# Encoding Medical Ontologies With Holographic Reduced Representations for Transformers

Bing Hu[1,*], Trevor Yu[1], Tia Tuinstra[1], Ryan Rezai[1], Harshit Bokadia[1], Rachel DiMaio[1], Thomas Fortin[1], Brian Vartian[1,2] and Bryan Tripp[1]

[1]University of Waterloo, Ontario, Canada
[2]McMaster University, Ontario, Canada

**Abstract**

Transformer models trained on NLP tasks with medical codes often have randomly initialized embeddings that are then adjusted based on training data. For terms appearing infrequently in the dataset, there is little opportunity to improve these representations and learn semantic similarity with other concepts. Medical ontologies represent many biomedical concepts and define a relationship structure between these concepts, making ontologies a valuable source of domain-specific information. Holographic Reduced Representations (HRR) are capable of encoding ontological structure by composing atomic vectors to create structured higher-level concept vectors. We developed an embedding layer that generates concept vectors for clinical diagnostic codes by applying HRR operations that compose atomic vectors based on the SNOMED CT ontology. This approach allows for learning the atomic vectors while maintaining structure in the concept vectors. We trained a Bidirectional Encoder Representations from the Transformers (BERT) model to process sequences of clinical diagnostic codes and used the resulting HRR concept vectors as the embedding matrix for the model. The HRR-based approach introduced interpretable structure into code embeddings while maintaining or modestly improving performance on the masked language modeling (MLM) pre-training task (particularly for rare codes) as well as the fine-tuning tasks of mortality and disease prediction. This approach also better maintains semantic similarity between medically related concept vectors, due to both shared atomic vectors and disentangling of code-frequency information.

**Keywords**

Deep Learning, Ontology, Knowledge-Integration

## 1. Introduction

Transformers [1] jointly optimize high-dimensional vector embeddings that represent input tokens, and a network that contextualizes and transforms these embeddings to perform a task. Originally designed for natural language processing (NLP) tasks, transformers are now widely used with other data modalities. In medical applications, one important modality consists of medical codes that are extensively used in electronic health records (EHR). A prominent example in this space is Med-BERT [2], which consumes a sequence of diagnosis codes. Tasks that Med-BERT and other EHR-transformers perform include disease and mortality prediction.

Deep networks have traditionally been alternatives to symbolic artificial intelligence with different advantages [3]. Deep networks use real-world data effectively, but symbolic approaches have completive properties, such as better transparency and capacity for incorporating structured information, inspiring many efforts to combine the two approaches in neuro-symbolic systems [4]. Additional transparency and ability to incorporate structured information are potential benefits of symbolic approaches in medical applications [5]. Standard large language models (LLMs) can be prone to biases in the training data, such as frequency bias, which can result in medical misinformation and potentially clinical harm [6, 7, 8].

Here we use a novel neuro-symbolic medical transformer architecture incorporating structured knowledge from an authoritative medical ontology into the embeddings. Specifically, we use vector-symbolic holographic reduced representations (HRRs) [9] to produce composite medical-code embeddings and backpropagate through the architecture to optimize the embeddings of atomic concepts. This approach produces optimized medical code embeddings with an explicit structure that incorporates medical knowledge.

We test our method, Holographic Reduced Representation Bi-directional Encoder Representations from Transformers (HRRBERT), on the Medical Information Mart for Intensive Care (MIMIC)-IV dataset [10] and show improvements in both pre-training and fine-tuning tasks. We also show that our embeddings of ontologically similar rare medical codes have high cosine similarity, in contrast with embeddings that are learned in the standard way. Finally, we investigate learned representations of medical-code frequency, in light of recent demonstration of frequency bias in EHR-transformers [6].

We contribute:

- A novel neuro-symbolic architecture, HRRBERT,

*KiL'24: Workshop on Knowledge-infused Learning co-located with 30th ACM KDD Conference, August 26, 2024, Barcelona, Spain*
*Corresponding author.
bingxu.hu@uwaterloo.ca (B. Hu)

that combines vector-symbolic embeddings with the BERT LLM architecture, leading to better performance in medical tasks.
- Efficient construction of vector-symbolic embeddings that leverage PyTorch autograd on GPUs.
- Optimized medical-code embeddings that better respect semantic similarity of medical terminology than standard embeddings for infrequently used codes.

We focus here on processing medical codes, but our methods would extend naturally to foundation models that combine medical codes and natural language. Specifically, the trained atomic vectors of our vector-symbolic embeddings could share a dictionary with language embeddings, so that training of each could improve the representation of the other.

## 1.1. Background and Related Works

The Vector-Symbolic Architectures (VSA) approach is a computing paradigm that relies on high dimensionality and randomness to represent concepts as unique vectors in a high dimensional space [11]. VSAs create and manipulate distributed representations of concepts by combining base vectors with bundling, binding, and permutation algebraic operators [12]. For example, a scene with a red box and a green ball could be described with the vector SCENE=RED⊗BOX+GREEN⊗BALL, where ⊗ indicates binding, and + indicates bundling. The atomic concepts of RED, GREEN, BOX, and BALL are represented by base vectors, which are typically random. VSAs also define an inverse operation that allows the decomposition of a composite representation. For example, the scene representation could be queried as SCENE⊗BOX$^{-1}$. This should return the representation of GREEN or an approximation of GREEN that is identifiable when compared to a dictionary. In a VSA, the similarity between concepts can be assessed by measuring the distance between the two corresponding vectors.

VSAs were proposed to address challenges in modelling cognition, particularly language [12]. However, VSAs have been successfully applied across a variety of domains and modalities outside of the area of language as well, including in vision [13, 14], biosignal processing [15], and time-series classification [16]. Regardless of the modality or application, VSAs provide value by enriching vectors with additional information, such as spatial semantic information in images and global time encoding in time series.

An early VSA framework was Smolensky's Tensor Product Representation [17], which addressed the need for compositionality, but suffered from exploding model dimensionality. The VSA framework introduced by Plate, Holographic Reduced Representations (HRR), improved upon Smolensky's by using circular convolution as the binding operator [9]. Circular convolution keeps the output in the same dimension, solving the problem of exploding dimensionality.

In the field of deep learning, HRRs have been used in previous work to recast self-attention for transformer models [18], to improve the efficiency of neural networks performing a multi-label classification task by using an HRR-based output layer [3], and as a learning model itself with a dynamic encoder that is updated through training [19]. In all of these works, the efficiency and simple arithmetic of HRRs are leveraged. Our work differs in that we also leverage the ability of HRRs to create structured vectors to represent complex concepts as inputs to a transformer model.

VSAs such as HRRs can effectively encode domain knowledge, including complex concepts and the relationships between them. For instance, Nickel et al. [20] propose holographic embeddings that make use of VSA properties to learn and represent knowledge graphs. Encoding domain knowledge is of interest in the field of deep learning, as it could improve, for example, a deep neural network's ability to leverage human knowledge and to communicate its results within a framework that humans understand [21]. Ontologies are a form of domain knowledge incorporated into machine learning models to use background knowledge to create embeddings with meaningful similarity metrics and for other purposes [22]. In our work, we use HRRs to encode domain knowledge in trainable embeddings for a transformer model. The domain knowledge we use comes from the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), which is a widely used clinical ontology system that includes definitions of relationships between clinical concepts [23].

To the best of our knowledge, HRRs have not been used before as embeddings for transformer models. Transformer models typically use learned embeddings with random initializations [1]. However, in the context of representing ontological concepts, using such unstructured embeddings can have undesirable effects. One problem is the inconsistency between the rate of co-occurrence or patterns of occurrence of medical concepts and their degree of semantic similarity described by the ontology. For example, the concepts of "Type I Diabetes" and "Type II Diabetes" are mutually exclusive in EHR data and do not follow the same patterns of occurrence due to differences in pathology and patient populations [24]. The differences in occurrence make it difficult for a transformer model to learn embeddings with accurate similarity metrics. The concepts should have relatively high similarity according to the ontology. They both share a common ancestor of "Diabetes Mellitus," they are both metabolic disorders that affect blood glucose levels, and they can both lead to similar health outcomes. Song et al.

[24] seeks to address this type of inconsistency by training multiple "multi-sense" embeddings for each non-leaf node in an ontology's knowledge graph via an attention mechanism. However, the "multi-sense" embeddings do not address the learned frequency-related bias that also arises from the co-occurrence of concepts. Frequency-related bias raises an explainability issue, as it leads to learned embeddings that do not reflect true similarity relationships between concepts, for example, as defined in an ontology, but instead reflect the frequency of the concepts in the dataset [6]. This bias particularly affects codes that are used less frequently.

Our proposed approach, HRRBERT, uses the structure from SNOMED CT to represent thousands of concepts with high-dim-ensional vectors such that each vector reflects a particular clinical meaning and can be compared to other vectors using the HRR similarity metric, cosine similarity. It also leverages the computing properties of HRRs to provide structured embeddings for a LLM that supports optimization through backpropagation.

# 2. Methods

## 2.1. MIMIC-IV Dataset

The data used in this study was derived from the Medical Information Mart for Intensive Care (MIMIC) v2.0 database, which is composed of de-identified EHRs from in-patient hospital visits between 2008 and 2019 [10]. MIMIC-IV is available through PhysioNet [25]. We used the ICD-9 and ICD-10 diagnostic codes from the *icd_diagnosis* table from the MIMIC-IV *hosp* module. We filtered patients who did not have at least one diagnostic code associated with their records. Sequences of codes were generated per patient by sorting their hospital visits by time. Within one visit, the order of codes from the MIMIC-IV database was used, since it represents the relative importance of the code for that visit. Each unique code was assigned a token. In total, there were 189,980 patient records in the dataset. We used 174,890 patient records for pre-training, on which we performed a 90–10 training-validation split. We reserved 15k records for fine-tuning tasks.

## 2.2. Model Architecture

We utilized a BERT-base model architecture with a post-layer norm position and a sequence length of 128 ICD codes [26]. A custom embedding class was used to support the functionality required for our HRR embeddings. We adapted the BERT segment embeddings to represent groups of codes from the same hospital visit, using up to 100 segment embeddings to encode visit sequencing. An embedding dimension of $d = 768$ was used, and all embeddings were initialized from $\sim \mathcal{N}_d(0, 0.02)$, as in [26], including the atomic vectors for HRR embeddings. Fine-tuning used a constant learning rate schedule with a weight decay of 4e-6. Fine-tuning lasted 10 epochs with a batch size of 80.

## 2.3. Encoding SNOMED Ontology with HRR Embeddings

In this section, we detail the methodologies of constructing vector embeddings for ICD disease codes using HRR operations based on the SNOMED CT structured clinical vocabulary. We first describe our mapping from ICD concepts to SNOMED CT terms. Next, we define how the atomic symbols present in the SNOMED CT ontology are combined using HRR operations to construct concept vectors for the ICD codes. Finally, we describe our method to efficiently compute the HRR embedding matrix using default PyTorch operations that are compatible with autograd.

### 2.3.1. Mapping ICD to SNOMED CT Ontology

Our data uses ICD-9 and ICD-10 disease codes while our symbolic ontology is defined in SNOMED CT, so we required a mapping from the ICD to the SNOMED CT system to build our symbolic architecture. We used the SNOMED CT International Release from May 31, 2022 [23] and only included SNOMED CT terms that were active at the time of that release. While SNOMED publishes a mapping tool from SNOMED CT to ICD-10, a majority of ICD-10 concepts have one-to-many mappings in the ICD-to-SNOMED CT direction [27]. To increase the fraction of one-to-one mappings, we used additional published mappings from the Observational Medical Outcomes Partnership (OMOP) [28], mappings from ICD-9 directly to SNOMED CT [29], and mappings from ICD-10 to ICD-9 [30].

Notably, after excluding ICD codes with no active SNOMED CT mapping, 671 out of the 26,164 unique ICD codes in the MIMIC-IV dataset were missing mappings. When those individual codes were removed, a data volume of 4.62% of codes was lost. This removed 58 out of 190,180 patients from the dataset, as they had no valid ICD codes in their history. Overall, the remaining 25,493 ICD codes mapped to a total of 12,263 SNOMED CT terms.

### 2.3.2. SNOMED CT vector symbolic architecture

Next, we define how the contents of the SNOMED CT ontology were used to construct a symbolic graph to represent ICD concepts. For a given SNOMED CT term, we used its descriptive words and its relationships to other SNOMED CT terms. A relationship is defined by a relationship type and a target term. In total, there

were 13,852 SNOMED CT target terms and 40 SNOMED CT relationship types used to represent all desired ICD concepts. In the ontology, many ICD concepts share SNOMED CT terms in their representations.

The set of relationships was not necessarily unique for each SNOMED CT term. To add more unique information, we used a term's "fully specified name" and any "synonyms" as an additional set of words describing that term. We set all text to lowercase, stripped punctuation, and split on spaces to create a vocabulary of words. We removed common English stopwords from a custom stopword list that was collected with assistance from a medical physician. The procedure resulted in a total of 8833 vocabulary words.

Overall, there were a total of 22,725 "atomic" symbols for the VSA which included the SNOMED CT terms, relationships, and the description vocabulary. Each symbol was assigned an "atomic vector". We built a "concept vector" for each of the target 25,493 ICD codes using HRR operations to combine atomic vectors according to the SNOMED CT ontology structure.

To build a $d$-dimensional concept vector for a given ICD concept, we first considered the set of all relationships that the concept maps to. We used the HRR operator for binding, circular convolution, to combine vectors representing the relationship type and destination term and defined the concept vector to be the bundling of these bound relationships. For the description words, we bundled the vectors representing each word together and bound this result with a new vector representing the relationship type "description," as shown in Equation 1.

$$\text{ICD concept} = \sum_{\text{SNOMED CT}} \text{rel} \circledast \text{term} + \sum_{\text{words}} \text{desc} \circledast \text{word} \quad (1)$$

Formally, let : $\{1, 2, ..., N_a\}$ be the set of integers enumerating the unique atomic symbols for SNOMED CT terms and description words. Let : $\{1, 2, ..., N_r\}$ be the set of integers enumerating unique relationships for SNOMED CT terms, including the description relationship and the binding identity. Let : $\{1, 2, ..., N_c\}$ be the set of integers enumerating the ICD-9 and ICD-10 disease concepts represented by the VSA.

has an associated embedding matrix $\in^{N_a \times d}$, where atomic vector $_k = _{[k,:]}$, $k \in$ is the $k$-th row the embedding matrix. Similarly, there is relationship embedding matrix, $\in^{N_r \times d}$ and $_j = _{[j,:]}$, $j \in$; and an ICD concept embedding matrix, $\in^{N_c \times d}$ and $_i = _{[i,:]}$, $i \in$. We described the VSA with the formula in Equation 2, where $\mathcal{G}_i$ is a graph representing the connections between ICD concept $i$ to atomic symbols $k$ by relationship $j$.

$$_i = \sum_{(j,k) \in \mathcal{G}_i} {}_j \circledast _k \quad (2)$$

Additional details on how to efficiently use PyTorch autograd to learn through these HRR operations are provided in Appendix A.1.

#### 2.3.3. Embedding Configurations

We call our method of constructing embeddings for ICD codes purely from HRR representations "HRRBase" and the standard method of creating transformer token embeddings from random vectors "unstructured". While the HRRBase configuration enforces the ontology structure, we wondered whether it would be too rigid and have difficulty representing information not present in SNOMED CT. As dataset frequency information for ICD medical codes is not present in the HRR structure, we tried adding an embedding that represented the empirical frequency of that ICD code in the dataset. We also tried adding fully learnable embeddings with no prior structure.

Given the wide range of ICD code frequencies in MIMIC, we log-transformed the empirical ICD code frequencies, and then discretized the resulting range. For our HRRFreq configuration, we used the sinusoidal frequency encoding as in [1] to encode the discretized log-frequency information. The frequency embeddings were normalized before being summed with the HRR embedding vectors.

We defined two additional configurations in which a standard embedding vector was integrated with the structured HRR concept vector. With "HRRAdd", a learnable embedding was added to the concept embedding, HRRAdd = $+_{\text{add}}$, $_{\text{add}} \in^{N_c \times d}$. However, this roughly doubled the number of learnable parameters compared to other formulations.

With "HRRCat", a learnable embedding of dimension $d/2$ was concatenated with the HRR concept embedding of dimension $d/2$. This keeps the total number of learnable parameters roughly the same as the unstructured configuration (25,493 $d$-dimensional vectors) and the HRRBase configuration (22,725 $d$-dimensional vectors). The final embedding matrix was defined as HRRCat = [ $_{\text{cat}}$], where $_{,\text{cat}} \in R^{N_c \times d/2}$.

### 2.4. Experiments

We pre-trained the unstructured, HRRBase, HRRCat, and HRRAdd embedding configurations of HRRBERT on the masked language modelling (MLM) task, for 3 trials each. For each of the 3 pre-trained models, 10 fine-tuning trials were conducted for a total of 30 trials per fine-tuning task. The best checkpoint from the 10 epochs of fine-tuning was saved based on validation performance. A test set containing 666 patient records was used to evaluate each of the fine-tuned models for both mortality and disease prediction. We report accuracy, precision, recall, and F1 scores averaged over the 30 trials for the fine-tuning tasks.

# 3. Experimental Results

## 3.1. Pre-training

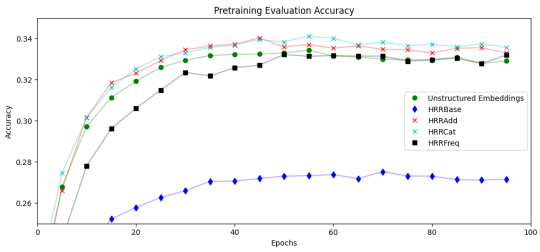


**Figure 1:** Pre-training validation set evaluation results for different configurations

MLM accuracy is evaluated on a validation set over the course of pre-training. Pre-training results for different configurations are shown in Figure 1. The pre-training results are averaged over 3 runs for each of the configurations except for HRRFreq where only 1 model run was completed.

The baseline of learned unstructured embeddings has a peak pre-training validation performance of around 33.4%. HRRBase embeddings perform around 17% worse compared to the baseline of learned unstructured embeddings. We hypothesize that this decrease in performance is due to a lack of embedded frequency information in HRRBase compared to learned unstructured embeddings. HRRFreq (which combines SNOMED CT information with frequency information) has a similar performance compared to unstructured embeddings, supporting this hypothesis. Compared to baseline, HRRAdd and HRRCat improve pre-training performance by a modest margin of around 2%. We posit that this almost 20% increase in performance of HRRCat and HRRAdd over HRRBase during pre-training is partly due to the fully learnable embedding used in HRRCat and HRRAdd learning frequency information.

## 3.2. Fine-tuning

We fine tuned the networks for mortality prediction and disease prediction. Across metrics and tasks, the best results were often seen in HRRBase (Table 1) with some being statistically significant.

### 3.2.1. Mortality Prediction Task

The mortality prediction task is defined as predicting patient mortality within 6 months after the last visit. Binary mortality labels were generated by comparing the time difference between the last visit and the mortality date. A training set of 13k patient records along with a validation set of 2k patient records were used to fine-tune each model on mortality prediction. Table 1 shows the evaluation results of mortality prediction for each of the configurations. We performed a two-sided Dunnett's test to compare our multiple experimental HRR embedding configurations to the control unstructured embeddings, with $p < 0.05$ significance level. HRRBase embeddings had a significantly greater mean F1-score ($p = 0.043$) and precision ($p = 0.042$) compared to unstructured embeddings.

### 3.2.2. Disease Prediction Task

The disease prediction task is defined as predicting which disease chapters were recorded in the patient's last visit using information from earlier visits. We converted all ICD codes in a patient's last visit into a multi-label binary vector of disease chapters. As there are 22 disease chapters defined in ICD-10, the multi-label binary vector has a size of 22 with binary values corresponding to the presence of a disease in each chapter. A training set of 4.5k patient records along with a validation set of 500 patient records were used to fine-tune each model on this task. Table 1 shows the evaluation results of disease prediction for each of the configurations. For the two-sided Dunnett test, Levene's test shows that the equal variance condition is satisfied, and the Shapiro-Wilk test suggests normal distributions except for HRRAdd accuracy. The test showed HRRBase embeddings had a significantly greater mean accuracy ($p = 0.033$) and precision ($p = 0.023$) compared to unstructured embeddings. No other comparisons of mean metrics for HRR embeddings were significantly greater than the control.

### 3.2.3. eICU Mortality Prediction

An additional experiment conducted on the Philips Electronic Intensive Care Unit (eICU) [31] shows corroborating results with the MIMIC-IV experiments. For our experiment, we applied our mortality prediction models that were fine-tuned on MIMIC-IV to eICU data to see if our results generalize. Table 1 shows that HRRBase embeddings had a significantly greater mean accuracy ($p = 0.046$) compared to unstructured embeddings when applied to the eICU dataset. These models are not optimized for mortality prediction for other hospitals where coding methodology and clinical practice may differ. For example, the most common code in the eICU dataset represents acute respiratory failure, whereas the most common code in the MIMIC-IV dataset represents hypertension.

**Table 1**
Finetuning mean test scores and standard deviations for mortality prediction, disease prediction, eICU mortality prediction, and both Really-Out-Of-Distribution (ROOD) Unseen and Overall disease prediction tasks. The best scores are bolded and are underlined if statistically significant.

| Finetuning Task | Configuration | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| ROOD | HRRBase | **94.9**±1.0 | **83.5**±4.6 | **76.8**±5.1 | **79.5**±4.9 |
| Unseen | Unstructured | 92.3±0.3 | 46.2±0.0 | 50.0±0.1 | 48.0±0.1 |
| ROOD | HRRBase | **81.9**±0.1 | 78.3±0.3 | **75.2**±0.8 | **76.4**±0.5 |
| Overall | Unstructured | 81.9±0.2 | **78.7**±0.7 | 74.4±1.2 | 76.0±0.8 |
| | HRRBase | **84.4**±2.3 | **65.8**±2.0 | 85.6±2.2 | **69.2**±2.7 |
| Mortality | HRRAdd | 84.0±2.2 | 65.7±1.9 | **85.7**±2.3 | 68.9±2.5 |
| Prediction | HRRCat | 83.9±2.3 | 65.6±1.7 | 84.9±2.8 | 68.8±2.5 |
| | Unstructured | 83.4±1.9 | 64.9±1.2 | 84.6±2.2 | 67.9±1.8 |
| | HRRBase | **79.9**±0.5 | **73.0**±1.2 | 67.2±0.7 | 69.0±0.6 |
| Disease | HRRAdd | 79.6±0.7 | 72.6±1.4 | 67.3±0.9 | 69.0±0.6 |
| Prediction | HRRCat | 79.6±0.8 | 72.5±1.7 | 67.3±1.0 | 68.9±0.8 |
| | Unstructured | 79.4±0.5 | 72.1±1.1 | **67.8**±1.0 | **69.2**±0.7 |
| eICU | HRRBase | **68.9**±1.3 | **75.0**±1.8 | **57.0**±5.8 | **64.5**±3.5 |
| Mortality | HRRAdd | 68.1±1.6 | 74.0±2.2 | 56.2±6.8 | 63.6±3.9 |
| Prediction | HRRCat | 68.2±1.2 | 73.8±2.6 | 57.0±7.2 | 64.0±3.7 |

#### 3.2.4. Really-Out-Of-Distribution (ROOD) Disease Prediction

We conducted an additional disease-prediction experiment to test generalization to patients with codes outside the training distribution. We found six patients with records that consisted of only 32 codes between them (see list of codes in Appendix A). We created a really-out-of-distribution (ROOD) dataset that consisted of all patients in MIMIC-IV (nearly 30K) with at least one of these codes. We used this as a validation set. The separate pre-training and fine-tuning dataset did not contain these codes. We also created a smaller validation dataset consisting of the six patients with only these codes. During pretraining, the HRRBase and unstructured models did not encounter any examples using the 32 ROOD codes and so did not explicitly learn representations for those codes. The trained models were then tested using the ROOD dataset.

Results from Table 1 on ROOD dataset disease prediction show that HRRBase outperforms the unstructured embedding model for contexts of entirely unseen codes. We assess statistical significance using two-tailed, independent t-test with unequal variance, as some measurements failed Levene's test for equal variance. The means of all the metrics for HRRBase are significantly greater than for unstructured when making inferences on patients with entirely unseen codes, $p < 0.001$ for all metrics. Given the embedded ontological structure, we hypothesize that HRRBase implicitly learns useful embeddings for the 32 unseen ROOD codes by learning any shared embedding components of the VSA when training on other codes. Unstructured embeddings cannot learn better representations for codes never seen in training.

### 3.3. t-SNE of Frequency Bias

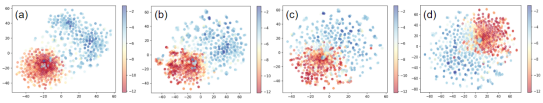


**Figure 2:** Comparing t-SNE of (a) unstructured embeddings, (b) HRRAdd, (c) HRRCat, and (d) HRRBase. The t-SNE graphs are color-coded by the frequency of the ICD codes in the dataset - highly frequent codes are colored blue while infrequent codes are colored red.

We computed t-SNE dimension reductions to visualize relationships among ICD code embeddings in the pre-trained models. Figure 2 shows that unstructured embeddings of common ICD codes are clustered together with a large separation from those of uncommon codes. This suggests that code-frequency information is prominently represented in these embeddings, consistent with frequency bias in related models [6]. Common and uncommon code clusters are less distinct in HRRBase, which does not explicitly encode frequency information.

As shown in Figure 1, adding code-frequency information to the structured HRRBase embeddings, i.e. the HRRFreq embeddings, improved the pre-training loss be similar to unstructured embeddings. This suggests that unstructured components in HRRAdd and HRRCat may

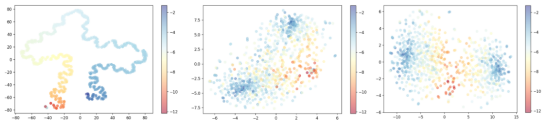

**Figure 3:** t-SNE representation of sinusoidal frequency embeddings (left), and unstructured embedding components of HRRAdd (middle) and HRRCat (right).

have learned some frequency information, since these losses are also similar to the loss of models with Unstructured embeddings. To investigate whether this occurred, we performed t-SNE dimension reductions of the unstructured components of HRRAdd and HRRCat and colored the points by code frequency, shown in Figure 3. This graph suggests that these additional unstructured embeddings learn some frequency information, due to clustering of high frequency codes. However, the frequency information learned by HRRCat and HRRAdd learnable embeddings influence overall embeddings less strongly in comparison to unstructured embeddings as seen in Figure 2, where low frequency embeddings are less distinctly separated from higher frequency embeddings.

## 3.4. Top-k Accuracy for MLM

Accurately predicting infrequently used disease codes is an important clinically relevant task. Given that the model trains and sees more common codes compared to rare codes, rare codes are naturally challenging to predict. Through promising empirical results on out-of-distribution mortality prediction for eICU and disease prediction on ROOD, we hypothesized that our HRR embedding models should have improved accuracy when predicting rare codes in the dataset compared to unstructured embedding models, since rare codes should share some atomic vectors in their representations with common codes.

To test this, we evaluated the accuracy of an MLM pre-trained model predicting a single masked code of a known frequency. We split the codes in the pre-training validation dataset into 7 bins from log frequency -14 to 0, such that each bin has a width of 2. The most common codes are in a bin with log frequencies between -2 and 0, while the rarest codes are from a bin with log frequencies between -14 and -12. From each bin, we selected 400 codes at random, repeating codes from that bin if there were fewer than 400. For each of these codes, we selected one patient that had that code in their history, masked that code as would be done in MLM, and created a dataset of these 2,800 patients to use for MLM inference.

Figure 4 and Figure 5, respectively, show the MLM top-10 and Top-100 accuracy on predicting codes in the different frequency bins, averaged across the three pre-training models per configuration. Significant comparisons to the unstructured control at a $p < 0.05$ level indicated with an asterisk. We assess statistical significance for each bin using a two-tailed Dunnett's test comparing mean accuracy scores of experimental HRR configurations against the control unstructured configuration. Notably, the top-100 accuracy in frequency bin -12 is non-zero for the HRR methods. These codes in the rarest bin occur only

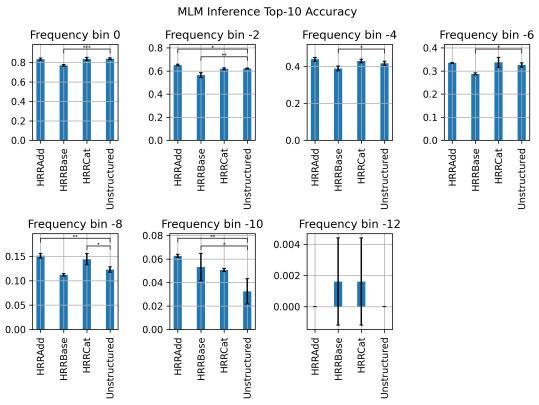


**Figure 4:** The top-10 MLM accuracy for binned code frequencies in log scale. Common codes are in frequency bin 0 with rarest codes being in frequency bin -12. 0.05, 0.01, and 0.001 significance levels comparing to unstructured embeddings are indicated with 1, 2, and 3 asterisks respectively. Note that HRRBase is expected to perform poorly in this test due to lack of code-frequency information.

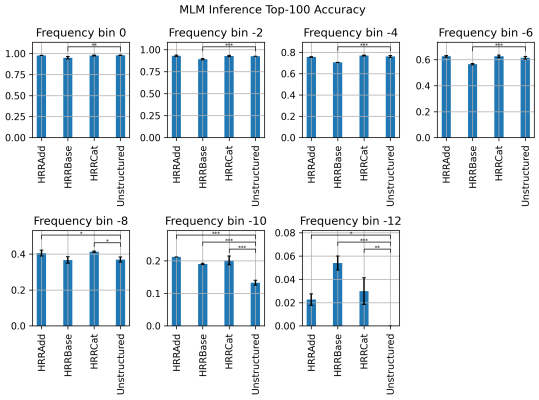


**Figure 5:** The top-100 MLM accuracy for binned code frequencies in log scale. Common codes are in frequency bin 0 with rarest codes being in frequency bin -12. 0.05, 0.01, and 0.001 significance levels comparing to unstructured embeddings are indicated with 1, 2, and 3 asterisks respectively.

**Table 2**
Three cosine similarity case studies looking at related ICD codes for unstructured and HRRBase. The top 4 cosine-similar ICD codes to the chosen code are listed (most to least similar) with their full description and similarity value.

| 2724-9 - Other and unspecified hyperlipidemia | | | |
|---|---|---|---|
| Unstructured | | HRRBase | |
| Pure hypercholesterolemia | 0.542 | Other hyperlipidemia | 1.000 |
| Hyperlipidemia, unspecified | 0.482 | Hyperlipidemia, unspecified | 1.000 |
| Esophageal reflux | 0.304 | Pure hypercholesterolemia | 0.463 |
| Anemia, unspecified | 0.279 | Mixed hyperlipidemia | 0.418 |
| 9916-9 - Hypothermia | | | |
| Unstructured | | HRRBase | |
| Frostbite of hand | 0.418 | Hypothermia, initial encounter | 0.794 |
| Frostbite of foot | 0.361 | Hypothermia not with low env. temp. | 0.592 |
| Drowning and nonfatal submersion | 0.352 | Effect of reduced temp., initial encounter | 0.590 |
| Immersion foot | 0.341 | Other specified effects of reduced temp. | 0.590 |
| K219-10 - Gastro-esophageal reflux disease without esophagitis | | | |
| Unstructured | | HRRBase | |
| Esophageal reflux | 0.565 | Esophageal reflux | 0.635 |
| Hyperlipidemia, unspecified | 0.335 | Gastro-eso. reflux d. with esophagitis | 0.512 |
| Anxiety disorder, unspecified | 0.332 | Reflux esophagitis | 0.512 |
| Essential (primary) hypertension | 0.326 | Hypothyroidism, unspecified | 0.268 |

once in the dataset and therefore have never been used by the model for gradient updates, since they are in the validation dataset. This suggests that the HRR methods have some ability to provide clinically relevant information about rare codes. However, accuracy with the rarest codes remains too low to be of practical value, perhaps due to limited overlap of these codes' atomic vectors with those of more common codes.

## 3.5. Medical Code Case Study

Table 2 shows case studies for codes *Other and unspecified hyperlipidemia* (2724-9), *Hypothermia* (9916-9), and *Gastro-esophageal Reflux disease without esophagitis* (K219-10). In the first case study for 2724-9, we observe highly ontologically similar codes, such as *Other hyperlipidemia* and *Hyperlipidemia, unspecified*, are encoded with high cosine similarity for HRRBase, which is not the case for unstructured embeddings. The co-occurrence problem can be seen in the second case study for 9916-9. The most similar codes for HRRBase are medically similar codes that would not usually co-occur, while for unstructured embeddings the most similar codes co-occur frequently. For the final case study on K219-10, frequency-related bias can be observed in the unstructured embeddings with frequent but mostly ontologically unrelated codes as part of the top list of cosine similar codes, whereas the top list of cosine similar codes for HRRBase contains medically similar codes.

We broadened this case study to test statistical differences in cosine and semantic embedding similarity between structured and unstructured embeddings. 30 ICD codes were selected from different frequency categories in the dataset, with 10 codes drawn randomly from the 300 most common codes, 10 codes drawn randomly by weighted frequency from codes appearing fewer than 30 times in the dataset, and 10 codes randomly selected by weighted frequency from the entire dataset. For each selected code, the top 4 cosine-similar ICD codes were assessed by a physician for ontological similarity.

For each frequency category, a one-tailed Fisher's exact test was conducted to determine whether a relationship existed between embedding type and clinical relatedness. We found that results in the case of the rare codes were statistically significant, with $p = 2.44{\times}10^{-8}$. With 10 rare codes and the top 4 cosine-similar ICD codes selected for each rare code, there are 40 top cosine-similar codes in total. In the case of unstructured embeddings, only 4 of the top 40 cosine-similar codes were deemed to be strongly ontologically related by our physician with the remaining codes deemed to be less related and unrelated. In the case of our structured HRRBase embeddings, 28 of the top 40 cosine-similar codes were deemed to be strongly ontologically related by our physician with the remaining codes deemed to be less related and unrelated. This suggests that knowledge-integrated structured embeddings are associated with greater clinical relevance of the top cosine-similar codes than unstructured embeddings for rare codes where little training data exists.

## 4. Discussion

Transformers have leading performance in many applications, but their internal processes are opaque, emerging from enormous parameter sets and data volumes beyond human experience. It is hard to know when they can be trusted. For example, generative transformers are prone to subtle confabulations. Transformers have a general-purpose architecture that performs as well in vision and other modalities as in language. They are a culmination of a key trend in artificial intelligence, away from problem-specific engineering, and toward massive data and computation. This trend is justified in terms of performance. However, given two models with equal performance, one with more explicit conceptual structure is preferable in terms of trust and explainability.

The work presented here is a step in this direction, with our HRRBase embeddings that have explicit conceptual structure and perform equivalently or better compared to typical transformer embeddings. The benefit of structured embeddings becomes more pronounced for tasks that involve codes that are rare or are not present in training data. HRR embeddings can also be relied on to represent medical meaning rather than co-occurrence in the training data. They also untangle the representation of code frequency, so that it can be included or not, and its effects on decisions understood. Importantly, despite this additional structure, the embeddings are thoroughly learned, suggesting that the approach will be consistent with high performance beyond the examples we have studied.

As our method scales with and leverages PyTorch autograd in the construction of the vector-symbolic embeddings, it is compatible with existing medical LLM architectures as an embedding component capable of encoding domain knowledge.

Future work could explore the potential of these structured embeddings for explaining and controlling the observed frequency bias. As HRRs can be queried with linear operations, future work could also explore whether transformers can learn to extract specific information from these composite embeddings. Limitations to address in future work include the complexity of processing knowledge graphs to be compatible with HRRs. Another important limitation is that our method relies on rare-code HRRs sharing atomic elements with common-code HRRs. However, in SNOMED CT, rare codes are likely to contain some rare atomic elements. To address this point, in addition to SNOMED CT, knowledge could be encoded from sources such as pre-trained medical embeddings, different medical ontologies, and other medical domain knowledge to further improve our proposed methodology. In LLMs that process both medical codes and text, it would make sense to share word embeddings between modalities. This would allow training of each modality to benefit from training of the other, and may help to align the representations of codes and text.

## 5. Conclusion

We proposed a novel hybrid neural-symbolic approach called HRR-BERT that integrates medical ontologies represented by HRR embeddings. In tests with the MIMIC-IV dataset, HRRBERT models modestly outperformed baseline models with unstructured embeddings for pretraining, disease prediction accuracy, mortality prediction F1, and fine-tuning tasks involving infrequently seen codes. HRRBERT models had pronounced performance advantages in MLM with rare codes and disease prediction for patients with no codes seen during training (ROOD - Unseen in Table 1). We also showed that HRRs can be used to create medical code embeddings that better respect ontological similarities for rare codes. A key benefit of our approach is that it facilitates explainability by disentangling token-frequency information, which is prominently represented but implicit in unstructured embeddings.

## References


[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762.

[2] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, npj Digital Medicine 4 (2021) 86. doi:10.1038/s41746-021-00455-y.

[3] A. Ganesan, H. Gao, S. Gandhi, E. Raff, T. Oates, J. Holt, M. McLean, Learning with holographic reduced representations, CoRR abs/2109.02157 (2021). URL: https://arxiv.org/abs/2109.02157. arXiv:2109.02157.

[4] M. K. Sarker, L. Zhou, A. Eberhart, P. Hitzler, Neuro-symbolic artificial intelligence, AI Communications 34 (2021) 197–209.

[5] S. Ramgopal, L. N. Sanchez-Pinto, C. M. Horvat, M. S. Carroll, Y. Luo, T. A. Florin, Artificial intelligence-based clinical decision support in pediatrics, Pediatric research 93 (2023) 334–341.

[6] T. Yu, T. Tuinstra, B. Hu, R. Rezai, T. Fortin, R. DiMaio, B. Vartian, B. Tripp, Frequency bias in mlm-trained bert embeddings for medical codes, CMBES Proceedings 45 (2023). URL: https://proceedings.cmbes.ca/index.php/proceedings/article/view/1050.

[7] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, et al., Pythia: A suite for analyzing large language models across

training and scaling, in: International Conference on Machine Learning, PMLR, 2023, pp. 2397–2430.
[8] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al., Large language models encode clinical knowledge, Nature 620 (2023) 172–180.
[9] T. Plate, Holographic reduced representations, IEEE Transactions on Neural Networks 6 (1995) 623–641. doi:10.1109/72.377968.
[10] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, R. Mark, Mimic-iv (version 2.0, 2022. URL: https://doi.org/10.13026/7vcr-e114. doi:10.13026/7vcr-e114.
[11] P. Kanerva, Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors, Cognitive Computation 1 (2009) 139–159. URL: https://api.semanticscholar.org/CorpusID:733980.
[12] R. W. Gayler, Vector symbolic architectures answer jackendoff's challenges for cognitive neuroscience, 2004. arXiv:cs/0412059.
[13] P. Neubert, S. Schubert, Hyperdimensional computing as a framework for systematic aggregation of image descriptors (2021). URL: http://arxiv.org/abs/2101.07720. doi:10.48550/arXiv.2101.07720, arXiv:2101.07720 [cs].
[14] P. Neubert, S. Schubert, K. Schlegel, P. Protzel, Vector semantic representations as descriptors for visual place recognition, in: Robotics: Science and Systems XVII, Robotics: Science and Systems Foundation, 2021. URL: http://www.roboticsproceedings.org/rss17/p083.pdf. doi:10.15607/RSS.2021.XVII.083.
[15] A. Rahimi, P. Kanerva, L. Benini, J. M. Rabaey, Efficient biosignal processing using hyperdimensional computing: Network templates for combined learning and classification of exg signals, Proceedings of the IEEE 107 (2019) 123–143. doi:10.1109/JPROC.2018.2871163.
[16] K. Schlegel, P. Neubert, P. Protzel, Hdc-minirocket: Explicit time encoding in time series classification with hyperdimensional computing (2022). URL: http://arxiv.org/abs/2202.08055. doi:10.48550/arXiv.2202.08055, arXiv:2202.08055 [cs].
[17] P. Smolensky, Tensor product variable binding and the representation of symbolic structures in connectionist systems, Artificial Intelligence 46 (1990) 159–216. URL: https://www.sciencedirect.com/science/article/pii/000437029090007M. doi:https://doi.org/10.1016/0004-3702(90)90007-M.
[18] M. M. Alam, E. Raff, S. Biderman, T. Oates, J. Holt, Recasting self-attention with holographic reduced representations, 2023. arXiv:2305.19534.
[19] J. Kim, H. Lee, M. Imani, Y. Kim, Efficient hyperdimensional learning with trainable, quantizable, and holistic data representation, in: 2023 Design, Automation Test in Europe Conference Exhibition (DATE), 2023, pp. 1–6. doi:10.23919/DATE56975.2023.10137134.
[20] M. Nickel, L. Rosasco, T. Poggio, Holographic embeddings of knowledge graphs (2015). URL: http://arxiv.org/abs/1510.04935. doi:10.48550/arXiv.1510.04935, arXiv:1510.04935 [cs, stat].
[21] T. Dash, A. Srinivasan, L. Vig, Incorporating symbolic domain knowledge into graph neural networks, Machine Learning 110 (2021) 1609–1636. URL: https://doi.org/10.1007%2Fs10994-021-05966-z. doi:10.1007/s10994-021-05966-z.
[22] M. Kulmanov, F. Z. Smaili, X. Gao, R. Hoehndorf, Semantic similarity and machine learning with ontologies, Briefings in Bioinformatics 22 (2020) bbaa199. URL: https://doi.org/10.1093/bib/bbaa199. doi:10.1093/bib/bbaa199. arXiv:https://academic.oup.com/bib/article-pdf/22/4/bbaa199/39132158/bbaa199.pdf.
[23] V. Riikka, V. Anne, P. Sari, Systematized nomenclature of medicine-clinical terminology (snomed ct) clinical use cases in the context of electronic health record systems: Systematic literature review, JMIR Med Inform (2023). doi:10.2196/43750.
[24] L. Song, C. W. Cheong, K. Yin, W. K. Cheung, B. C. M. Fung, J. Poon, Medical concept embedding with multiple ontological representations, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 4613–4619. URL: https://doi.org/10.24963/ijcai.2019/641. doi:10.24963/ijcai.2019/641.
[25] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation 101 (2000) e215–e220. Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.
[26] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.
[27] NLM, Snomed ct to icd-10-cm map, 2022. URL: https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html.
[28] OHDSI, Ohdsi standardized vocabularies, 2019. URL: https://github.com/OHDSI/Vocabulary-v5.0/wiki.

[29] NLM, Icd-9-cm diagnostic codes to snomed ct map, 2022. URL: https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html.
[30] NCHS, Diagnosis code set general equivalence mappings, 2018. URL: https://ftp.cdc.gov/pub/health_statistics/nchs/Publications/ICD10CM/2018/Dxgem_guide_2018.pdf.
[31] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, O. Badawi, The eICU Collaborative Research Database, a freely available multi-center database for critical care research, Scientific data 5 (2018) 1–13.

# A. List of 32 ROOD Codes

The following is the list of 32 ROOD codes:

1. G248-10: Other dystonia
2. E8498-9: Accidents occurring in other specified places
3. E9688-9: Assault by other specified means
4. Z681-10: Body mass index (BMI) 19.9 or less, adult
5. 30550-9: Opioid abuse, unspecified
6. R262-10: Difficulty in walking, not elsewhere classified
7. E887-9: Fracture, cause unspecified
8. R471-10: Dysarthria and anarthria
9. 9916-9: Hypothermia
10. E9010-9: Accident due to excessive cold due to weather conditions
11. F10129-10: Alcohol abuse with intoxication, unspecified
12. E8499-9: Accidents occurring in unspecified place
13. R636-10: Underweight
14. 920-9: Contusion of face, scalp, and neck except eye(s)
15. R4182-10: Altered mental status, unspecified
16. 95901-9: Head injury, unspecified
17. 78097-9: Altered mental status
18. F29-10: Unspecified psychosis not due to a substance or known physiological condition
19. Z880-10: Allergy status to penicillin
20. Z818-10: Family history of other mental and behavioral disorders
21. 81600-9: Closed fracture of phalanx or phalanges of hand, unspecified
22. 87341-9: Open wound of cheek, without mention of complication
23. H9222-10: Otorrhagia, left ear
24. Z978-10: Presence of other specified devices
25. G20-10: Parkinson's disease
26. G249-10: Dystonia, unspecified
27. 9100-9: Abrasion or friction burn of face, neck, and scalp except eye, without mention of infection
28. 78906-9: Abdominal pain, epigastric
29. E8889-9: Unspecified fall
30. 30500-9: Alcohol abuse, unspecified
31. G520-10: Disorders of olfactory nerve
32. 8020-9: Closed fracture of nasal bones

## A.1. Learning through HRR Operations Efficiently

To make the HRR concept embeddings useful for a deep neural network, the operations used to form the embeddings need to be compatible with backpropagation so that gradient descent can update the lower-level atomic vectors. We desired a function that produced the ICD concept embedding matrix, , given the inputs of the VSA knowledge graphs, $\mathcal{G}_i$, and symbol embedding matrices, and .

We attempted three approaches to computing through VSA operations. First, we naively tried to compute each concept vector in one at a time. However, this approach was too slow in both forward and backward pass, requiring more than 1 second for each pass. Our second approach was using slices of along the relationship dimension as a sparse binary matrix, which, when multiplied with , would perform the indexing and summing of atomic vectors for each concept. This result can be convolved with the relationship vector and added to the concept embedding matrix. This approach was much faster and used a moderate amount of memory for one of our less complex VSA formulations. However, when dealing with our most complex formulation, it used ~15 GB of memory.

Our final approach took advantage of the fact that many disease concepts use relationship, but to different atomic symbols. Also, number of times a concept uses a particular relationship is relatively low, except for the SNOMED "isA" relationship and our defined "description" relationship. Thus, for a particular relationship, we can contribute to building many disease concept vectors at once by selecting many atomic vectors, doing a vectorized convolution with the relationship vector, and distributing the results to be added with the appropriate concept embedding rows. This step needs to be repeated at most $m$ times for a particular relationship, where $m$ is the maximum multiplicity of that relationship among all concepts. We improved memory efficiency by performing fast Fourier transforms (FFTs) on the atomic vector embeddings and construct the concept vectors by performing binding via element-wise multiplication in the Fourier domain. Due to the linearity of the HRR opera-

tions, we performed a final FFT on the complex-valued concept embedding to convert back to the real domain.

The final approach is much faster than the first approach since it takes advantage of vectorized operations to contribute to many concept vectors at once. It is also more memory efficient than the second approach since all the intermediate results are dense, so allocations are not wasted on creating mostly sparse results. On our most complex formulation, this approach uses ~3.5 GB of memory, and takes ~80 ms and ~550 ms for forward and backward pass respectively.