0.23 | 56.42 \pm 0.21 | 69.79 \pm 0.30 | 10.68 \pm 0.06 |
| DAPSNet | 41.14 \pm 0.28 | 56.44 \pm 0.30 | 69.99 \pm 0.31 | 10.87 \pm 0.06 |
| PROMISE | 40.89 \pm 0.25 | 56.41 \pm 0.26 | 73.54 \pm 0.32 | 7.16 \pm 0.05 |
| COGNet | 41.19 \pm 0.42 | 56.20 \pm 0.43 | 67.68 \pm 0.41 | 20.14 \pm 0.10 |
| DGCL | 38.23 \pm 0.14 | 53.38 \pm 0.15 | 69.63 \pm 0.13 | 8.70 \pm 0.02 |
| VITA | 41.46 \pm 0.26 | 56.38 \pm 0.29 | 67.39 \pm 0.37 | 18.47 \pm 0.03 |
| MoleRec | 41.32 \pm 0.20 | 56.63 \pm 0.24 | 70.14 \pm 0.27 | 10.44 \pm 0.05 |
| HypeMed | 41.95 \pm 0.23 | 56.97 \pm 0.23 | 68.71 \pm 0.27 | 16.32 \pm 0.02 |

consistency ensures semantically aligned visit matching and contributes to the observed performance gains across multiple evaluation metrics.

Consequently, HypeMed surpasses the strongest baselines on the primary recommendation metrics across all three datasets, achieving Jaccard score improvements of +0.61%, +0.41%, and +0.49% on MIMIC-III, MIMIC-IV, and eICU, respectively. Moreover, HypeMed maintains DDI rates below or comparable to real-world prescriptions (8.68% for MIMIC-III and 7.24% for MIMIC-IV), indicating that its superior accuracy does not compromise safety (on MIMIC-III/IV where DDI graphs are available). Notably, the results on the **eICU** dataset further demonstrate HypeMed’s robustness and generalization capability on cross-domain, cross-institution, and cross-population datasets. Overall, HypeMed achieves a favorable balance between accuracy and safety, underscoring its potential as a reliable clinical decision-support system.

5.3 Ablation Study

To evaluate the contribution of each component in HypeMed, we conduct comprehensive ablation studies on the **MIMIC-III** [20] and **MIMIC-IV** [19] datasets. We construct several model variants as follows: (1) HypeMed–MedRep–None initializes entity and visit representations randomly, removing the pretrained entity and visit representations from MedRep; (2) HypeMed–MedRep–Fixed keeps the embeddings generated by MedRep frozen during SimMR training, preventing task-specific adaptation; (3) HypeMed–MedRep–HGNCN replaces the KHGE encoder in MedRep with a standard HGNCN [9]; (4) HypeMed–MedRep–GCN replaces the KHGE encoder with a conventional GCN [36]; (5) HypeMed–SimMR–w/o Hist. removes the historical information encoder, relying solely on the similarity channel; and (6) HypeMed–SimMR–w/o Sim. disables the similarity channel, preserving only the historical encoder.

From the results, several key observations can be drawn. First, all ablated variants show degraded performance compared to the full HypeMed, confirming the overall effectiveness and complementarity of its components. Second, removing MedRep (HypeMed–MedRep–None) leads to a notable decline (e.g., Jaccard decreases from 53.10 \rightarrow 51.70 in MIMIC-III), highlighting the essential role of pretrained representations in encoding medical knowledge. Notably, the HypeMed–MedRep–Fixed variant, which freezes the pretrained embeddings without further optimization, exhibits a more pronounced decrease compared to the jointly optimized model. This confirms that fine-tuning the pretrained representations during SimMR training is crucial for aligning them with the downstream recommendation objective.

Table 7. Ablation study on **MIMIC-III** and **MIMIC-IV**. Results are reported for both MedRep and SimMR ablations. Jaccard, F1, PRAUC, and DDI are expressed as percentages (%), while #Med. retains its original scale. Values denote mean \pm standard deviation over ten bootstrap sampling.

| MIMIC-III | | | | | |
|--------------------|--------------------|------------------|------------------|------------------|--------------------|
| Model | Jaccard \uparrow | F1 \uparrow | PRAUC \uparrow | DDI \downarrow | #Med. \downarrow |
| HypeMed | 53.10 \pm 0.43 | 68.49 \pm 0.37 | 77.32 \pm 0.37 | 6.67 \pm 0.06 | 22.77 \pm 0.13 |
| – SimMR– w/o Sim. | 52.91 \pm 0.42 | 68.33 \pm 0.37 | 77.44 \pm 0.34 | 6.40 \pm 0.06 | 22.50 \pm 0.13 |
| – SimMR– w/o Hist. | 51.90 \pm 0.41 | 67.40 \pm 0.37 | 76.40 \pm 0.44 | 6.12 \pm 0.06 | 23.16 \pm 0.17 |
| – MedRep–None | 51.70 \pm 0.46 | 67.23 \pm 0.42 | 76.28 \pm 0.36 | 6.52 \pm 0.06 | 23.51 \pm 0.14 |
| – MedRep–Fixed | 50.49 \pm 0.23 | 66.24 \pm 0.21 | 75.44 \pm 0.33 | 6.45 \pm 0.04 | 22.74 \pm 0.13 |
| – MedRep–HGNC | 52.17 \pm 0.45 | 67.67 \pm 0.40 | 75.66 \pm 0.35 | 6.01 \pm 0.05 | 22.37 \pm 0.13 |
| – MedRep–GCN | 51.40 \pm 0.40 | 66.96 \pm 0.36 | 76.31 \pm 0.38 | 6.05 \pm 0.05 | 23.50 \pm 0.12 |

| MIMIC-IV | | | | | |
|-------------------|--------------------|------------------|------------------|------------------|--------------------|
| Model | Jaccard \uparrow | F1 \uparrow | PRAUC \uparrow | DDI \downarrow | #Med. \downarrow |
| HypeMed | 50.92 \pm 0.25 | 66.35 \pm 0.24 | 75.96 \pm 0.26 | 6.60 \pm 0.13 | 16.68 \pm 0.11 |
| – SimMR–w/o Sim. | 50.52 \pm 0.26 | 66.03 \pm 0.23 | 75.91 \pm 0.24 | 6.79 \pm 0.15 | 16.61 \pm 0.12 |
| – SimMR–w/o Hist. | 49.98 \pm 0.23 | 65.52 \pm 0.21 | 75.19 \pm 0.25 | 6.47 \pm 0.11 | 16.26 \pm 0.11 |
| – MedRep–None | 50.15 \pm 0.25 | 65.68 \pm 0.23 | 75.47 \pm 0.25 | 6.65 \pm 0.11 | 16.61 \pm 0.11 |
| – MedRep–Fixed | 46.68 \pm 0.21 | 62.58 \pm 0.22 | 72.99 \pm 0.27 | 6.01 \pm 0.07 | 21.11 \pm 0.12 |
| – MedRep–HGNC | 50.21 \pm 0.23 | 65.76 \pm 0.22 | 75.54 \pm 0.30 | 6.48 \pm 0.12 | 16.63 \pm 0.12 |
| – MedRep–GCN | 50.25 \pm 0.20 | 65.77 \pm 0.19 | 75.61 \pm 0.23 | 6.48 \pm 0.12 | 16.95 \pm 0.11 |

Furthermore, substituting the KHGE encoder with HGCN or GCN (HypeMed–MedRep–HGNC / HypeMed–MedRep–GCN) produces only marginal gains (Jaccard: 50.15 \rightarrow 50.25 in MIMIC-IV) over the random baseline, indicating that conventional graph encoders are less capable of capturing intra-visit combinatorial semantics among medications. Finally, removing the historical channel (HypeMed–SimMR–w/o Hist.) results in the largest degradation (e.g., F1: 68.49 \rightarrow 67.40 / 66.35 \rightarrow 65.52 in MIMIC-III/IV respectively), underscoring the critical importance of longitudinal information. Nevertheless, the two channels are complementary, and their integration yields the best overall performance.

5.4 Hyperparameters Analysis

In HypeMed, three hyperparameters potentially influence the final performance: (1) the number of pretraining epochs in the MedRep stage, (2) the number of retrieved similar visits (Top- n) in SimMR, and (3) the number of augmented hypergraph views used for contrastive learning. We conduct a comprehensive sensitivity analysis to evaluate their effects, with results summarized in Tab. 8, Tab. 9, and Tab. 10. All experiments evaluate five metrics—Jaccard, F1, PRAUC, DDI rate, and average number of medications—while keeping all other settings fixed.

(1) Number of Pretraining Epochs. We vary the number of pretraining epochs over $\{0, 100, 300, 1000\}$ to examine the effect of contrastive pretraining duration. As shown in Tab. 8, performance on both datasets first improves and then slightly declines as the number of epochs increases, indicating that moderate pretraining enhances representation quality, whereas excessive training may lead to overfitting or representation collapse.

(2) Number of Retrieved Similar Visits. To assess the impact of inter-visit reference retrieval, we vary the number of retrieved similar visits (Top- n) in $\{0, 1, 10, 100, 500, 1000\}$. As shown in Tab. 9, incorporating a small number of similar visits significantly enhances predictive performance, whereas a large n introduces noise or less relevant samples,

Table 8. Pretraining-epoch sensitivity analysis on **MIMIC-III** and **MIMIC-IV**. Jaccard, F1, PRAUC, and DDI are reported in percentage (%), while #Med. retains its original scale. Values denote mean \pm standard deviation over ten bootstrap sampling.

| Dataset | Epoch | Jaccard \uparrow | F1 \uparrow | PRAUC \uparrow | DDI \downarrow | #Med. \downarrow |
|------------------|-------|--------------------|------------------|------------------|------------------|--------------------|
| MIMIC-III | 0 | 52.27 \pm 0.25 | 67.75 \pm 0.22 | 77.08 \pm 0.32 | 7.93 \pm 0.10 | 21.90 \pm 0.17 |
| | 100 | 52.48 \pm 0.23 | 67.97 \pm 0.20 | 76.10 \pm 0.29 | 6.00 \pm 0.04 | 21.93 \pm 0.10 |
| | 300 | 52.26 \pm 0.25 | 67.78 \pm 0.22 | 76.27 \pm 0.39 | 6.06 \pm 0.04 | 22.74 \pm 0.11 |
| | 1000 | 51.91 \pm 0.22 | 67.41 \pm 0.19 | 75.95 \pm 0.37 | 6.28 \pm 0.04 | 22.95 \pm 0.14 |
| MIMIC-IV | 0 | 49.87 \pm 0.23 | 65.48 \pm 0.23 | 75.60 \pm 0.27 | 6.30 \pm 0.10 | 16.50 \pm 0.14 |
| | 100 | 50.03 \pm 0.23 | 65.65 \pm 0.22 | 75.87 \pm 0.27 | 6.34 \pm 0.07 | 16.81 \pm 0.17 |
| | 300 | 50.17 \pm 0.33 | 65.75 \pm 0.32 | 75.92 \pm 0.27 | 6.39 \pm 0.09 | 16.32 \pm 0.15 |
| | 1000 | 49.66 \pm 0.26 | 65.30 \pm 0.26 | 75.67 \pm 0.25 | 6.46 \pm 0.09 | 16.20 \pm 0.16 |

Table 9. Top- n sensitivity analysis on **MIMIC-III** and **MIMIC-IV**. Jaccard, F1, PRAUC, and DDI are expressed as percentages (%), while #Med. retains its original scale. Values denote mean \pm standard deviation over ten bootstrap sampling.

| Dataset | Top- n | Jaccard \uparrow | F1 \uparrow | PRAUC \uparrow | DDI \downarrow | #Med. \downarrow |
|------------------|----------|--------------------|------------------|------------------|------------------|--------------------|
| MIMIC-III | 0 | 53.10 \pm 0.43 | 68.49 \pm 0.37 | 77.32 \pm 0.37 | 6.67 \pm 0.06 | 22.77 \pm 0.13 |
| | 1 | 53.23 \pm 0.40 | 68.59 \pm 0.35 | 77.50 \pm 0.35 | 6.52 \pm 0.04 | 22.58 \pm 0.13 |
| | 10 | 52.93 \pm 0.37 | 68.33 \pm 0.33 | 77.43 \pm 0.38 | 6.25 \pm 0.05 | 22.63 \pm 0.13 |
| | 100 | 52.88 \pm 0.40 | 68.27 \pm 0.35 | 77.20 \pm 0.36 | 6.44 \pm 0.06 | 23.31 \pm 0.13 |
| | 500 | 52.80 \pm 0.38 | 68.22 \pm 0.34 | 77.13 \pm 0.36 | 6.39 \pm 0.06 | 23.20 \pm 0.12 |
| | 1000 | 52.86 \pm 0.37 | 68.26 \pm 0.33 | 77.16 \pm 0.39 | 6.40 \pm 0.05 | 22.97 \pm 0.12 |
| MIMIC-IV | 0 | 50.52 \pm 0.26 | 66.03 \pm 0.23 | 75.91 \pm 0.24 | 6.79 \pm 0.15 | 16.61 \pm 0.12 |
| | 1 | 50.51 \pm 0.25 | 66.01 \pm 0.24 | 75.79 \pm 0.21 | 6.69 \pm 0.12 | 16.93 \pm 0.11 |
| | 10 | 50.39 \pm 0.23 | 65.90 \pm 0.21 | 75.98 \pm 0.23 | 6.49 \pm 0.12 | 16.69 \pm 0.11 |
| | 100 | 50.70 \pm 0.18 | 66.23 \pm 0.17 | 76.18 \pm 0.23 | 6.49 \pm 0.13 | 16.80 \pm 0.12 |
| | 500 | 50.77 \pm 0.21 | 66.29 \pm 0.19 | 76.30 \pm 0.23 | 6.52 \pm 0.13 | 16.51 \pm 0.12 |
| | 1000 | 50.72 \pm 0.15 | 66.22 \pm 0.14 | 76.25 \pm 0.22 | 6.54 \pm 0.10 | 16.42 \pm 0.12 |

leading to marginal declines in Jaccard and F1. Nevertheless, the DDI rate and average medication count remain stable, suggesting that HypeMed achieves a robust balance between accuracy and safety.

Table 10. Performance and GPU memory usage under different numbers of views on **MIMIC-III** and **MIMIC-IV**. Jaccard, F1, PRAUC, and DDI are reported in percentage (%); #Med. and GPU Memory (GiB) keep their original scales.

| Dataset | #Views | Jaccard \uparrow | F1 \uparrow | PRAUC \uparrow | DDI \downarrow | #Med. \downarrow | GPU Mem. |
|------------------|--------|--------------------|------------------|------------------|------------------|--------------------|----------|
| MIMIC-III | 2 | 51.13 \pm 0.17 | 66.77 \pm 0.16 | 75.73 \pm 0.34 | 6.50 \pm 0.05 | 23.80 \pm 0.14 | 9.76 |
| | 3 | 51.26 \pm 0.22 | 66.85 \pm 0.18 | 75.86 \pm 0.40 | 6.23 \pm 0.05 | 24.60 \pm 0.12 | 14.09 |
| | 4 | 51.05 \pm 0.20 | 66.67 \pm 0.19 | 75.89 \pm 0.43 | 6.33 \pm 0.04 | 23.49 \pm 0.12 | 24.53 |
| MIMIC-IV | 2 | 48.67 \pm 0.23 | 64.40 \pm 0.22 | 73.57 \pm 0.32 | 5.90 \pm 0.07 | 20.93 \pm 0.17 | 14.46 |
| | 3 | 48.61 \pm 0.21 | 64.34 \pm 0.20 | 73.44 \pm 0.27 | 5.99 \pm 0.06 | 21.19 \pm 0.18 | 23.75 |
| | 4 | 48.66 \pm 0.25 | 64.39 \pm 0.25 | 73.31 \pm 0.29 | 6.03 \pm 0.07 | 20.87 \pm 0.17 | 38.61 |

(3) **Number of Augmented Views.** In lines 5–6 of Alg. 2, HypeMed generates two augmented hypergraph views for contrastive pretraining in the MedRep stage. This design follows standard contrastive frameworks [23, 41], where two

correlated views are typically sufficient to learn discriminative yet invariant representations. To validate this choice, we vary the number of augmented views over $\{2, 3, 4\}$. As shown in Tab. 10, adding more views provides negligible improvements or even slight degradation, while GPU memory usage increases sharply (from 9.8→24.5 GiB on MIMIC-III and 14.5→38.6 GiB on MIMIC-IV). These findings suggest that two views strike the best balance between performance and computational efficiency, supporting the design adopted in Alg. 2.

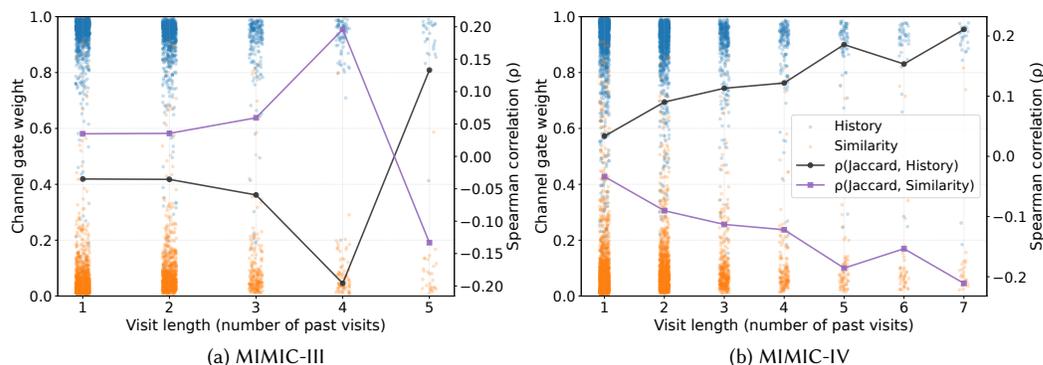


Fig. 5. Visualization of adaptive channel weighting and its association with model performance across different visit lengths. Blue and orange dots represent per-visit gate weights for the history and similarity channels, respectively. Black and purple lines indicate the Spearman correlations (ρ) between Jaccard performance and each channel’s weight, computed within visit-length groups. To ensure statistical reliability, visit-length groups with fewer than 20 samples were excluded.

5.5 Interpretability analysis

To better understand how HypeMed balances information from different sources, we visualize the per-visit gate weights of the two channels learned by the fusion module (blue: history channel; orange: similarity channel) in Fig. 5, along with their Spearman correlations (ρ) with predictive performance (Jaccard). Each dot represents a single visit.

In both datasets, the similarity channel generally receives smaller gate weights than the history channel, since the latter always incorporates the current visit’s diagnosis and procedure information—even for single-visit patients—whereas the similarity channel depends on retrieved visits and provides supplementary context when longitudinal data are sparse. This finding aligns with our ablation results, where removing the history channel led to a larger performance degradation. However, the strength of association between each channel and predictive accuracy varies across datasets. In MIMIC-III, where patient trajectories are shorter and sparser, the similarity-channel weight correlates positively with performance during early visits (1–4), suggesting that retrieved visits compensate for limited historical context. Once sufficient history accumulates (≥ 5 visits), the correlation shifts toward the history channel, indicating increasing reliance on personal longitudinal information. In contrast, MIMIC-IV contains richer and longer patient trajectories, and its performance correlates more strongly with the history channel across all visit lengths.

Overall, the correlation between predictive accuracy and the similarity-channel weight exhibits a general downward trend as visit length increases, indicating that HypeMed gradually shifts its reliance from retrieved similar visits to accumulated patient-specific history. This visualization thus provides an intrinsic, model-level diagnostic perspective on how the relative contributions of different information sources evolve with increasing data completeness.

Table 11. Performance comparison under the cold-start (first-visit) setting on **MIMIC-III** and **MIMIC-IV**. All metrics are reported as percentages (%), except for the average number of medications (#Med.). Values denote mean \pm standard deviation over ten bootstrapped runs. A total of 192 and 307 patients have only one visit in MIMIC-III and MIMIC-IV, respectively.

| Dataset | Model | Jaccard \uparrow | F1 \uparrow | PRAUC \uparrow | DDI \downarrow | #Med. \downarrow |
|-----------|---------------------------|-------------------------|-------------------------|-------------------------|------------------|--------------------|
| MIMIC-III | SafeDrug | 49.47 \pm 1.24 | 65.22 \pm 1.03 | 74.92 \pm 0.91 | 6.22 \pm 0.20 | 18.01 \pm 0.26 |
| | DAPSNet | 50.68 \pm 0.95 | 66.20 \pm 0.79 | 74.90 \pm 1.00 | 6.12 \pm 0.11 | 19.92 \pm 0.33 |
| | COGNet | 51.19 \pm 1.20 | 66.64 \pm 1.11 | 75.62 \pm 0.69 | 8.62 \pm 0.15 | 24.90 \pm 0.27 |
| | HypeMed | 51.60 \pm 0.86 | 67.14 \pm 0.73 | 75.83 \pm 0.98 | 8.65 \pm 0.29 | 20.04 \pm 0.31 |
| | HypeMed -SimMR- w/o Hist. | 49.93 \pm 1.00 | 65.46 \pm 0.92 | 74.78 \pm 1.04 | 6.19 \pm 0.25 | 23.44 \pm 0.32 |
| | HypeMed -SimMR- w/o Sim. | 49.52 \pm 1.08 | 65.24 \pm 0.98 | 74.62 \pm 0.96 | 6.35 \pm 0.25 | 22.43 \pm 0.28 |
| MIMIC-IV | SafeDrug | 44.89 \pm 1.16 | 60.61 \pm 1.14 | 66.83 \pm 1.34 | 6.41 \pm 0.24 | 17.58 \pm 0.31 |
| | DAPSNet | 49.49 \pm 1.10 | 65.23 \pm 1.01 | 73.71 \pm 1.05 | 6.07 \pm 0.22 | 17.49 \pm 0.25 |
| | COGNet | 46.91 \pm 0.48 | 62.64 \pm 0.46 | 72.82 \pm 0.70 | 8.90 \pm 0.17 | 18.67 \pm 0.39 |
| | HypeMed | 50.30 \pm 0.94 | 65.78 \pm 0.86 | 75.19 \pm 1.11 | 6.81 \pm 0.37 | 16.31 \pm 0.27 |
| | HypeMed -SimMR- w/o Hist. | 49.84 \pm 0.98 | 65.39 \pm 0.89 | 74.94 \pm 1.13 | 6.61 \pm 0.34 | 16.30 \pm 0.25 |
| | HypeMed -SimMR- w/o Sim. | 49.62 \pm 1.00 | 65.15 \pm 0.94 | 74.89 \pm 1.12 | 7.18 \pm 0.26 | 16.86 \pm 0.36 |

5.6 Cold-start performance analysis

To evaluate whether the proposed similar-visit retrieval mechanism effectively alleviates the cold-start problem, we perform a targeted assessment under the no-history (first-visit) setting, as shown in Tab. 11. Across both **MIMIC-III** and **MIMIC-IV**, HypeMed achieves the best performance among all baselines (*SafeDrug*, *DAPSNet*, and *COGNet*) in terms of Jaccard, F1, and PRAUC, demonstrating strong generalization capability even when patient history is unavailable.

Compared with its ablated variants, the complete model exhibits clear advantages. Removing the similarity channel (*w/o Sim.*) leads to consistent declines in all major metrics—on **MIMIC-III**, Jaccard decreases by approximately 2% and F1 by 1.5%—indicating that retrieved visits provide valuable external priors that compensate for the absence of personal history. In contrast, removing the historical channel (*w/o Hist.*) results in slightly lower performance, suggesting that under the single-visit condition, the model mainly relies on current-visit information while still benefiting from the structural representation of the historical channel. Together with the observations from Fig. 5, these results highlight a complementary effect between the two channels: the similarity channel is crucial in true cold-start cases, whereas the historical channel becomes increasingly influential as longitudinal information accumulates.

These findings quantitatively verify that the dual-channel design enables HypeMed to adaptively balance retrieved population-level knowledge and patient-specific information, thereby achieving superior robustness and predictive accuracy in cold-start scenarios.

5.7 Retrieval-quality correlation analysis

To quantify the advantage of *joint* representation–retrieval learning over a decoupled retrieval paradigm, we report the retrieval-quality correlation analysis in Fig. 6.

From Fig. 6, we observe two key trends: First, for both HypeMed and HypeMed-MedRep-None, higher retrieval quality generally leads to better prediction performance, confirming that relevant references provide useful complementary signals [21, 40, 43]. Second, under the *same* retrieval-quality bin, HypeMed consistently achieves higher accuracy than HypeMed-MedRep-None, indicating that the performance gain is not solely explained by retrieving “more correct” cases.

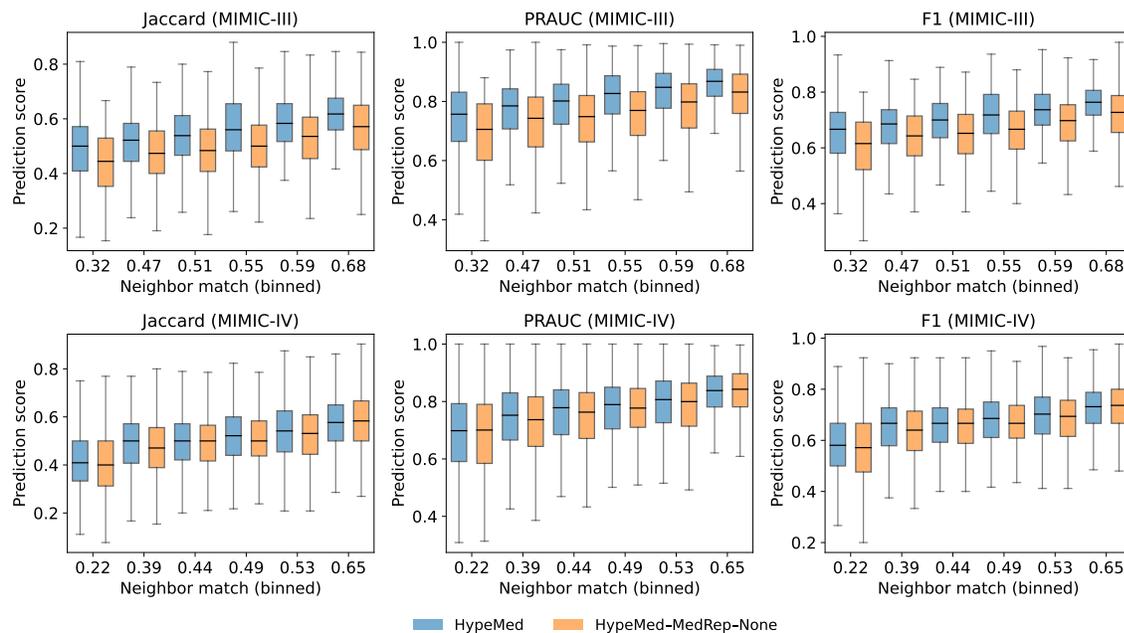


Fig. 6. Retrieval-quality correlation analysis. We compare HypeMed with HypeMed-MedRep-None, a controlled variant that keeps the same prediction architecture but removes MedRep pre-training, resulting in retrieval within a non-pretrained embedding space. Test visits are grouped into bins by retrieval quality, measured as the F1 score between medications in the retrieved visits and the ground-truth medications. The x-axis denotes retrieval-quality bins (higher is better), and the y-axis reports the final prediction performance.

This result supports our hypothesis that *coupling representation learning with retrieval improves how retrieved context is represented and used*. In HypeMed-None, retrieval is performed in a randomly initialized or weakly regularized embedding space; even when the retrieved visits exhibit comparable label overlap, their embeddings may be poorly organized with respect to visit semantics, making the fused context less predictable and less exploitable by the downstream predictor. In contrast, MedRep pre-trains a hyperedge-aware metric space where visit embeddings are explicitly optimized to encode high-order clinical interactions and preserve set-level semantics. Therefore, retrieved neighbors become not only label-relevant but also *geometrically compatible* with the query visit, providing a more structured and informative signal for fusion, which ultimately translates into consistently better prediction performance.

5.8 Time complexity comparison

We perform a time complexity analysis by comparing our approach with state-of-the-art methods, using floating point operations (FLOPs) as the metric. Specifically, we measure the FLOPs for each model during the inference phase. The results are illustrated in Fig. 7. Thanks to its two-stage design, HypeMed reuses MedRep embeddings and avoids additional GNN computations during inference. This significantly reduces its computational load, making its FLOPs count second only to SafeDrug. On the other hand, MoleRec executes a 4-layer GNN across multiple graphs, which results in the highest FLOPs during inference among the evaluated models.

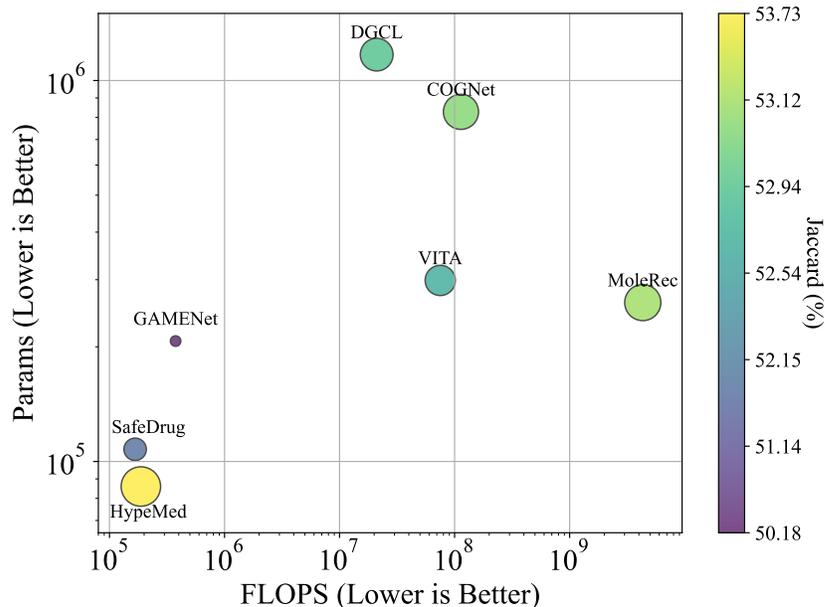


Fig. 7. Comparative time complexity across models (FLOPs vs. Parameters)

5.9 Case study

To illustrate the effectiveness of HypeMed’s recommendations more vividly, we conduct a case study with *Patient-908* as an example. *Patient-908* from the MIMIC-III test set has two medical visits. During the first visit, the patient was diagnosed with diseases such as cirrhosis, liver cancer, and chronic hepatitis. In the second visit, new conditions were added, including complications from liver transplantation, renal failure, and confirmed sepsis. Tab. 12 displays detailed information. Due to space constraints, we use the ICD codes to represent diagnoses and ATC codes to represent medications.

Fig. 8 demonstrates the visualization of **diagnosis** entities using HGCN and KHGE as encoders in MedRep. Points in the same color correspond to the same ICD category. The KHGE group embeddings show apparent clustering by ICD categories. Nodes within the same category display closer distances. In contrast, the HGCN group shows scattered distribution without clear category distinction. To better demonstrate the effectiveness of KHGE, we select three groups of diagnosis entities from Visit-2, each belonging to the same category in the ICD classification system. The groups are as follows:

- ★ *Persons With Potential Health Hazards Related To Personal And Family History* (V1083, V168, and V1582);
- ► *Symptoms* (78060, 78552);
- + *Varicose veins of other sites* (4568, 4561).

Different markers are used to highlight entities belonging to different groups. We find that in the HGCN group, the distribution of these three entity groups was dispersed, while in the KHGE group, entities belonging to the same category are clustered together. This suggests that by incorporating the ICD classification system into the modeling process, KHGE can translate the relative distances of entities in ICD into distances in embedding representations,

granting similar representations to entities of the same category. In contrast, HGCN, lacking ICD information, fails to capture such information.

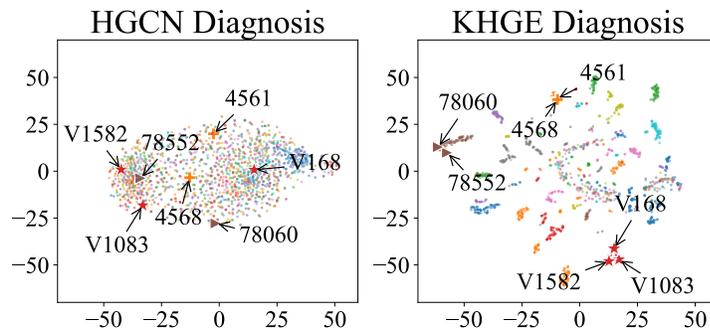


Fig. 8. Comparative Visualization of Diagnosis Entity Representations in MedRep using HGCN and KHGE Encoders.

We then compare the effectiveness of medication recommendations for *Patient-908* under two scenarios: with and without similar visit information (refer to Tab. 12). During the first visit, where the patient has no prior medical history, direct recommendations are prone to omissions and errors. To address this, HypeMed assigns a higher weight (90.43%) to similar visit information, thereby mitigating these inaccuracies. In the second visit, owing to the continuity of the patient’s condition, HypeMed can recommend the most suitable medications by relying on historical information, which is assigned a weight of 67.55%. However, errors in recommendations persist. At this juncture, information from similar visits plays a role in eliminating inappropriate medications. Ultimately, using similar visit information in the model increases the Jaccard similarity of medication recommendations from 46.15% / 55.17% to 77.27% / 64.00% in Visit-1 / Visit-2. This indicates that information on similar visits can significantly improve the accuracy of recommendations, especially when there is no history of prior medical consultations.

Table 12. The example of Patient-908 with two visits for the case study. Medications recommended correctly without using similar visit information are highlighted in **bold**. **Blue** and **orange** are used respectively to denote additional correct medications introduced and erroneous medications eliminated after incorporating similar visit information.

| | Diagnoses | Medications |
|---------|--|---|
| Visit-1 | V8741, V153, 79029, V1083, 5715, 1550, 07054 | N02B, A01A, A02B , A06A , B05C , J05A , A12C , A07A , J01E , N01A , C03C , J01C , N02A , B01A , N05C , D01A , J01D , L04A , D11A , A04A , A07E , D07A , D04A , C03D |
| Visit-2 | 99682, 78552, 0389, 99592, 5848, 570, 28419, 4561, 5589, 78060, 07070, 5715, V1083, E8780, 4568, V168, V1582 | N02B , A01A , A02B , A06A , B05C , A12A , A12C , C01C , A07A , J02A , J05A , V03A , J01E , J01C, N02A, A02A , J01M , B01A , N05C , J01D , B02B, C08C , D11A , A04A, A07E |

6 Conclusion and Future Work

In conclusion, we propose HypeMed, a hypergraph-based medication recommendation framework that reconstructs a patient’s latent clinical condition by unifying *intra-visit* set-level coherence modeling with *inter-visit* reference augmentation. Through a two-stage design, MedRep encodes high-order clinical co-occurrence into a globally consistent, retrieval-friendly embedding space, while SimMR performs visit-conditioned retrieval of similar visits and integrates them with longitudinal history for robust medication prediction. Extensive experiments on three real-world benchmarks (MIMIC-III, MIMIC-IV, and eICU) show that HypeMed consistently achieves strong overall performance and outperforms state-of-the-art baselines on key accuracy metrics while maintaining competitive safety. Ablation and case studies further verify the complementary contributions of MedRep and SimMR and highlight the benefit of representation–retrieval consistency in utilizing retrieved references.

Nevertheless, similar to prior methods [43, 48], HypeMed may recommend slightly more medications than the ground truth, leading to suboptimal performance on the #Med. metric. In addition, embedding-based retrieval can occasionally introduce less relevant visits, especially when the current visit is under-specified. In future work, we plan to improve retrieval robustness (e.g., by incorporating stronger relevance filtering or uncertainty-aware retrieval) and enhance prescription compactness to further strengthen the clinical applicability of HypeMed.

Acknowledgments

This research was supported by the Public Computing Cloud of Renmin University of China and by the Fund for Building World-Class Universities (Disciplines) at Renmin University of China.

References

- [1] Suman Bhoi, Mong Li Lee, Wynne Hsu, Hao Sen Andrew Fang, and Ngaiap Chuan Tan. 2021. Personalizing medication recommendation with a graph-based approach. *ACM Transactions on Information Systems (TOIS)* 40, 3 (2021), 1–23.
- [2] David Blumenthal, Elizabeth J Fowler, Melinda Abrams, and Sara R Collins. 2020. Covid-19—implications for the health care system. , 1483–1488 pages.
- [3] Paula Braveman and Laura Gottlieb. 2014. The social determinants of health: it’s time to consider the causes of the causes. *Public health reports* 129, 1_suppl2 (2014), 19–31.
- [4] Lee R Brooks, Geoffrey R Norman, and Scott W Allen. 1991. Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General* 120, 3 (1991), 278.
- [5] Sherry-Ann Brown. 2016. Patient similarity: emerging concepts in systems and precision medicine. *Frontiers in physiology* 7 (2016), 561.
- [6] Qianyu Chen, Xin Li, Kunnan Geng, and Mingzhong Wang. 2023. Context-Aware Safe Medication Recommendations with Molecular Graph and DDI Graph Embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 7053–7060.
- [7] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems* 29 (2016).
- [8] R Scott Evans. 2016. Electronic health records: then, now, and in the future. *Yearbook of medical informatics* 25, S 01 (2016), S48–S61.
- [9] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 3558–3565.
- [10] Jean Louis Frossard and Florent Bonvin. 2011. Charcot’s triad. *International journal of emergency medicine* 4, 1 (2011), 18.
- [11] Fan Gong, Meng Wang, Haofen Wang, Sen Wang, and Mengyue Liu. 2021. SMR: medical knowledge graph embedding for safe medicine recommendation. *Big Data Research* 23 (2021), 100174.
- [12] Hanze Guo, Jianxun Lian, and Xiao Zhou. 2026. Why not Collaborative Filtering in Dual View? Bridging Sparse and Dense Models. *ACM Transactions on Information Systems* (2026).
- [13] Hanze Guo, Yijun Ma, and Xiao Zhou. 2025. Sorex: Towards self-explainable social recommendation with relevant ego-path extraction. *ACM Transactions on Information Systems* (2025).
- [14] Hanze Guo, Jing Yao, Xiao Zhou, Xiaoyuan Yi, and Xing Xie. 2025. Counterfactual Reasoning for Steerable Pluralistic Value Alignment of Large Language Models. *arXiv preprint arXiv:2510.18526* (2025).
- [15] Tengfei He and Xiao Zhou. 2025. MotifGPL: Motif-Enhanced Graph Prototype Learning for Deciphering Urban Social Segregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 28079–28087.

- [16] Qian Hong, Cheng Bian, Xiao Zhou, Xiaoyu Li, Yelei Li, and Zijing Zeng. 2025. Lost in Time? A Meta-Learning Framework for Time-Shift-Tolerant Physiological Signal Transformation. *arXiv preprint arXiv:2511.21500* (2025).
- [17] Qian Hong, Siyuan Chang, and Xiao Zhou. 2026. WED-Net: A Weather-Effect Disentanglement Network with Causal Augmentation for Urban Flow Prediction. *arXiv preprint arXiv:2601.22586* (2026).
- [18] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. 2017. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology* 2, 4 (2017).
- [19] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data* 10, 1 (2023), 1.
- [20] 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. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.
- [21] Taeri Kim, Jiho Heo, Hongil Kim, Kijung Shin, and Sang-Wook Kim. 2024. VITA: ‘Carefully Chosen and Weighted Less’ Is Better in Medication Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8600–8607.
- [22] Hung Le, Truyen Tran, and Svetha Venkatesh. 2018. Dual memory neural computer for asynchronous two-view sequential learning. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1637–1645.
- [23] Dongjin Lee and Kijung Shin. 2023. I’m me, we’re us, and i’m us: Tri-directional contrastive learning on hypergraphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 8456–8464.
- [24] Jinlin Li, Xintong Li, and Xiao Zhou. 2025. FAP-CD: Fairness-driven age-friendly community planning via conditional diffusion generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 28168–28176.
- [25] Lei Li and Xiao Zhou. 2025. Leave no one Behind: Enhancing Diversity While Maintaining Accuracy in Social Recommendation. In *International Conference on Database Systems for Advanced Applications*. Springer, 51–67.
- [26] Lei Li, Xiao Zhou, and Zheng Liu. 2025. R2med: A benchmark for reasoning-driven medical retrieval. *arXiv preprint arXiv:2505.14558* (2025).
- [27] Xintong Li, Haoran Zhang, and Xiao Zhou. 2025. Spatio-Temporal Hierarchical Causal Models. *arXiv preprint arXiv:2511.20558* (2025).
- [28] Xingwang Li, Yijia Zhang, Xiaobo Li, Hao Wei, and Mingyu Lu. 2023. Dgcl: Distance-wise and graph contrastive learning for medication recommendation. *Journal of Biomedical Informatics* 139 (2023), 104301.
- [29] Yijun Ma, Chaozhuo Li, and Xiao Zhou. 2024. Tail-steak: Improve friend recommendation for tail users via self-training enhanced knowledge distillation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 8895–8903.
- [30] Nir Menachemi and Taleah H Collum. 2011. Benefits and drawbacks of electronic health record systems. *Risk management and healthcare policy* (2011), 47–55.
- [31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
- [32] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. 2018. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific data* 5, 1 (2018), 1–13.
- [33] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine learning* 85 (2011), 333–359.
- [34] Steva Rumsey, Joel Winders, and Andrew D MacCormick. 2017. Diagnostic accuracy of Charcot’s triad: a systematic review. *ANZ journal of surgery* 87, 4 (2017), 232–238.
- [35] Deepti Saraswat, Pronaya Bhattacharya, Ashwin Verma, Vivek Kumar Prasad, Sudeep Tanwar, Gulshan Sharma, Pitshou N Bokoro, and Ravi Sharma. 2022. Explainable AI for healthcare 5.0: opportunities and challenges. *IEE Access* 10 (2022), 84486–84517.
- [36] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks* 20, 1 (2008), 61–80.
- [37] Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. 2019. Gamenet: Graph augmented memory networks for recommending medication combination. In *proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1126–1133.
- [38] Julie Solomon. 2006. Case Studies: why are they important? *Nature Clinical Practice Cardiovascular Medicine* 3, 11 (2006), 579–579.
- [39] Hongda Sun, Shufang Xie, Shuqi Li, Yuhan Chen, Ji-Rong Wen, and Rui Yan. 2022. Debaised, Longitudinal and Coordinated Drug Recommendation through Multi-Visit Clinic Records. *Advances in Neural Information Processing Systems* 35 (2022), 27837–27849.
- [40] Jialun Wu, Yuxin Dong, Zeyu Gao, Tieliang Gong, and Chen Li. 2023. Dual attention and patient similarity network for drug recommendation. *Bioinformatics* 39, 1 (2023), btad003.
- [41] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-Supervised Graph Learning for Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (, Virtual Event, Canada,) (SIGIR ’21). Association for Computing Machinery, New York, NY, USA, 726–735. <https://doi.org/10.1145/3404835.3462862>
- [42] Jialun Wu, Xinyao Yu, Kai He, Zeyu Gao, and Tieliang Gong. 2024. PROMISE: A pre-trained knowledge-infused multimodal representation learning framework for medication recommendation. *Information Processing & Management* 61, 4 (2024), 103758.
- [43] Rui Wu, Zhaopeng Qiu, Jiacheng Jiang, Guilin Qi, and Xian Wu. 2022. Conditional generation net for medication recommendation. In *Proceedings of the ACM Web Conference 2022*. 935–945.
- [44] Lianghao Xia, Chao Huang, Yong Xu, Jiashu Zhao, Dawei Yin, and Jimmy Huang. 2022. Hypergraph contrastive collaborative filtering. In *Proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval*. 70–79.

- [45] Xin Xia, Hongzhi Yin, Junliang Yu, Qinyong Wang, Lizhen Cui, and Xiangliang Zhang. 2021. Self-supervised hypergraph convolutional networks for session-based recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 4503–4511.
- [46] Chaoqi Yang, Cao Xiao, Lucas Glass, and Jimeng Sun. 2021. Change matters: Medication change prediction with recurrent residual networks. *arXiv preprint arXiv:2105.01876* (2021).
- [47] Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. 2021. SafeDrug: Dual Molecular Graph Encoders for Safe Drug Recommendations. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021*.
- [48] Nianzu Yang, Kaipeng Zeng, Qitian Wu, and Junchi Yan. 2023. Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning. In *Proceedings of the ACM Web Conference 2023*. 4075–4085.
- [49] Xixian Yong, Jianxun Lian, Xiaoyuan Yi, Xiao Zhou, and Xing Xie. 2025. MOTIVEBENCH: How Far Are We From Human-Like Motivational Reasoning in Large Language Models?. In *Findings of the Association for Computational Linguistics: ACL 2025*. 20059–20089.
- [50] Xixian Yong, Peilin Sun, Ziheng Wang, and Xiao Zhou. 2026. Intelli-Planner: Towards Customized Urban Planning via Large Language Model Empowered Reinforcement Learning. *arXiv preprint arXiv:2601.21212* (2026).
- [51] Xixian Yong, Xiao Zhou, Yingying Zhang, Jinlin Li, Yefeng Zheng, and Xian Wu. [n. d.]. Think or Not? Exploring Thinking Efficiency in Large Reasoning Models via an Information-Theoretic Lens. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- [52] Junliang Yu, Hongzhi Yin, Jundong Li, Qinyong Wang, Nguyen Quoc Viet Hung, and Xiangliang Zhang. 2021. Self-supervised multi-channel hypergraph convolutional network for social recommendation. In *Proceedings of the web conference 2021*. 413–424.
- [53] Haoran Zhang, Youjin Wang, Yi Duan, Rong Fu, Dianyu Zhao, Sicheng Fan, Shuaishuai Cao, Wentao Guo, and Xiao Zhou. 2026. Social-JEPA: Emergent Geometric Isomorphism. *arXiv preprint arXiv:2603.02263* (2026).
- [54] Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart, and Jimeng Sun. 2017. LEAP: learning to prescribe effective and safe treatment combinations for multimorbidity. In *proceedings of the 23rd ACM SIGKDD international conference on knowledge Discovery and data Mining*. 1315–1324.
- [55] Yingying Zhang, Xian Wu, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2023. Knowledge-enhanced attributed multi-task learning for medicine recommendation. *ACM Transactions on Information Systems* 41, 1 (2023), 1–24.
- [56] Xiao Zhou, Zhongxiang Zhao, and Hanze Guo. 2025. Tricolore: Multi-Behavior User Profiling for Enhanced Candidate Generation in Recommender Systems. *IEEE Transactions on Knowledge and Data Engineering* (2025).
- [57] Guangpu Zhu, Xiaoyu Li, Xin Sui, Jie Hawes, Ramy Hussein, Qian Hong, Xiao Zhou, Zijing Zeng, and Yelei Li. 2025. Hypertension Risk Screening Using Long-Term Photoplethysmogram and Ballistocardiograph Measurements from a Smartwatch. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 1–6.
- [58] Yanxu Zhu, Shitong Duan, Xiangxu Zhang, Jitao Sang, Peng Zhang, Tun Lu, Xiao Zhou, Jing Yao, Xiaoyuan Yi, and Xing Xie. 2025. MoHoBench: Assessing Honesty of Multimodal Large Language Models via Unanswerable Visual Questions. *arXiv preprint arXiv:2507.21503* (2025).