KEEP: Integrating Medical Ontologies with Clinical Data for Robust Code Embeddings

Ahmed Elhussein[1,2] AE2722@CUMC.COLUMBIA.EDU
Paul Meddeb[2] PAUL.MEDDEB@ETU.MINESPARIS.PSL.EU
Abigail Newbury[1,2] ABIGAIL.NEWBURY@COLUMBIA.EDU
Jeanne Mirone[2] JEANNE.MIRONE@ETU.MINESPARIS.PSL.EU
Martin Stoll[2] MARTIN.STOLL@ETU.MINESPARIS.PSL.EU
Gamze Gürsoy[1,2] GAMZE.GURSOY@COLUMBIA.EDU
[1] *Columbia University, USA*
[2] *New York Genome Center, USA*

Abstract

Machine learning in healthcare requires effective representation of structured medical codes, but current methods face a trade-off: knowledge graph-based approaches capture formal relationships but miss real-world patterns, while data-driven methods learn empirical associations but often overlook structured knowledge in medical terminologies. We present KEEP (Knowledge-preserving and Empirically refined Embedding Process), an efficient framework that bridges this gap by combining knowledge graph embeddings with adaptive learning from clinical data. KEEP first generates embeddings from knowledge graphs, then employs regularized training on patient records to adaptively integrate empirical patterns while preserving ontological relationships. Importantly, KEEP produces final embeddings without task-specific axillary or end-to-end training enabling KEEP to support multiple downstream applications and model architectures. Evaluations on structured EHR from UK Biobank and MIMIC-IV demonstrate that KEEP outperforms both traditional and Language Model-based approaches in capturing semantic relationships and predicting clinical outcomes. Moreover, KEEP's minimal computational requirements make it particularly suitable for resource-constrained environments.

Data and Code Availability This research has been conducted using data from UK Biobank (Sudlow et al., 2015) and MIMIC-IV Johnson et al. (2021). Researchers can request access via https://www.ukbiobank.ac.uk/ and https://physionet.org/content/mimiciv/3.1/, respectively. Implementation code is available at https://github.com/G2Lab/keep

Institutional Review Board (IRB) This study does not require IRB approval as all data used are publicly available.

1. Introduction

Structured electronic health records (EHRs) contain extensive data across multiple domains including diagnoses, medications, procedures, and clinical observations. These datasets are a major resource for developing machine learning (ML) models across various healthcare applications including predictive modeling, phenotyping, and drug repurposing (Tang et al., 2024). However, the discrete nature of medical codes within these datasets presents challenges for effective representation learning in ML. Traditional discrete representations like one-hot encoding produce extremely high-dimensional and sparse representations (Johnson and Khoshgoftaar, 2021). These high-dimensional, sparse representations create several challenges for gradient descent algorithms: they slow convergence rates, dilute the strength of relevant signals, and ultimately restrict the model's ability to generalize effectively (Guo and Berkhahn, 2016). Additionally, one-hot encoding treats each code as independent, ignoring the relationships defined in their source medical terminologies. However, these terminologies contain knowledge graphs that encode essential clinical and biological relationships. By failing to capture these relationships, one-hot encoding discards valuable context that is essential for more

sophisticated tasks (Chang et al., 2020; Choi et al., 2016).

Representation learning addresses these limitations by transforming high-dimensional medical data into embeddings, which are compact real-valued vector representations (Bengio et al., 2013). Research in Natural Language Processing (NLP) has shown that sequence-based representation learning can transform high dimensional discrete data into dense vector spaces that capture semantic relationships through distance metrics (Mikolov, 2013; Pennington et al., 2014). In healthcare applications, ML models using real-valued embeddings have demonstrated superior performance in tasks such as risk prediction and patient stratification (Choi et al., 2018; Rasmy et al., 2021).

Recently, language models (LMs) have been used for medical code representation, offering several advantages. First, unlike traditional methods that produce static embeddings, LMs generate dynamic, context-aware representations that adapt to the surrounding information. This capability allows them to capture nuanced relationships between medical codes more effectively (Alsentzer et al., 2019; Pang et al., 2021; Rasmy et al., 2021). Second, these models benefit from pre-training on extensive unstructured biomedical corpora, including PubMed and clinical notes, allowing them to learn valuable domain knowledge from many sources (Lee et al., 2020; Luo et al., 2022). Third, LMs excel at transfer learning, efficiently adapting to new tasks with limited data through their few-shot learning capabilities (Brown, 2020). However, applying LMs to structured medical data also has limitations. The subword tokenization approach fragments the unit meanings of medical codes, introducing semantic ambiguity and failing to preserve hierarchical structures (Dwivedi et al., 2024; Yuan et al., 2022). This fragmentation is particularly problematic because medical codes are designed as complete, atomic units with precise meanings. LMs also lack mechanisms to directly integrate knowledge graphs from medical ontologies, resulting in embeddings that inadequately capture known biomedical relationships (Chang et al., 2024). The practical impact of these limitations is clear: without fine-tuning, even GPT-4 achieves less than 50% accuracy on basic medical code matching tasks (Soroush et al., 2024). While fine-tuning can partially address these challenges, it demands labeled data and substantial computational resources (Zhang et al., 2024).

Prior work in medical code representation reveals three critical requirements: (1) capturing semantic relationships from medical knowledge graphs, (2) reflecting empirical patterns observed in patient records, and (3) demonstrating robust utility across diverse healthcare tasks (Si et al., 2021). We propose KEEP (Knowledge-preserving and Empirically refined Embedding Process), an efficient framework that addresses these requirements through a two-step approach. First, KEEP utilizes knowledge graphs to generate initial embeddings that preserve semantic relationships. Next, it refines these embeddings through regularized training on empirical co-occurrence data, ensuring rare codes retain meaningful ontological representations while frequently co-occurring codes benefit from data-driven adjustments. With its computational efficiency and compatibility with existing data formats, KEEP enables broad generalizability across healthcare settings. Comprehensive empirical evaluations demonstrate KEEP's robust performance, consistently outperforming traditional and LM-based models in both intrinsic tasks (semantic relationship encoding) and extrinsic tasks (clinical prediction).

Our framework complements recent advances in LMs in two key ways by enabling more effective bridging between structured medical knowledge and unstructured data. First, it generates structured data representations that can be integrated with LM-derived embeddings through multimodal integration techniques (Ebrahimi et al., 2023). Second, it could enhance LM performance on medical codes by providing domain-aware initialization for fine-tuning, enabling more effective incorporation of medical ontologies and relationships (Hewitt, 2021; Fatemi et al., 2023).

2. Related Works

Early efforts to create medical code embeddings built directly on advances in NLP, treating medical codes as tokens and patient records as sequences (De Vine et al., 2014). This approach evolved with Med2Vec, which incorporated visit-level information to better capture co-occurrence patterns (Choi et al., 2016). However, recognizing that medical codes differ fundamentally from natural language through their well-defined ontological structure, researchers developed approaches that leverage knowledge graphs to generate more semantically meaningful embeddings (Grover and Leskovec, 2016; Agarwal et al., 2019).

Yuan et al. (2022) further advanced this direction by incorporating medical knowledge graphs through contrastive learning techniques.

The development of LMs marked a significant shift in medical code representation. BERT-based models such as BioBERT and ClinicalBERT introduced context-dependent embeddings and leveraged pre-training on unstructured biomedical corpora like PubMed and MIMIC-III (Alsentzer et al., 2019; Lee et al., 2020). These innovations significantly improved performance on tasks like named entity recognition and relation extraction. However, their application to structured data was constrained by tokenization strategies ill-suited for medical codes. Models like Med-BERT and CEHR-BERT tried to address this limitation by introducing visit-level tokenization tailored for structured EHR data (Rasmy et al., 2021; Pang et al., 2021). Recent work has introduced purpose-built foundation models specifically designed for structured longitudinal health data (Dwivedi et al., 2024; Steinberg et al., 2021). Alternatives to code tokenization have explored generating embeddings by using descriptions rather than the code itself, achieving notable gains in predictive performance (Kane et al., 2023; Lin et al., 2020; Lee et al., 2024).

Despite these advancements, significant challenges remain. The non-Identically Independently Distributed (IID) nature of medical data necessitates institution-specific model fine-tuning for pre-trained model, requiring substantial computational resources and introducing deployment complexity. This presents a significant barrier to widespread implementation, particularly for institutions with limited resources (Wornow et al., 2023). Practical implementation is further hindered by the fact that embeddings from recent structured EHR models are rarely made publicly available, which limits external validation of the models (Wornow et al., 2023). Current methods also struggle to fully capture the hierarchical nature of medical coding systems, where codes exist in complex parent-child relationships that influence their semantic meaning.

3. Background

3.1. OMOP knowledge graph

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is a standardized data model that transforms heterogeneous structured EHR data into a consistent format. Like most medical terminologies, OMOP maintains an associated knowledge graph that captures relationships between clinical concepts through various semantic connections. The graph structure encodes both hierarchical relationships (*e.g.,* "Type 2 Diabetes" is-a "Diabetes Mellitus") and non-hierarchical clinical associations (*e.g.,* "may-prevent", "has-finding-site"). These relationships provide valuable domain knowledge but are limited to pre-defined medical associations. They do not capture the full complexity of relationships or empirically observed patterns, which can vary across healthcare settings.

3.2. Resnik similarity

Resnik similarity measures semantic similarity between two nodes in a graph by evaluating the information content (IC) of their least common ancestor (LCA). For any concept c, IC is defined as $IC(c) := -\log p(c)$, where $p(c)$ represents the proportion of nodes that are descendants of c in the hierarchy. This formulation assigns higher IC to specific, rare concepts and lower IC to common, general concepts. Nodes sharing a more specific LCA are considered more semantically similar.

3.3. Models

Below we discuss two complementary approaches to embedding construction leveraged by our framework. We also discuss LMs to contextualize current approaches in medical code representation.

GloVe (Global Vectors for Word Representation) learns semantic relationships by analyzing co-occurrence patterns in data (Pennington et al., 2014). It uses a co-occurrence matrix X, where X_{ij} represents how often concept i appears with concept j to minimize:

$$\sum_{i,j=1}^{V} f(X_{ij}) \left(w_i^\top \tilde{w}_j - \log X_{ij}\right)^2, \qquad (1)$$

where V denotes vocabulary size, w_i and $\tilde{w}_j$ are the learned concept vectors, and $f(X_{ij})$ weights frequent co-occurrences to prevent them from dominating the optimization.

node2Vec generates embeddings that preserve graph structure by simulating biased random walks through a knowledge graph (Grover and Leskovec, 2016). For each node u, it maximizes the probability

of observing its network neighborhood $\mathcal{N}_S(u)$ given its embedding $f(u)$:

$$\max_f \sum_{u \in V} \log \Pr(\mathcal{N}_S(u) \mid f(u)), \tag{2}$$

where V represents the set of nodes, $f(u)$ maps node u to a vector in $\mathbb{R}^d$ and $\Pr(v \mid f(u))$ is the probability of node v being sampled as part of $\mathcal{N}_S(u)$ given $f(u)$. Assuming conditional independence, this probability factorizes as:

$$\Pr(\mathcal{N}_S(u) \mid f(u)) = \prod_{v \in \mathcal{N}_S(u)} \Pr(v \mid f(u)). \tag{3}$$

Language Models (LMs) are the current state-of-the-art in text representation but face limitations with structured medical data. BERT uses bidirectional transformers to generate context-aware embeddings (Kenton and Toutanova, 2019). GPT models process text unidirectionally and with embeddings usually averaged from a decoder layer in the absence of a [CLS] token (Achiam et al., 2023).

3.4. Code representation

Recent advances in medical code representation have been dominated by data-driven approaches, particularly foundation models that employ autoregressive techniques like masked token prediction. Despite their success in general language tasks, these models face distinct challenges in the medical domain. They struggle to capture the explicit hierarchical relationships within medical ontologies and often generate poor representations of rare conditions due to limited clinical examples (Banerjee et al., 2023). Furthermore, by learning from individual patient sequences in isolation, these models cannot explicitly leverage patterns shared across patients. Both limitations increase the risk of overfitting to institution-specific data. This dependence could compromise their generalizability across different healthcare settings.

Knowledge graphs offer an alternative foundation for medical code representation by explicitly encoding relationships between clinical concepts (Lee et al., 2021). These structured graphs preserve semantic relationships and enable models to incorporate established medical knowledge. However, a major limitation of knowledge graphs is their inability to comprehensively capture the full complexity of biomedical relationships. Even the most detailed knowledge graphs have fundamental limitations: they cannot encode all relationships, especially those that emerge from real-world data, and their encoded relationships fail to capture the full nature of relationships, particularly the context-dependence (see Section 3.1). These inherent constraints mean that representations based solely on knowledge graphs fail to reflect all associations observed in practice.

4. Methods

4.1. Problem setup

Our objective is to develop medical code embeddings that integrate structured knowledge graphs with real-world medical data. This approach captures both the taxonomic design of medical codes and their practical clinical usage, enabling the embeddings to reflect biological relationships and real-world comorbidity patterns. We believe such integrated representations can enhance downstream analytical tasks. Additionally, we aim to ensure these embeddings can be generated with minimal computational requirements, making them accessible to a wide range of institutions.

4.2. Data

In our analysis, we use the UK Biobank (UKBB) dataset (Papez et al., 2023), comprising structured outpatient and hospital records for approximately 500,000 patients and the MIMIC-IV dataset comprising of over 200,000 patients admitted to the ICU. Both datasets were converted to the OMOP Common Data Model. We used OMOP CDM due to its richer graph structure and multi-domain integration. Note, however, that any terminology with an associated knowledge graph can be used (*e.g.,* SNOMED, ICD) instead.

To generate embeddings, we require two key components derived from the EHR and their associated terminology: a knowledge graph and a co-occurrence matrix. We construct the OMOP knowledge graph using only condition concepts connected through hierarchical 'is-a' relationships. To maintain computational efficiency and focus on useful concept abstractions, we limit the hierarchical depth of our knowledge graph to five levels from the root node. Codes further away from this are aggregated to their parent node (see Section A.1.1 for more details). We found that a depth of 5 provided a level of specificity comparable to ICD codes, making it well-suited for our task. While our current implementation utilizes only 'is-a' relationships, the framework can accommodate more complex relationships and additional con-

cept types, such as drugs and procedures, enabling multi-domain knowledge integration as enriching the knowledge graph has been shown to improve embedding quality (Shen et al., 2019).

For the co-occurrence matrix, we analyze complete patient histories rather than individual visits to reduce matrix sparsity, though visit-level analysis could provide more granular comorbidity relationships given sufficient data volume. To ensure reliable phenotyping, we require two occurrences of a diagnosis code in a patient's record to label a patient with that condition (see Section A.4 for more details).

4.3. Algorithm

Our method, KEEP, is summarized in Figure 1 and Algorithm 1. KEEP addresses the limitations discussed in Section 3.4 by integrating knowledge graphs with real-world clinical data. This integration creates embeddings that are both theoretically grounded and empirically validated, promoting robustness and generalizability across institutions. KEEP takes a more intuitive approach to integrating knowledge graphs with EHR data. Rather than treating graph knowledge as an additional modeled input, we approach it as a prior that regularizes learning from real-world data—more closely mimicking how physicians learn. Further, by prioritizing computational efficiency, we aim to support representation learning even in resource constrained hospitals.

KEEP follows a two-stage training procedure that combines established lightweight algorithms through regularized training. The first stage employs node2vec to generate initial embeddings based on a knowledge graph. This step, independent of institution-specific data, captures the fundamental biological relationships encoded in medical ontologies. Recent research has demonstrated that these knowledge graph-based embeddings alone effectively capture phenotype-aligned relationships (Lee et al., 2021).

In the second stage, KEEP uses the node2vec embeddings as initialization values for a modified GloVe model. To preserve the valuable medical knowledge while incorporating empirical patterns, we augment the standard GloVe objective function with a regularization term. This approach prevents catastrophic forgetting during training (Kirkpatrick et al., 2017), ensuring the final embeddings maintain their ontological foundation while incorporating real-world associations. In our experiments, both node2vec and GloVe models produced embeddings of identical dimensionality. While maintaining consistent dimensions between both models is recommended, if working with pretrained embeddings of different dimensions, we suggest using autoencoders to either reduce or expand the node2vec embedding dimensions to match those required by the GloVe model.

The objective function for KEEP combines the GloVe loss with a regularization term:

$$\underbrace{\sum_{i,j=1}^{V} f(X_{ij})\left(w_i^\top \tilde{w}_j - \log X_{ij}\right)^2}_{\text{GloVe}} + \underbrace{\lambda \sum_{i=1}^{V} |w_i - w_i^{\text{n2v}}|^2}_{\text{Regularization Term}} \tag{4}$$

Here, λ controls the regularization strength and w_i^{n2v} represents the initial node2vec embedding for code i. This formulation provides two key technical advantages. It maintains robust representations for rare codes by anchoring their embeddings to biologically meaningful dimensions established through the node2vec initialization. Additionally, the regularization term enables controlled adaptation of the knowledge graph-based initialization as more comorbidity data becomes available, with λ determining the degree of permissible deviation. This approach effectively creates a continuum between purely knowledge-based and purely data-driven representations, with λ serving as the control parameter.

5. Experiment setup

We evaluate KEEP's embeddings against several baseline techniques through both intrinsic and extrinsic assessments. The intrinsic evaluation examines the relationships captured within the embedding space, while the extrinsic evaluation measures their effectiveness in downstream clinical prediction tasks.

5.1. Comparator embeddings

We evaluate KEEP's embeddings against two categories of existing approaches: pretrained language model embeddings and specialized embeddings trained on this data. This allows us to assess the relative advantages of our method across different use cases and computational requirements.

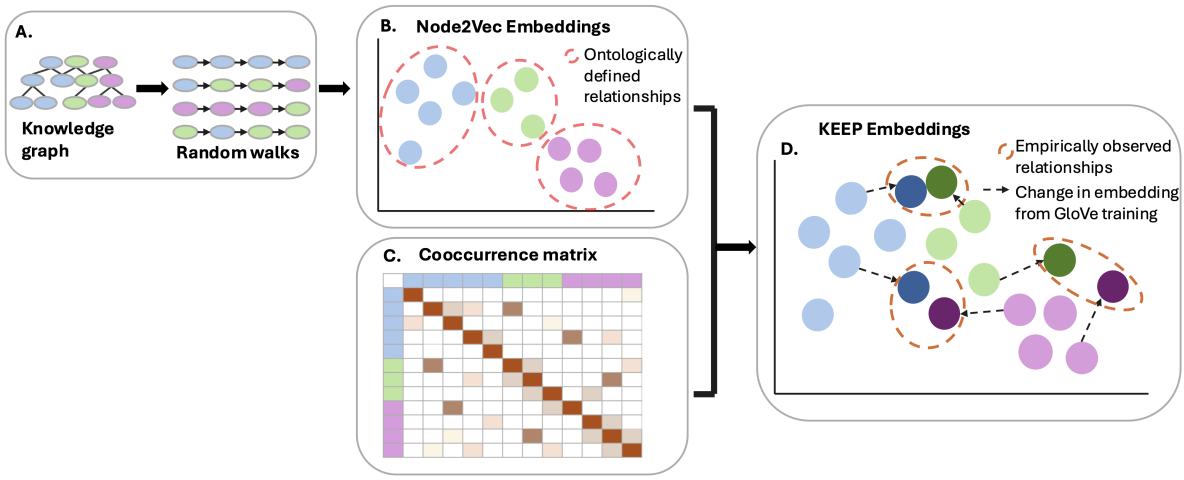

Figure 1: Overview of KEEP's approach: (A) Generate random walks on knowledge graph. (B) Walks used to create initial embeddings whose dimensions align with the ontology. (C) A co-occurrence matrix is constructed from EHR data. (D) GloVe model is initialized with the embeddings from (B) and regularized to incorporate empirical relationships from (C) while preserving ontologically-aligned dimensions. That is embeddings are adjusted based on the strength of observed associations.

5.1.1. Pre-trained Language Model Embeddings

We generated embeddings from pre-trained LMs using OMOP code descriptions, following the methodology of Kane et al. (2023). For BERT architectures, we extracted the `[CLS]` token from the final layer. For BioGPT, we used the mean of the final hidden state, as this can capture semantics in a continuous latent space (Hao et al., 2024).

We evaluated three leading pre-trained models for medical tasks: BioClinBERT (Alsentzer et al., 2019), fine-tuned on MIMIC clinical notes after PubMed training; ClinBERT (Wang et al., 2023), trained on extensive EHR data from over 3 million patients; and BioGPT (Luo et al., 2022), a transformer model trained on PubMed biomedical literature.

To capture hierarchical relationships, we generated embeddings that incorporated that also include the descriptions of parent and grandparent codes. Hierarchical information was explicitly encoded by appending 'is a' followed by the description. This provided the models with a contextual understanding of the taxonomic structure when generating embeddings. This is a similar approach as proposed by Deng et al. (2024). We refer to these embeddings as Hierarchy Aware (HA). For more details see Section A.5

5.1.2. Dataset-Specific Embeddings

We evaluated several embedding approaches that derive representations directly from structured medical data. To establish baselines for KEEP, we separately implemented node2vec and GloVe (see Sections A.1 and A.3 for training details). We included Cui2Vec (Beam et al., 2020) in our comparison due to its demonstrated ability to capture semantic relationships and enhance performance across various downstream tasks. Additionally, we also added a Graph Attention Network (GAT) baseline, following the approach of (Piya et al., 2024), with Node2Vec-based initialization and a co-occurrence-based graph construction. However, our implementation differs from the original HealthGAT in two key aspects. First, we do not integrate a supervised auxiliary task during embedding generation, as this is designated for extrinsic evaluation. Second, we use only diagnoses data, excluding procedures.

5.2. Assessing the Impact of Regularization

We first assess KEEP's ability to effectively combine knowledge graph structure with co-occurrence patterns. Our analysis aims to verify that KEEP effectively combines both data sources, demonstrating stronger co-occurrence relationships than node2vec and better graph relationships than GloVe. While

not direct measures of quality, these assessments help validate the effective integration of both training objectives in our embeddings.

We measured the correlation between embedding cosine similarities and co-occurrence values to assess preservation of empirical relationships (See Section B.1). For hierarchical relationships, we computed correlations between embedding cosine similarities and Resnik similarity scores between node pairs. Comparing these metrics across different embedding approaches provides insight into how each method balances graph-based and empirical relationships. Note, we only used UK Biobank dataset that contains a more diverse range of diseases.

5.3. Intrinsic Evaluation

Our intrinsic evaluation framework assesses how effectively embeddings capture known relationships. Following the methodology of (Beam et al., 2020), we focus on discriminating between genuine relationships and random associations. To do this, we curated a set of test codes that span both data-driven relationships (*e.g.,* obesity) and medical relationships derived from medical ontologies (*e.g.,* type 1 diabetes) (see Table 8 for list of concepts). For each disease, we evaluate its representation by comparing the cosine similarity between its embedding and those of its known condition set against a null distribution generated from similarity scores with randomly selected conditions. A relationship is defined as correctly identified if its similarity score exceeds the 95th percentile of this disease-specific null distribution. Given our analysis spans 6 diseases, we employ the Wilcoxon signed-rank test to assess systematic differences in identification performance across embedding methods. Note, for this task we only used UK Biobank dataset that contains a more diverse range of diseases. For a full description of the method see Section B.1.

5.4. Extrinsic Evaluation

5.4.1. Clinical tasks

We evaluated our embeddings' practical utility through clinical prediction tasks focused on two distinct datasets—the UK Biobank (UKBB) and MIMIC-IV - focusing on clinically diverse and consequential outcomes. In the UKBB cohort we evaluated rare complications in chronic disease populations (see Table 1). Our prediction tasks examined: diabetic sequelae (including eye complications, peripheral vascular disease, neuropathy, and kidney disease) in patients with type 2 diabetes; myocardial infarction in patients with angina; and acute renal failure in patients with chronic kidney disease (see Table 9 for list of codes used). These complications were selected as improved risk stratification in these areas can significantly impact patient care decisions (Ndjaboue et al., 2022; Ramspek et al., 2020; Reeh et al., 2019). Additionally, these complications involve the interaction of multiple diseases, providing a robust framework for evaluating the ability of embeddings to capture complex relationships. Complementing this, we analyzed 30-day readmission/representation risk in the MIMIC-IV dataset for patients discharged directly from emergency departments. By spanning chronic disease complications (UKBB) and acute care transitions (MIMIC-IV), our evaluation framework rigorously probes embedding capabilities.

Table 1: Prevalence of Complications (UK Biobank) and Readmission (MIMIC-IV)

Cohort	Complications	Prevalence
Type 2 DM	All Diabetes	17.2%
Type 2 DM	Peripheral vascular	1.7%
Type 2 DM	Kidney	2.8%
Type 2 DM	Eye	9.7%
Type 2 DM	Neuropathy	3.4%
Angina	MI	9.6%
CKD	Renal failure	7.0%
ER discharge	Readmission	3.0%

5.4.2. Model training and evaluation

For our prediction tasks, we implemented a single-layer encoder model with 4 attention heads to process sequences of patient diagnoses. The final representation is passed through a feedforward layer for prediction. Each diagnosis code was represented using the various embedding approaches under evaluation. We retained the original embedding dimensions for each method to preserve information content, though this resulted in varying model sizes, particularly for the pre-trained LM embeddings which typically have larger dimensions (see Table 7 for dimensions). For the trained embeddings, we use a fixed dimension size of 100 as preliminary analysis using the reconstruction losses for GloVe and cui2vec models showed a plateau at 100 dimensions. Given that the pre-trained models have fixed dimensions,

we standardized embedding dimensions to ensure fair performance comparison.

We conducted a learning rate sweep across evenly spaced values ranging from $1e{-}1$ to $5e^{-5}$, using the AdamW optimizer for all experiments. For each learning rate, we performed five training runs and selected the learning rate yielding the lowest median loss. Model performance was evaluated through 100 independent runs, with bootstrap sampling across these runs to generate 95% confidence intervals. See Section B.3 for full training details.

For model evaluation, we use Test loss, AUPRC, AUC, Mathews Correlation Coefficient (MCC), and F1 score. To make generalizable statements across all tasks, we evaluated differences in the overall rank between embeddings using the Wilcoxon signed-rank test. AUPRC is presented in the main manuscript, rest are found at Section C.

Table 2: Resnik similarity and co-occurrence similarity scores for different models (UK Biobank).

Model	Resnik Sim.	Co-occur. Sim.
BioClinBERT	0.40	0.13
BioClinBERT$_{HA}$	0.50	0.15
BioGPT	0.48	0.25
BioGPT$_{HA}$	0.67	0.36
ClinBERT	0.40	0.18
ClinBERT$_{HA}$	0.51	0.21
Cui2Vec	0.37	0.55
GloVe	0.55	0.57
Node2Vec	**0.70**	0.46
GAT	0.43	0.31
KEEP	0.68	**0.62**

Model$_{HA}$ refers to Hierarchy Aware embeddings. Column names: "Sim." = Similarity, "Co-occur." = Co-occurrence

6. Experiment Results

6.1. Assessing the Impact of Regularization

Our combined method effectively balanced both co-occurrence and hierarchical relationships, demonstrating strong performance on both metrics. Specifically, we improved upon node2vec's co-occurrence correlation and surpassed GloVe's Resnik similarity correlation. These results validate our approach's ability to preserve both graph-based structure and empirical relationships in the final embeddings.

Notably, enriching language model inputs with hierarchical descriptions improved their ability to capture taxonomic relationships without compromising co-occurrence pattern recognition. However, language models showed substantially lower co-occurrence similarity compared to embeddings trained directly on the dataset. This performance discrepancy likely stems from differences in training data sources—our embeddings were trained on UK health system data, while the language models were trained on US intensive care and Chinese healthcare datasets. These geographic and healthcare system variations influence the captured relationships, reflecting distinct clinical practices, coding patterns, and patient populations.

6.2. Intrinsic Evaluation

Table 3 presents performance comparisons across eight clinical conditions. KEEP demonstrated superior performance in capturing clinically meaningful relationships, outperforming both language models and traditional embedding methods. Our method achieved the highest average rank of 1.19 across all evaluated conditions (p=0.01). The advantage of combining structural and empirical relationships is particularly evident in conditions like Type 1 diabetes mellitus, where our method achieved a performance score of 0.93, substantially outperforming both GloVe (0.82) and Node2Vec (0.27).

Language models demonstrated notably weaker performance, particularly for specific conditions such as Prematurity (scores ranging from 0.00-0.11) and Obesity (scores ranging from 0.00-0.20). The addition of hierarchical information to these models did not yield consistent improvements in performance.

6.3. Extrinsic Evaluation

Table 4 presents AUPRC results for the downstream clinical prediction tasks. KEEP demonstrated superior performance compared to baseline methods, achieving a mean rank of 1.62 across multiple clinical outcomes (p=0.02). Additional metrics reported in Tables 10-13 show that KEEP achieved the lowest test loss, second-highest AUC, highest MCC, and highest F1 score. All performance differences were statistically significant ($p < 0.05$).

Among the baseline approaches evaluated, traditional methods demonstrated consistency in their performance. Cui2Vec and node2vec achieved strong results across the prediction tasks. In contrast, LM performance showed greater variation. While BioGPT emerged as the second-strongest performer

Table 3: **Intrinsic evaluation**: Accuracy of embedding models in identifying comorbid conditions across eight test diseases. Values represent the proportion of correctly identified comorbid conditions (UK Biobank).

	Asthma	Obesity	Pre-maturity	Renal Tx Reject.	Schizo-phrenia	Resp. Tumor	T1DM	T2DM	Mean Rank	p-val
BioClinBERT	0.54	0.12	0.00	0.35	0.43	0.00	0.27	0.20	7.63	0.03
BioClinBERT$_{HA}$	0.12	0.02	0.14	0.00	0.40	0.31	0.02	0.08	9.75	0.01
ClinBERT	0.24	0.09	0.00	**0.72**	0.46	0.25	0.52	0.33	6.75	0.64
ClinBERT$_{HA}$	0.22	0.07	0.14	0.02	0.49	0.45	0.05	0.00	8.50	0.04
BioGPT	0.37	0.16	0.14	0.30	0.23	0.54	0.27	0.36	7.43	0.04
BioGPT$_{HA}$	0.39	0.22	0.57	0.43	0.43	0.66	0.18	0.16	6.13	0.87
Cui2vec	0.61	0.74	**0.90**	0.43	0.69	**0.72**	0.91	0.69	2.63	0.01
GloVe	0.54	0.64	0.76	0.70	**0.86**	0.68	0.82	0.69	2.63	0.01
Node2Vec	0.61	0.09	0.62	0.50	0.31	0.34	0.27	0.15	6.56	0.64
GAT	0.65	0.26	0.19	0.0	0.40	0.28	0.64	0.69	5.88	0.93
KEEP	**0.78**	**0.93**	0.57	0.70	0.69	0.68	**0.93**	**0.79**	**2.13**	0.01

Model$_{HA}$ refers to Hierarchy Aware embeddings. Column names: "Tx" = Transplant, "Resp." = Respiratory, "T1DM" = Type 1 Diabetes Mellitus, "T2DM" = Type 2 Diabetes Mellitus, "p-val" = p-value.

overall—likely attributable to its modern architecture and comprehensive training dataset—ClinBERT consistently delivered suboptimal results. Our attempts to enhance language model performance through hierarchical descriptions did not yield consistent improvements, suggesting that simple concatenation may not capture the complexity of clinical relationships in prediction tasks.

6.4. Runtime

KEEP offers a significant advantage in computational efficiency. Generating embeddings for the entire disease terminology using data from 500,000 patients was completed in under 2 hours on a single NVIDIA L40S GPU with 46 GB of memory (see Section A.2 for training details). In contrast, fine-tuning language models (LMs) often requires several hours and multiple GPUs, making KEEP a more practical and resource-efficient solution.

7. Discussion

Our results demonstrate two key findings about medical code embeddings. First, embeddings trained directly on institutional data can outperform LMs when fine-tuning is not feasible. This advantage likely stems from their ability to learn dataset-specific relationships. Second, our approach shows that carefully combining structural knowledge with empirical patterns through regularization creates more effective embeddings for real-world clinical prediction tasks. KEEP produces final embeddings without auxiliary or end-to-end training for specific tasks. This makes KEEP complementary to existing methods. For instance, GRAM (Choi et al., 2016) and HealthGAT (Piya et al., 2024) initialize embeddings with GloVe and node2vec, respectively, before conducting task-specific training. This suggests KEEP embeddings could serve as an initialization for such methods.

These findings have important implications for healthcare institutions. KEEP provides a practical approach for generating high-quality disease representations without requiring extensive computational resources or massive datasets. This efficiency makes our method particularly suitable for resource-constrained healthcare settings.

KEEP's flexibility is another key advantage. Through the regularization parameter λ, institutions can tune the contribution of knowledge graphs and empirical patterns based on their specific needs and data characteristics. Healthcare systems with limited patient data can emphasize knowledge graphs through a higher λ value, while those with rich clinical data might benefit from a lower λ to prioritize empirical patterns. By integrating both knowledge graph structures and local clinical patterns, institutions can create embeddings that better reflect their specific patient populations while maintaining alignment with established medical knowledge. This bal-

Table 4: **Extrinsic evaluation**: AUPRC scores

	All DM	Periph. Vasc.	Kidney	Eye	Neuro-pathy	MI	Renal Fail.	Re-admit	Mean Rank
BioClin BERT	0.59±0.003	0.21±0.013	0.29±0.002	0.45±0.002	0.25±0.002	0.20±0.003	0.35±0.002	0.073±0.005	5.88
BioClin BERT$_H$	0.55±0.002	0.14±0.016	0.30±0.006	0.44±0.001	0.24±0.001	0.20±0.001	0.34±0.001	0.062±0.011	7.81
Clin BERT	0.51±0.010	0.10±0.011	0.28±0.010	0.40±0.010	0.20±0.013	0.20±0.005	0.18±0.025	0.045±0.001	10.19*
Clin BERT$_H$	0.52±0.003	0.15±0.006	0.27±0.008	0.41±0.009	0.23±0.005	0.20±0.001	0.28±0.018	0.045±0.002	9.44*
Bio GPT	0.60±0.001	0.20±0.006	0.33±0.002	**0.48**±0.001	0.22±0.001	0.21±0.000	**0.36**±0.001	0.079±0.007	4.12
Bio GPT$_H$	**0.61**±0.001	0.18±0.005	0.32±0.002	0.41±0.006	0.26±0.005	**0.22**±0.001	0.31±0.019	0.077±0.017	5.88
Cui2vec	0.57±0.002	0.17±0.009	0.32±0.003	0.47±0.001	0.26±0.004	**0.22**±0.001	**0.36**±0.002	0.095±0.004	4.00*
Node2 Vec	0.60±0.001	0.20±0.010	0.34±0.002	0.46±0.001	0.31±0.002	**0.22**±0.001	0.34±0.001	0.103±0.004	4.00*
GloVe	0.54±0.001	0.13±0.012	0.32±0.002	0.46±0.001	0.27±0.003	**0.22**±0.001	**0.36**±0.001	0.113±0.003	5.19
GAT	0.57±0.003	0.22±0.011	0.26±0.013	0.47±0.004	0.18±0.006	0.18±0.002	0.33±0.005	0.113±0.003	7.88
KEEP	**0.61**±0.001	**0.28**±0.003	**0.36**±0.002	**0.48**±0.001	**0.32**±0.002	**0.22**±0.001	0.35±0.001	**0.122**±0.002	**1.62***

Model$_H$ refers to Hierarchy Aware embeddings. Column names: "DM" = Diabetes Mellitus, "Periph. Vasc." = Peripheral Vascular Disease, "Neuro." = Neuropathy, "MI" = Myocardial Infarction, "Renal Fail." = Renal Failure, "Readmit" = 30d readmission (MIMIC-IV), * = p< 0.05

ance is particularly valuable for clinical prediction tasks and rare diseases, where understanding both theoretical disease relationships and real-world manifestations is crucial.

7.1. Complementarity LM's

Rather than competing with LM-based approaches, our approach serves as a complementary tool that addresses specific limitations of language models in handling structured medical data. Our framework bridges structured medical knowledge and unstructured data representations in two key ways. First, our embeddings can be integrated with LM-derived representations through multimodal integration techniques (Ebrahimi et al., 2023). Second, our method could enhance LM performance by providing domain-aware initialization for fine-tuning, potentially improving convergence and reducing data requirements while enabling more effective incorporation of medical ontologies and relationships (Hewitt, 2021; Fatemi et al., 2023).

7.2. Limitations and Future Work

Our knowledge graph implementation currently utilizes only hierarchical relationships between diseases, omitting other important clinical connections such as causal and associative relationships. Incorporating these additional relationship types could provide richer representations of disease interactions (Shen et al., 2019). Also, the current approach does not capture the temporal dynamics of disease progression. Future implementations could address this limitation by incorporating visit-level information rather than aggregating across entire patient histories. Finally, we focus only on disease codes which provides a limited view of patient health. Future work could explore methods for generating comprehensive patient representations that use multiple domains such as medications, labs, and observations.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Khushbu Agarwal, Tome Eftimov, Raghavendra Addanki, Sutanay Choudhury, Suzanne Tamang, and Robert Rallo. Snomed2vec: Random walk and poincar\'e embeddings of a clinical knowledge base for healthcare analytics. *arXiv preprint arXiv:1907.08650*, 2019.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Jineta Banerjee, Jaclyn N Taroni, Robert J Allaway, Deepashree Venkatesh Prasad, Justin Guinney, and Casey Greene. Machine learning in rare disease. *Nature Methods*, 20(6):803–814, 2023.

Andrew L Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S Kohane. Clinical concept embeddings learned from massive sources of multimodal medical data. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 25, page 295, 2020.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

David Chang, Ivana Balažević, Carl Allen, Daniel Chawla, Cynthia Brandt, and Richard Andrew Taylor. Benchmark and best practices for biomedical knowledge graph embeddings. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 167. NIH Public Access, 2020.

Eunsuk Chang, Sumi Sung, et al. Use of snomed ct in large language models: Scoping review. *JMIR Medical Informatics*, 12(1):e62924, 2024.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. In *proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1495–1504, 2016.

Edward Choi, Cao Xiao, Walter Stewart, and Jimeng Sun. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. *Advances in neural information processing systems*, 31, 2018.

Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 1819–1822, 2014.

Yihe Deng, Chenchen Ye, Zijie Huang, Mingyu Derek Ma, Yiwen Kou, and Wei Wang. Graphvis: Boosting llms with visual knowledge graph integration. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Vijay Prakash Dwivedi, Viktor Schlegel, Andy T Liu, Thanh-Tung Nguyen, Abhinav Ramesh Kashyap, Jeng Wei, Wei-Hsian Yin, Stefan Winkler, and Robby T Tan. Representation learning of structured data for medical foundation models. *arXiv preprint arXiv:2410.13351*, 2024.

Sayna Ebrahimi, Sercan O Arik, Yihe Dong, and Tomas Pfister. Lanistr: Multimodal learning from structured and unstructured data. *arXiv preprint arXiv:2305.16556*, 2023.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. Talk like a graph: Encoding graphs for large language models. *arXiv preprint arXiv:2310.04560*, 2023.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

Cheng Guo and Felix Berkhahn. Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*, 2016.

Shibo Hao, Sainbayar Sukhbaatar, Dijia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv [cs.CL]*, December 2024.

John Hewitt. Initializing new word embeddings for pretrained language models, 2021. URL https:/nlp.stanford.edu/~johnhew//vocab-expansion.html.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv (version 1.0). *PhysioNet*, 101:e215–e220, 2021.

Justin M Johnson and Taghi M Khoshgoftaar. Encoding techniques for high-cardinality features and ensemble learners. In *2021 IEEE 22nd international conference on information reuse and integration for data science (IRI)*, pages 355–361. IEEE, 2021.

Michael J Kane, Casey King, Denise Esserman, Nancy K Latham, Erich J Greene, and David A Ganz. A compressed large language model embedding dataset of icd 10 cm descriptions. *BMC bioinformatics*, 24(1):482, 2023.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Junghwan Lee, Cong Liu, Jae Hyun Kim, Alex Butler, Ning Shang, Chao Pang, Karthik Natarajan, Patrick Ryan, Casey Ta, and Chunhua Weng. Comparative effectiveness of medical concept embedding for feature engineering in phenotyping. *JAMIA open*, 4(2):ooab028, 2021.

Simon A Lee, Sujay Jain, Alex Chen, Kyoka Ono, Jennifer Fang, Akos Rudas, and Jeffrey N Chiang. Emergency department decision support using clinical pseudo-notes. *arXiv preprint arXiv:2402.00160*, 2024.

Zhihuang Lin, Dan Yang, and Xiaochun Yin. Patient similarity via joint embeddings of medical knowledge graph and medical entity descriptions. *IEEE Access*, 8:156663–156676, 2020.

Ilya Loshchilov, Cheng-Ping Hsieh, Simeng Sun, and Boris Ginsburg. ngpt: Normalized transformer with representation learning on the hypersphere. *arXiv preprint arXiv:2410.01131*, 2024.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022.

Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781, 2013.

Ruth Ndjaboue, Gérard Ngueta, Charlotte Rochefort-Brihay, Sasha Delorme, Daniel Guay, Noah Ivers, Baiju R Shah, Sharon E Straus, Catherine Yu, Sandrine Comeau, et al. Prediction models of diabetes complications: a scoping review. *J Epidemiol Community Health*, 76(10): 896–904, 2022.

OpenAI. Embeddings frequently asked questions, 2025. URL https://help.openai.com/en/articles/6824809-embeddings-frequently-asked-questions. Accessed: 2025-01-09.

Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Machine Learning for Health*, pages 239–260. PMLR, 2021.

Vaclav Papez, Maxim Moinat, Erica A Voss, Sofia Bazakou, Anne Van Winzum, Alessia Peviani, Stefan Payralbe, Elena Garcia Lara, Michael Kallfelz, Folkert W Asselbergs, et al. Transforming and evaluating the uk biobank to the omop common data model for covid-19 research and beyond. *Journal of the American Medical Informatics Association*, 30(1):103–111, 2023.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Fahmida Liza Piya, Mehak Gupta, and Rahmatollah Beheshti. Healthgat: Node classifications in electronic health records using graph attention networks. In *2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 132–141. IEEE, 2024.

Chava L Ramspek, Ype de Jong, Friedo W Dekker, and Merel van Diepen. Towards the best kidney failure prediction tool: a systematic review and selection aid. *Nephrology Dialysis Transplantation*, 35(9):1527–1538, 2020.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

Jacob Reeh, Christina Bachmann Therming, Merete Heitmann, Søren Højberg, Charlotte Sørum, Jan Bech, Dorte Husum, Helena Dominguez, Thomas Sehestedt, Thomas Hermann, et al. Prediction of obstructive coronary artery disease and prognosis in patients with suspected stable angina. *European heart journal*, 40(18):1426–1435, 2019.

Feichen Shen, Suyuan Peng, Yadan Fan, Andrew Wen, Sijia Liu, Yanshan Wang, Liwei Wang, and Hongfang Liu. Hpo2vec+: Leveraging heterogeneous knowledge resources to enrich node embeddings for the human phenotype ontology. *Journal of biomedical informatics*, 96:103246, 2019.

Yuqi Si, Jingcheng Du, Zhao Li, Xiaoqian Jiang, Timothy Miller, Fei Wang, W Jim Zheng, and Kirk Roberts. Deep representation learning of patient data from electronic health records (ehr): A systematic review. *Journal of biomedical informatics*, 115:103671, 2021.

Ali Soroush, Benjamin S Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W Charney, Girish N Nadkarni, and Eyal Klang. Large language models are poor medical coders—benchmarking of medical code querying. *NEJM AI*, 1(5):AIdbp2300040, 2024.

Ethan Steinberg, Ken Jung, Jason A Fries, Conor K Corbin, Stephen R Pfohl, and Nigam H Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of biomedical informatics*, 113: 103637, 2021.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3): e1001779, 2015.

Alice S Tang, Sarah R Woldemariam, Silvia Miramontes, Beau Norgeot, Tomiko T Oskotsky, and Marina Sirota. Harnessing ehr data for health research. *Nature Medicine*, 30(7):1847–1855, 2024.

Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10):2633–2642, 2023.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135, 2023.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of biomedical informatics*, 126: 103983, 2022.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets llm finetuning: The effect of data, model and finetuning method. *arXiv preprint arXiv:2402.17193*, 2024.

Appendix A. Embedding Creation

Algorithm 1 KEEP

Input: Knowledge Graph $G = (V, E)$, EHR data D, Regularization parameter λ, Learning rate η, Number of epochs T

Output: Final embeddings $W = \{w_i\}$ for all $i \in V$

Stage 1: Knowledge Graph Embedding

Initialize node2vec parameters: *window size, walk length, number of walks per node*

foreach *node* $v \in V$ **do**
| Generate random walks starting from v
end

Train a skip-gram model using random walks to obtain node2vec embeddings:

$$W^{\text{n2v}} = \{w_i^{\text{n2v}}\}$$

Stage 2: EHR-Enhanced Representation Learning

Construct co-occurrence matrix X from EHR data D

Initialize GloVe embeddings with node2vec:

$$w_i \leftarrow w_i^{\text{n2v}}, \quad \forall i \in V$$

Training Loop:

for $t = 1$ *to* T **do**

Compute the **GloVe loss**:

$$L_{\text{GloVe}} = \sum_{i,j=1}^{V} f(X_{ij}) \left(w_i^\top \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2$$

Compute the **Regularization loss**:

$$L_{\text{reg}} = \lambda \sum_{i=1}^{V} ||w_i - w_i^{\text{n2v}}||^2$$

Compute the total objective:

$$J(W) = L_{\text{GloVe}} + L_{\text{reg}}$$

Update embeddings using AdamW:

$$w_i \leftarrow w_i - \eta \cdot \nabla J(W), \quad \forall i \in V$$

end

return *Final learned embeddings* $W = \{w_i\}$

A.1. Node2Vec

Node2vec embeddings were generated using random walks on the OMOP knowledge graph.

A.1.1. Graph Creation

We construct the graph by first filtering for disease-related concepts. Disease concepts are identified using the "CONCEPT" table, selecting entries with "domain_name" "Condition". To limit graph complexity, we filter concepts based on their hierarchical distance from the root node, "Disease" (concept ID: 4274025). Using the "CONCEPT_ANCESTOR" table, we calculate the minimum number of hierarchical levels separating each disease concept from the root node. Concepts with more than five levels of separation are excluded from the graph.

A.2. Model parameters

We used the following parameters to train Node2Vec:

Table 5: Hyperparameters for Node2Vec Training

Hyperparameter	**Value**
Embedding dimension	100
Walk length	30
Number of walks	750
p, q	1
Window size	10
Minimum count	1
Batch size	4096

A.3. GloVe

GloVe embeddings were generated using co-occurrence data from UKBB in OMOP format.

A.4. Co-occurrence matrix

We construct the co-occurrence matrix using the same codes from our previous graph analysis. Instead of excluding codes more than five levels away from the root node, we implement a roll-up procedure that maps each code to its parent codes present in the graph. Adopting a dense roll-up approach, we map every code to *all* of its parents, creating multiple entries when a code has multiple parent nodes. Using this dataset we create a co-occurrence matrix X where each entry X_{ij} represents the frequency of diseases i and j occurring together in a patient's records. To establish the presence of a disease, we require at least two occurrences in a patient's history. Co-occurrence is determined based on the patient's complete medical history, rather than being restricted to individual visits.

A.4.1. Model parameters

We used the following parameters to train GloVe:

Note the regularization parameter λ is only used in KEEP.

A.5. Pre-trained embeddings

We utilized three pre-trained biomedical language models to extract disease embeddings in two ways: **Basic** and **Hierarchy-aware** description embeddings. The basic version uses only the description for the disease code. To incorporate hierarchical context, we constructed ancestry paths for each disease concept by combining descriptions from its hierarchy, specifically parent and grandparent concepts. The format for the hierarchical path was:

```
"[Disease]; is a [Parent]; is a [Grandparent]"
```

Table 6: Hyperparameters for GloVe Training

Hyperparameter	Value
Embedding dimension	100
Learning rate	0.05
Number of epochs	300
Batch size	1024
X_{max}	75^{th} percentile
α	0.75
λ	1×10^{-3}

As we maintain the embedding dimensions for the BERT and GPT-based models, they have the following dimensions:

Table 7: Embedding sizes for Pre-trained models.

Model Name	Embedding Size
BioClinicalBERT	768
ClinicalBERT	768
BioGPT	1024

To mitigate the risks overfitting and ensure numerical stability, we applied L2 normalization to these embeddings, following the recommendations of (OpenAI, 2025) and (Loshchilov et al., 2024).

Appendix B. Experiment setup

B.1. Assessing the Impact of Regularization

For each code, we identified the ten most similar concepts based on cosine similarity, along with 150 randomly sampled concepts. For each selected concept, we computed its Resnik similarity (refer to Section 3.1 for calculation details) and its co-occurrence frequency with the original code. We the measured the correlation between cosine similarity and both Resnik similarity and co-occurrence values. To ensure statistical robustness, we repeated the experiment 250 times report the median correlation values across all runs.

B.2. Intrinsic evaluation

To evalute the quality of the embedding space, we assessed the embeddings' ability to identify known medical relationships, adapting the approach described by (Beam et al., 2020). This benchmark evaluates how well the embeddings capture established medical knowledge and real-world disease associations by measuring their ability to discriminate between known disease relationships.

For each core concept, we constructed positive pairs that included known relationships, such as complications, comorbidities, child diseases, and synonyms. Negative pairs were generated by randomly sampling concept pairs from the terminology, ensuring no overlap with the positive set. Discrimination performance was quantified as the proportion of known relationships achieving higher cosine similarity than 95% of random pairs for each core concept. Formally, this can be represented as $P\left(\text{sim}(+\text{pair}) > \text{sim}(-\text{pair})\right)$ where $\text{sim}(\cdot)$ denotes the cosine similarity between embeddings. To ensure robust estimates, we employed repeated the experiment times and used bootstrap sampling (1000 iteration) to calculate median scores, confidence intervals, and statistical significance. Table 8 shows the OMOP concept IDs for the core concept and the known relationship concepts used in the intrinsic evaluation.

Table 8: **Intrinsic Evaluation**: OMOP concept IDs for core concepts along with their associated synonyms, child diseases, complications, and comorbidities used for intrinsic evaluation.

Core Concept	Synonyms	Child Diseases	Complications	Comorbidities
Asthma (317009)	·	4123253, 4051466, 4110051, 4233784, 40483397, 4145497, 312950, 443801, 4191479, 4155469, 4155470, 4119298, 4155468	257581, 45769438	4148368, 260123, 255573, 4214438, 4280726, 255841, 4182370, 4335888, 4112367, 4305500, 4283893, 256439, 256448, 259848, 4110489, 4110492, 4329087, 4223595, 380111, 4138403, 4195007, 257007, 133835, 257012, 42537251, 42537252
Obesity (433736)	·	434005, 4189665, 4217557, 4087487	4059290, 4026131, 436940, 4100857, 40482277, 40484532	4079750, 314378, 4291025, 321318, 4083696, 313459, 373503, 201820, 44809569, 381591, 321588, 315286, 4079749, 314666, 4112022, 442604, 44809026, 4209293, 73553, 374384, 4110196, 320128, 442588, 4147779, 4098302, 381316, 4045734, 443454, 4149320, 4186397, 439150, 435524, 443732, 201826, 4185932, 319844, 312327, 80180, 316866, 4043734, 432867, 319034, 4314692, 40481943, 46271022, 4170226, 4151170, 317576
Premature rupture of membranes (194702)	·	4064296, 45772076, 4060691	·	36712695, 434111, 4058243, 43530976, 4049790, 438815, 435656, 4042220, 437623, 72693, 435875, 4146482, 433823, 4118910, 443247, 4304781, 4024659, 4062791
Renal transplant rejection (4128369)	4309006, 4324887, 4309320, 4127554	·	201461, 140362, 4324887, 36715574	443612, 443614, 4242411, 443731, 4030664, 4056478, 4264718, 198185, 4059452, 4153654, 443597, 443601, 192359, 437247, 4126305, 312358, 4206115, 4118795, 4128221, 196455, 606956, 200687, 4030518, 192964, 201313, 4269363, 45757752, 4131748, 443919, 35623051, 252365, 197320, 45757772, 4220238, 193253, 42536547, 435308, 195556, 443611, 35624383, 46271022, 192279
Schizophrenia (435783)	4100365	4286201, 435219, 4100365, 4153292	4286201, 4101149, 432590, 4100247, 4102670, 433450, 618641, 436073, 4335169	4152280, 4159691, 434613, 4308866, 4004672, 438028, 434223, 440383, 4098302, 40481346, 442077, 4149320, 4103574, 4149321, 4282316, 36713290, 4314692, 435520, 434819, 4239381, 4338031, 4151170
Tumor of respiratory system (40491439)	·	4180795, 45769033, 4054511, 4112735, 4093957, 4242982, 78093, 4247331, 4128888, 4113116, 252840, 4114353, 4240153, 40493428, 4055270, 261528, 4054836	192568, 4129382, 434875, 4317284, 319049, 317003, 4246451, 441258, 4114341, 72266, 320342, 4256228, 443588, 4131304, 439751, 4102360, 4180795, 432851, 4233244, 318096, 4130839	4124677, 257011, 255573, 37206139, 4307774, 44807895, 255841, 4317284, 4112341, 4170143, 4000938, 4115044, 4119786, 4110056, 4208807, 4195694, 4110479, 253506, 4131304, 40491473, 132797, 4175297, 260139, 4112357, 4148204, 256451, 257004
Type 1 diabetes mellitus (201254)	4130161, 201820, 4034959, 44808373, 4308509, 4008576, 4311629	·	42536605, 443727, 4058243, 4029423, 443735, 376979, 4174262, 4030664, 42538169, 377821, 380097, 4082346, 4034959, 4206115, 443730, 4128221, 318712, 200687, 4016047, 321822, 435216, 442793, 4044391, 376112, 24609, 4214376, 4048028, 4209145, 4174977, 30361, 443767, 192279	315286, 4185932, 319844, 376686, 443612, 4217557, 443614, 433736, 443597, 443601, 443611, 46271022
Type 2 diabetes mellitus (201826)	44808385, 201820, 4034959, 44808373, 4308509, 4008576, 4311629	·	376979, 443731, 443729, 380097, 376065, 4082346, 4206115, 443730, 321822, 442793, 4044391, 376112, 443733, 443732, 4174977, 4223739, 443767, 192279, 4221487	434005, 319826, 443612, 443614, 373503, 321318, 4119612, 44809569, 4087487, 4200991, 321588, 315286, 314666, 4217557, 442604, 44809026, 4209293, 443597, 443601, 374384, 374034, 320128, 4098302, 4149320, 4186397, 43021237, 4185932, 319844, 4149321, 312327, 316866, 433736, 4043734, 432867, 4124836, 3655355, 4314692, 443611, 46271022, 4170226, 4151170, 317576

All child concepts of codes shown are also used

B.3. Extrinsic evaluation

Below we detail the evaluation methodology. Implementation code is available at https://github.com/G2Lab/keep

B.3.1. Data Preprocessing and Cohort Creation

We required disease concepts to appear a minimum of two times in a patient's history for inclusion. Table 9 shows the OMOP concept IDs used to define the cohort and outcome. For patients with positive outcomes, we required two instances of the outcome code and censored all data after the first occurrence to ensure only preceding medical history was considered. Sequence lengths were limited to a maximum of 20 codes, with random sampling applied to the small subset of sequences exceeding this limit. The final dataset was partitioned using label stratification to maintain outcome distributions: an initial 80-20 train-test split was performed, followed by an 80-20 subdivision of the training data, yielding final proportions of 64% training, 16% validation, and 20% testing data.

Table 9: **Extrinsic evaluation**: OMOP concept IDs used to define cohort, outcome and exclusions

Inclusion	Outcome	Exclusion
201826	442793	201254
201826	321822	201254
201826	1992279	201254
201826	443767	201254
201826	443730	201254
321318	4329847	314666*
46271022	197320, 432961, 444044	·

B.3.2. Model Architecture

We implemented a transformer model with one encoder layer and four attention heads. Input dimensions match those of the corresponding embeddings. Regularization was applied through dropout (p=0.2). We employed attention-based pooling for sequence representation, followed by a binary classification layer.

B.3.3. Training Protocol

Models were trained with a batch size of 32 using AdamW optimization (weight decay=0.01) and cross-entropy loss. Learning rates were evaluated across nine values ranging from $1e^{-5}$ to $1e^{-1}$ in logarithmic increments using grid search. We trained for a maximum of 500 epochs with early stopping triggered after 5 epochs without validation loss improvement. We repeated each experiment 5 times and selected the learning rate and epoch that produced the lowest median validation loss. Final evaluation comprised 100 independent runs, with confidence intervals computed via bootstrap resampling (1000 iterations).

B.3.4. Evaluation Metrics

Model performance was primarily assessed using AUPRC AUC, MCC, F1 score, and Test loss. For statistical comparison between embedding methods, we employed the Wilcoxon signed-rank test on the performance ranks across all tasks.

Appendix C. Experiment results

Tables 10-13 show the results for Test loss, AUC, MCC, and F1 score in the extrinsic evaluation tasks.

Table 10: **Extrinsic evaluation**: Test loss (lowest is best)

	All DM	Periph. Vasc.	Kidney	Eye	Neuro-pathy	MI	Renal Fail.	Re-admit	Mean Rank
BioClin BERT	0.53±0.003	0.63±0.008	0.59±0.001	0.55±0.001	0.61±0.001	0.61±0.003	0.38±0.008	0.67±0.002	6.12
BioClin BERT$_H$	0.55±0.001	0.65±0.007	0.57±0.003	0.55±0.001	0.62±0.001	0.62±0.002	0.37±0.002	0.70±0.003	7.75
Clin BERT	0.57±0.005	0.63±0.008	0.57±0.006	0.57±0.004	0.62±0.006	0.62±0.005	0.56±0.029	0.68±0.001	9.12
Clin BERT$_H$	0.57±0.001	0.63±0.004	0.59±0.004	0.57±0.005	0.61±0.004	0.62±0.001	0.43±0.024	0.68±0.001	8.50
Bio GPT	0.52±0.001	0.65±0.007	0.56±0.001	0.53±0.001	0.59±0.001	0.60±0.000	0.37±0.001	0.68±0.010	4.38
Bio GPT$_H$	0.52±0.001	0.64±0.006	0.56±0.001	0.55±0.002	0.58±0.003	0.61±0.002	0.44±0.023	0.70±0.008	6.06
Cui2vec	0.54±0.001	0.60±0.006	0.60±0.005	0.53±0.001	0.59±0.001	0.59±0.000	0.38±0.002	0.67±0.002	4.31
GloVe	0.55±0.001	0.62±0.008	0.56±0.002	0.56±0.002	0.58±0.001	0.59±0.000	0.37±0.001	0.67±0.003	4.56
Node2 Vec	0.53±0.001	0.60±0.005	0.58±0.002	0.55±0.001	0.58±0.001	0.60±0.001	0.38±0.001	0.67±0.002	4.56
GAT	0.55±0.002	0.60±0.008	0.60±0.006	0.55±0.005	0.64±0.020	0.62±0.001	0.39±0.005	0.69±0.006	7.88
KEEP	0.52±0.001	0.60±0.002	0.57±0.002	0.54±0.001	0.57±0.001	0.59±0.000	0.38±0.001	0.65±0.001	**2.75**

Table 11: **Extrinsic Evaluation**: AUC scores

	All DM	Periph. Vasc.	Kidney	Eye	Neuro-pathy	MI	Renal Fail.	Re-admit	Mean Rank
BioClin BERT	0.79±0.001	0.70±0.009	0.73±0.001	0.78±0.001	0.71±0.001	0.74±0.002	0.90±0.002	0.65±0.003	6.31
BioClin BERT$_H$	0.77±0.001	0.60±0.016	0.75±0.002	0.78±0.001	0.70±0.001	0.74±0.001	0.90±0.000	0.62±0.026	7.25
Clin BERT	0.75±0.005	0.63±0.018	0.73±0.005	0.76±0.004	0.69±0.006	0.73±0.007	0.67±0.040	0.61±0.001	10.0
Clin BERT$_H$	0.76±0.001	0.72±0.007	0.73±0.005	0.76±0.004	0.70±0.005	0.73±0.001	0.83±0.029	0.61±0.001	8.94
Bio GPT	0.80±0.000	0.72±0.003	0.75±0.001	0.79±0.001	0.73±0.001	0.75±0.000	0.90±0.000	0.67±0.009	**2.69**
Bio GPT$_H$	**0.81**±0.000	**0.75**±0.002	**0.75**±0.001	0.78±0.001	0.73±0.003	0.75±0.001	0.85±0.024	0.63±0.032	4.00
Cui2vec	0.78±0.001	0.71±0.012	0.74±0.003	**0.79**±0.001	0.71±0.002	**0.76**±0.000	0.90±0.000	0.66±0.002	3.69
Node2 Vec	0.79±0.001	0.69±0.008	0.74±0.002	0.77±0.001	0.71±0.002	0.75±0.001	0.90±0.000	0.66±0.003	6.44
GloVe	0.77±0.001	0.69±0.018	0.74±0.002	0.78±0.001	0.71±0.002	0.75±0.000	**0.90**±0.000	0.66±0.003	4.50
GAT	0.77±0.001	0.69±0.018	0.74±0.002	0.78±0.001	0.71±0.002	0.75±0.000	**0.90**±0.000	0.58±0.029	9.19
KEEP	0.80±0.001	0.74±0.002	0.74±0.001	0.79±0.001	**0.74**±0.002	0.75±0.000	0.90±0.000	**0.67**±0.001	3.00

Table 12: **Extrinsic Evaluation**: Mathews Correlation Coefficient

	All DM	Periph. Vasc.	Kidney	Eye	Neuro-pathy	MI	Renal Fail.	Re-admit	Mean Rank
BioClin BERT	0.39±0.009	0.19±0.009	0.25±0.002	0.29±0.004	0.11±0.001	0.20±0.007	0.31±0.008	0.09±0.005	5.69
BioClin BERT$_{HA}$	0.37±0.004	0.14±0.021	0.18±0.005	0.30±0.004	0.08±0.001	0.21±0.001	0.32±0.001	0.00±0.000	7.38
Clin BERT	0.33±0.013	0.13±0.020	0.17±0.009	0.24±0.007	0.14±0.010	0.18±0.009	0.14±0.031	0.06±0.006	9.75
Clin BERT$_{HA}$	0.35±0.005	0.17±0.011	0.25±0.011	0.29±0.008	0.24±0.010	0.19±0.002	0.26±0.023	0.07±0.000	6.38
BioGPT	0.44±0.003	0.15±0.003	0.17±0.002	0.32±0.004	0.15±0.002	0.21±0.001	0.33±0.001	0.08±0.015	5.25
BioGPT$_{HA}$	0.47±0.004	0.19±0.003	0.17±0.002	0.27±0.004	0.14±0.003	0.19±0.004	0.27±0.024	0.03±0.032	7.31
Cui2vec	0.43±0.005	0.13±0.007	0.22±0.004	0.29±0.003	0.17±0.002	0.22±0.001	0.33±0.002	0.09±0.002	4.75
GloVe	0.48±0.003	0.13±0.008	0.21±0.003	0.35±0.002	0.21±0.004	0.20±0.001	0.31±0.001	0.11±0.004	3.94
Node2Vec	0.39±0.005	0.13±0.013	0.21±0.003	0.27±0.005	0.19±0.002	0.22±0.001	0.34±0.001	0.11±0.004	4.94
GAT	0.36±0.008	0.12±0.013	0.17±0.009	0.33±0.011	0.20±0.084	0.17±0.005	0.32±0.004	0.03±0.023	7.75
KEEP	0.41±0.002	0.20±0.003	0.26±0.004	0.35±0.003	0.18±0.003	0.21±0.001	0.33±0.001	0.11±0.004	**2.88**

Table 13: **Extrinsic Evaluation**: F1 score

	All DM	Periph. Vasc.	Kidney	Eye	Neuro-pathy	MI	Renal Fail.	Re-admit	Mean Rank
BioClin BERT	0.69±0.010	0.58±0.005	0.62±0.001	0.63±0.004	0.47±0.004	0.55±0.005	0.57±0.003	0.48±0.008	5.12
BioClin BERT$_{H}$	0.68±0.003	0.56±0.010	0.55±0.003	0.64±0.004	0.42±0.003	0.58±0.003	0.56±0.002	0.49±0.000	6.38
Clin BERT	0.65±0.011	0.56±0.009	0.55±0.005	0.59±0.007	0.56±0.005	0.50±0.018	0.53±0.011	0.40±0.017	9.00
Clin BERT$_{H}$	0.67±0.003	0.57±0.005	0.62±0.006	0.63±0.006	0.62±0.005	0.52±0.002	0.56±0.008	0.38±0.000	6.19
Bio GPT	0.72±0.002	0.54±0.003	0.54±0.002	0.65±0.003	0.54±0.002	0.54±0.002	0.57±0.003	0.38±0.060	6.75
Bio GPT$_{H}$	0.73±0.002	0.57±0.002	0.55±0.003	0.61±0.003	0.50±0.004	0.48±0.008	0.57±0.012	0.50±0.009	6.25
Cui2vec	0.72±0.002	0.52±0.004	0.60±0.003	0.62±0.003	0.56±0.002	0.54±0.001	0.60±0.003	0.46±0.004	5.62
GloVe	0.69±0.002	0.54±0.007	0.59±0.002	0.60±0.006	0.58±0.002	0.54±0.001	0.59±0.002	0.50±0.004	5.25
Node2 Vec	0.73±0.001	0.53±0.006	0.59±0.002	0.67±0.001	0.59±0.002	0.51±0.002	0.57±0.001	0.49±0.004	4.88
GAT	0.68±0.006	0.52±0.012	0.57±0.008	0.66±0.009	0.59±0.044	0.47±0.009	0.57±0.003	0.47±0.021	6.88
KEEP	0.70±0.001	0.59±0.002	0.62±0.002	0.67±0.002	0.56±0.002	0.53±0.001	0.59±0.001	0.47±0.004	**3.69**

Model$_{\text{H}}$ refers to Hierarchy Aware embeddings. Column names: "DM" = Diabetes Mellitus, "Periph. Vasc." = Peripheral Vascular Disease, "Neuro." = Neuropathy, "MI" = Myocardial Infarction, "Renal Fail." = Renal Failure, "Readmit" = 30d readmission (MIMIC-IV), * = p< 0.05