Proceedings of Machine Learning Research 225:52–81, 2023 Machine Learning for Health (ML4H) 2023

Multi-modal Graph Learning over UMLS Knowledge Graphs

Manuel Burger MANUEL.BURGER@INF.ETHZ.CH
Gunnar Rätsch RAETSCH@INF.ETHZ.CH
Rita Kuznetsova RITA.KUZNETSOVA@INF.ETHZ.CH
Department of Computer Science, ETH Zürich, Switzerland

Abstract

Clinicians are increasingly looking towards machine learning to gain insights about patient progression. We propose a novel approach named *Multi-Modal UMLS Graph Learning (MMUGL)* for learning meaningful representations of medical concepts using graph neural networks over knowledge graphs based on the unified medical language system. These concept representations are aggregated to represent a patient visit and then fed into a sequence model to perform predictions at the granularity of multiple hospital visits of a patient. We improve performance by incorporating prior medical knowledge and considering multiple modalities. We compare our method to existing architectures proposed to learn representations at different granularities on the MIMIC-III dataset and show that our approach outperforms these methods. The results demonstrate the significance of multi-modal medical concept representations based on prior medical knowledge.

We provide our code on GitHub[1].

Keywords: Knowledge Graphs, EHR, UMLS, Multi-Modal

1. Introduction

Modern healthcare facilities record patient information as Electronic Health Records (EHR). EHR Datasets such as MIMIC-III (Johnson et al., 2016), HiRID (Faltys et al., 2021), and eICU (Pollard et al., 2018) enable modeling of disease progressions within a single hospital visit, for example in Intensive Care Units (ICU) (Harutyunyan et al., 2019; Yèche et al., 2021), or progressions across multiple patient visits (Choi et al., 2018). These progressions can be meaningfully encoded into patient representations using deep learning as shown by numerous prior works (Choi et al., 2016, 2018, 2017; Bai et al., 2018; Lu et al., 2021a, 2022). This large body of work highlights the value of strong patient representations which aggregate information across entire patient histories from multiple hospital stays, enabling clinicians to model potential risks in various predictive tasks regarding patients' evolution.

We see advantages of recent multi-modal approaches in the ICU setting (Khadanga et al., 2019; Husmann et al., 2022) and visit sequence modeling (Lu et al., 2021a). In multi-modal EHR representation learning (Park et al., 2022), we benefit from two modalities: structured EHR data (e.g., billing codes) and unstructured text information stored in rich clinical reports. Other modalities of medical data exist outside of in-hospital datasets, where a vast amount of prior medical knowledge is stored in static form in databases such as the Unified Medical Language System (UMLS) (Bodenreider, 2004).

We identify two drawbacks of current UMLS-based approaches (Beam et al., 2020; Skreta et al., 2021; Mao and Fung, 2020). First, the approaches do not consider a complete set of relational information stored in UMLS (not considering multiple vocabularies) and solely use UMLS as a unified concept space. Second, prior solutions (Skreta et al., 2021; Mao and Fung, 2020) specify the usage of hierarchical relations, which implies the use of an underlying graph in the form of a tree (single vocabulary). More complex graph structures inside and across vocabularies are thus omitted.

We introduce *Multi-Modal UMLS Graph Learning (MMUGL)* to overcome the previously stated limitations. *MMUGL* is a novel approach for learning representations over medical concepts extracted from the UMLS Metathesaurus in the form of a complex knowledge graph and relations; extracted using a simple and ambitious procedure including a considerable set of vocabularies and all the relations across and within them.

[1] https://github.com/ratschlab/mmugl

© 2023 M. Burger, G. Rätsch & R. Kuznetsova.

We train a shared latent space (Liu et al., 2021b) using auto-encoder pretraining techniques (Shang et al., 2019) and we bridge the modality gap between structured EHR codes and unstructured text. The proposed approach includes rich prior knowledge, which is important in the medical domain. It deals with sample scarcity by relying on prior knowledge structure and pretraining techniques, and it leverages multiple modalities as inputs.

Contributions

- In Section 4.1, we present a new medical knowledge representation technique using graph neural networks (GNN) on the UMLS Metathesaurus of unprecedented complexity. Previous research has explored this knowledge, but we go a step further and demonstrate that we can extract large and intricate knowledge graphs by taking into account a substantial portion of the entire UMLS Metathesaurus and incorporating a strong structural prior into our machine learning model, resulting in improved performance.

- We introduce a shared latent *Concept Embedding* (Sec. 4.1) space and a shared *Visit Encoder* (Sec. 4.3) to optimize the single latent space from any modality jointly in a parameter efficient manner. Opposed to the idea of separate model inputs per modality, we show the benefits of grounding all modalities with the same prior knowledge and training a single latent space for all input modalities (structured and unstructured EHR).

- In Section 6 we demonstrate, that we strongly outperform prior graph-based works in pretraining and downstream tasks and can perform competitively with or outperform prior work trained at a much larger scale of data. We show increased relevance of our approach at longer prediction horizons for readmission predictions.

2. Related Work

In the following, we introduce related work in EHR modeling, knowledge graph learning, and graph learning in the context of EHRs.

EHR Various types of deep learning architectures have been proposed to learn representations at different granularities (patients, visits, histories, etc.) in EHR datasets. Choi et al. (2016); Bai et al. (2018) propose EHR-specific visit sequence models. Choi et al. (2018) propose to focus on the inherent structure of EHRs w.r.t. treatments, diagnosis, visits, and patients. Rasmy et al. (2021) adapt the masked language modeling approach to learn medical concept embeddings.

Multi-Modality Prior work has considered learning representations from either structured components of EHR data (Choi et al., 2018, 2017; Rasmy et al., 2021) or from unstructured clinical text reports (Alsentzer et al., 2019; Peng et al., 2019). Meng et al. (2021); Gong and Guttag (2018); Suresh et al. (2017) have proposed multi-modal architectures and Park et al. (2022) go a step further and introduce even stronger structural priors, while considering the two modalities of structured EHR data, as well as unstructured clinical reports.

Knowledge Graphs and GNNs A vast amount of static prior medical knowledge often remains untouched in current modeling approaches. This prior knowledge can be extracted and transformed into knowledge graphs (Rotmensch et al., 2017; Harnoune et al., 2021). Existing work in natural language processing has established the benefits of knowledge graph representations to various downstream applications (Zhang et al., 2019; Sun et al., 2020; He et al., 2020; Wang et al., 2021); where the most recent approaches include GNNs (Yasunaga et al., 2021, 2022). We aim to leverage the success of GNNs to learn node (and edge) representations (Kipf and Welling, 2017; Hamilton et al., 2017; Veličković et al., 2018).

Graph Learning in EHR *GRAM* (Choi et al., 2017) proposed to include prior knowledge from medical ontologies such as the International Classification of Diseases (ICD). To model structural and relational data explicitly, approaches have started to use GNNs. Shang et al. (2019) proposed to use the Graph Attention (Veličković et al., 2018) operator together with an architecture to pretrain embeddings over two ontologies.

Other works learn over heterogeneous graphs with different types of nodes (Lu et al., 2021a; Mao et al., 2022; Gong et al., 2021). Lu et al. (2022) construct a global graph of diseases, as well as dynamic local (within a single visit) subgraphs. Choi et al. (2020) focus on the EHR structure within a single visit. Finally, Lu et al. (2021b, 2019) consider hyperbolic embeddings for medical ontologies. The learned embed-

dings can then be incorporated into task-specific architectures (Rasmy et al., 2021; Shang et al., 2019; Lu et al., 2021a; Ma et al., 2018) to improve outcome predictions in different healthcare settings.

Previous approaches do consider dataset-specific structures such as the hierarchical organization of EHRs (patients, visits, etc.) and co-occurrence information or structure coming from ontologies. However, the explored set of ontologies is usually kept small and most of them are tree-like structures. To the best of our knowledge, no prior work has considered using a GNN directly on top of a complex large-scale ontology such as the UMLS Metathesaurus and the complete set of unstructured relational information within it.

Further, while previous work considered multiple modalities, they use fusion approaches to join modalities, which tend to rely on larger architectures and thus require larger amounts of data to train effectively. Our work proposes to use the learned knowledge representations over the UMLS Metathesaurus as a single shared latent space for information coming from both the structured (billing codes) and unstructured modalities (clinical reports).

3. Glossary

We consider an EHR dataset of multiple patients and present the following terminology:

- *Patient*: p_i indexed by i
- *Visit*: a patient p_i has one or multiple visits $v_{i,t}$ indexed by t. A visit contains a set of medical concepts $c \in \mathcal{C}_{i,t}$, the total set of medical concepts over the dataset is then $\mathcal{C} = \cup_{\forall i,t}\mathcal{C}_{i,t}$. A medical concept can be of different *types* and we distinguish them by index $\mathcal{C}(*)$:
  - *Disease*: indexed by d s.t. $\mathcal{C}_{i,t}(d)$ and $\mathcal{C}(d) = \cup_{\forall i,t}\mathcal{C}_{i,t}(d)$ the total set of disease concepts
  - *Medication*: (or prescriptions) with type m, similar to diseases we introduce $\mathcal{C}(m)$ and $\mathcal{C}_{i,t}(m)$
  - *Concept from clinical reports*: a set of medical concepts extracted from text data (clinical reports, Sec. 4.2). The total set of considered medical concepts from text $\mathcal{C}(n) = \cup_{\forall i,t}\mathcal{C}_{i,t}(n)$ where the set $\mathcal{C}_{i,t}(n)$ is collected from all reports at a specific visit t of patient i. The type is n for text *note*.

  The vector representation of a visit considering data of a specific type $*$ is $\mathbf{v_{i,t}}(*) \in \mathbb{R}^k$.
- *Ontology*: each ontology has a vocabulary $\mathcal{V}_{Ont}$ of medical concepts c and defines some relation amongst the concepts of the vocabulary using an edge set $\mathcal{E}_{Ont}$, which constitutes the ontology graph $\mathcal{G}_{Ont} = (\mathcal{V}_{Ont}, \mathcal{E}_{Ont})$. We consider the following ontologies/databases:
  - $\mathcal{G}_{ICD}$ (International Classification of Diseases) where $\mathcal{C}(d) \subseteq \mathcal{V}_{ICD}$
  - $\mathcal{G}_{ATC}$ (Anatomical Therapeutic Chemical Classification) where $\mathcal{C}(m) \subseteq \mathcal{V}_{ATC}$
  - $\mathcal{G}_{UMLS}$ (Unified Medical Language System) where $\{\mathcal{C}(d) \cup \mathcal{C}(m) \cup \mathcal{C}(n)\} = \mathcal{C} \subseteq \mathcal{V}_{UMLS}$
- *Modality*: We leverage two EHR data modalities in this work: *Structured* and *unstructured* EHRs. We consider billing code information provided in a tabular fashion to be *structured* EHR and clinical notes to be a form of *unstructured* EHR.

4. Method

The architecture consists of three main components (Figure 1 provides an overview).

- **Concept embedding module** $f_\theta(c) : \mathcal{C} \mapsto \mathbb{R}^k$ (parametrized by θ), which computes a representation for any given medical concept c.
- **Visit encoding module** Assume $q = |\mathcal{C}_{i,t}|$ and $r \in \{2, 3\}$ the number of concept types considered (diseases and medications with or without concepts from text) then $g_\psi(v) : \mathbb{R}^{q\times k} \mapsto \mathbb{R}^{r\times k}$ (parametrized by ψ), which, given all concept token representations of a single visit v computes single representations for each different type of tokens thereof.
- **Predictor module** which performs either a pretraining task on a single visit or a downstream fine-tuning task across a sequence of visits. In either case, this module receives representations for each visit of a patient from the previous visit encoding module.

In the following subsections, we introduce the *Concept Embedding* module (Sec. 4.1), present how we extract richer concepts from clinical reports (Sec. 4.2), encode the information (Sec. 4.3), and perform predictions (Sec. 4.4).

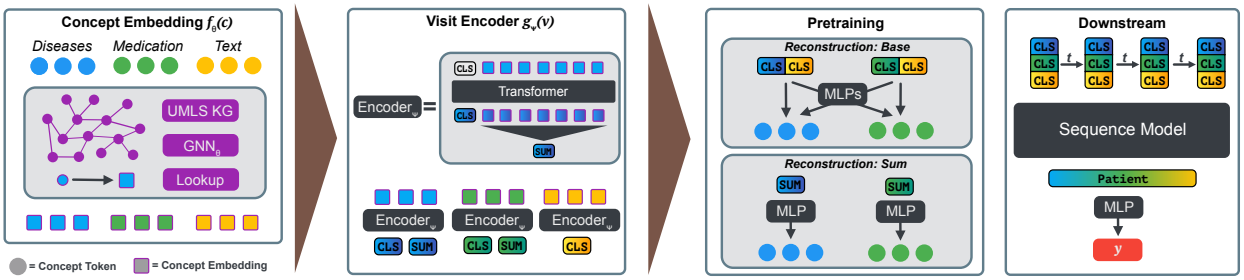


Figure 1: *Model Architecture* brown arrows indicate information flow (Sec. 4); visit information $v_{i,t}$ goes through the concept embedding f_θ (Sec. 4.1), the representations get aggregated by g_ψ (Sec. 4.3) and then used by the pretraining single visit model and the downstream sequence models (Sec. 4.4). Data extraction illustrated in App. A, Fig. 4.

4.1. Concept Embeddings

We present our novel approach to rely on the UMLS Metathesaurus as a unified concept space to learn representations for any general medical concept present in the database based on multiple modalities. Given that, we refer to our approach as *Multi-Modal UMLS Graph Learning (MMUGL)*.

To constrain the number of concepts we consider from the database we use the set of clinical reports present in EHR datasets such as MIMIC-III (Johnson et al., 2016). Using an extraction pipeline (Sec. 4.2) we collect the set of medical concepts $\mathcal{C}(n)$. The final vocabulary includes diseases and medications $\{\mathcal{C}(d) \cup \mathcal{C}(m) \cup \mathcal{C}(n)\} = \mathcal{C} = \mathcal{V}_{UMLS}$ and is used to construct $\mathcal{G}_{UMLS}$ by extracting all the edges in UMLS fully contained within the vocabulary.

In UMLS many concepts are annotated with a short natural language description. We use *SapBERT* (Liu et al., 2021a), a fine-tuned language model to represent UMLS concepts, to initialize the node embeddings from these descriptions (ablation in Appendix B.2). This contributes in two ways: (i) by not using trainable embeddings, we reduce the otherwise huge amount of free parameters given the large vocabulary $\mathcal{V}_{UMLS}$ (ii) we incorporate prior medical knowledge by considering the concept descriptions. We then train a multi-layer GNN on top of the extracted graph:

$$f_\theta(c) = \big(GNN_\theta(\mathcal{G}_{UMLS})\big)[c] \tag{1}$$

To retrieve a concept, we return its computed node embedding. We additionally found it to be beneficial for performance to consider two distinct stacks of GNN layers over the same graph and perform a *Max-Pooling* operation after the final layer across the two stacks.

We consider two baseline graphs. First the tuple of the ICD and ATC hierarchy ICD/ATC (Shang et al., 2019) and second an extension thereof including co-occurrence information $ICD/ATC - CO$ (e.g. Lu et al. (2021a)). More details are in Appendix A.3.2. Additionally we consider a structureless baseline by replacing the concept embedding with an *Embedding Matrix* (see App. A.6.6) which learns a representation for each concept without graph connectivity.

4.2. Concept Extraction

The goal of our approach is to include data from additional modalities such as the clinical reports found in EHR datasets. *MMUGL* learns modality agnostic representations of medical concepts based on UMLS knowledge. It fuses discrete code information (e.g., ICD codes) with medical concepts extracted from text. The extraction with *QuickUMLS* (Soldaini and Goharian, 2016) yields a *set* of medical concepts $\mathcal{C}_{i,t}(n)$ based on the collection of clinical reports of that particular visit. Further, we perform a rule-based negation extraction using *NegEx* (Chapman et al., 2001); for each concept, we extract a binary feature, whether it is negated or not, and concatenate it with its learned concept embedding (Eqn. 1). This is a crucial piece of information as clinical reports can both mention the existence or the absence of a certain condition.

4.3. Visit Encoder

The visit encoder $g_\psi(v)$ relies on the established Transformer architecture (Vaswani et al., 2017). Us-

ing a learned CLS token and without positional encoding, this serves as a set aggregation function. For each concept type in a given visit, we encode a separate representation using the same (weight-sharing) $Transformer_\psi$ with parameters ψ where $* \in \{d, m, n\}$. The concatenated representations $\mathbf{v_{i,t}}(*)$ encode the visit:

$$\mathbf{v_{i,t}}(*) = Transformer_\psi\Big(f_\theta\big(\mathcal{C}(*)_{i,t}\big)\Big)[\texttt{CLS}], \quad (2)$$

$$g_\psi(v) = (\mathbf{v_{i,t}}(d), \mathbf{v_{i,t}}(m), \mathbf{v_{i,t}}(n)) \quad (3)$$

we can omit the multi-modal information from clinical reports (Eqn. 4), e.g., in cases where we use a simpler baseline graph such as *ICD/ATC* (Appendix Eqn. 12) or in *MMUGL without* $\mathcal{C}(n)$.

$$g_\psi(v) = (\mathbf{v_{i,t}}(d), \mathbf{v_{i,t}}(m)) \quad (4)$$

4.4. Predictors and Training

In the following, we introduce the pretraining module and downstream fine-tuning modules.

4.4.1. Pretraining Module

We extend the auto-encoding pretraining approach developed by Shang et al. (2019) with the reconstruction loss $\mathcal{L}_{recon}$ and perform four different predictions (from each of the two *types*, disease and prescription, as a source to either as the label) using distinct Multi-layer Perceptrons ($MLP_{\bullet\to*}$ predicting type $*$ from representations of type $\bullet$) and attach a binary cross-entropy loss $\mathcal{L}_{BCE}$ to model multi-label classification.

$$\mathcal{L}_{recon} = \sum_{\bullet,*\in\{d,m\}} \mathcal{L}'(\bullet,*), \quad (5)$$

$$\mathcal{L}'(\bullet,*) = \mathcal{L}_{BCE}\Big(MLP_{\bullet\to*}\big(\mathbf{v}(\bullet)\big), \mathcal{C}(*)\Big)$$

During pretraining we randomly mask and replace tokens at the input in Eqn. 2 (inspired by masked language modeling (Devlin et al., 2019)).

Weighted reconstruction pretraining We consider a weighted version of Eqn. 5:

$$\mathcal{L}_{recon} = \sum_{\bullet,*\in\{d,m\}} w_{\bullet,*}\, \mathcal{L}'(\bullet,*) \quad (6)$$

Some downstream tasks focus on disease diagnosis. Thus, we consider a tailored *disease-focused* pretraining approach. In this setting, we omit the predictions (and loss signal) to medications and only predict diseases from either the visits aggregated disease or medication representation. Meaning we set $w_{\bullet,d} = 1 \wedge w_{\bullet,m} = 0$.

The contributions to the performance of this adaption are presented in Section 6 and Appendix B.5.

Sum Aggregation Loss Due to the strong imbalance in the distribution of diseases and medications, we explore additional loss components to prevent the attention mechanism from overfitting to the most common tokens. Instead of taking the CLS token representation we take the sum over all tokens excluding CLS and again decode this unbiased aggregate using an *MLP* to predict the set of diseases or prescriptions ($\setminus$ for set difference):

$$\mathcal{L}_{sum} = \sum_{*\in\{d,m\}} \mathcal{L}'(t), \quad (7)$$

$$\mathcal{L}'(*) = \mathcal{L}_{BCE}\Big(MLP^{sum}_{*\to*}\big(\mathbf{v}^{sum}(*)\big), \mathcal{C}(*)\Big),$$

$$\mathbf{v}^{sum}(*) = \sum\Big(Transf._\psi\Big(f_\theta\big(\mathcal{C}(*)\big)\Big) \setminus \{\texttt{CLS}\}\Big)$$

the idea is to ensure a more unbiased aggregation while still allowing the tokens to interact and impute masked or missing information. With this approach, we can induce a more dispersed distribution in the attention mechanism (App. B.4).

Concepts from clinical reports In our approach *MMUGL* we consider additional medical concepts extracted from text (clinical reports) and we concatenate the aggregated representation of these concepts for the respective visit $\mathbf{v}(n)$ to each of the two modalities at the input to the predictor *MLP*. For example in the case of $\mathcal{L}_{recon}$:

$$\mathcal{L}_{recon} = \sum_{\bullet,*\in\{d,m\}} \mathcal{L}'(\bullet,*), \quad (8)$$

$$\mathcal{L}'(\bullet,*) = \mathcal{L}_{BCE}\Big(MLP_{\bullet\to*}\big(\mathbf{v}(\bullet) \oplus \mathbf{v}(n)\big), \mathcal{C}(*)\Big)$$

The final loss for pretraining $\mathcal{L}_{pre}$ is a combination of $\mathcal{L}_{recon}$ (Eqn. 5, 6, 8) and $\mathcal{L}_{sum}$ (Eqn. 7):

$$\mathcal{L}_{pre} = \mathcal{L}_{recon} + \lambda\mathcal{L}_{sum}, \quad (9)$$

Table 1: Performance comparison on *Medication Recommendation* comparing to G-BERT (Shang et al., 2019); best in bold, $w_{\bullet,d}$ refers to Eqn. 6. $\mathcal{C}(n)$ refers to med. concepts from text. Details in Section 6.

Method	*AuPRC*	*F1*
G-BERT	69.60	61.52
MMUGL w/o $\mathcal{C}(n)$	**72.27±0.21**	63.39±0.04
MMUGL	72.10±0.17	**63.43±0.06**
MMUGL, $w_{\bullet,d} = 0$	72.10±0.19	63.34±0.23

where $\mathcal{L}_{sum}$ is configured as a regularizer with hyperparameter λ (for which we provide an ablation in Sec. B.4).

4.4.2. Downstream Modules

This work's contribution lies in learning concept representations over a knowledge graph from multiple modalities. We consider two prior architectures to perform time-series modeling and leave them mostly unchanged. It is intentional, that we do not propose a novel downstream architecture, but aim to show performance improvements alone through learning more robust and meaningful medical knowledge graph representations and aggregations thereof.

For medication recommendation, we employ an *average pooling* scheme over the history used by Shang et al. (2019). For all other tasks, we use an *RNN* based model proposed by Lu et al. (2021a). More details are in Appendix A.3.1.

5. Experimental Setup

We perform our experiments on the MIMIC-III (Johnson et al., 2016) dataset (version 1.4). Medications are mapped to the ATC hierarchy using the approach shared by Shang et al. (2019). For pretraining we consider the training splits of the respective baselines as well as unused patients with only a single visit and thus not suitable for fine-tuning sequence tasks. We consider five different downstream tasks all trained using (binary) cross-entropy (binary/multi-label/multi-class). The tables show test set performance with standard deviations over three seeded training runs and we highlight the best results in bold font. In Appendix A.1, A.3, A.4, and A.5 we share data, training, architecture, and task details.

Training Procedure The proposed pretraining is task-agnostic and the same for all downstream tasks. After pretraining, the parameters of the concept embedding (Sec. 4.1, GNN over the knowledge graph) are not fine-tuned for specific tasks. This contributes to the efficiency of our approach, as we can achieve all of the proposed performance gains without the need to retrain the large knowledge graph GNN for specific tasks. Only task-specific architectural components and the visit encoder (Sec. 4.3) are fine-tuned.

Medication Recommendation To compare to the work by Shang et al. (2019) (who have shown improvements over any previously published results on this task) we benchmark the medication recommendation task. We use their provided preprocessed patient data derived from MIMIC-III. The multi-label prediction task was evaluated on a sample-averaged Area under the precision-recall curve *AuPRC*, as well as sample-averaged macro *F1* score.

Heart Failure This task has been benchmarked in *CGL* (Lu et al., 2021a), *Chet* (Lu et al., 2022), and *Sherbet* (Lu et al., 2021b); who have performed extensive benchmarking against prior work. We run their provided preprocessing and extract the used target code sets, as well as the computed patient splits. The binary classification is evaluated using *F1* score and area under the receiver-operator curve *AuROC*.

Diagnosis Similar to the previous heart failure task we compare to the results of *CGL*, *Chet*, and *Sherbet*. We extract the target code sets and patient splits by running the provided preprocessing in each of the repositories to ensure comparability. We consider thresholded *weighted F1* (*w-F1*) score, and to be comparable to Lu et al. (2021a) we consider their adapted computation of *F1*. The variant is slightly inflated by considering the number of ground truth positive labels for each sample. This avoids the need to set a threshold, but leaks the number of ground-truth positives to the evaluation; we refer to it as *w-F1 (infl.)*. We also report recall at top k predictions (according to model confidence); referred to as *R@k* (e.g. *R@20*).

Readmission We define a readmission task at a horizon h for a given patient history at time t and a subsequent admission at time $t_{readm.}$ the target is defined as $y_{readm.@h} = (t_{readm.} - t) < h$. (Emergency) readmissions are clinically highly relevant as shown by nationwide deployments of such systems in e.g. Scotland (Liley et al., 2021). We evaluate the performance using *AuROC*.

Table 2: Performance comparison on disease-related tasks with structural priors of EHR. Each section refers to results on the split and target code sets provided by the respective reference publication. *MMUGL* shows the performance of our approach using a UMLS knowledge graph and information from clinical reports; $w_{\bullet,m}$ refers to Eqn. 6; best in bold. Details in Section 6 and Appendix A.6.

Method	**Heart Failure**		**Diagnosis**		
	AuROC	*F1*	*w-F1 (infl.)*	*R@20*	*R@40*
Reference: CGL Lu et al. (2021a)					
GRAM Choi et al. (2017)	82.82±0.06	71.43±0.05	21.06±0.19	36.37±0.16	45.61±0.27
Timeline Bai et al. (2018)	80.75±0.46	69.81±0.34	16.83±0.62	32.08±0.66	41.97±0.74
MedGCN Mao et al. (2022)	81.25±0.15	70.86±0.18	20.93±0.25	35.69±0.50	43.36±0.46
RETAIN Choi et al. (2016)	82.73±0.21	71.12±0.37	19.66±0.58	33.90±0.47	42.93±0.39
CGL Lu et al. (2021a)	85.66±0.19	72.68±0.22	22.97±0.19	38.19±0.16	48.26±0.15
MedPath Ye et al. (2021)	82.90±0.46	67.32±0.57	-	-	-
GraphCare Jiang et al. (2023)	84.34±0.05	74.15±0.56	14.80±0.15	29.97±0.22	41.31±0.22
MMUGL	87.19±0.21	74.46±0.41	25.81±0.17	41.02±0.10	52.41±0.14
MMUGL, $w_{\bullet,m} = 0$	**87.60±0.40**	**74.71±0.72**	**26.40±0.02**	**41.54±0.17**	**53.02±0.37**
Reference: Chet Lu et al. (2022)					
Chet Lu et al. (2022)	86.14±0.14	73.08±0.09	22.63±0.08	37.87±0.09	-
MMUGL	**87.67±0.36**	**75.49±0.48**	**25.55±0.48**	**40.95±0.13**	52.51±0.33
Reference: Sherbet Lu et al. (2021b)					
G-BERT Shang et al. (2019)	83.61±0.18	72.37±0.46	22.28±0.25	36.46±0.15	-
Sherbet Lu et al. (2021b)	86.04±0.16	74.27±0.07	25.74±0.04	41.08±0.08	-
MMUGL	**87.58±0.32**	**75.65±1.04**	25.78±0.08	41.00±0.08	52.81±0.31
MMUGL, $w_{\bullet,m} = 0$	87.47±0.31	75.25±0.26	**26.79±0.18**	**42.03±0.04**	53.55±0.32

Length of Stay We perform a length of stay prediction (*LoS*) using a multi-class approach with 10 categories (Yang et al. (2023)). 0 for stays under 1 day, 1-7 for stays with the respective length in number of days, 8 for stays from 1 to two weeks, and 9 for stays over 2 weeks). To evaluate we use *weighted-AuROC*.

6. Results and Discussion

Medication Recommendation Our method and training approach can outperform the previously published state-of-the-art results by Shang et al. (2019) (see Table 1). We note that the multi-modal approach with medical concepts from clinical reports cannot provide improvements on this task and data split (patients have high variation w.r.t. the richness of available clinical reports); also see Appendix B.1.

Disease Tasks In Table 2 we present benchmarking results on two disease-related tasks (Sec. 5). We train and evaluate our models on the patient splits and code sets extracted by considering three different prior work implementations, which have performed extensive benchmarking on previous state-of-the-art methods.

Overall we can conclude improved performance against approaches using pretraining schemes (Lu et al., 2021b), including text data (Lu et al., 2021a), hyperbolic embeddings (Lu et al., 2021b), ontology graphs (Choi et al., 2017; Shang et al., 2019; Mao et al., 2022), temporally localized graphs (Lu et al., 2022), personalized patient graphs (Ye et al., 2021), and language model knowledge graphs (Jiang et al., 2023).

Concept Embedding Ablation In Table 3 we show ablations over different types of concept embeddings (Sec. 4.1). All methods are trained with our proposed pipeline and only the concept embedding module is replaced. Our *MMUGL* strongly benefits from richer multi-modal information coming from clinical reports and thus outperforms prior work (the multi-modal approach can also increase robustness w.r.t. missing and erroneous information, Appendix F). We can see further improvements by tailoring our pretraining towards the downstream task by using *disease-focused pretraining* (Eqn. 6 with $w_{\bullet,m} = 0$).

Table 3: Ablation of different concept embeddings (Sec. 4.1) considering data splits and target code sets of CGL (Lu et al., 2021a). $w_{\bullet,m}$ refers to Eqn. 6. $\mathcal{C}(n)$ refers to med. concepts from text. Best in bold. Details in Sec. 6 and Appendix A.6.

Method	**Heart Failure**		**Diagnosis**		**Readm.@1y**	**LoS**
	AuROC	*F1*	*w-F1 (infl.)*	*R@20*	*AuROC*	*w-AuROC*
Embedding Mat.	87.02±0.49	74.26±1.00	24.87±0.24	40.17±0.12	70.27±1.55	78.12±0.52
Node2Vec *w/o* $\mathcal{C}(n)$	86.07±0.21	73.38±0.20	24.69±0.30	40.10±0.23	66.20±3.62	78.77±0.33
Cui2Vec *w/o* $\mathcal{C}(n)$	86.15±1.66	73.20±1.62	25.03±0.09	40.57±0.34	70.59±0.57	78.30±0.37
GNN ICD/ATC	87.03±0.10	74.07±0.18	24.75±0.29	40.15±0.09	71.59±0.23	78.80±0.36
GNN ICD/ATC-CO	87.05±0.15	74.16±0.04	24.59±0.41	39.95±0.22	71.64±0.97	78.90±0.28
MMUGL w/o $\mathcal{C}(n)$	86.27±0.18	73.36±0.70	25.01±0.46	40.42±0.43	71.59±0.31	78.88±0.52
Node2Vec *with* $\mathcal{C}(n)$	87.03±0.35	74.25±0.38	24.75±0.87	40.02±0.77	71.23±0.50	81.20±0.26
Cui2Vec *with* $\mathcal{C}(n)$	87.51±0.10	**74.71±0.24**	25.84±0.11	41.02±0.15	72.57±0.20	81.96±0.11
MMUGL	87.19±0.21	74.46±0.41	25.81±0.17	41.02±0.10	**73.26±0.18**	**82.18±0.37**
MMUGL, $w_{\bullet,m} = 0$	**87.60±0.40**	**74.71±0.72**	**26.40±0.02**	**41.54±0.17**	72.27±0.60	81.12±0.45

We compare to two alternative approaches without GNNs for learning concept embeddings by replacing the *Concept Embedding* 4.1 module and performing the same proposed training procedure. As presented by Lee et al. (2021) we pretrain our concept embeddings using *Node2Vec* (Grover and Leskovec, 2016). Secondly, we compare to *Cui2Vec* (Beam et al., 2020). *Cui2Vec* consists of medical concept embeddings pretrained on a large-scale corpus using a *Word2Vec* (Mikolov et al., 2013) style objective function. We show, that using our graph on the scale of 100,000 nodes and around 30,000 patients for pretraining, we can compete with or outperform an approach that used training data on the order of 60 million patients, 20 million clinical notes, and 1.7 million biomedical journal articles.

Our approach uses richer information from clinical reports and a larger concept vocabulary without introducing new parameters. Further, our approach is grounded in prior knowledge and is more general than previous work based solely on individual ontologies.

Readmission We benchmark readmission at different horizons and observe an increasing impact of our multi-modal knowledge graph at larger horizons (Figure 2). *MMUGL* combined graph-structured and multi-modal unified latent space is thus increasingly relevant for long-term risk predictions.

Table 3 shows performance at a relevant 1 year horizon (Liley et al., 2021). At a 15-day horizon we outperform very recent results by GraphCare (Jiang et al., 2023) using language model enhanced knowledge graphs who report 69.0 *AuROC* on MIMIC-III, where we achieve average *AuROC* of 73.69 with *MMUGL* (and 71.93 with *MMUGL w/o* $\mathcal{C}(n)$). MedGTX by Park et al. (2022) is a large fusion architecture and at a 30-day horizon they report 41.9 *AuPRC*, we outperform this result with our efficient shared latent space and achieve 45.27 with *MMUGL* (and 45.57 with *MMUGL w/o* $\mathcal{C}(n)$).

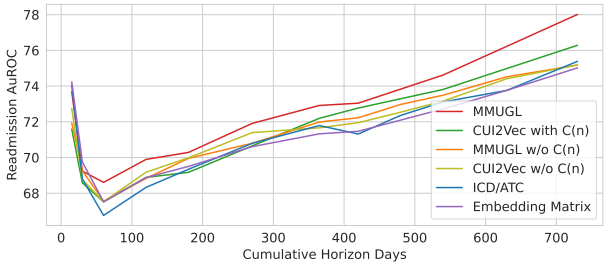


Figure 2: Readmission prediction at increasing cumulative horizons for different concept embeddings.

Summary In Table 4 we summarize the performance benefits of different components for each benchmarked task. We can conclude that both pretraining (Sec. 4.4.1), as well as our proposed knowledge graph concept embedding (Sec. 4.1) improve performance across the set of considered tasks. We further note the relevance of disease-specific pretraining for tasks such as *heart failure* or *diagnosis*.

Interpretability Using the Transformer in our visit encoder (Sec. 4.3) over the shared latent space we can provide insights into how a trained model leverages the different modalities to perform predictions by analyzing attention scores. These scores can highlight to clinicians relevant medical concepts used

Table 4: Performance contribution of different components: *Base Emb. Matrix* uses an embedding matrix as *concept embedding* (Sec. 4.1, App. A.6.6) and is not pretrained. *Pretr. Emb* pretrains the same embedding matrix (using Eqn. 9). *MMUGL* is our introduced knowledge graph based approach and $w_{\bullet,m} = 0$ refers to disease-specific pretraining (Eqn. 6). Best in bold, brackets indicate improvements over the baseline.

Component	**Heart Fail.** *AuROC*	**Diag.** *w-F1 (infl.)*	**Med. Rec.** *AuPRC*	**Readm.@1y** *AuROC*	**LoS** *w-AuROC*
Base Emb. Mat.	85.85	22.10	69.73	69.18	75.52
Pretr. Emb. Mat.	87.02 (+1.17)	24.87 (+2.77)	71.62 (+1.89)	70.27 (+1.09)	78.12 (+2.60)
MMUGL	87.19 (+1.34)	25.81 (+3.71)	**72.10 (+2.37)**	**73.26 (+4.08)**	**82.18 (+6.66)**
MMUGL, $w_{\bullet,m} = 0$	**87.60 (+1.75)**	**26.40 (+4.30)**	-	72.27 (+3.09)	81.12 (+5.60)

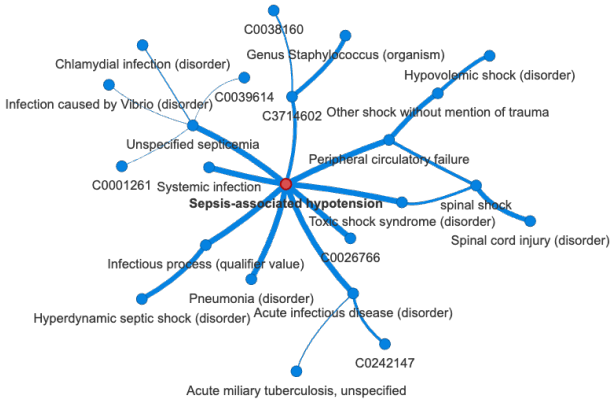


Figure 3: Subgraph relevant for the concept of *Septic Shock* (*Sepsis-associated hypotension*, UMLS `C0036983`) extracted using *GNNExplainer* (Ying et al., 2019). Edge thickness corresponds to the edge weight computed by *GNNExplainer* and the explanation is a reduction of 9932 nodes in a 3-hop neighborhood to 23 related nodes.

for the predictions. In Appendices E, C, and F we show examples of such an analysis.

We can leverage the *trained* knowledge graph to provide richer explanations to clinicians centered around important concepts. For example, the concept of *Septic Shock* (UMLS `C0036983`) is connected to 9932 nodes in a 3-hop neighborhood. Using *GNNExplainer* (Ying et al., 2019) we can retrieve a soft mask on the adjacency matrix (entries in $[0, 1]$), which represents an edge-weighted explanation subgraph. By thresholding this mask, we can retrieve an explicit subgraph explanation, where a user (e.g. a clinician) can control the size of the subgraph by adjusting the thresholding parameter. Figure 3 shows an explanation graph around *Septic Shock* with 23 nodes. The explanation subgraph aims to reduce the reconstruction error between the concept embedding $f_\theta(c)$ using the subgraph or the complete knowledge graph under the regularization constraints employed by *GNNExplainer*.

7. Conclusion

We have introduced a novel parameter-efficient way to train a unified latent space for general medical knowledge from multiple modalities. We have demonstrated improved performance on downstream tasks by grounding our representations with prior knowledge from the UMLS Metathesaurus. Our extended pretraining approach and the corresponding results emphasize its importance in tackling the supervised label scarcity in the medical domain. The more generalized approach to medical concept representations can accommodate heterogeneous multi-modal EHR without complex fusion architectures. Future work could explore the integration of lab tests or procedure codes. Finally, more general medical concept embeddings can remove the need to perform mappings between different systems, as long as a map to UMLS exists.

Our results pave the way for future research to bridge the gap between within-visit modeling (e.g., ICU time-series models (Harutyunyan et al., 2019)) and across-visit modeling, such as we benchmarked against in this work. Whereas disease and medication codes are usually assigned post-visit (for billing or archival purposes), many clinical reports are generated during the patient stays. To provide richer context information, future within-visit models might include patient histories and the multi-modal knowledge captured in our global concept representations.

Acknowledgments

This project was supported by grant #2022-278 of the Strategic Focus Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (Swiss Federal Institutes of Technology).

Further, we would like to thank Hugo Yèche for his feedback during the revision process. Thanks go to Jonas Bokstaller and Severin Husmann whose theses have provided relevant insights.

Institutional Review Board (IRB) This research does not require IRB approval in the country in which it was performed.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL https://aclanthology.org/W19-1909.

Tian Bai, Shanshan Zhang, Brian L. Egleston, and Slobodan Vucetic. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, page 43–51, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219904. URL https://doi.org/10.1145/3219819.3219904.

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. *Pac Symp Biocomput*, 25:295–306, 2020.

Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database issue):D267–70, January 2004.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310, 2001. ISSN 1532-0464. doi: https://doi.org/10.1006/jbin.2001.1029. URL https://www.sciencedirect.com/science/article/pii/S1532046401910299.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL https://aclanthology.org/W14-4012.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/231141b34c82aa95e48810a9d1b33a79-Paper.pdf.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. Gram: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 787–795, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098126. URL https://doi.org/10.1145/3097983.3098126.

Edward Choi, Cao Xiao, Walter Stewart, and Jimeng Sun. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/934b535800b1cba8f96a5d72f72f1611-Paper.pdf.

Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. Learning the graphical structure of electronic health records with graph convolutional transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):606–613, Apr. 2020. doi: 10.1609/aaai.v34i01.5400. URL https://ojs.aaai.org/index.php/AAAI/article/view/5400.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

M. Faltys, M. Zimmermann, X. Lyu, M. Hüser, S. Hyland, G. Rätsch, and T. Merz. HiRID, a high time-resolution icu dataset (version 1.1.1). *PhysioNet*, 2021. doi: 10.13026/nkwc-js72. URL https://doi.org/10.13026/nkwc-js72.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Fan Gong, Meng Wang, Haofen Wang, Sen Wang, and Mengyue Liu. Smr: Medical knowledge graph embedding for safe medicine recommendation. *Big Data Research*, 23:100174, 2021. ISSN 2214-5796. doi: https://doi.org/10.1016/j.bdr.2020.100174. URL https://www.sciencedirect.com/science/article/pii/S2214579620300423.

Jen J. Gong and John V. Guttag. Learning to summarize electronic health records using cross-modality correspondences. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, pages 551–570. PMLR, 17–18 Aug 2018. URL https://proceedings.mlr.press/v85/gong18a.html.

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*, KDD '16, page 855–864, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939754. URL https://doi.org/10.1145/2939672.2939754.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf.

Ayoub Harnoune, Maryem Rhanoui, Mounia Mikram, Siham Yousfi, Zineb Elkaimbillah, and Bouchra El Asri. Bert based clinical knowledge extraction for biomedical knowledge graph construction and analysis. *Computer Methods and Programs in Biomedicine Update*, 1:100042, 2021. ISSN 2666-9900. doi: https://doi.org/10.1016/j.cmpbup.2021.100042. URL https://www.sciencedirect.com/science/article/pii/S2666990021000410.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, Jun 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0103-9. URL https://doi.org/10.1038/s41597-019-0103-9.

Bin He, Di Zhou, Jinghui Xiao, Xin Jiang, Qun Liu, Nicholas Jing Yuan, and Tong Xu. BERT-MK: Integrating graph contextualized knowledge into pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2281–2290, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.207. URL https://aclanthology.org/2020.findings-emnlp.207.

Severin Husmann, Hugo Yèche, Gunnar Rätsch, and Rita Kuznetsova. On the importance of clinical notes in multi-modal learning for ehr data. *arXiv preprint arXiv:2212.03044*, 2022.

Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. Graphcare: Enhancing healthcare predictions with open-world personalized knowledge graphs, 2023.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035, May 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. URL https://doi.org/10.1038/sdata.2016.35.

Swaraj Khadanga, Karan Aggarwal, Shafiq Joty, and Jaideep Srivastava. Using clinical notes with time series data for ICU management. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6432–6437, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1678. URL https://aclanthology.org/D19-1678.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

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), 06 2021. ISSN 2574-2531. doi: 10.1093/jamiaopen/ooab028. URL https://doi.org/10.1093/jamiaopen/ooab028. ooab028.

James Liley, Gergo Bohner, Samuel R. Emerson, Bilal A. Mateen, Katie Borland, David Carr, Scott Heald, Samuel D. Oduro, Jill Ireland, Keith Moffat, Rachel Porteous, Stephen Riddell, Nathan Cunningham, Chris Holmes, Katrina Payne, Sebastian J. Vollmer, Catalina A. Vallejos, and Louis J. M. Aslett. Development and assessment of a machine learning tool for predicting emergency admission in scotland. *medRxiv*, 2021. doi: 10.1101/2021.08.06.21261593. URL https://www.medrxiv.org/content/early/2021/08/10/2021.08.06.21261593.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, June 2021a.

Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Sheng Wang, and Xu Sun. Auto-encoding knowledge graph for unsupervised medical report generation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021b. URL https://openreview.net/forum?id=nIL7Q-p7-Sh.

Fenglin Liu, Bang Yang, Chenyu You, Xian Wu, Shen Ge, Zhangdaihong Liu, Xu Sun, Yang Yang, and David A. Clifton. Retrieve, reason, and refine: Generating accurate and faithful patient instructions. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=dp0zWsdOV1h.

Chang Lu, Chandan K Reddy, Prithwish Chakraborty, Samantha Kleinberg, and Yue Ning. Collaborative graph learning with auxiliary text for temporal event prediction in healthcare. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3529–3535. International Joint Conferences on Artificial Intelligence Organization, 8 2021a. doi: 10.24963/ijcai.2021/486. URL https://doi.org/10.24963/ijcai.2021/486. Main Track.

Chang Lu, Chandan K. Reddy, and Yue Ning. Self-supervised graph learning with hyperbolic embedding for temporal health event prediction. *IEEE Transactions on Cybernetics*, pages 1–13, 2021b. doi: 10.1109/TCYB.2021.3109881.

Chang Lu, Tian Han, and Yue Ning. Context-aware health event prediction via transition functions on dynamic disease graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 (4):4567–4574, Jun. 2022. doi: 10.1609/aaai.v36i4.20380. URL https://ojs.aaai.org/index.php/AAAI/article/view/20380.

Qiuhao Lu, Nisansa de Silva, Sabin Kafle, Jiazhen Cao, Dejing Dou, Thien Huu Nguyen, Prithviraj Sen, Brent Hailpern, Berthold Reinwald, and Yunyao Li. Learning electronic health records through hyperbolic embedding of medical ontologies. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, page 338–346, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366663. doi: 10.1145/3307339.3342148. URL https://doi.org/10.1145/3307339.3342148.

Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '20, page 647–656, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403107. URL https://doi.org/10.1145/3394486.3403107.

Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 743–752, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3271701. URL https://doi.org/10.1145/3269206.3271701.

Chengsheng Mao, Liang Yao, and Yuan Luo. Medgcn: Medication recommendation and lab test imputation via graph convolutional networks. *Journal of Biomedical Informatics*, 127:104000, 2022. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2022.104000. URL https://www.sciencedirect.com/science/article/pii/S1532046422000168.

Yuqing Mao and Kin Wah Fung. Use of word and graph embedding to measure semantic relatedness between Unified Medical Language System concepts. *Journal of the American Medical Informatics Association*, 27(10):1538–1546, 10 2020. ISSN 1527-974X. doi: 10.1093/jamia/ocaa136. URL https://doi.org/10.1093/jamia/ocaa136.

Yiwen Meng, William Speier, Michael Ong, and Corey W Arnold. HCET: Hierarchical clinical embedding with topic modeling on electronic health records for predicting future depression. *IEEE J Biomed Health Inform*, 25(4):1265–1272, April 2021.

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. URL http://arxiv.org/abs/1301.3781.

Sungjin Park, Seongsu Bae, Jiho Kim, Tackeun Kim, and Edward Choi. Graph-text multi-modal pretraining for medical representation learning. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 261–281. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/park22a.html.

Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65, 2019.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):180178, Sep 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.178. URL https://doi.org/10.1038/sdata.2018.178.

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, May 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00455-y. URL https://doi.org/10.1038/s41746-021-00455-y.

Thomas C. Rindflesch, Halil Kilicoglu, Marcelo Fiszman, Graciela Rosemblat, and Dongwook Shin. Semantic medline: An advanced information management application for biomedicine. *Inf. Serv. Use*, 31(1–2):15–21, jan 2011. ISSN 0167-5265.

Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng, and David Sontag. Learning a health knowledge graph from electronic medical records. *Scientific Reports*, 7(1):5994, Jul 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-05778-z. URL https://doi.org/10.1038/s41598-017-05778-z.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, *The Semantic Web*, pages 593–607, Cham, 2018. Springer International Publishing. ISBN 978-3-319-93417-4.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of graph augmented transformers for medication recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5953–5959. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/825. URL https://doi.org/10.24963/ijcai.2019/825.

Marta Skreta, Aryan Arbabi, Jixuan Wang, Erik Drysdale, Jacob Kelly, Devin Singh, and Michael Brudno. Automatically disambiguating medical acronyms with ontology-aware deep learning. *Nature Communications*, 12(1):5319, Sep 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25578-4. URL https://doi.org/10.1038/s41467-021-25578-4.

Luca Soldaini and Nazli Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction. MedIR Workshop, SIGIR, 2016. URL https://github.com/Georgetown-IR-Lab/QuickUMLS.

Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. CoLAKE: Contextualized language and knowledge embedding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3660–3670, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.327. URL https://aclanthology.org/2020.coling-main.327.

Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding with deep neural networks. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 322–337. PMLR, 18–19 Aug 2017. URL https://proceedings.mlr.press/v68/suresh17a.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4808. URL https://aclanthology.org/W19-4808.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 2021. doi: 10.1162/

tacl_a_00360. URL https://aclanthology.org/2021.tacl-1.11.

Chaoqi Yang, Zhenbang Wu, Patrick Jiang, Zhen Lin, Junyi Gao, Benjamin Danek, and Jimeng Sun. PyHealth: A deep learning toolkit for healthcare predictive modeling. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) 2023*, 2023. URL https://github.com/sunlabuiuc/PyHealth.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.45. URL https://aclanthology.org/2021.naacl-main.45.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=4NpoSrT8uU-.

Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. Medpath: Augmenting health risk prediction via medical knowledge paths. In *Proceedings of the Web Conference 2021*, WWW '21, page 1397–1409, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3449860. URL https://doi.org/10.1145/3442381.3449860.

Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Ratsch. HiRID-ICU-benchmark — a comprehensive machine learning benchmark on high-resolution ICU data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=SnC9rUeqiqd.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/d80b7040b773199015de6d3b4293c8ff-Paper.pdf.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1139. URL https://aclanthology.org/P19-1139.

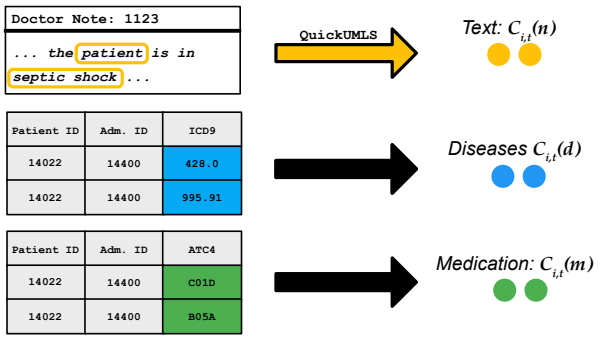


Figure 4: Illustration of the data extraction process (data is arbitrary). We extract structured EHR data in the form of billing codes from tables and use `QuickUMLS` to extract medical concepts from text (unstructured EHR) and match them against the UMLS database.

Appendix A. Experimental Details

A.1. Dataset and Split details

A small overview of data and task statistics are provided in Table 5. Splits and target code sets have been extracted from the respective repositories[2].

In Figure 4 we provide a small illustration how we extract the different inputs to our model from the data.

A.2. Knowledge Graph Statistics

The extracted knowledge graph contains 87'445 nodes, 261'212 edges with node degrees of 5.97 ± 20.91. The total vocabulary of all considered medical concepts is a subset of 21 *UMLS Metathesaurus Vocabularies* (percentages in brackets, some concepts belong to multiple): SNOMEDCT_US (46.75%), ICD9_CM (10.44%), CCPSS (7.71%), CSP (6.75%), FMA (6.15%), RXNORM (5.25%), DXP (4.21%), NCI_CDISC (4.13%), WHO (2.40%), ATC (2.12%), DRUGBANK (1.80%), CPT, NOC, BI, CCS, ICNP, NIC, ICF, CCC, PCDS, RAM.

Given a patient split we compute the coverage over our vocabulary during pretraining and downstream training. Inclusion criterias causing differences between the two are availability of medication (req. for pretraining) and multiple visits (req. for downstream training).

- *All splits*: 91.59% (pre), 71.24% (down)
- *Train*: 90.30% (pre), 68.21% (down)
- *Validation*: 16.45% (pre), 16.97% (down)
- *Test*: 37.68% (pre), 38.47% (down)

A percentage of concepts in the validation and test splits are unseen during training. Because of the graph structure, we can still learn meaningful representations for them:

- *Validation*: 0.78% (pre), 1.93% (down)
- *Test*: 3.11% (pre), 7.07% (down)

A.3. Architecture and Training

We perform early stopping based on the validation set loss both during pretraining and fine-tuning. The network is first fully pretrained until early stopped, the *concept embedding* (Sec. 4.1) backend is then frozen, the *visit encoder* (Sec. 4.3) is left trainable together with the downstream architecture to allow the attention mechanism to be fine-tuned to perform task-specific aggregations. We find a larger batch size (e.g. 32 or more) to be beneficial for better training stability. Appendix A.4 shows an overview of the hyperparameters, which have been tuned w.r.t. validation set performance.

A.3.1. Downstream Models

We provide some details about the two downstream models used for sequence modeling based on prior work.

Average Pooling To compare to work by Shang et al. (2019) in medication recommendation, we consider their downstream architecture. Given a patient history of visits (of which we get the representations using modules from Sec. 4.3 and 4.1), we perform the same pooling scheme over the past and current visit to get a final representation which is used as input to an *MLP* to perform a predictive task.

[2] https://github.com/jshang123/G-Bert, https://github.com/LuChang-CS/CGL, https://github.com/LuChang-CS/Chet, https://github.com/LuChang-CS/sherbet

Table 5: Dataset and task statistics for the disease-related tasks (Heart Failure Sec. 5 and Diagnosis Sec. 5) and the medication recommendation task Sec. 5. We compared to CGL by Lu et al. (2021a), Chet by Lu et al. (2022), and Sherbet by Lu et al. (2021b), as well as G-BERT (Shang et al., 2019) using the GitHub repository at https://github.com/jshang123/G-Bert. The splits and target code sets have been extracted by running the preprocessing code in their respective repositories; slight deviations from numbers reported in the referenced publications might be caused by a change of seed of the respective author or code modification post-publication which we can't reproduce. *Pretraining* (Pre.) refers to the pretraining auto-encoder setup and *Training* to the downstream medication recommendation task.

	CGL	Chet	Sherbet	G-BERT
# Pretraining patients	37'813	37'674	37674	34'965
# Training patients	6'155	6'009	6009	3'697
# Validation patients	125	493	493	1'058
# Test patients	1'221	1000	1000	1'059
# Unique Diseases Pretrain	6'984	6'984	6'984	1'997
# Unique Diseases Training	4'825	4'880	4'880	1'958
# Unique Med. Pretrain	348	348	348	468
# Unique Med. Training	348	348	348	145
# UMLS Graph Nodes	87'445	87'445	87'445	87'445

RNN Based on the architecture by Lu et al. (2021a) given a patient and sequence of past visits (obtained by encoding in Sec. 4.3), we feed them through a *GRU* (Cho et al., 2014). The hidden states at the output of the *GRU* are aggregated using a temporal attention mechanism where the query is a trainable embedding. We perform a minor modification here w.r.t. to the architecture by Lu et al. (2021a) and introduce a hyperparameter n_q, which refers to the number of trainable queries. If more than one query is used, we aggregate the different temporal aggregations of each query to get a single representation of the entire past of the patient. This representation is used to perform a prediction into the future using a *MLP*.

A.3.2. Baseline Graphs

The following two paragraphs introduce the baseline graphs considered, namely a combination of the ICD (diseases) and ATC (treatments) hierarchies, and an extension with co-occurrence information.

ICD/ATC Hierarchies Based on the work done by Shang et al. (2019), we consider the two tree hierarchies ICD[3] for diseases and ATC for medications. In this case, we consider $c \in \{\mathcal{C}(d) \cup \mathcal{C}(m)\}$. We compute the node embeddings $\mathbf{N}_*$ ($\oplus$ for concatenation):

$$\mathbf{N}_{\mathcal{C}(d)} = GNN_{\theta_1}(\mathcal{G}_{ICD}), \tag{10}$$

$$\mathbf{N}_{\mathcal{C}(m)} = GNN_{\theta_2}(\mathcal{G}_{ATC}), \tag{11}$$

$$f_\theta(c) = Lookup(\mathbf{N}_{\mathcal{C}(d)} \oplus \mathbf{N}_{\mathcal{C}(m)})(c) \tag{12}$$

where we use a distinct (parametrized by θ_1 and θ_2) multi-layer GNN for each of the two hierarchies (Eqns. 10, 11) and then perform a *lookup* (retrieve nodes by index) against the resulting node embeddings (Eqn. 12). In this case, we initialize all of the nodes with randomly initialized trainable embeddings. We refer to this approach to learn concept embeddings with *ICD/ATC*.

We can additionally consider co-occurrence information (e.g., Lu et al. (2021a); Liu et al. (2022)) to connect the two hierarchies. We refer to this approach with *ICD/ATC-CO* and provide details in the following paragraph.

ICD/ATC with Co-Occurrence Similar to work done by Lu et al. (2021a) or Liu et al. (2022) we can additionally consider co-occurrence information present in our dataset. We construct a new graph $\mathcal{G}_{ICD/ATC-CO}$ which contains multiple sets of nodes and edges. The node sets are the ICD and ATC tree

[3] We consider the 9th revision, as of working on MIMIC-III

hierarchy nodes, while the edge sets consist of the two ontologies and four co-occurrence edge sets; one for co-occurrence within each of the two ontologies and one (directed) from each of the two to the other. We then compute a heterogeneous (nodes of different types) multi-layer GNN (see Schlichtkrull et al. (2018); Mao et al. (2022)) over these node and edge sets, where each edge set is associated with its own parametrized graph convolution operator. As a result, we compute multiple different embeddings for a given node in each layer, which are summed. Co-Occurrence edges can additionally be weighted by computing a count over the dataset (training split) and normalizing s.t. incoming edges sum to one. Such weights can be considered by the GNN by multiplying messages from neighboring nodes with the corresponding weight. Again we have $c \in \{\mathcal{C}(d) \cup \mathcal{C}(m)\}$:

$$f_\theta(c) = GNN_{hetero}(\mathcal{G}_{ICD/ATC-CO})(c) \tag{13}$$

A.3.3. GNN Architecture

In Section 4.1 we use a parametrized GNN in Eqns. 10, 11, and 1. We use Pytorch Geometric (Fey and Lenssen, 2019) to implement these networks and based on our hyperparameter searches in Appendix A.4 we settled on using the graph convolution operator GraphSAGE as introduced by Hamilton et al. (2017).

The ICD and ATC hierarchical ontologies or our complex UMLS based knowledge graph are passed to the GNN considering all edges as undirected. In the case of multiple GNN layers we use a non-linear ReLU activation after all but the last layer. The representations for each medical concept of an ontology or the knowledge graph at the GNN output are cached and used to retrieve concept embeddings for further processing by the Visit Encoder (Sec. 4.3) module.

A.4. Hyperparameters

In Tables 6, 7, 8 and 9 we present an overview of the model hyperparameters. Final choices based on validation set performances have been marked in bold font.

Hardware A typical training is finished in under a day. Depending on the task and set of considered input modalities it can be much faster. We trained our models using mostly `Nvidia RTX2080Ti` GPUs with 11GB of dedicated GPU memory; some larger models, which included medical concepts extracted from text have been trained on `Nvidia Titan RTX` GPUs with 24GB of dedicated GPU memory. We use 2-6 worker processes and around 32-64GB of main memory.

A.5. Tasks and Evaluation

In the following, we provide a more detailed overview of the benchmarked downstream tasks (Sec. 5) and the evaluation thereof.

A.5.1. Medication Recommendation

We benchmark the medication recommendation task based on preprocessed data by Shang et al. (2019). The task is to predict a set of medications (ATC level 4 codes) given a patient's history and the current diagnosis (assigned ICD codes). Given a patient i and a trained predictor $\hat{h}$ we can formalize as follows:

$$\hat{\mathcal{C}}_{i,t}(m) = \hat{h}\big(\mathcal{C}_{i,0\ldots t-1}(*),\, \mathcal{C}_{i,t}(d)\big) \quad \text{where} * \in \{d, m, n\} \tag{14}$$

Given that this is a multi-label prediction we consider sample-averaged scores. Due to a significant imbalance in the distribution of medication codes, we use the *F1* score for thresholded hard predictions and the area under the precision-recall curve (*AuPRC*) for unthresholded confidence scores. This is in line with the evaluation by Shang et al. (2019).

A.5.2. Heart Failure

This is a binary prediction task as already benchmarked by many prior works on the MIMIC-III (Johnson et al., 2016) dataset. The task is to predict the risk of heart failure for a patient in a future visit given the patient's history. The label is extracted from the set of assigned ICD codes by matching with the prefix 428 after stripping the codes of any special characters. Let $y_{i,t}$ be the target label and it is 1 if there exists a code $c \in \mathcal{C}_{i,t}(d)$ which has the prefix 428. For a patient i and trained predictor $\hat{h}$ we can formalize as follows:

$$\hat{y}_{i,t} = \hat{h}\big(\mathcal{C}_{i,0\ldots t-1}(*)\big) \quad \text{where} * \in \{d, m, n\} \tag{15}$$

The task with mild label imbalance is evaluated using *F1* score and area under the receiver-operator

Table 6: Tested hyperparameters to optimize pretraining performance. Values that we have settled on (due to performance or training feasibility) are marked in bold. Graph depth of 0 means we have used a simple embedding layer (i.e. lookup table with trainable embeddings) instead of running a GNN over the node embeddings.

Parameter	**Values**
Effective Batch Size	(1, 2, 4, 8, 16, **32**, **64**, 128)
Hidden Dim.	(32, 64, 128, 196, **256**)
GNN Depth	(0, 1, **2**, **3**, 4, 5, 6)
GNN Activation Function	ReLU
Graph Convolution	(GCN, GAT, GINConv, **GraphSAGE**)
Graph Staged (like Shang et al. (2019))	(yes, **no**)
Graph Stacks	(**1**, **2**, 4); 2 for UMLS graphs
Transformer # Att. Heads	(1, **2**, 4)
Transformer # Layers	(**1**, 2)
Learning Rate (Adam (Kingma and Ba, 2015))	(1e-4, **5e-4**, 7.5e-4)
Visit Decoder Hidden	(**128**, 256)
Sum Loss λ_{Pool}	[0.0, 100.0]: **0.25**
Co-Occurrence Links (weighted)	(**yes**, no)

Table 7: Tested hyperparameters to optimize the medication recommendation task (Sec. 5). If a relevant parameter is not mentioned explicitly here, we have defaulted to the choices presented in Table 6. Graph depth of 0 means we have used a simple embedding layer instead of running a GNN over the node embeddings.

Parameter	**Values**
Effective Batch Size	(1, 2, 4, 8, 16, **32**, 64)
Freeze Graph	(**yes**, no)
Freeze Encoder	(yes, **no**)
Graph Depth	(0, 1, **2**, 3, 4, 5, 6)
Learning Rate (Adam (Kingma and Ba, 2015))	(1e-3, 5e-4, **1e-4**)

Table 8: Tested hyperparameters to optimize the heart failure task (Sec. 5). If a relevant parameter is not mentioned explicitly here, we have defaulted to the choices presented in Table 6.

Parameter	**Values**
Effective Batch Size	**32**
Freeze Graph	(**yes**, no)
Freeze Encoder	(yes, **no**)
Graph Depth	**2**
Learning Rate (Adam (Kingma and Ba, 2015))	(**1e-5**, 2e-5, 5e-5, 1e-4, 5e-4)
# Temporal query vectors n_q	(**1**, 2, 4)
Classification MLP Hidden Layers	**2x128**

curve (*AuROC*) for untresholded performance evaluation; this is in line with work by Lu et al. (2021a,b, 2022); Choi et al. (2016) and others.

A.5.3. Diagnosis

This is a multi-label prediction over a set of diseases. Given a patient's history we predict the set of potential diseases for an upcoming visit . For a patient i and trained predictor $\hat{h}$ we can formalize as follows:

$$\hat{\mathcal{C}}_{i,t}(d) = \hat{h}\big(\mathcal{C}_{i,0\ldots t-1}(*)\big) \quad \text{where} * \in \{d, m, n\} \tag{16}$$

This task might not seem very sensible at first as we cannot expect to reliably predict accidents that cause a hospital visit based on past EHR records. However, this is useful to catch chronic diseases and re-occurring patient patterns. Such a model's predictions could serve as a high-level aggregation of all EHR records for a specific patient. A doctor can get a very quick assessment of the potential risks for a patient upon admission and can tailor further investigations to this.

Due to the extreme imbalance over the very large set of potential labels we use *weighted-F1* score. To assess the unthresholded model confidence scores we use a popular metric from information retrieval. Recall at top k predictions (ranked by model confidence scores) can give an intuitive indication if the model can retrieve the desired ground truth diseases. The evaluation is in line with prior work e.g. by Lu et al. (2021a,b, 2022).

A.5.4. Readmission

See details in Sec. 5. We define a readmission task at a horizon h for a given patient history at time t and a subsequent admission at time $t_{readm.}$ the target is defined as $y_{readm.@h} = (t_{readm.} - t) < h$. (Emergency) readmissions are clinically highly relevant as shown by nation-wide deployments of such systems in e.g. Scotland (Liley et al., 2021). We evaluate the performance using *AuROC*.

$$days(v_{t+1} - v_t) < horizon = model\big(\mathcal{C}_{i,0\ldots t}(*)\big) \quad \text{where} * \in \{d, m, n\} \tag{17}$$

A.5.5. Length of Stay

We perform a length of stay prediction (*LOS*) using a multi-class approach with 10 categories (Yang et al. (2023). 0 for stays under 1 day, 1-7 for stays with the respective length in number of days, 8 for stays from 1 to two weeks, and 9 for stays over 2 weeks). To evaluate the imbalanced multi-class prediction task we use *weighted-AuROC*.

$$duration(v_t) = model\big(\mathcal{C}_{i,0\ldots t}(*)\big) \quad \text{where} * \in \{d, m, n\} \tag{18}$$

A.6. Baselines

In this section, we provide a summary overview of the presented baselines and the key points of their architectures.

A.6.1. CGL: Collaborative Graph Learning

In this work, Lu et al. (2021a) propose a collaborative graph learning approach. They consider two graphs, one where patients and diseases are connected based on co-occurrence and one where only diseases are connected amongst each other based on the ICD ontology. GNN layers over the two edge sets and the shared set of nodes are run in an interleaved fashion (collaboratively). The computed embeddings for a certain disease are aggregated to represent patient visits and a sequence model performs task predictions.

A.6.2. Chet: Context aware Health Event Prediction via Transition Functions

The core contribution of this work by Lu et al. (2022) is to consider a global disease graph, which connects diseases by co-occurrence and ontology relations, as well as a local graph (for each visit), which models the interactions of assigned disease codes within this specific visit. The architecture includes aggregation functions and sequence modeling to perform task-specific predictions.

A.6.3. Sherbet: Self- Supervised Graph Learning With Hyperbolic Embedding for Temporal Health Event Prediction

With *Sherbet* (Lu et al., 2021b) propose to encode the structure of a disease ontology in hyperbolic space. The hyperbolic embeddings for the respective diseases are used to pretrain (using a patient history reconstruction task) and fine-tune a sequence model architecture to perform task-specific predictions.

A.6.4. MedPath: Augmenting Health Risk Prediction via Medical Knowledge Paths

With *MedPath* Ye et al. (2021) propose to enhance the performance of existing EHR representation learning architectures by incorporating a *personalized* graph extracted using knowledge from *Semantic MEDLINE* (Rindflesch et al., 2011). The extracted graph is dataset and task-specific and can improve the performance of the backbone architecture. We transformed our data to adapt to their published pipeline and performed the heart failure prediction task using their implementations. We use *HiTANet* (Luo et al., 2020) as the backbone architecture, because it performed the best on the validation set in our hyperparameter search.

A.6.5. G-BERT: Pre-training of Graph Augmented Transformers for Medication Recommendation

Shang et al. (2019) show performance improvements on a medication recommendation task by pretraining disease and medication code embeddings using GNNs over two ontologies. The pretraining objective is a reconstruction task of observed codes during a patient visit and borrows ideas from masked language modeling. The pretrained architecture includes a Transformer-based encoder, which outputs a `CLS` encoding for each patient visit. The proposed downstream architecture performs a pooling scheme over patient histories and recommends medications for a current patient visit given the patient's history and the current diagnosis of diseases.

A.6.6. Embedding Matrix

In Table 3 we show a concept embedding ablation using an *Embedding Matrix*. This refers to a matrix of trainable parameters $\mathbf{E} \in \mathbb{R}^{|\mathcal{C}| \times k}$ where $|C|$ is the total number of considered medical concepts and k the embedding dimension. The embedding matrix replaces the *Concept Embedding* (Sec. 4.1) module and is pretrained and fine-tuned using the same procedure.

A.6.7. Concept Embeddings using Node2Vec

In Table 3 we show a concept embedding ablation using Node2Vec (Grover and Leskovec, 2016). We consider our extracted complex UMLS-based knowledge graph and perform Node2Vec-style pretraining to obtain embeddings for each concept in our knowledge graph. We then initialize an embedding matrix (which is used to retrieve concept embeddings by index lookup) and use it to replace our proposed GNN-based concept embeddings. To ensure fair comparison we then perform the same reconstruction pretraining as our proposed approach *MMUGL* to ensure the parameters of the Visit Encoder (Sec. 4.3) module are well pretrained too. Similarly, we apply the same pipeline as for our approach during fine-tuning for downstream tasks.

A.6.8. Concept Embeddings using Cui2Vec

Cui2Vec as introduced by Beam et al. (2020) is a collection of pretrained medical concept embeddings mapped to the space of UMLS. Their training optimizes a *Word2Vec* (Mikolov et al., 2013) style objective over a large-scale corpus (60 million patient records, 20 million clinical notes, and 1.7 million full-text biomedical journal articles). We use the *Cui2Vec* embeddings to initialize a lookup matrix from which concept embeddings are retrieved by index and replace our GNN-based concept embeddings. To ensure fair comparison we apply the same pretraining (reconstruction) and fine-tuning procedure to obtain downstream task performance results.

Appendix B. Training and Architecture Ablations

B.1. Clinical Reports Performance Contribution

In this section, we would like to clarify our findings about why the additional modality of extracted concepts from unstructured text (i.e. clinical reports) cannot yield a performance improvement in all cases.

Overall, the billing codes represent an aggregate of information for an entire patient's visit to the hospital and the labels are defined based on them. Thus, the billing codes (ICD, ATC codes) are the strongest signal for our predictions. The additional medical concepts from clinical reports can help in two ways. First, they can help to deal with missing or noisy information from the billing codes (see also Appendix F). Second, they can help the model to make more fine-grained predictions due to the higher level of detail.

Heart Failure Here the additional concepts from clinical reports do seem to help, but in most cases

only marginally. We hypothesize this is due to the fact, that we are only performing a binary prediction and the finer details of the clinical reports cannot yield enough additional information in most cases to significantly improve our predictions.

Diagnosis This is a very complex classification task and here we see the strongest improvement after adding the concepts extracted from clinical reports. For this task, we can benefit from the higher level of detail present in the clinical reports compared to the billing codes.

Medication Recommendation On this task the strongest signal comes from the current set of diseases. The additional concepts from clinical reports are only present in the representation of the patient's history, where we do not seem to benefit from the more detailed content of the clinical reports. To avoid information leakage we cannot directly use all concepts from all reports of the current visit when performing the medication recommendation. To accommodate for this we would have to adapt the task to a within-visit online medication recommendation; predicting medication based on the patient's global (past hospital visits) as well as local (past time within current visit) history. This would enable the inclusion of already accumulated clinical reports in the local (current visit) context.

B.2. Ablation: SapBERT

We ablate the use of *SapBERT* (Liu et al., 2021a) compared to training randomly initialized node embeddings. *SapBERT* performs better in pretraining (the selection criteria), where we see an increase from 49.38±0.49 to 61.77±0.44 *AuPRC*. The improvement carries over to the downstream performance, where for the diagnosis prediction we see an improvement of 25.46 ± 0.50 to 26.19 ± 0.30 in the *F1 (inflated)* score.

B.3. Pretraining Ablation

In Table 2 we can see, that prior work including pretraining schemes performs much stronger than the ones that don't. In Table 10 we perform an ablation w.r.t. pretraining different concept embeddings and report performance on the *Diagnosis* task (Sec. 5) on pretrained and on randomly initialized networks. We note, the more structure bias we provide, the better the performance without pretraining. In Appendix B.4 and B.5 we present further results on exploring modifications to the pretraining loss function.

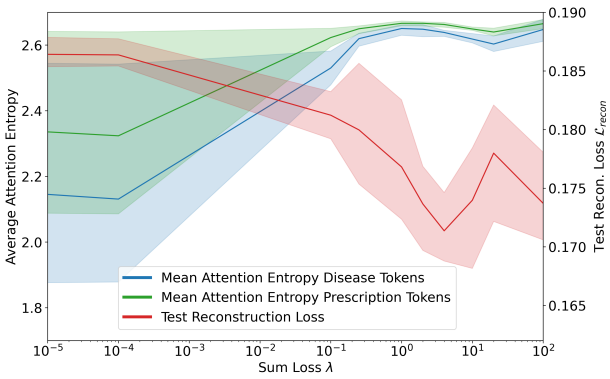


Figure 5: Behavior of average CLS attention entropy and pretraining test reconstruction loss $\mathcal{L}_{recon}$ w.r.t. λ parameter of $\mathcal{L}_{sum}$ (Eqn. 7). Details in Sec. B.4.

B.4. Sum Aggregation Loss

In Figure 5 we perform an ablation w.r.t. to the hyperparameter λ controlling the contribution of $\mathcal{L}_{sum}$ (Eqn. 7) to the total pretraining loss. Vig and Belinkov (2019) have computed the entropy of the distribution induced by the attention mechanism to analyze Transformer behavior. Similarly, we show the average (test set) entropy of the distribution induced by attention from the CLS token to all the other tokens. For larger λ the entropy increases, hence the distribution is more dispersed, and we can see an improvement in pretraining performance (shown by the test set reconstruction loss $\mathcal{L}_{recon}$ corresponding to improved test log-likelihood of our model). The idea is, that a more dispersed distribution is a better aggregator and generalizes better to rare diseases, which might otherwise be overlooked by a pointy (overfitted) attention distribution. In Appendix B.4 and B.3 we provide further experimental results ablating pretraining and the different loss terms.

We provide further empirical evidence for the contribution of the additional loss term introduced in Eqn. 7 in Table 11.

Table 11 shows results on the *Diagnosis* downstream task across different *Concept Embedding* implementations and with different pretraining regimes. We show results without pretraining, pretraining on only the default reconstruction loss $\mathcal{L}_{recon}$ (Eqn. 5)

and including the additional introduced loss term $\mathcal{L}_{sum}$ (Eqn. 7).

We can see that the additional loss component $\mathcal{L}_{sum}$ during pretraining contributes to better pretrained representations as across different downstream models we can see either at least the same performance or increased performance. This difference is especially notable and important for the best-performing model implementation *MMUGL*, where $w_{\bullet,m} = 0$ (Eqn. 6, pretraining focused on recovering diseases only). We hypothesize, that without the additional loss regularization, we experience stronger overfitting to the training distribution during pretraining, as we have more data available (given that *MMUGL* includes additional rich information coming from medical concepts in clinical reports) and we have reduced the task complexity (as we set $w_{\bullet,m} = 0$ in the pretraining loss, Eqn. 6). We also observe a tendency to more consistent results under pretraining including the $\mathcal{L}_{sum}$ loss component as standard deviations tend to be lower.

This stays consistent also on a further task e.g. *Heart Failure*. For *MMUGL* with $w_{\bullet,m} = 0$ including $\mathcal{L}_{sum}$ in pretraining we observe a downstream heart failure prediction performance (on the CGL Lu et al. (2021a) patient split) of 87.60 ± 0.40 where this drops to 86.93 ± 0.13 if we pretrain without $\mathcal{L}_{sum}$.

B.5. Reconstruction Loss

In Tables 12 and 13 we perform an ablation with respect to the different weights in the weighted version of the pretraining reconstruction loss $\mathcal{L}_{recon}$ (Eqn. 6). The base version as introduced by Shang et al. (2019) considers all weights $w_{\bullet,*} = 1$. This is flexible in the sense that it does not enforce a bias towards encoding information relevant for disease or medication predictions. However, by weighting (or fully disabling) the different terms, we can tailor our pretraining to different downstream scenarios. Please also note, that the following experiments have been performed without the additional loss component $\mathcal{L}_{sum}$ (Eqn. 7) to focus purely on the effects within the reconstruction loss term $\mathcal{L}_{recon}$ (Eqn. 5, 6).

Downstream Diagnosis Table 12 shows this effect on the downstream *Diagnosis* task. We can see that while having all loss terms active yields strong performance, in the case of a diagnosis prediction it is beneficial to only pretrain on loss terms that are predictive for diseases i.e. $w_{\bullet,d} = 1 \wedge w_{\bullet,m} = 0$. This is further supported by results shown in Table 2 and Table 3, where results on the full *MMUGL* model (including medical concepts from clinical reports) improve by pretraining with $w_{\bullet,m} = 0$.

Downstream Medication Recommendation Table 13 shows the exact same behaviour when performing downstream *medication recommendation*. The best performance is achieved by only considering loss terms towards predicting the modality relevant for the downstream prediction task. We can conclude, that cross-modality pretraining is beneficial to learn embeddings that can be useful for a yet unspecified downstream application. However, if the nature of the target modality of the downstream task is known and the cost of pretraining affordable, we can achieve better performance by adapting the pretraining to the downstream scenario.

B.6. Knowledge Graph Scaling

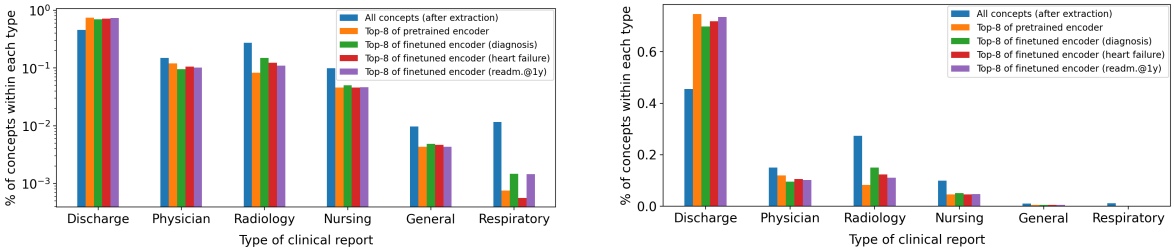
We perform a small set of experiments to investigate the effect of reducing the size of our knowledge graph. We reduce the graph to 50% of its original size (App. A.1) by dropping concepts with lower frequencies first.

We observe for diagnosis 26.41 ± 0.12 (full size graph: 26.40 ± 0.02, Table 2). However, for readmission at a 1 year horizon we observe 73.11 ± 1.87 (full size graph: 73.26 ± 0.18, Table 3), which is again similar but with much higher variance (over 3 random seeds). For readmission at a 2 year horizon we observe 77.27 ± 1.36 (full size graph: 78.01 ± 0.91, Figure 2). Some task seem to remain stable at smaller scale, whereas others show slight decrease in performance or higher variance.

Appendix C. Clinical Report Concept Category Distribution

We can use the attention mechanism to interpret the results on a patient level to rank diagnosis and medications, as well as general medical concepts from clinical reports w.r.t. their importance for the prediction. An example of such an analysis is in Appendix E.

We can also perform various dataset global analyses. We analyze the distribution of medical concepts extracted from clinical reports w.r.t. MIMIC-III report type and present the results in Figure 6(a); please mind the logarithmic y-scale (Appendix C also shows a linear scale). After pretraining, we can see a very strong shift from the dataset's type distribution



(a) Logarithmic y-axis (b) Linear y-axis

Figure 6: Medical concepts from clinical reports (Sec. 4.2) ranked by the visit encoders attention scores. We extract the top-8 tokens, group them by note type and compare the distribution of all tokens in the dataset with the ranked top-k distribution. Details in Section 6 and Appendix C.

toward discharge summaries. This is sensible given the pretraining task is an auto-encoder, essentially training for *summarizing* the visit. By fine-tuning for specific tasks we can see slight shifts towards more specific report types, which can help provide more detailed insights for a given task; note for example how the focus in the respiratory category increases as we fine-tune for a general diagnosis, but decreases below the pretraining level for a heart failure prediction.

The plot with logarithmic scale in Figure 6(a) is better suited to highlight the fine details and changes in categories such as *Respiratory* or *Radiology*. The linear scale in Figure 6(b) shows the strong changes caused by the pretraining (compared to the actual token distribution per category) in e.g. the *discharge summary* type of reports.

One might notice a particularly large drop in tokens from reports of the respiratory type. First we would like to highlight that Figure 6(a) uses a logarithmic y-axis and thus the absolute number of tokens found in the respective report type is comparatively low. Still, we can observe a change over one order of magnitude. This can be explained by looking more in-depth at the reports of this specific type.

In MIMIC-III the clinical reports of type *Respiratory* are mostly highly structured status reports assessing a patient's state w.r.t. the respiratory system. Being a structured report, there is a large set medical concepts matched, which correspond to the field names of the structured report to be filled with patient information and further most of the provided assessments in the form do not vary much across patients. As such, many of the extracted medical concepts from these reports are not discriminative across patients and thus we observe a drop in attention to the tokens extracted from these reports after training the model.

Appendix D. Heart Failure Performance Disentanglement

Due to the chronic nature of heart failure, we disentangle the performance on the test set with a fixed model for patients with and without reported histories of heart failure (the target codes have appeared in the patient history). The results are shown in Table 14.

The model is naturally performing much better on the subset of patients with a reported history of heart failure and can exploit the chronic nature of the disease. However, we note that with our proposed multimodal approach we see a notable performance improvement on the hard cases of patients without a reported history of heart failure. We conclude, using clinical report concepts backed by a knowledge graph, not just billing codes, aids in understanding disease progressions.

Appendix E. Single Patient Interpretability

In Figure 7 we present various ways how attention scores of our visit encoder (Sec. 4.3) can be used to provide interpretability of our predictions. We provide an example score analysis of visit 121518 by patient 1784 in the MIMIC-III (Johnson et al., 2016) dataset. The patient was assigned the following set of codes:

- *ICD*: 519.1, 496.0, 414.01, 401.9, 443.9, V45.82
- *ATC4*: N05CD, A02BC, B01AB, A06AD, C07AB, B05CX, G04CA, A07EA

The scores can be used to highlight the most relevant diseases and medications (Figure 7(a)). By grouping scores of individual codes and computing an aggregate for each group (e.g. 90th-percentile of scores) we can highlight the most relevant disease and medication categories for this patient at the given visit.

We can further extract which of the reports collected during the entire visit contain the most predictive identifiers by computing an aggregated score over the scores of all the matched concepts within each report (Figure 7(a)).

In (Figure 7(b)) we then highlight the concepts within the two highest-ranked reports with the largest attention scores.

We can see that the scores are consistent across different modalities, considering for example the high score given to the Respiratory category for the disease (ICD) codes (Figure 7(a)), as well as high scores for concepts found in clinical reports (e.g. `C0948187` (Tracheomalacia) in `Note: Radiology 0` or `C0189436` (Carinal reconstruction) in `Note: Discharge Summary: 0`; Figure 7(b)) related to respiratory conditions. We can conclude that for this sample the unified concept latent space promotes consistency across modalities and can improve interpretability.

Appendix F. Robustness w.r.t. missing information

In Figure 8 we show the results of an experiment, where we progressively mask a larger percentage of input tokens of different modalities. This is done by replacing the respective token identifier with the `MASK` token used during masked language modeling style pretraining (Shang et al., 2019; Devlin et al., 2019).

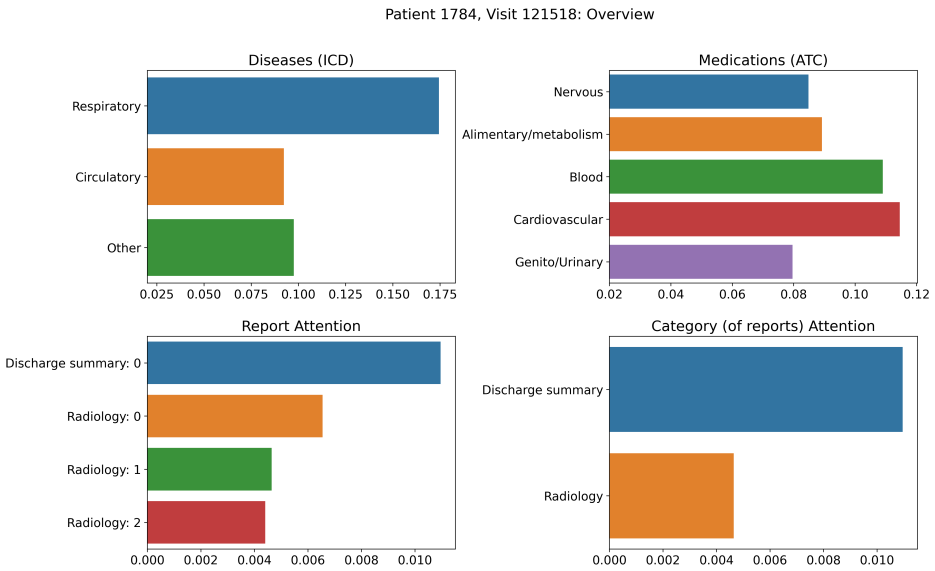
Tokens can either be masked randomly or we sort them with respect to the attention score assigned to them in the visit encoder. The y-axis shows the pretraining performance w.r.t to Eqn. 5; decoding to any of the two modalities (diseases, medications) from the visit representation of either.

The results show, that although the auto-encoding objective is only formulated w.r.t. the disease and medications tokens, the additional text information can successfully prevent stronger decay in performance and help impute the missing or incorrect information.

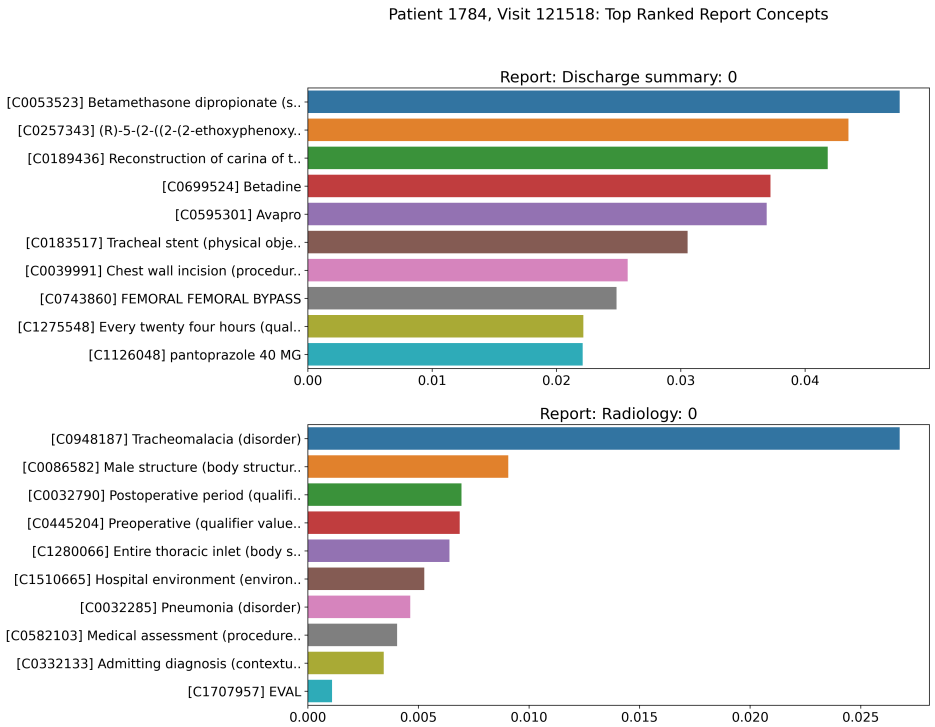
We can further see that masking tokens according to their attention scores results in a faster overall decrease in performance, highlighting the benefits of using an attention-based encoder, that can focus on relevant medical concepts when encoding a patient's current state.

Table 9: Tested hyperparameters to optimize the diagnosis prediction task (Sec. 5). If a relevant parameter is not mentioned explicitly here, we have defaulted to the choices presented in Table 6. 0 hidden layers in the classification MLP essentially means there is a single linear classification layer, so no non-linear activations are involved and there are no hidden layers.

Parameter	**Values**
Effective Batch Size	**32**
Freeze Graph	(**yes**, no)
Freeze Encoder	(yes, **no**)
Graph Depth	(2, **3**)
Learning Rate (Adam (Kingma and Ba, 2015))	(1e-5, 2e-5, 5e-5, **1e-4**, 5e-4)
# Temporal query vectors n_q	(1, 2, **4**)
Classification MLP # Hidden Layers	(**0**, 1, 2)
Classification MLP Hidden Layer Dim.	(64, 128, 256, 512)



(a) Overview of attention scores for different modalities. We show categories for ICD and ATC codes, aggregated scores for each clinical report within the visit and an aggregate for each category of reports.



(b) We show the 10 top ranked UMLS concepts within the two notes with highest aggregated attention score (see Subfigure 7(a)).

Figure 7: Attention score analysis for a single visit of a single patient (MIMIC-III (Johnson et al., 2016) dataset). We extract the scores for the different modalities from our visit encoder $g_\psi(v)$. Details in Appendix E.

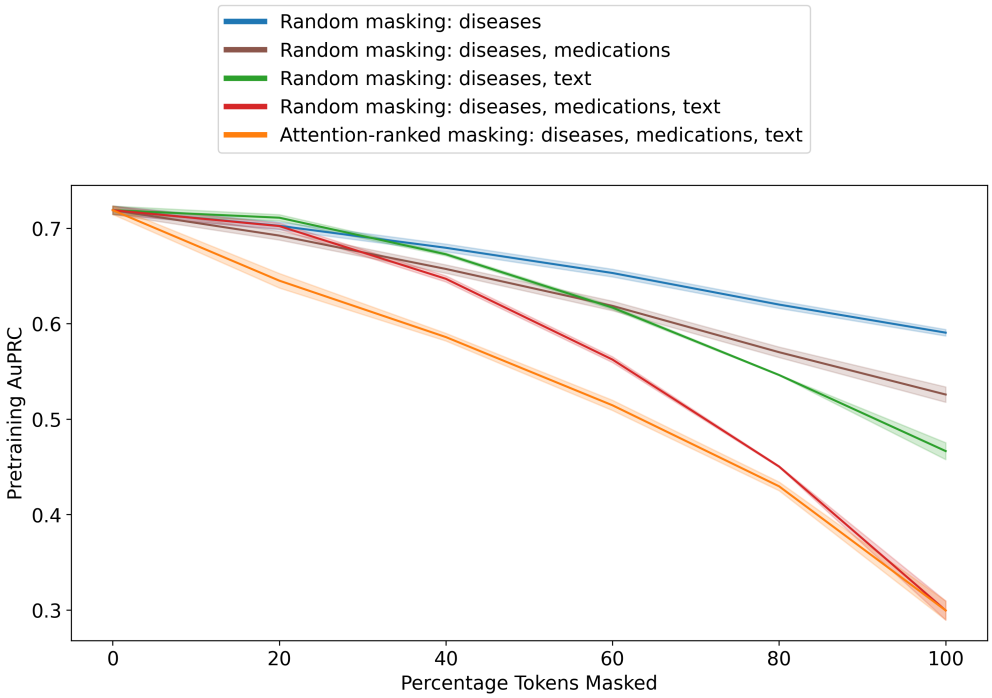


Figure 8: We analyze the performance decay caused by progressively masking a larger percentage of tokens of certain input types. We observe, that the additional information provided by text notes can have a large impact on performance, and can help prevent stronger decays in performance under missing information in the discrete token identifiers (ICD and ATC codes). Details in Appendix F.

Table 10: Ablation for pretraining, data splits, and target code sets of CGL (Lu et al., 2021a) on the *Diagnosis* task; best in bold. $\mathcal{C}(n)$ refers to med. concepts from text. Details in Section B.3.

Method	***w-F1*** random init.	***w-F1*** pretrained
ICD/ATC	13.68±0.62	24.18±0.24
ICD/ATC-CO	16.69±0.39	24.11±0.24
MMUGL w/o $\mathcal{C}(n)$	**20.78±0.21**	24.24±0.34
MMUGL	20.18±0.52	**24.69±0.21**

Table 11: Ablation for pretraining considering data splits and target code sets of CGL (Lu et al., 2021a) on the *Diagnosis* task. We show results on the downstream network without pretraining, with pretraining only using $\mathcal{L}_{recon}$ Eqn. 5, and with pretraining including $\mathcal{L}_{sum}$ Eqn. 7. $w_{\bullet,m}$ refers to Eqn. 6. Bold font highlights the best in each row. Details in Appendix B.4.

Method	***w-F1*** random init.	***w-F1*** pretrain w/o $\mathcal{L}_{sum}$	***w-F1*** pretrain w. $\mathcal{L}_{sum}$
ICD/ATC	13.68±0.62	23.84±0.39	**24.18±0.24**
ICD/ATC-CO	16.69±0.39	23.30±0.79	**24.11±0.24**
MMUGL w/o Text	20.78±0.21	24.05±0.34	**24.24±0.34**
MMUGL	20.18±0.52	24.62±0.22	**24.69±0.21**
MMUGL, $w_{\bullet,m} = 0$	-	24.42±0.86	**25.39±0.09**

Table 12: Ablation for the pretraining reconstruction loss $\mathcal{L}_{recon}$ (Eqn. 5) considering data splits and target code sets of CGL (Lu et al., 2021a) on the *Diagnosis* task. The weights $w_{\bullet,*}$ refer to Eqn. 6. The results shown are computed on the model variant *MMUGL w/o Text*. Best in bold. Details in Appendix B.5.

$w_{d,d}$	$w_{m,d}$	$w_{m,m}$	$w_{d,m}$	*w-F1*	*R@20*	*R@40*
1	0	0	0	23.18±0.20	39.30±0.42	51.03±0.26
1	1	0	0	**24.46±0.18**	**40.76±0.10**	52.14±0.31
1	1	1	0	24.15±0.56	40.39±0.07	51.61±0.17
1	1	1	1	24.24±0.26	40.49±0.02	**52.15±0.23**

Table 13: Ablation for the pretraining reconstruction loss $\mathcal{L}_{recon}$ (Eqn. 5) considering data splits and target code sets of G-BERT (Shang et al., 2019) on the *Medication Recommendation* task. The weights $w_{\bullet,*}$ refer to Eqn. 6. The results shown are computed on the model variant *MMUGL w/o Text*. Best in bold. Details in Appendix B.5.

$w_{d,d}$	$w_{m,d}$	$w_{m,m}$	$w_{d,m}$	*AuPRC*	*F1*
0	0	1	0	68.00±0.24	59.97±0.43
0	0	1	1	**72.44±0.07**	**63.58±0.15**
1	0	1	1	72.37±0.11	63.51±0.11
1	1	1	1	72.09±0.02	63.34±0.04

Table 14: Performance disentanglement for the heart failure prediction task on patients with and without reported history of heart failure. *AuROC* with text shows the full model performance, and *AuROC* w/o text the performance when all concepts extracted from clinical reports are masked.

Patient Set	# Patients	Prevalence	*F1*		
			w/o text	w. text	Δ text
All	1221	37%	73.29	75.08	+1.79
W. HF History	425	74%	85.52	86.28	+0.76
W/o HF History	796	17%	44.02	47.68	+3.66