BMC Medical Informatics and
Decision Making

RESEARCH

KAMPNet: multi-source medical knowledge augmented medication prediction network with multi-level graph contrastive learning

Yang An[1], Haocheng Tang[2], Bo Jin[3*], Yi Xu[2*] and Xiaopeng Wei[4]

Abstract

Backgrounds Predicting medications is a crucial task in intelligent healthcare systems, aiding doctors in making informed decisions based on electronic medical records (EMR). However, medication prediction faces challenges due to complex relations within heterogeneous medical data. Existing studies primarily focus on the supervised mining of hierarchical relations between homogeneous codes in medical ontology graphs, such as diagnosis codes. Few studies consider the valuable relations, including synergistic relations between medications, concurrent relations between diseases, and therapeutic relations between medications and diseases from historical EMR. This limitation restricts prediction performance and application scenarios.

Methods To address these limitations, we propose KAMPNet, a multi-sourced medical knowledge augmented medication prediction network. KAMPNet captures diverse relations between medical codes using a multi-level graph contrastive learning framework. Firstly, unsupervised graph contrastive learning with a graph attention network encoder captures implicit relations within homogeneous medical codes from the medical ontology graph, generating knowledge augmented medical code embedding vectors. Then, unsupervised graph contrastive learning with a weighted graph convolutional network encoder captures correlative relations between homogeneous or heterogeneous medical codes from the constructed medical codes relation graph, producing relation augmented medical code embedding vectors. Finally, the augmented medical code embedding vectors, along with supervised medical code embedding vectors, are fed into a sequential learning network to capture temporal relations of medical codes and predict medications for patients.

Results Experimental results on the public MIMIC-III dataset demonstrate the superior performance of our KAMPNet model over several baseline models, as measured by Jaccard, F1 score, and PR-AUC for medication prediction.

Conclusions Our KAMPNet model can effectively capture the valuable relations between medical codes inherent in multi-sourced medical knowledge using the proposed multi-level graph contrastive learning framework. Moreover, The multi-channel sequence learning network facilitates capturing temporal relations between medical codes, enabling comprehensive patient representations for downstream tasks such as medication prediction.

Keywords Electronic medical records, Medication prediction, Intelligent healthcare system, Multi-source medical knowledge, Graph contrastive learning

*Correspondence:
Bo Jin
jinbo@dlut.edu.cn
Yi Xu
yi.xu@ia.ac.cn
Full list of author information is available at the end of the article

Background

The availability of immense accumulation of electronic medical records (EMR) data, coupled with advancements in deep computational methods, has provided a solid foundation for intelligent healthcare applications, including disease risk prediction [1–3] and medication prediction task [4–6]. Among these applications, the prediction of medications for patients plays a crucial role in assisting doctors in making efficient clinical decisions, thereby enabling more time for doctor-patient communication and improving the quality of medical services. Thus, it will be conducive to improving the medical service quality. Consequently, there has been a growing demand for deep learning-based medication prediction models.

However, the majority of existing methods are not specifically tailored to address scenarios where multiple medical experts collaborate in joint consultations for patients. In such situations, various types of medical knowledge, including common-sense medical ontology and empirical medical knowledge derived from historical EMR data, are taken into consideration during the medication decision-making process. As a result, conventional methods often yield suboptimal performance in these complex decision-making scenarios. Therefore, effectively capturing the intricate and diverse relationships between medical codes from multi-source medical knowledge to enhance medication prediction becomes a highly challenging yet significant task.

As depicted in Fig. 1, the hierarchical structures inherent in the diagnosis ontology graph and medication ontology graph (representing common sense medical domain knowledge) imply relationships between homogeneous medical codes, which will contribute to the representation learning. The existing studies, such as GRAM [7] and KAME [8] utilize the diagnosis ontology graph to enhance the representations of diagnosis codes by incorporating information from relational medical codes using supervised methods. Meanwhile, Shang et al. [5] propose G-BIRT, combining a graph neural network and bidirectional encoder representation from transformers (BERT), to enhance medical code representations through a pretraining approach. These models consider the inherent relations between homogeneous medical codes in medical domain ontology graphs in a supervised or self-supervised manner to augment the medical code representations. Furthermore, MK-GNN [9] and COGNet [10] predominantly emphasize the incorporation of medical code relations, particularly between drug codes, using EHR graph or DDI graph. However, they tend to overlook concurrent relationships between diseases and the heterogeneous therapeutic relations between medications and diseases, as present in historical EMR data.
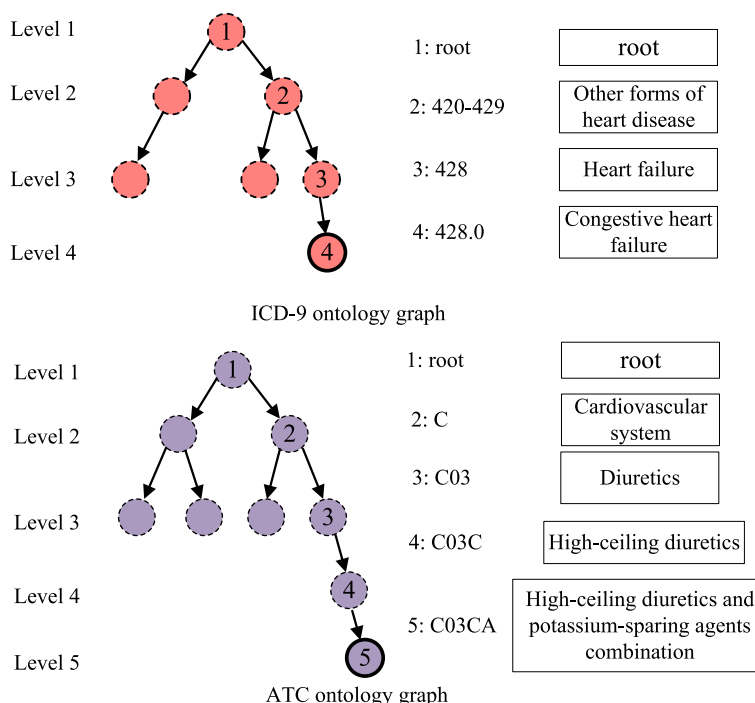

Fig. 1 Illustration of Medical ontology graph: ICD-9 ontology for diagnosis and ATC ontology for medications. For each diagnosis or medication, we can obtain a unique parent path containing hierarchical ontology concepts from the root to the leaf levels. The ontology graph will be utilized to cultivate the inherent relations of homogeneous codes

However, a limitation arises when attempting to transfer such forms of learned representations from one predictive task to another, such as from diagnosis prediction to medication prediction, despite using the same dataset. This necessitates repetitive model training to obtain medical code representations for each different downstream task. Moreover, purely supervised methods are ineffective in acquiring medical code representations when the downstream task is unknown. Therefore, there is an urgent need to develop a novel unsupervised method for learning the embeddings of medical codes based on the medical ontology graph. This approach would facilitate downstream tasks in diverse clinical scenarios, alleviating the limitations associated with the transferability and applicability of learned medical code representations.

Furthermore, the aforementioned models primarily focus on capturing the inherent relationships among homogeneous medical codes within the hierarchical structures of medical ontology graphs. However, they tend to overlook the valuable correlative relationships between both homogeneous and heterogeneous medical codes that are implicitly present in historical EMR data. This historical EMR data is typically regarded as a valuable source of empirical medical knowledge. For instance, clinicians commonly administer multiple medications simultaneously to patients to enhance the therapeutic effect, indicating the presence of synergistic relationships between medications. Additionally, major diseases frequently co-occur with inevitable concurrent diseases or symptoms, highlighting the concurrent relationships

between diseases. Moreover, in prescriptions, medications are prescribed for specific diseases or symptoms, reflecting therapeutic causal relationships between medications and diseases. Unfortunately, only a limited number of studies have explicitly represented and captured these meaningful relationships hidden within EMR data for medication prediction.

To address the aforementioned limitations and enhance the performance of medication prediction, we present a novel multi-sourced medical knowledge augmented network, named KAMPNet. The network leverages multi-level graph contrastive learning to capture the diverse relations between homogeneous or heterogeneous medical codes and improve their representations. The main workflow of medication prediction using the proposed model is depicted in Fig. 2: Firstly, similar to the existing method G-BIRT [5], we consider the inherent common-sense medical ontology graph (medical domain knowledge, shown in Fig. 1) to capture the local relations between medical codes. Additionally, we construct a medical codes relation graph, shown in Fig. 4, based on the co-occurrence of diagnosis codes and medication codes in a single visit from historical EMR data. This graph allows us to capture the global relations between medical codes. Secondly, to address the problems of label dependency and repetitive model training in existing ontology graph representation learning methods, we incorporate an improved graph contrastive learning framework based on DGI [11], which can provide better representations for downstream tasks without supervised
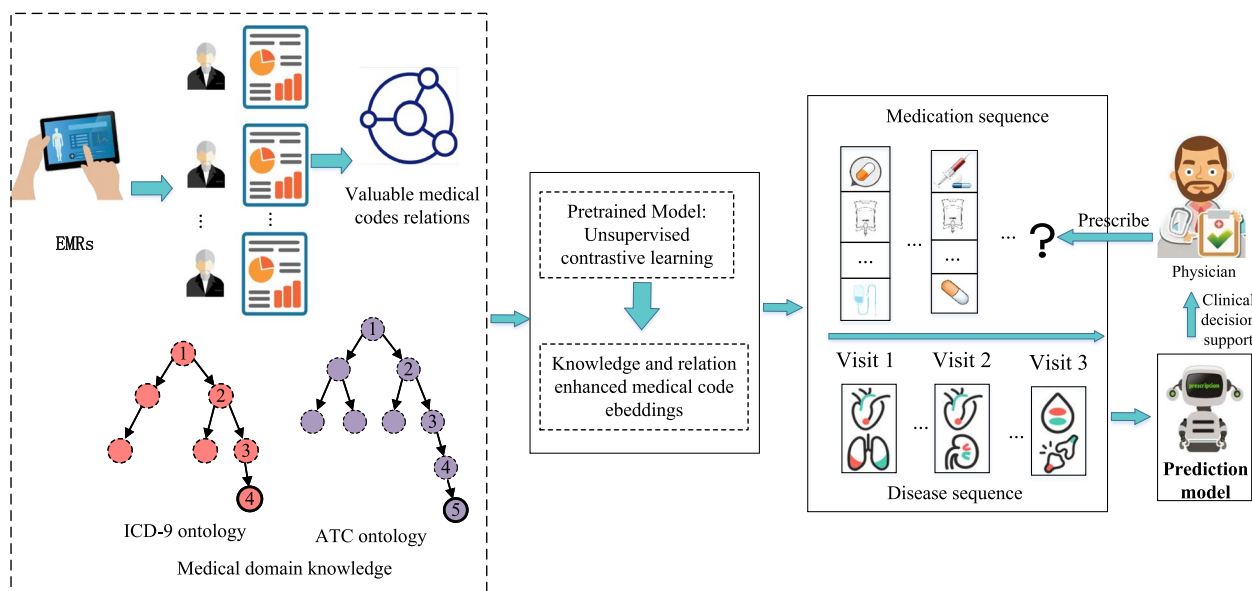

Fig. 2 Main workflow of medication prediction using proposed model. It consists of three main steps: medical knowledge-related graph construction, unsupervised contrastive learning for medical code embeddings, and supervised sequential learning for medication prediction task

labels and can facilitate to improve the model robustness [12]. This unsupervised learning approach enables us to obtain representations of medical codes from the multi-source knowledge graph. By infusing the information from relational medical codes, the augmented medical code representations are mutually enhanced, capturing both the local relations from the ontology graph and the global relations from the medical codes relation graph. Finally, the retrieved augmented medical code representations are fed into a supervised sequential learning model to capture the temporal relations between medical codes. This enables us to obtain a comprehensive patient representation, which can be utilized for medication prediction and assist in clinical decision-making.

In summary, the technical contributions of this paper are as follows: (1) We propose KAMPNet, a multi-source medical knowledge augmented medication prediction network, which incorporates a multi-level graph contrastive learning framework. To the best of our knowledge, our model is the first to capture valuable relations between medical codes and augment their representations using a cascaded unsupervised approach on both ontology graphs and constructed medical codes relation graph. (2) We integrate different graph encoders, such as the graph attention network and weighted graph convolutional network, into the multi-level graph contrastive learning framework to consider the meaningful relation weights between heterogeneous or homogeneous medical codes. (3) We present a sequential learning network that combines the multi-source embedding vectors of medical codes into the patient representation, enabling the capture of temporal relations between medical codes for medication prediction.

Related works
Deep learning in medication prediction
Medication prediction is a significant application of deep learning in intelligent healthcare systems, garnering considerable attention from researchers due to its practical importance. Researchers [4] have categorized medication prediction algorithms into instance-based and longitudinal sequential prediction methods.

Instance-based methods primarily focus on capturing the nonlinear relations between diagnosed disease status and the prescribed medications. For example, Zhang et al. [13] formulate the medication problem as a sequential decision-making problem and employ recurrent neural networks to encode the label dependency. Wang et al. [14] propose three linear models that combine multi-source patient information, including demographic data, laboratory indicators, and diagnosis outcomes, for personalized medication prediction. Wang et al. [15] transform the medication prediction task into

an unordered Markov decision process, predicting prescription medications step by step. However, these methods overlook the critical temporal information present in historical EMR data, leading to suboptimal prediction performance.

Nowadays, longitudinal sequential prediction models that consider the temporal relations between historical medical records have become prevalent in medication prediction tasks. Jin et al. [16] present three different heterogeneous LSTM models to capture the interaction between heterogeneous temporal sequence data and incorporate two heterogeneous sequence information into the patient representation to predict the next stage of treatment medications. Shang et al. [4] incorporate sequential information, including diagnosis and procedure information, through multi-channel sequence learning networks to learn comprehensive patient representations for medication prediction. DMNC [17] and AMANet [6] integrate attention networks to capture the interactions between procedure and diagnosis sequences and model sequential dependencies. MeSIN [18] models both temporal dependencies between sequential medical records and the relations between hierarchical sequences for medication prediction. However, these models primarily focus on mining inherent relations between multiple medical sequences in EMR data and often neglect the empirical knowledge implicit in EMR data and external common-sense medical knowledge.

Graph neural networks in healthcare applications
Graph neural networks (GNNs) [19] have emerged as effective frameworks for graph representation learning. GNNs leverage a neighbourhood aggregation mechanism to recursively aggregate and transform the representation vectors of adjacent nodes, effectively utilizing the topological relationships between graph nodes and enhancing the representation ability of nodes [20, 21]. As a result, GNNs have been widely used in biological and health informatics to model valuable relationships between multiple entities [4, 22–26]. The learned medical code embedding vectors in health informatics can be augmented to facilitate downstream predictive tasks. For example, GRAM [7] and KAME [8] leverage the diagnosis ontology graph to enhance the representations of diagnosis codes by incorporating information from relational diagnosis codes in the ontology graph using attention mechanisms. Zhang et al. [27] and Ye et al. [24] incorporate medical knowledge into sequential networks to enhance representation learning and obtain interpretable disease risk prediction results. Lu et al. [26] employ a patient-disease bipartite graph to create a weighted patient network (WPN) for learning robust patient representations in chronic disease prediction. Shang et al. [5]

An *et al. BMC Medical Informatics and Decision Making*   (2023) 23:243

Page 5 of 19

and Wang et al. [28] combine graph neural networks with bidirectional encoder representation from transformers (BERT) to capture inherent relationships between homogeneous medical codes in medical ontology graphs. Liu et al. [29] use temporal medical event graphs to represent complex relationships among different types of medical information for next-period prescription prediction. Mao et al. [30] construct medical graphs using historical EMR data to incorporate information from correlative entities into code representation for medication prediction and lab test imputation. Su et al. [31] construct dynamic co-occurrence graphs for each patient admission record and employ a graph-attention augmented sequential network to model inherent structural and temporal information for medication prediction. However, these models do not fully leverage the significant relations between medical codes due to a lack of global graph representation learning and the extraction of only local sub-graphs from the global guidance relation graph.

Although GNN-based models have achieved good performance in various tasks, they still face certain challenges. For instance, obtaining supervised labels can be complex and laborious, and models may require retraining for new tasks or feature changes. To address these issues, researchers have explored novel training methods for GNNs, such as graph self-supervised learning. Graph self-supervised learning is an unsupervised graph representation learning approach that relies solely on the topology and node information of the graph itself, without depending on explicit labels. For instance, Petar [11] incorporates Deep InfoMax [32] into the graph learning domain and models a general self-supervised learning framework based on the mutual information maximization. Similarly, Kaveh et al. [33] train the graph by maximizing the representation graph encoding of different graph structure perspectives. In the biomedical domain, Sun et al. [34] propose a novel molecular graph contrastive learning framework that incorporates local and global domain knowledge to enhance graph representation learning. Therefore, inspired by the advantages of contrastive learning methods, we incorporate a graph contrastive learning framework with two different graph encoders to learn the embeddings of medical codes from the medical ontology graph and medical codes relation graph in an unsupervised manner.

Materials and methods
Dataset and dataset preprocessing
To validate the effectiveness of the proposed KAMP-Net, in this paper, we conduct the experiments on a publicly available dataset MIMIC-III [35] which is a large, freely-available database comprising deidentified health-related information associated with over 40,000

patients who were admitted to critical care units of the Beth Israel Deaconess Medical Center over the 12-year period between 2001 and 2012 and had relatively complete multi-sourced EMR. In particular, the dataset contains various kinds of heterogeneous information, such as diagnosed diseases, treatment medications and patient demographics etc., which satisfy the data requirement of our model. And we mainly extract useful information from tables PRESCRIPTIONS and DIAGNOSES_ICD.

The critical selection process of the experimental cohort is shown in Fig. 3. Firstly, the patients in the experimental cohort should have at least one complete visit record consisting of diagnosed diseases and treatment medications. Specifically, the patients with only one hospitalized visit record are mainly utilized to construct the medical knowledge graphs, including the ontology and relation graphs. While the patients with multiple hospitalized visit records are divided into training and testing sets and are harnessed for the sequence learning of KAMPNet. It is worth noting that the patients in the training set are also used to assist in constructing the multi-sourced medical knowledge graph.

Besides, similar to [4], the medications prescribed by doctors for each patient within the first 24 h are selected as the medication set since it belongs to a crucial period for each patient to get rapid and accurate treatment [36]. In addition, the medicine codes from NDC are transformed to ATC Level 3 for integration with MIMIC-III, and predicting category information not only guarantees the sufficient granularity of all the diagnoses but also improves the training speed, and predictive performance [7, 37]. Table 1 provides more information about the patient cohort from the dataset.

Proposed model: KAMPNet
The proposed model KAMPNet mainly consists of three main substructures: medical knowledge-related graph construction, unsupervised contrastive learning for obtaining the enhanced medical code embeddings, and a supervised sequence learning network for medication prediction task.

Medical graph construction

Medical ontology graph construction The hierarchical structures of the diagnosis classification system ICD-9 and medication classification system ATC imply the meaningful relations between medical codes, which have been constructed as the medical domain knowledge in previous studies [5, 7]. Similarly, the leaf nodes of the tree structure based graph comprise the medical codes of history EMR, and the non-leaf nodes mainly come from the medical codes classification system ICD-9 or ATC. By the above existing approach, the ICD-9 ontology graph
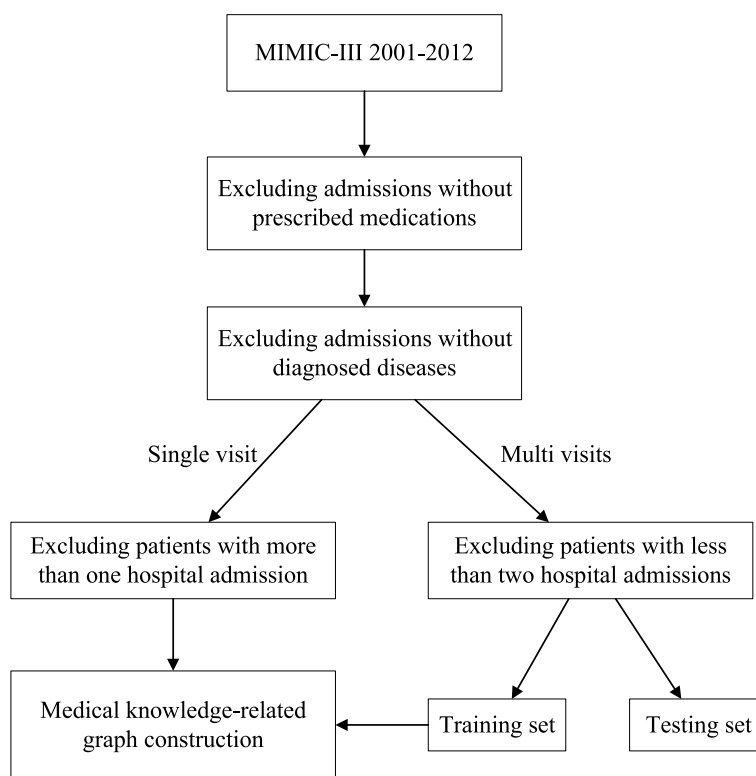
Fig. 3 The flow chart of patients selection

Table 1 Statistics of the MIMIC-III datasets

MIMIC III	Quantity
# of patients (Single visit)	30745
avg # of diagnosis	39
avg # of medication	52
# of unique diagnosis	1997
# of unique medication	323
# of patients (Multi visit)	6350
avg # of diagnosis	10.16
avg # of medication	7.33
avg # of visits	2.36
# of unique diagnosis	1958
# of unique medication	145

$\mathcal{G}_d = \{\mathcal{V}_d, \mathcal{E}_d\}$ consisting of diagnosed disease codes and the hierarchical relations, the ATC ontology graph $\mathcal{G}_m = \{\mathcal{V}_m, \mathcal{E}_m\}$ consisting of treatment medication codes and the hierarchical relations are constructed respectively. And the unified indication of the above different medical ontology graphs is represented as $\mathcal{G}_* = \{\mathcal{V}_*, \mathcal{E}_*\}$, where $\mathcal{V}_*$ is the set of graph nodes (medical codes), and $\mathcal{E}_*$ is the set of edges (the relations of medical codes). Among them, the set of medical codes $\mathcal{C}_*$ constitutes the leaf nodes of medical ontology graphs, and the set of all the graph nodes of $\mathcal{G}_*$ satisfies $\mathcal{V}_* = \mathcal{C}_* \cup \mathcal{C}'$, where $\mathcal{C}'$ denotes the set of non-leaf nodes.

Medical codes relation graph construction In the actual clinical application scenario, patients will generally be diagnosed with a variety of diseases and given a variety of treatment medications during a single medical treatment, which also indirectly illustrates that there exist specific relations between diseases, medications, and between diseases and medications in the electronic medical records data. Therefore, based on the historical EMRs data generated by patients during each visit, and inspired by the dynamic weighted graph built on the history purchase records [38], we will construct the medical codes relation graph based on the co-occurrent medical codes including diagnosis codes and medication codes.

Figure 4 shows the detailed process of the medical codes relation graph construction based on the history EMR of patients. Since the patient with a single hospitalized visit has only one visit record, and the patient with multiple hospitalized visits has multiple visit records, in this paper, the construction of the medical codes relation graph does not consider the temporal sequence information but only considers the implicit relation between medical codes
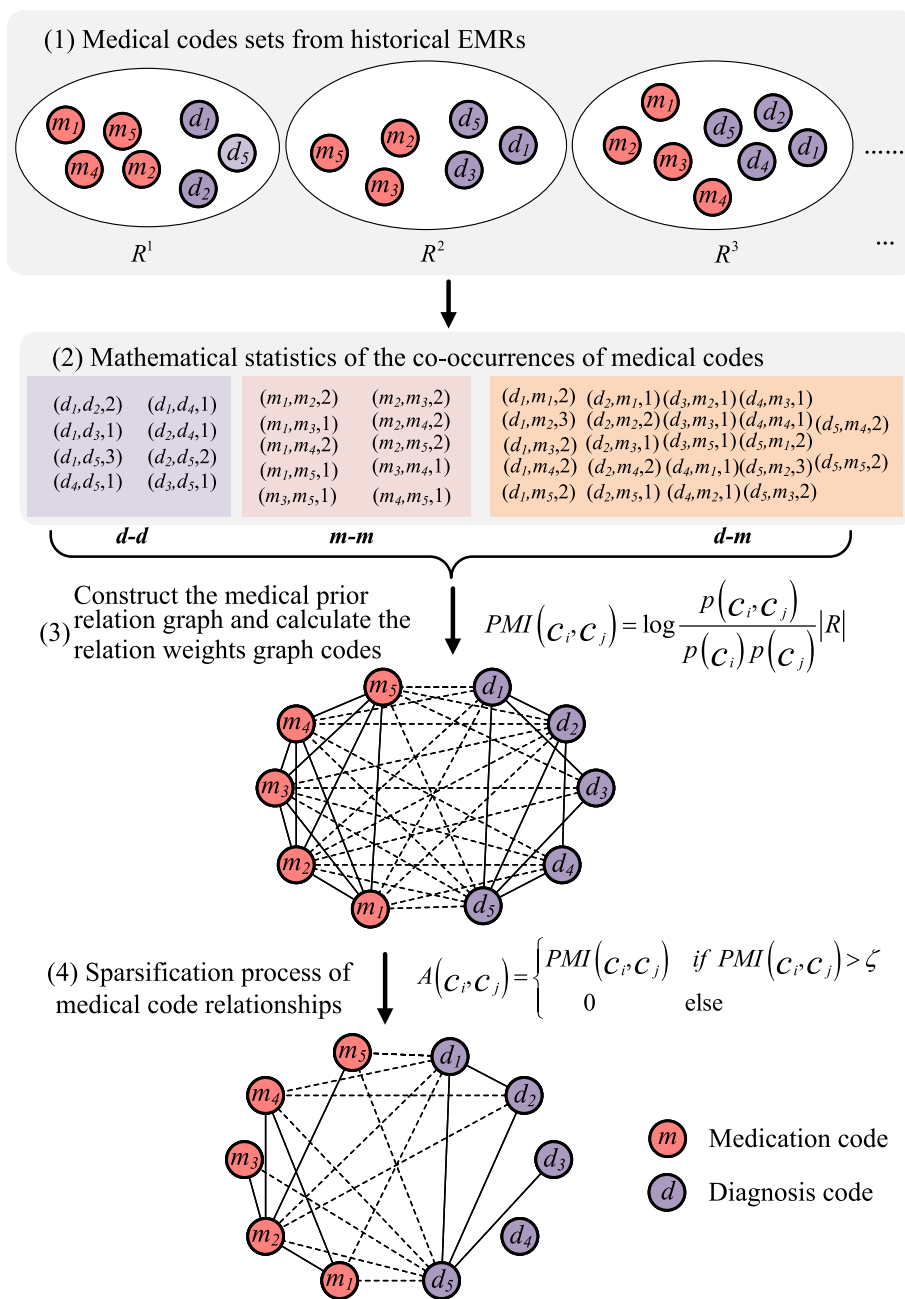
Fig. 4 Construction process of medical codes relation graph

of history EMR data. Therefore, the visit-based medical record extracted from the patient's history EMR is represented as R^i, and every hospitalized visit could produce a diagnosis code and medication code. As shown in Fig. 4, there are three historical records: R^1, R^2, R^3, and each record R^i includes diagnosis code m_* and medication code d_*. Based on the visit records shown in Fig. 4 (1), the corresponding medical code pairs in each visit record can be generated such as $(d_1, d_2), (d_1, m_1), (m_1, m_2), \cdots$.

After mathematical statistics, the generated medical code pairs and the number of co-occurrence are shown in Fig. 4 (2). There are three implicit relations: the concurrent relation between diseases $d - d$, the synergistic relation between medications $m - m$, and the therapeutic relation between diseases and medication $d - m$. Then, considering that the relations between medical codes are not only related to the number of co-occurrence but also related to the frequency of medical codes, the pointwise

An *et al. BMC Medical Informatics and Decision Making*   (2023) 23:243

Page 8 of 19

mutual information (PMI) [39] commonly used in natural language processing to measure the relevance of words is introduced to calculate the relation weights between medical codes:

$$
\begin{aligned}
PMI(c_i, c_j) &= \log \frac{P(c_i, c_j)}{P(c_i)P(c_j)} \\
&= \log \frac{\frac{p(c_i, c_j)}{|\boldsymbol{R}|}}{\frac{p(c_i)}{|\boldsymbol{R}|} \frac{p(c_j)}{|\boldsymbol{R}|}} \\
&= \log \frac{p(c_i, c_j)}{p(c_i)p(c_j)} |\boldsymbol{R}|,
\end{aligned} \tag{1}
$$

where $|\boldsymbol{R}|$ denotes the total number of visit records of all patients in the training dataset, $P(c_i, c_j)$ and $p(c_i, c_j)$ respectively denotes the probability and number of co-occurrence of medical codes c_i and c_j in single visit record, $P(c_i)$ and $p(c_i)$ indicates the probability and number of occurrence of medical code c_i in visit records, $P(c_j)$ and $p(c_j)$ indicates the probability and number of occurrence of medical code c_j in visit records. The mutual information between the above medical code pairs is for all the code pairs, including $d - d, m - m$ and $d - m$ forms, without distinguishing whether they are homogeneous medical codes or heterogeneous medical codes. The main reason is that the relation degree between the medical code pairs is obtained by calculating statistical mutual information, which belongs to the scope of quantitative analysis without involving any medical background. Therefore, the qualitative relation is ignored in the quantitative calculation, even though the heterogeneity between medical codes does exist in reality.

Then, as shown in Fig. 4 (3), the medical codes relation graph $\mathcal{G}_r = \{\mathcal{V}_r, \mathcal{E}_r, \mathcal{W}_r\}$ can be built through the triplets obtained by the previous step, where $\mathcal{V}_r$ represents the graph nodes set, also called the medical codes set $\mathcal{C} = [\mathcal{C}_d, \mathcal{C}_m]$, $\mathcal{E}_r$ is the graph edges set, $\mathcal{W}_r$ represents the relation weight. Moreover, the heterogeneity between diagnosis code and medication code would be neglected for the reason that the relation strength between medical codes is the critical consideration factor in this paper. The medical code pairs obtained in Fig. 4 (2) determine the edges between nodes in the relation graph. Thus, the edge weights in adjacency matrix $\boldsymbol{A}$ of the medical codes relation graph $\mathcal{G}_r$ can be further calculated as follows:

$$
A(c_i, c_j) = \begin{cases} PMI(c_i, c_j) & \text{if } PMI(c_i, c_j) > \zeta, \\ 0 & \text{else.} \end{cases} \tag{2}
$$

Different from the calculation method in [31], here, we incorporate a graph sparsity factor $0 < \zeta < PMI_{max}$ in Eq. (2), which aims to mitigate the effects of noise that might be introduced by relying solely on statistical

quantitative computation method and ignoring medical expertise. Moreover, the relation graph belongs to a symmetric matrix, i.e., $A(c_j, c_i) = A(c_i, c_j)$. In this way, the ultimately complete medical codes relation graph is obtained as shown in Fig. 4 (4) In the end, another initial representation method of medical codes relation graph $\mathcal{G}_r$ can be obtained, i.e., $\mathcal{G}_r = (\mathcal{C}, \boldsymbol{A})$.

Unsupervised contrastive learning on medical ontology graphs

To avoid the dependency on the labels and make the learned medical codes representations directly applied in the downstream tasks such as medication prediction, the unsupervised graph contrastive learning method based on DGI [11] are incorporated to learn the medical codes representations by maximizing the mutual information of graph representation in the following two cascaded sections.

As illustrated in Fig. 5, (1) and (2) respectively describe the introduced unsupervised contrastive learning processes on the medication code ontology graph and diagnosis code ontology graph. Such a learning process consists of four critical substructures: graph data augmentation, GNN-based encoder, graph pooling layer, and contrastive loss function.

Similar to previous knowledge-enhanced methods presented in G-BIRT [5], the medical codes c_i of ontology graph $\mathcal{G}_*$ can be randomly assigned an initialized embedding vector $\boldsymbol{v}_i$ which could be optimized and updated through a learned embedding matrix $\boldsymbol{W}_e \in \mathbb{R}^{\mathcal{C}_* \times d}$, where d indicates the dimension of the medical code embedding vector. Then the constructed medical ontology graph $\mathcal{G}_* = \{\mathcal{V}_*, \mathcal{E}_*\}$ can be indicated as $(\boldsymbol{X}_*, \boldsymbol{A}_*)$, where $\boldsymbol{X}_*$ represents the set of initialized medical codes representations of the ontology graph $\mathcal{G}_*$ and $\boldsymbol{A}_*$ represents the corresponding adjacency matrix.

Graph data augmentation   Then as shown in Fig. 5 (1) and (2), the initialized constructed medical ontology graph $(\boldsymbol{X}_*, \boldsymbol{A}_*)$ undergoes graph data augmentation to obtain the negative sample via corruption function such as the graph nodes permutation. In detail, the $\mathcal{C}$ is used to randomly perturb the nodes without changing the topology structure of the ontology graph; namely, the adjacency matrix $\boldsymbol{A}_*$, which aims to obtain the augmented ontology graph: $(\hat{\boldsymbol{X}}_*, \hat{\boldsymbol{A}}_*) = \mathcal{C}(\boldsymbol{X}_*, \boldsymbol{A}_*)$.

GNN-based encoder   In fact, there should be no restrictions on the choice of graph encoders. Here, to fully utilize the inherent topological hierarchical structure of the medical ontology graph for further capturing the implicit relations between medical codes, we directly incorporate the graph attention network (GAT) [40] used in previous study G-BIRT [5] as the graph encoder $\mathcal{F}$ to obtain
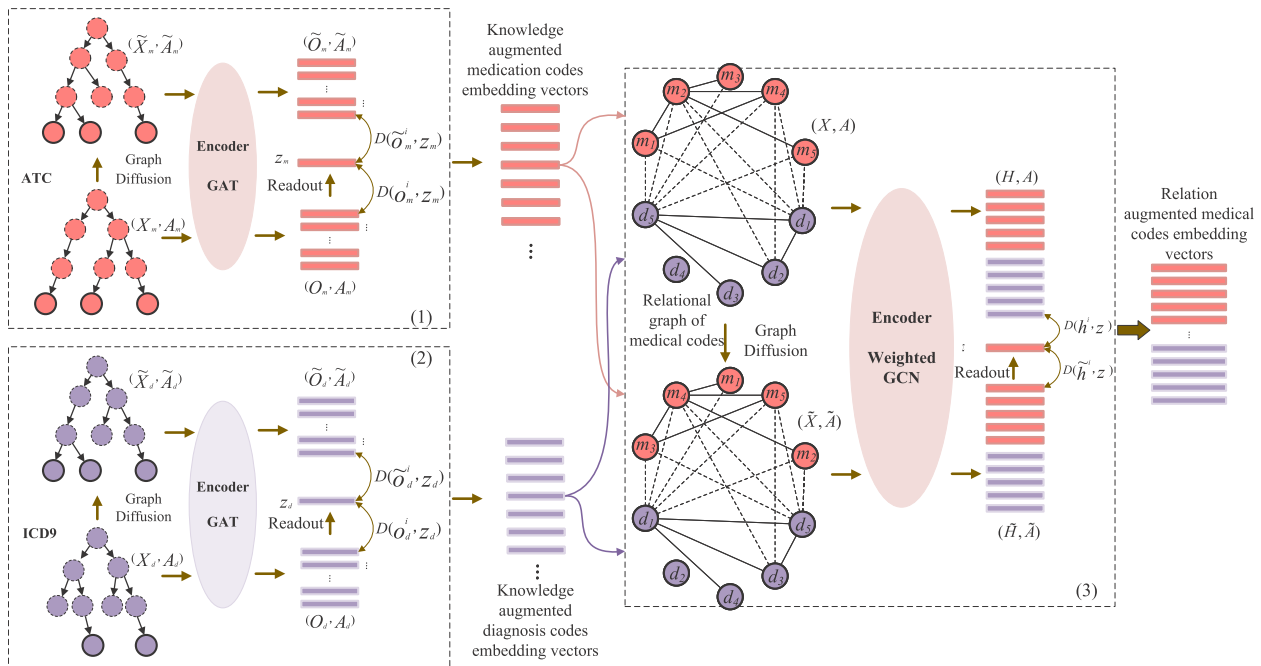
Fig. 5 Multi-level unsupervised contrastive learning framework. It consists of three different contrastive learning processes: (1) Contrastive learning on medication code ontology graph; (2) Contrastive learning on diagnosis code ontology graph; (3) Contrastive learning on medical codes relation graph

the medical codes embedding representations of medical ontology graphs.

Given the medical ontology graph $\mathcal{G}_*$ is a hierarchical structural graph with a parent-child substructure, the relation between medical codes can be captured from two different paths. That is, on the one hand, the medical codes corresponding to the parent nodes should infuse the information corresponding to the medical code of the child node; on the other hand, the embedding representation information corresponding to the medical code of the parent node also should be transmitted to the leaf node. For each non-leaf node in the medical domain knowledge graph $c' \in \mathcal{C}'$ and the leaf node $c_* \in \mathcal{C}_*$, the corresponding augmented embedding representations $\boldsymbol{v}_{c_*} \in \mathbb{R}^d$ and $\boldsymbol{h}_{o_*} \in \mathbb{R}^d$ can be computed as follows:

$$
\begin{aligned}
\boldsymbol{v}_{c'} &= f(c', ch(c'), \boldsymbol{W}_e), \\
\boldsymbol{o}_{c_*} &= f(c_*, pa(c_*), \boldsymbol{V}_e),
\end{aligned}
\tag{3}
$$

where $f(\cdot, \cdot, \cdot)$ indicates the graph information aggregation function, $ch(c')$ is a function to extract all direct child nodes of non-leaf medical code c', while $pa(c_*)$ represents a function which can extract all parent nodes of the leaf medical code c_*. The above method realizes the two-way information transmission (from top to bottom and from bottom to top) through the hierarchical structure.

In this way, the implicit relations between the medical codes can be fully captured; that is, the medical codes representations are augmented, which can further alleviate the sudden decline of prediction accuracy caused by the insufficient learning problems of the tail codes in electronic medical records.

Due to the graph information aggregation in above two-way information transmission process requires considering the relation difference between medical codes, in this paper, we incorporate the graph attention network (GAT) [40] as the aggregation function $f(\cdot, \cdot, \cdot)$, and it is also the shared graph encoder of the graph contrastive learning framework for calculating the graph nodes representations. Specifically, the representation vectors of each medical code c_i is computed using the graph attention network aggregation function $f(\cdot, \cdot, \cdot)$ as follows:

$$
f(c_i, p(c_i), \boldsymbol{H}_e) = \|_{k=1}^{K} \sigma \left(\sum_{j \in \mathcal{N}_i} \alpha_{i,j}^k \boldsymbol{W}^k \boldsymbol{v}_j \right),
\tag{4}
$$

where $\mathcal{N}_i = \{c_i\} \cup pa(c_i)$ denotes the first order neighborhood nodes set of medical code c_i in the medical ontology graph, $\|$ is the concatenation operation of the embedding representations computed by multi-head attention, K is the number of attention heads, σ is the nonlinear activation function Sigmoid, $\boldsymbol{W}^k \in \mathbb{R}^{m \times d}$ is the transformation matrix to be learned, where $m = d/k$.

While, $\alpha_{i,j}^k$ indicates the k-th standardized relevance score, which can be calculated as follows:

$$\alpha_{i,j}^k = \frac{e^{\left(\text{LeakyReLU}\left(\boldsymbol{a}^{\mathsf{T}}\left[\boldsymbol{W}^k\boldsymbol{v}_i \| \boldsymbol{W}^k\boldsymbol{v}_j\right]\right)\right)}}{\sum_{k \in \mathcal{N}_i} e^{\left(\text{LeakyReLU}\left(\boldsymbol{a}^{\mathsf{T}}\left[\boldsymbol{W}^k\boldsymbol{v}_i \| \boldsymbol{W}^k\boldsymbol{v}_k\right]\right)\right)}} \quad (5)$$

where $\boldsymbol{a} \in \mathcal{R}^{2m}$ is the learned weight parameter. And we propose to use LeakyReLU [41] as the nonlinear activation function. The reason is Leaky ReLU maintains a small slope (typically a small positive value, such as 0.01) for negative inputs, unlike traditional ReLU which has a gradient of zero for negative inputs and the occurrence of "dead" or "dying" neurons. This non-zero gradient property is beneficial as it provides a continuous and differentiable activation function, allowing gradient-based optimization methods to be applied effectively during backpropagation. The continuous gradient enables smoother and more stable learning, aiding convergence and improving the overall training process.

Therefore, we can utilize the above graph encoder $\mathcal{F}$ to respectively compute the embedding representations of the medical codes from medical ontology graph $\mathcal{G}_*$ and augmented medical ontology graph $\hat{\mathcal{G}}_*$ as follows:

$$\begin{aligned}\boldsymbol{O}_{c_*} &= \mathcal{F}(\mathcal{V}_*, \mathcal{E}_*) = \{\boldsymbol{o}_{c_*^1}, \boldsymbol{o}_{c_*^2}, ..., \boldsymbol{o}_{c_*^{|\mathcal{C}|}}\}, \\ \hat{\boldsymbol{O}}_{c_*} &= \mathcal{F}(\hat{\mathcal{V}}_*, \hat{\mathcal{E}}_*) = \{\hat{\boldsymbol{o}}_{c_*^1}, \hat{\boldsymbol{o}}_{c_*^2}, ..., \hat{\boldsymbol{o}}_{c_*^{|\mathcal{C}|}}\},\end{aligned} \quad (6)$$

where $\boldsymbol{O}_{c_*}$ and $\hat{\boldsymbol{O}}_{c_*}$ respectively denotes the embedding representations sets of medical codes from $\mathcal{G}_*$ and $\hat{\mathcal{G}}_*$.

Graph pooling layer   The graph pooling layer is mainly leveraged to compute the global feature vector $\boldsymbol{z}_*$ through readout function $\mathcal{R}$:

$$\boldsymbol{z}_* = \mathcal{R}(\boldsymbol{O}_{c_*}) = \frac{1}{|\mathcal{C}_*|} \sum_{i=1}^{|\mathcal{C}_*|} \boldsymbol{o}_{c_*^i}, \quad (7)$$

where $\boldsymbol{O}$ indicates the unified symbol of embedding representation of medical codes of the medical ontology graph (diagnosis and medication).

Contrastive loss function   In order to train the graph encoder end-to-end and learn to obtain the informative medical codes embedding vectors and the medical ontology graph embedding representation, we still utilize the maximization of mutual information [32] between medical code embedding vector $\boldsymbol{o}_{c_*^i}$ and the medical ontology graph embedding representation $\boldsymbol{z}_*$ as the objective loss function. First, the negative and positive sample pairs $(\boldsymbol{o}_{c_*^i}, \boldsymbol{z}_*)$ and $(\hat{\boldsymbol{o}}_{c_*^i}, \boldsymbol{z}_*)$ can be obtained through graph encoder and graph pooling layer. Then, the discriminator $\mathcal{D}$ is introduced to score the positive and negative sample pairs:

$$\begin{aligned}\mathcal{D}(\boldsymbol{o}_{c_*^i}, \boldsymbol{z}_*) &= \sigma(\boldsymbol{z}_*^{\mathsf{T}} \boldsymbol{W}_D \boldsymbol{o}_{c_*^i}), \\ \mathcal{D}(\hat{\boldsymbol{o}}_{c_*^i}, \boldsymbol{z}_*) &= \sigma(\boldsymbol{z}_*^{\mathsf{T}} \boldsymbol{W}_D \hat{\boldsymbol{o}}_{c_*^i}).\end{aligned} \quad (8)$$

Finally, the overall objective function is to maximize the mutual information in the form of JS divergence as follows:

$$\begin{aligned}\mathcal{L} = &\frac{1}{|\mathcal{C}_*|} \sum_{i=1}^{|\mathcal{C}_*|} \mathbb{E}_{(\mathcal{V}, \mathcal{E})}\left[\log \mathcal{D}\left(\boldsymbol{o}_{c_*^i}, \boldsymbol{z}_*\right)\right] \\ &+ \frac{1}{|\mathcal{C}_*|} \sum_{j=1}^{|\mathcal{C}_*|} \mathbb{E}_{(\tilde{\mathcal{V}}, \tilde{\mathcal{E}})}\left[\log\left(1 - \mathcal{D}\left(\hat{\boldsymbol{o}}_{c_*^i}, \boldsymbol{z}_*\right)\right)\right]\end{aligned} \quad (9)$$

Through the above detailed graph contrastive learning on medical ontology graphs, we can obtain the knowledge augmented medical codes embedding vectors $\boldsymbol{O}_c$ including knowledge augmented medication codes embedding vectors $\boldsymbol{O}_{c_m}$ and knowledge augmented diagnosis codes embedding vectors $\boldsymbol{O}_{c_d}$ (as shown in Fig. 5).

Unsupervised contrastive learning on medical relation graph
Factually, the obtained medical codes embedding vectors have infused the relational information from correlative medical codes, which can provide the initial embedding vectors for the nodes of the medical codes relation graph in this section. However, the medical ontology graphs do not explicitly provide the practically meaningful relations between medical codes and the implicit relations between homogeneous medical codes in the hierarchical topology structure (belonging to a local relation). Therefore, in this section, the medical codes relation graph constructed on the basis of history EMR can directly models the explicit relations (belonging to a global relation) between homogeneous and heterogeneous medical codes through quantitative statistics and then uses the graph contrastive learning framework shown in Fig. 5 (3) to learn the relation augmented embedding vectors of the medical codes of the medical codes relation graph. The learning process will fully leverage the global relation between medical codes to model the interrelations of the medical codes embedding vectors.

As shown in Fig. 5 (3), the unsupervised contrastive learning on medical codes relation graph also includes four critical steps. First, the initialization vector of the corresponding node of medical codes from the medical codes relation graph can be retrieved directly from the obtained knowledge augmented medical code embedding vectors set $\boldsymbol{O}_c$, and denoted as $\boldsymbol{X} \subset \boldsymbol{O}_c$. Then, the medical codes relation graph can be indicated as $\mathcal{G}_r = (\mathcal{C}, \boldsymbol{A}) : (\boldsymbol{X}, \boldsymbol{A})$. Subsequently, we can obtain the negative graph sample through graph nodes perturbation, namely augmented medical codes relation graph

$(\widetilde{X}, \widetilde{A}) \curvearrowleft (X, A)$. After that, the incorporated shared graph encoder is incorporated to encode the above two relation graphs. Different from the selected graph encoder GAT, the specially weighted graph convolutional network (GCN) [20] is incorporated as the graph encoder in the graph contrastive learning framework in this section to infuse the relation weights between medical codes. Taking the medical codes relation graph (X, A) as an example, we calculate the corresponding embedding vectors as follows:

$$H = \hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}X\Theta \qquad (10)$$

where $\hat{A} = A + I, I$ is the identity matrix that aims to avoid the information loss caused by the small number of neighborhood nodes. The diagonal matrix $\hat{D}_{ii} = \sum_{j=0} A_{ij}$ is used to normalize the weights of the connected edges of each node from the relation graph according to the edge weights from the adjacency matrix $\hat{A}$. $\Theta \in \mathbb{R}^{d_x \times d_h}$ is the learned network parameter. Considering the edge weights, the weighted graph encoder realizes the mutual infusion of correlative medical codes information in the medical codes embedding vectors learning process according to the relevance degree. Similarly, with the help of the above graph encoder weighted GCN, we can further encode the negative graph sample, i.e., augmented medical codes relation graph $(\widetilde{X}, \widetilde{A})$, and obtain the corresponding medical codes embedding matrix $(\widetilde{H}, \widetilde{A})$ (shown in Fig. 5 (3)). Afterwards, the readout function $\hat{\mathcal{R}}$ is used to compute the global embedding vector z of medical codes relation graph $H = h_1, h_2, \ldots, h_{|\mathcal{C}|}$:

$$z = \hat{\mathcal{R}}(O_{c_*}) = \frac{1}{|\mathcal{C}|}\sum_{i=1}^{|\mathcal{C}|} h_i \qquad (11)$$

Finally, the constructed objective function based on the discriminant equation is introduced to optimize the graph contrastive learning process for the medical codes relation graph. Similar to the contrastive learning framework for medical ontology graph, we still utilize the maximization of mutual information [32] between the embedding representation vector h_i of medical code on the medical codes relation graph and the global relation graph embedding vector z. Firstly, the negative and positive sample pairs (h_i, z) and $(\hat{h}_i, z)$ can be obtained through graph encoder and graph pooling layer. Then, the discriminator $\mathcal{D}$ is introduced to score the positive and negative sample pairs:

$$\begin{aligned} \mathcal{D}(h_i, z) &= \sigma(z^T W_D h_i) \\ \mathcal{D}(\hat{h}_i, z) &= \sigma(z^T W_D \hat{h}_i). \end{aligned} \qquad (12)$$

Then, the overall objective function is to maximize the mutual information in the form of JS divergence as follows:

$$\begin{aligned} \mathcal{L} &= \frac{1}{|\mathcal{C}|}\sum_{i=1}^{|\mathcal{C}|} \mathbb{E}_{(X,A)}[\log \mathcal{D}(h_i, z)] \\ &+ \frac{1}{|\mathcal{C}|}\sum_{j=1}^{|\mathcal{C}|} \mathbb{E}_{(\widetilde{X}, \widetilde{A})}\left[\log\left(1 - \mathcal{D}\left(\hat{h}_i, z\right)\right)\right]. \end{aligned} \qquad (13)$$

After continuous optimization and iterative calculation, the final relation augmented medical codes embedding representation vectors as shown in Fig. 5 (3) can be obtained, which, in the end, not only integrates the information from globally correlative heterogeneous and homogeneous medical codes embedding vectors in the medical codes relation graph but also infuses the information from locally correlative homogeneous medical codes embedding vectors implied in the medical domain knowledge graphs (the medical ontology graphs).

Sequential learning network for medication prediction

As illustrated in Fig. 6, the sequential medication prediction framework comprise three critical substructures, i.e., multi-sourced medical codes are embedding vectors fusion, sequence learning network, and the comprehensive medication prediction.

Multi-sourced medical codes embedding vectors fusion The knowledge augmented medical codes embedding vectors $O_c = \{O_{c_m}, O_{c_d}\}$ and the relation augmented medical codes embedding vectors H are respectively obtained through corresponding graph contrastive learning network. While for each specific prediction task, it has different task-specific requirements for the medical codes embedding vectors. Therefore, in the medication prediction task, the learned medical codes embedding representation vector $e^i \in \mathbb{R}^d$ that can be supervised and fine-tuned in the specific prediction task is also incorporated:

$$e^i = W^e c^i, \qquad (14)$$

where $W^e \in \mathbb{R}^{|\mathcal{C}| \times d}$ is the medical codes embedding matrix to be learned, $\mathcal{C} = [\mathcal{C}_d, \mathcal{C}_m]$ is the union set of medical codes, and $\mathcal{C}_d$ and $\mathcal{C}_m$ respectively denote the sets of diagnosis codes and medication codes. In this way, we can get the dense embedding vector matrix of medical codes that can be learned, i.e. $E = [e^1, \ldots, e^{|\mathcal{C}|}] \in \mathbb{R}^{|\mathcal{C}| \times d}$.

Each hospitalized visit will generate different medical codes corresponding to different diagnoses and medications. The corresponding medical codes embedding vectors could be retrieved from the obtained medical codes
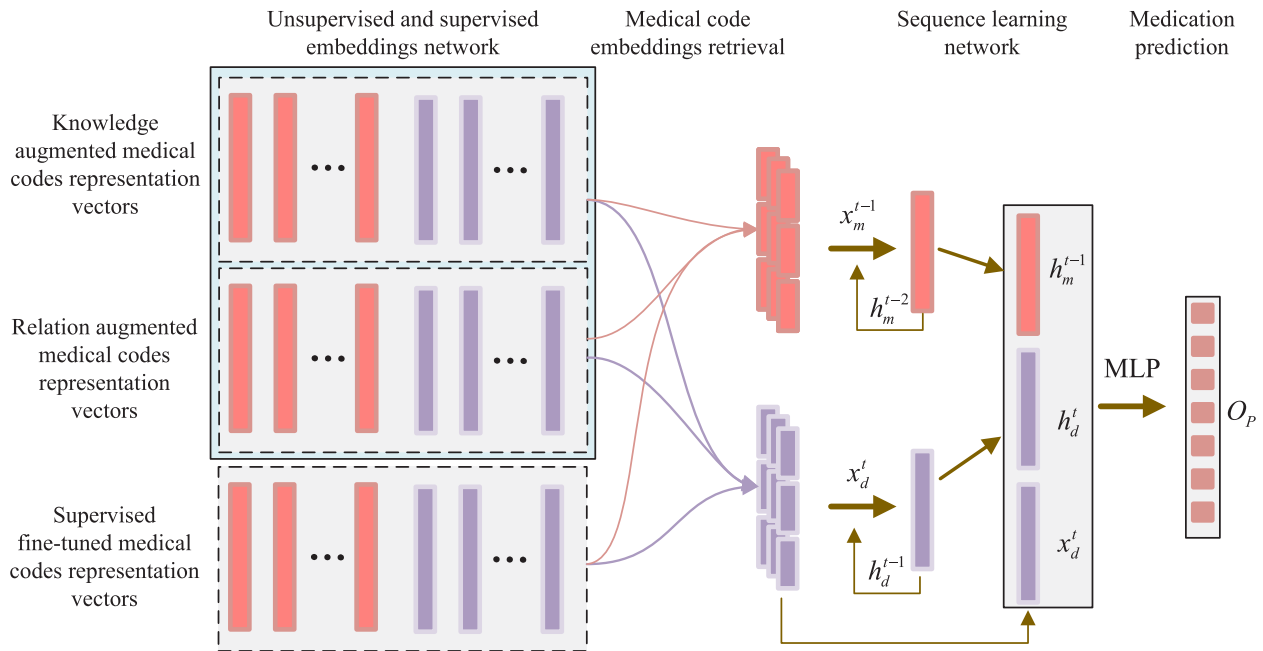
Fig. 6 Medication prediction framework

embedding vectors sets, including knowledge augmented medical codes embedding vectors set O_c, relation augmented medical codes embedding vectors set H, and the supervised learning medical codes embedding vectors set E. For instance, we can first retrieve the corresponding embedding representations of medical codes produced in i-th visit record from the medical codes embedding vectors sets, i.e. the medical diagnosis codes embedding representations including $O_d^i \subset O_c$, $H_d^i \subset H$ and $E_d^i \subset E$, the medical medication codes embedding representations including $O_m^i \subset O_c$, $H_m^i \subset H$ and $E_m^i \subset E$. Then the corresponding mean values of medical codes representation vectors sets are calculated to obtain the visit-level diagnosis codes embedding vectors containing o_d^i, h_d^i and e_d^i, and medication codes embedding vectors containing o_m^i, h_m^i and e_m^i; Finally, the calculated embedding vectors are respectively concatenated together to obtain the input of diagnosis codes sequence learning network and medication sequence learning network, i.e. $x_d^i = [o_d^i, h_d^i, e_d^i]$ and $x_m^i = [o_m^i, h_m^i, e_m^i]$.

Sequence learning network When patient possesses multiple hospitalized visit records and needs to predict the treatment medications at current timestamp t, firstly, we require integrating described multi-sourced medical codes embedding vectors together at history timestamps, including the history diagnosis codes embedding representation sequence $\{x_d^1, x_d^2, \ldots, x_d^t\}$ and the history medication codes embedding representation

sequence $\{x_m^1, x_m^2, \ldots, x^{(t-1)}_m\}$; afterwards, the temporal sequential learning network such as the recurrent neural networks (RNNs) are respectively utilized to capture the temporal dependencies of sequential medical codes as follows:

$$
\begin{aligned}
h_m^t &= RNN_m(x_m^1, x_m^2, \ldots, x_m^{t-1}), \\
h_d^t &= RNN_d(x_d^1, x_d^2, \ldots, x_d^t).
\end{aligned}
\tag{15}
$$

Medication prediction The hidden state vectors h_m^t and h_d^t that incorporates the history information are obtained through the sequence learning network (Eq. (15)). However, considering the importance of current diagnosis information for the medication prediction at the current timestamp, the current diagnosis code embedding vector x_d^t is also integrated into the patient representation. Therefore, the comprehensive patient representation vector O_P can be calculated as follows:

$$
O_P = W_P \cdot [h_m^t, h_d^t, x_d^t],
\tag{16}
$$

where $x_d^t \in \mathbb{R}^{d+d_o+d_h}$, $h_m^t, h_d^t \in \mathbb{R}^{2d}$, and $W_P \in \mathbb{R}^{(5d+d_o+d_h) \times d_e}$ is the parameter to be learned. According to the comprehensive patient representation vector O_P, the current treatment medication $\hat{y}_t^m$ can be predicted as follows:

$$
\hat{y}_t^m = \text{softmax}(W_O \cdot O_P + b_o),
\tag{17}
$$

where $\hat{\boldsymbol{y}}_t^m$ denotes the predicted multi-label medications set. $\boldsymbol{W}_O \in \mathbb{R}^{|\mathcal{C}_m| \times d_e}$ and $\boldsymbol{b}_o \in \mathbb{R}^{|\mathcal{C}_m|}$ are the parameters to be learned, where $\mathcal{C}_m$ denotes the medication codes set, and $|\mathcal{C}_m|$ is the size of set.

Due to the medication prediction task belonging to the domain of sequential multilabel prediction, we utilize the binary cross-entropy loss $\mathcal{L}$ as the objective function. According to the prediction result $\hat{\boldsymbol{y}}_t^m$ at each timestamp t and the real label $\boldsymbol{y}_t^m$, the predictive function binary cross-entropy loss is formulated as follows:

$$\mathcal{L} = -\frac{1}{T-1}\sum_{t=2}^{T}\boldsymbol{y}_t^m\log\sigma\left(\hat{\boldsymbol{y}}_t^m\right) + (1-\boldsymbol{y}_t^m)\log\left(1-\sigma\left(\hat{\boldsymbol{y}}_t^m\right)\right) \tag{18}$$

Experiments
Experimental details
The experiment details include three parts: evaluation metrics, benchmark models and experimental setting.

Evaluation metrics
To evaluate the performance of the proposed KAMPNet, the Jaccard similarity score (Jaccard), precision-recall AUC (PR-AUC), and average F1 (F1) are adopted as the evaluation metrics. In practice, KAMPNet cannot wholly replace doctors and only screen possible medications as much as possible to assist physicians in prescribing medications for patients. Therefore, Jaccard should be one of the appropriate evaluation metrics. And it is defined as the size of the intersection divided by the size of the union of the predicted set $\hat{Y}_t^m$ and ground truth set y_t^m as follows:

$$\text{Jaccard} = \frac{1}{\sum_i^N \sum_t^{T_i} 1}\sum_i^N \sum_t^{T_i}\frac{\left|Y_t^i \cap \hat{Y}_t^i\right|}{\left|Y_t^i \cup \hat{Y}_t^i\right|}, \tag{19}$$

where T_i is the number of visits for the ith patient, and N denotes the number of patients in the test set. And recall can be utilized to measure the predicted medications' completeness:

$$\text{Recall} = \frac{1}{\sum_i^N \sum_t^{T_i} 1}\sum_i^N \sum_t^{T_i}\frac{\left|Y_t^i \cap \hat{Y}_t^i\right|}{\left|Y_t^i\right|} \tag{20}$$

Furthermore, for the medication prediction task, due to the number of positive and negative labels being imbalanced, the precision-recall curve utilized to calculate the PR-AUC has proved to be an appropriate evaluation metric.

In addition, the predicted medications' correctness should also be evaluated. Thus, the evaluation metric

F1 is incorporated to evaluate the multilabel classification task comprehensively. Firstly, precision is generally adopted to measure the prediction correctness and can be calculated as follows:

$$\text{Precision} = \frac{1}{\sum_i^N \sum_t^{T_i} 1}\sum_i^N \sum_t^{T_i}\frac{\left|Y_t^i \cap \hat{Y}_t^i\right|}{\left|\hat{Y}_t^i\right|} \tag{21}$$

Then, the metric F1 can be computed as follows:

$$\text{F1} = \frac{1}{\sum_i^N \sum_t^{T_i} 1}\sum_i^N \sum_t^{T_i}\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{22}$$

Benchmark models
To verify the superiority of the proposed model KAMP-Net on medication prediction tasks, we compare it with the following baseline methods used for medication prediction, which include one machine learning-based method and six deep learning-based algorithms:

- *LR* [42]. It is a logistic regression with L1/L2 regularization. We sum the multi-hot vector of each visit together and apply the binary relevance technique [42] to handle the multilabel output.
- *Retain* [43]. RETAIN is an interpretable model with a two-level reverse time attention mechanism to predict diagnoses, which can detect significant past visits and associated clinical variables. It can be used for similar sequential prediction tasks, such as predicting treatment medicines.
- *LEAP* [13]. Leap formulates the medicine prediction problem as a multi-instance multilabel learning problem, mainly using a recurrent neural network (RNN) to recommend medicines.
- *GRAM* [7]. It utilizes the diagnosis ontology graph to enhance the diagnosis code representation by infusing the information from relational medical codes in a supervised method.
- *GAMENet* [4]. It employs a dynamic memory network to save encoded historical medication information, and further utilizes a query representation formed by encoding sequential diagnosis and procedure codes to retrieve medications from the memory bank.
- *G-BIRT* [5]. G-BIRT combines the graph neural network and bidirectional encoder representation from transformers (BERT) to enhance medical code representations through a pre-training method.
- *GATE* [31]. GATE constructs the dynamic co-occurrence graph at each admission record for every

patient. It then introduces a graph-attention augmented sequential network to model the inherent structural and temporal information for medication prediction.

Experimental settings

To ensure the rationality of experimental verification, the pretreated clinical patients' EMR data are still randomly divided into training, validation, and test sets with 2/3 : 1/6 : 1/6 ratios and the experimental results are the average values across five runs of random grouping and training. Moreover, the dimension of hidden layers and hyper-parameters in KAMPNet is set as follows: in the unsupervised graph contrastive learning framework for medical ontology graph, the dimension of hidden layers are set to 128, the node embedding size of graph encoder GAT is set to 32, the attention head is set to 4; in the unsupervised graph contrastive learning framework for medical codes relation graph, the dimension of hidden layers and the node embedding size of encoder GCN are all set to 64; the dimension of hidden layers in temporal sequential learning network is set to 256. In addition, the training is performed using Adam [44] at a learning rate of 5e-4, and we report the model performance in the test set within 40 epochs. All methods are trained on a Windows with 11GB memory and an Nvidia 2080Ti GPU using the deep learning computation platform Pytorch 1.6.

Discussion
Discussion of prediction results

As demonstrated in Table 2, the treatment medication prediction results show that the performance of our proposed method KAMPNet is better than the existing state-of-the-art predictive models in health informatics in most cases. In detail, compared with baseline model LEAP, KAMPNet achieves about 10.52%, 9.46%,

13.4% higher performance concerning Jaccard, F1, and PR-AUC, respectively. We think the prominent reason might be that LEAP models the medication prediction problem as an instance-based medication prediction process which directly neglects the temporal dependency and does not consider the importance of implicit domain knowledge. The medication prediction performance of Retain is relatively better than LEAP, which could be attributed to its two-level attention-based model, which can capture the temporal relations and the relation between input features and output labels. Such a two-level attention-based sequential prediction model makes Retain's performance even better than the GRAM model, which firstly introduces the hierarchical knowledge into the healthcare predictive model through the attention mechanism.

Secondly, the similarity between GAMENet and KAMPNet is that the obtained diagnoses sequence and treatment medications sequence all utilize a sequence learning model to capture the temporal dependencies between medical codes. The difference is that GAMENet does not construct the relations between medical codes directly but constructs the medication-visit graph for capturing the indirect relations between medications for later retrieval using the attention mechanism. In other words, GAMENet considers modelling the co-occurrence relations between medications, while our KAMPNet takes the multiple relations between heterogeneous or homogeneous medical codes into consideration. Therefore, on the medication prediction task, our KAMPNet outperforms GAMENet by 4.18%, 3.28%, 3.41% on Jaccard, F1, PR-AUC, respectively. Unlike GAMENet, G-BIRT does not directly construct the relation between medical codes. Instead, it enhances the relation between medical codes using a pre-training method on the medical ontology graphs, including diagnosis code ontology graph and medication code ontology graph, which can make full use of the EMR data of single hospitalized patients. However, medical code embedding representation learning in G-BIRT mainly relies on the medication or diagnosis labels provided in history EMR and ignores the inherent relations between medical codes implied in EMR data. Although GATE also considers the relations between medical codes by building the co-occurrence graphs for each patient from the global guidance relation graph, it still neglects the infusion of valuable information of correlative medical codes from the medical codes relation graph and relies on the supervised label of the medication prediction task. Therefore, KAMPNet performs better than the latest knowledge-enhanced algorithm G-BIRT and GATE, and its Jaccard, F1 and PR-AUC improve at least by 2.31% on Jaccard, by 1.39% on F1 and by 1.08% on PR-AUC, respectively.

Table 2 Performance comparison on medication prediction task

Methods	Jaccard	F1	PR-AUC
LR	0.4075	0.5658	0.6716
LEAP	0.3921	0.5508	0.5855
Retain	0.4456	0.6064	0.6838
GRAM	0.4176	0.5788	0.6638
GAMENet⁻	0.4401	0.5996	0.6672
GAMENet	0.4555	0.6126	0.6854
G-BIRT	0.4565	0.6152	0.6960
GATE	0.4742	0.6315	0.7087
KAMPNet	**0.4973**	**0.6454**	**0.7195**

The critical reason that our proposed KAMPNet achieves the best performance compared with baseline models could be summarized as follows: (1) With the help of a multi-level unsupervised contrastive learning framework, it can capture the relations between medical codes and augments the medical codes representations based on the medical ontology graphs. (2) Then, the relations between medical codes implicit in the constructed medical code relations graph are further captured to learn more informative medical codes embedding representation vectors, contributing to the downstream tasks such as medication prediction. (3) The incorporated sequential learning network can further combine the supervised medical codes representations with the learned knowledge and relation augmented medical codes representations and then captures the temporal relations between medical codes for downstream medication prediction task.

Ablation study on model components

To verify the effectiveness of the critical components of KAMPNet and analyze their influence on the performance of medication prediction tasks, we conduct an ablation study to explore further the necessity of the proposed model components in our multi-level graph contrastive learning framework on the multi-sourced medical knowledge for the medication prediction task.

As illustrated in Table 3, compared with KAMPNet, the performance of five model variants decreased to varying degrees. We think the reason might be that the lack of model components leads to the failure of effective mining of the valuable relations between medical codes implicit in the multi-sourced medical knowledge. The details of the five variants are as follows:

- KAMPNet$_{RG-}$. The variant KAMPNet$_{RG-}$ does not consider the medical codes relation graph constructed based on the empirical knowledge from the history EMR data. That is, the relation augmented medical codes embedding representation

vectors $\boldsymbol{h}_d^i$ and $\boldsymbol{h}_m^i$ are respectively excluded from the input of the corresponding sequence learning network. the performance of variant KAMPNet$_{RG-}$ decreases by 1.48% on Jaccard, 1.01% on F1, and 0.22% on PR-AUC. The main reason is that the variant considers the locally inherent relations between medical codes in the medical ontology graph and does not consider capturing the valuable global relations between homogeneous or heterogeneous medical codes in the medical codes relation graph.

- KAMPNet$_{HG-}$. The variant KAMPNet$_{HG-}$ does not consider the medical domain knowledge graph, including the diagnosis code ontology graph and medication code ontology graph. That is, the knowledge augmented medical codes embedding representation vectors $\boldsymbol{o}_d^i$ and $\boldsymbol{o}_m^i$ are respectively excluded from the input of the corresponding sequence learning network. The performance of the variant declines by 1.81% on Jaccard, 1.41% on F1, 0.51% on PR-AUC, respectively, which is mainly because the medical codes relation graph can not acquire the meaningful initialization vectors for its nodes from the medical ontology graph. Moreover, the valuable relations between medical codes embodied in the inherent hierarchical structures in medical ontology graphs would not be captured for augmenting the medical codes embedding vectors.

- KAMPNet$_{HGRG-}$. The variant KAMPNet$_{HGRG-}$ concurrently neglects the multi-sourced knowledge augmented medical codes embedding representation vectors and considers the supervised learnable medical codes embedding representations $\boldsymbol{e}_d^i$ and $\boldsymbol{e}_m^i$ in a sequence learning network. It achieves lower by 1.64%, 1.29%, 0.57% than KAMPNet respectively on Jaccard, F1, PR-AUC, but has relatively better performance than KAMPNet$_{HG-}$, which demonstrates that there exists specific noise in the medical ontology graphs. In the future, we will explore how to effectively reduce the adverse effect of noise in multi-sourced medical knowledge.

- KAMPNet$_{R-}$. The variant KAMPNet$_{R-}$ indicates that the medical codes relation graph would not utilize the learned knowledge augmented medical codes embedding representation vectors to initialize the graph nodes (medical codes). It further validates the importance of such implicit information in the multi-level graph contrastive learning framework, which is proposed based on the relations between the medical ontology graph and the medical codes relation graph. It also indirectly explains why the performance of KAMPNet is relatively optimal when using the knowledge augmented medical codes embedding representation vectors as the initialization vector of the nodes of the medical codes relation graph.

Table 3 Performance comparison of the variants of KAMPNet on MIMIC-III dataset

Model	Jaccard	F1	PR-AUC
KAMPNet$_{RG-}$	0.4825	0.6353	0.7173
KAMPNet$_{HG-}$	0.4792	0.6313	0.7144
KAMPNet$_{HGRG-}$	0.4809	0.6325	0.7138
KAMPNet$_{R-}$	0.4866	0.6357	0.7147
KAMPNet$_{RGW-}$	0.4911	0.6396	0.7165
KAMPNet	**0.4973**	**0.6454**	**0.7195**

An *et al. BMC Medical Informatics and Decision Making* (2023) 23:243

Page 16 of 19

- KAMPNet$_{RGW-}$. The variant KAMPNet$_{RGW-}$ considers whether there are relations between medical codes and does not consider the relation weights reflected in the medical codes relation graph. In this way, the performance of variant KAMPNet$_{RGW-}$ decreases compared with presented KAMPNet, which further testifies that the relation weights based on co-occurrence probability obtained via the statistical method in the medical codes relation graph have a positive promoting effect on the performance of medication prediction.

Therefore, the proposed KAMPNet in this paper achieves the best performance only when the model's components complement each other.

Analysis on the graph sparsity factor ζ

In the construction process of the medical codes relation graph, considering the adverse effects of the noise introduced by relying solely on the statistical quantitative computation method and ignoring medical expertise, the graph sparsity factor $0 < \zeta < PMI_{max}$ (Eq. 2) is incorporated in the Eq. 2. This section will explore the effect of graph sparsity factor $0 < \zeta < PMI_{max}$ on medication prediction performance.

Table 4 shows the prediction results of treatment medications when the medical codes relation graph is constructed based on different graph sparsity factors. It can be seen from the table that when $\zeta = 0.07$, the prediction performance of model KAMPNet is the best, and the performance decreases in varying degrees with the increase or decrease of ζ value. When $\zeta < 0.07$, the redundancy of relations between medical codes might result in relatively more noise, which leads to a decline in prediction performance; when $\zeta > 0.07$, with the increasing value of the graph sparsity factor, some meaningful edges with beneficial relation might be neglected due to the sparsity,

which would bring about the incomplete captures of the relations between medical codes and would further cause the decline of model performance. In general, the value of the sparsity factor has a relatively small impact on the prediction results compared with the critical components studied in the Ablation study on model components section. The prominent reason may be that the noise influence in the medical codes relation graph is relatively small, and the valuable relations between medical codes and the augmentation of the embedding representation vectors of medical codes are still dominant. In addition, the final representation of medical code is concatenated by multi-source medical codes embedding representations. The incomplete learning of one source medical code embedding representation can not have a noticeable adverse impact on the prediction performance of the model KAMPNet. Factually, it further shows that the proposed KAMPNet is more robust in predicting treatment medications.

Effects of graph encoders in contrastive learning networks

The multi-level unsupervised contrastive learning framework described in Materials and methods section, including the medical domain knowledge graph-based contrastive learning framework and the medical codes relation graph-based unsupervised contrastive learning framework, possesses respectively different graph encoders because of distinct reasons. In this section, we will analyze the selection of different graph encoders in the graph contrastive learning framework and explore their effects on downstream tasks such as medication prediction results. In Table 5, "Encoder in HG" indicates the applied graph encoder in the contrastive learning framework based on the medical domain knowledge graph (HG), while "Encoder in RG" represents the applied graph encoder in the contrastive learning on the medical codes relation graph (RG). $\sqrt{}$ denotes using the corresponding graph encoder, while $\times$ denotes not using the corresponding graph encoder.

KAMPNet$_{AC}$ indicates that the graph encoders in HG and RG based contrastive learning frameworks are graph attention network (GAT) and weighted graph convolutional network (GCN), respectively, which achieves the relatively optimal prediction result. While in KAMPNet$_{CC}$, the graph encoder in HG based contrastive learning framework is replaced with general GCN, which results in a decrease in the medication prediction performance. We think the main reason is that GAT can aggregate the information of correlative neighbourhood codes to the leaf codes according to the learned relevance scores between connected medical codes, while general GCN does not consider the relation weights between medical codes and aggregate the information of connected (or

Table 4 The effect of graph sparsity factor ζ on model performance

ζ	Jaccard	F1	PR-AUC
0.01	0.4917	0.6398	0.7176
0.02	0.4908	0.6392	0.7163
0.03	0.4895	0.6384	0.7187
0.04	0.4948	0.6431	0.7199
0.05	0.4916	0.6407	0.7176
0.06	0.4924	0.6407	0.7196
0.07	**0.4973**	**0.6454**	0.7195
0.08	0.4922	0.6404	**0.7201**
0.09	0.4933	0.6413	0.7194
0.10	0.4903	0.6382	0.7175

Table 5 The effect on prediction performance of different graph encoders

Model	Encoder in HG		Encoder in RG		Prediction Performance		
	GAT	GCN	GAT	GCN	Jaccard	F1	PR-AUC
KAMPNet$_{AC}$	✓	✗	✗	✓	0.4973	0.6454	0.7195
KAMPNet$_{AA}$	✓	✗	✓	✗	0.4878	0.6381	0.7160
KAMPNet$_{CC}$	✗	✓	✗	✓	0.4869	0.6373	0.7176
KAMPNet$_{CA}$	✗	✓	✓	✗	0.4853	0.6358	0.7122

correlative) medical codes equally. Compared with variant KAMPNet$_{AC}$, variant KAMPNet$_{AA}$ directly uses GAT as the graph encoder, which could relearn the relevance score and does not consider the relation weights representing the empirical knowledge from medical codes relation graph. The result is a decrease in prediction performance. Using GAT as the graph encoder in the medical codes relation graph-based contrastive learning framework is not appropriate. We think the main reason is that the learned normalized relation score between medical codes in GAT belongs to an uncertain relevance score. In contrast, the empirical relation weight between medical codes computed based on the co-occurrence probability in the medical codes relation graph is a relatively specific relevance score. Therefore, compared to the above other variants, KAMPNet$_{CA}$ performs worst on the medication prediction task.

Limitations of the medication prediction model

Though the above extensive experiments have testified the efficacy of KAMPNet for reasonable medication prediction, there are still some limitations on the results due to the complexity of the healthcare system. First, the real clinical decision-making process is full of uncertainty caused by various factors such as the professional level of doctors, the social environment, and the economic conditions of patients, which might cause medication prediction bias. Thus, in the future, we would consider cultivating more advanced methods to reduce the impact of bias on the prediction outcomes. Second, the medication prediction model is essential for hospitals, clinics, and retail pharmacies. Since we only utilized the MIMIC-III dataset to testify the efficacy of our proposed model, the prediction outcomes of the model are not guaranteed on the data from other sources. However, provided the data has similar composition and structure defined in the previous section, the model should be applicable after fine-tuning. In the future, we will further evaluate our model on the datasets from clinics or retail pharmacies.

Additionally, the use of machine learning models in healthcare presents several ethical implications, such as privacy and data security, transparency and

interpretability, informed consent and autonomy, human oversight and decision-making, etc. And here are some corresponding potential approaches to address them: (1) Cultivate encryption techniques and secure data storage protocols to safeguard sensitive data; (2) Develop explainable and interpretable machine learning models that provide insights into the factors influencing predictions; (3) Educate patients about the use of machine learning models in their healthcare and provide clear explanations of the benefits, risks, and limitations; (4) Offer patients the option to opt out of using machine learning models in their treatment decisions if they have concerns or preferences; (5) Encourage interdisciplinary collaboration between healthcare providers and data scientists to ensure a holistic approach to patient care. Thus, in the future, collaboration, transparency, and an ongoing commitment to ethical practices are essential to ensure the responsible and beneficial use of machine learning models in healthcare.

For the practical applications of model KAMPNet in the future, clinically, it could be utilized to assist doctors in making informed medication decisions for patients according to electronic medical records (EMR). In addition, the model could also be used to assist the hospital pharmacy management department, clinic, or retail pharmacies to predict the medications in advance for stocking. However, the model needs to be retrained using the datasets from different application scenarios. Thus, in the future, a user study pertaining to the practical application scenario should be undertaken for validating the model's feasibility in practice. Additionally, owing to the limited size of the population cohort itself, and the experimental outcomes of existing approaches, the effect of the size of experimental samples on the prediction outcomes is neglected in our manuscript. In the future, the above factors will be taken into account once applied in the new dataset or private dataset for the generalization.

Conclusion and future work

In this study, we propose a multi-sourced medical knowledge-augmented medication prediction network. Expressly, we incorporate a novel multi-level graph contrastive learning framework for fully capturing the valuable relations between medical codes implicit in

An *et al. BMC Medical Informatics and Decision Making*   (2023) 23:243

Page 18 of 19

the multi-sourced medical knowledge. The framework firstly leverages the local graph contrastive learning on the medical ontology graphs to learn the knowledge augmented embedding vectors of diagnosis codes and medication codes, which factually have infused the information of correlative medical codes into each other in the learning process. Then, the medical codes relation graph is constructed and utilized to learn the relation augmented medical codes embedding vectors using the graph contrastive learning framework, which aims to capture the global relations between homogeneous and heterogeneous medical codes. Finally, the multi-channel sequence learning network is presented to capture the temporal relations between medical codes, by which we can get a comprehensive patient representation for downstream tasks such as medication prediction. We evaluate the performance of the proposed KAMPNet on a real-world clinical dataset, and the experimental results show that our model achieves the best medication prediction performance against baseline models in terms of Jaccard, F1 score, and PR-AUC.

With the help of the presented framework, in the future, we can introduce more related medical domain knowledge, such as the medication-related molecular graph and the bipartite graph representing adverse medication reactions. In addition, we will cultivate more advanced algorithms to better mine the insightful information implicit in the multi-sourced medical knowledge and explore how to determine the importance of multi-sourced medical knowledge automatically.

Abbreviations

KAMPNet	Knowledge augmented medication prediction network
EMR	Electronic medical records
ICD	International classification of diseases
ATC	Anatomical therapeutic chemical classification system
MIMIC	Medical information mart for intensive care
BERT	bidirectional encoder representation from transformers
DGI	Deep graph infomax
LSTM	Long short-term memory
GCN	Graph convolutional network
GNN	Graph neural network
GAT	Graph attention network
LR	Logistic regression
LEAP	Learn to prescribe
GRAM	Graph-based attention model
GAMENet	Graph augmented memory networks
GATE	Graph-Attention augmented temporal neural network
RETAIN	Reverse time attention

Authors' contributions
YA developed the conceptual framework and research protocol for the study and testified the proposed methods via the experiments. HT put forward constructive suggestions on the revision of this draft. BJ read the final manuscript and gave some useful suggestions. YX read and approved the final manuscript. XW read and approved the final manuscript.

Availability of data and materials
The data used in the paper is from a publicly available dataset MIMIC-III [35]. The processed data can be obtained on Github (https://github.com/uctor onto/KAMPNet/tree/main/Data).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]School of Software, North University of China, No.3 Xueyuan Road, Jiancaoping District, 030051 Taiyuan, Shanxi, China. [2]Institute of Automation Chinese Academy of Sciences, 95 Zhongguancun East Road, 100190 Beijing, China. [3]School of Innovation and Entrepreneurship, Dalian University of Technology, No.2 Linggong Road, Ganjingzi District, 116024 Dalian, Liaoning, China. [4]School of Computer Science and Technology, Dalian University of Technology, No.2 Linggong Road, Ganjingzi District, 116024 Dalian, Liaoning, China.

References
1. Ye M, Luo J, Xiao C, Ma F. LSAN: Modeling Long-term Dependencies and Short-term Correlations with Hierarchical Attention for Risk Prediction. CIKM. 2020. p. 1753–62. https://doi.org/10.1145/3340531.3411864.
2. Choi E, Xiao C, Stewart WF, Sun J. MiME: Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare. In: NeurIPS. 2018. p. 4547–57. https://doi.org/10.5555/3327345.3327366.
3. Zhang Y, Yang X, Ivy JS, Chi M. ATTAIN: Attention-based Time-Aware LSTM Networks for Disease Progression Modeling. In: IJCAI. 2019. p. 4369–75. https://doi.org/10.5555/3367471.3367649.
4. Shang J, Xiao C, Ma T, Li H, Sun J. GAMENet: Graph Augmented MEmory Networks for Recommending Medication Combination. In: AAAI. 2019. p. 1126–33. https://doi.org/10.1609/aaai.v33i01.33011126.
5. Shang J, Ma T, Xiao C, Sun J. Pre-training of Graph Augmented Transformers for Medication Recommendation. In: IJCAI. 2019. p. 5953–9. https://doi.org/10.1145/3498851.3498968.
6. He Y, Wang C, Li N, Zeng Z. Attention and Memory-Augmented Networks for Dual-View Sequential Learning. In: SIGKDD. 2020. p. 125–34. https://doi.org/10.1145/3394486.3403055.
7. Choi E, Bahadori MT, Song L, Stewart WF, Sun J. GRAM: Graph-based Attention Model for Healthcare Representation Learning. In: SIGKDD. 2017. p. 787–95. https://doi.org/10.1145/3097983.3098126.
8. Ma F, You Q, Xiao H, Chitta R, Zhou J, Gao J. KAME: Knowledge-based Attention Model for Diagnosis Prediction in Healthcare. In: CIKM. 2018. p. 743–52. https://doi.org/10.1145/3269206.3271701.
9. Gao C, Yin S, Wang H, Wang Z, Du Z, Li X. Medical-Knowledge-Based Graph Neural Network for Medication Combination Prediction. IEEE Trans Neural Netw Learn Syst. 2023;1–12. https://doi.org/10.1109/TNNLS.2023.3266490.

10. Wu R, Qiu Z, Jiang J, Qi G, Wu X. Conditional Generation Net for Medication Recommendation. In: WWW. April 2022. p. 935–45. https://doi.org/10.1145/3485447.3511936. Accessed April 2022.

11. Veličković P, Fedus W, Hamilton WL, Liò P, Bengio Y, Hjelm RD. Deep Graph Infomax. In: ICLR. New Orleans; 2019. p. 1–17.

12. Hendrycks D, Mazeika M, Kadavath S, Song D. Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. In: NeurIPS. vol. 32. Vancouver: Curran Associates, Inc.; 2019. p. 15663–74. https://doi.org/10.5555/3454287.3455690.

13. Zhang Y, Chen R, Tang J, Stewart WF, Sun J. LEAP: Learning to Prescribe Effective and Safe Treatment Combinations for Multimorbidity. In: SIGKDD. 2017. p. 1315–24. https://doi.org/10.1145/3097983.3098109.

14. Wang L, Zhang W, He X, Zha H. Personalized Prescription for Comorbidity. In: DASFAA. Gold Coast; 2018. p. 3–19. https://doi.org/10.1007/978331 9914589_1.

15. Wang S, Ren P, Chen Z, Ren Z, Ma J, de Rijke M. Order-free Medicine Combination Prediction with Graph Convolutional Reinforcement Learning. In: CIKM. 2019. p. 1623–32. https://doi.org/10.1145/3357384.3357965.

16. Jin B, Yang H, Sun L, Liu C, Qu Y, Tong J. A Treatment Engine by Predicting Next-Period Prescriptions. In: SIGKDD. ACM; 2018. p. 1608–16. https://doi.org/10.1145/3219819.3220095.

17. Le H, Tran T, Venkatesh S. Dual Memory Neural Computer for Asynchronous Two-view Sequential Learning. In: SIGKDD. ACM; 2018. p. 1637–45. https://doi.org/10.1145/3219819.3219981.

18. An Y, Zhang L, You M, Tian X, Jin B, Wei X. MeSIN: Multilevel selective and interactive network for medication recommendation. Knowl-Based Syst. 2021;233:107534. https://doi.org/10.1016/j.knosys.2021.107534.

19. Scarselli F, Gori M, Tsoi A, Hagenbuchner M, Monfardini G. The Graph Neural Network Model. IEEE Trans Neural Netw. 2009;20:61–80. https://doi.org/10.1109/TNN.2008.2005605.

20. Kipf T, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. In: ICLR. Toulon; 2017. p. 1–14.

21. Xu K, Hu W, Leskovec J, Jegelka S. How Powerful are Graph Neural Networks? In: ICLR. New Orleans; 2019. p. 1–17.

22. Ruiz C, Zitnik M, Leskovec J. Identification of disease treatment mechanisms through the multiscale interactome. Nat Commun. 2021;12. https://doi.org/10.1038/s41467-021-21770-8.

23. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics. 2018;34:i457–66. https://doi.org/10.1093/bioinformatics/bty294.

24. Ye M, Cui S, Wang Y, Luo J, Xiao C, Ma F. Medpath: Augmenting health risk prediction via medical knowledge paths. In: WWW. 2021. p. 1397–409. https://doi.org/10.1145/3442381.3449860.

25. Zeng X, Tu X, Liu Y, Fu X, Su Y. Toward better drug discovery with knowledge graph. Curr Opin Struct Biol. 2022;72:114–26. https://doi.org/10.1016/j.sbi.2021.09.003.

26. Lu H, Uddin S. A weighted patient network-based framework for predicting chronic diseases using graph neural networks. Sci Rep. 2021;11:1–12.

27. Zhang X, Qian B, Li Y, Yin C, Wang X, Zheng Q. KnowRisk: an interpretable knowledge-guided model for disease risk prediction. In: ICDM. IEEE; 2019. p. 1492–7. https://doi.org/10.1109/ICDM.2019.00196.

28. Wang M, Chen J, Lin S. Medication Recommendation Based on a Knowledge-enhanced Pre-training Model. In: IEEE/WIC/ACM International Conference on Web Intelligence. 2021. p. 290–4. https://doi.org/10.1145/3498851.3498968.

29. Liu S, Li T, Ding H, Tang B, Wang X, Chen Q, et al. A hybrid method of recurrent neural network and graph neural network for next-period prescription prediction. Int J Mach Learn Cybern. 2020;11:2849–56.

30. Mao C, Yao L, Luo Y. MedGCN: Medication recommendation and lab test imputation via graph convolutional networks. J Biomed Inform. 2022;127:104000. https://doi.org/10.1016/j.jbi.2022.104000.

31. Su C, Gao S, Li S. GATE: graph-attention augmented temporal neural network for medication recommendation. IEEE Access. 2020;8:125447–58. https://doi.org/10.1109/ACCESS.2020.3007835.

32. Hjelm RD, Fedorov A, Lavoie-Marchildon S, Grewal K, Trischler A, Bengio Y. Learning deep representations by mutual information estimation and maximization. In: ICLR. New Orleans; 2019. p. 1–24. https://doi.org/10.1109/TPAMI.2022.3147886.

33. Hassani K, Khasahmadi AH. Contrastive Multi-View Representation Learning on Graphs. In: PMLR. Online. 2020;119:4116–26.

34. Sun M, Xing J, Wang H, Chen B, Zhou J. MoCL: Contrastive Learning on Molecular Graphs with Multi-level Domain Knowledge. In: SIGKDD. Virtual Event; 2021. p. 3585–94. https://doi.org/10.1145/3447548.3467186.

35. Johnson AEW, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016. https://doi.org/10.1038/sdata.2016.35.

36. Fonarow GC, Wright RS, Spencer FA, Fredrick PD, Dong W, Every N, et al. Effect of Statin Use Within the First 24 Hours of Admission for Acute Myocardial Infarction on Early Morbidity and Mortality. Am J Cardiol. 2005;96:611–6. https://doi.org/10.1016/j.amjcard.2005.04.029.

37. Ma F, Chitta R, Zhou J, You Q, Sun T, Gao J. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. In: SIGKDD. 2017. p. 1903–11. https://doi.org/10.1145/3097983.3098088.

38. Yu L, Sun L, Du B, Liu C, Xiong H, Lv W. Predicting Temporal Sets with Deep Neural Networks. In: SIGKDD. Virtual Conference; 2020. p. 1083–91. https://doi.org/10.1145/3394486.3403152.

39. Church KW, Hanks P. Word Association Norms, Mutual Information and Lexicography. In: the Association for Computational Linguistics. Vancouver; 1989. p. 22–30. https://doi.org/10.5555/89086.89095.

40. Velickovic P, Cucurull G, Casanova A, Romero A, Lio' P, Bengio Y. Graph Attention Networks. In: ICLR. Vancouver; 2018. p. 1–12.

41. Smith J, Johnson S. Rectifier Nonlinearities Improve Neural Network Acoustic Models. J Mach Learn Res. 2013;28:1–6.

42. Luaces O, Díez J, Barranquero J, del Coz JJ, Bahamonde A. Binary relevance efficacy for multilabel classification. Prog Artif Intell. 2012;1:303–13. https://doi.org/10.1007/s13748-012-0030-x.

43. Choi E, Bahadori MT, Sun J, Kulas JA, Schuetz A, Stewart WF. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In: NeurIPS. Barcelona, Spain; 2016. p. 3504–12. https://doi.org/10.5555/3157382.3157490.

44. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. CoRR. 2014. abs/1412.6980.

Publisher's Note