TACCO: Task-guided Co-clustering of Clinical Concepts and Patient Visits for Disease Subtyping based on EHR Data

Ziyang Zhang
Emory University
Atlanta, Georgia, USA
ziyang.zhang2@emory.edu

Hejie Cui
Emory University
Atlanta, Georgia, USA
hejie.cui@emory.edu

Ran Xu
Emory University
Atlanta, Georgia, USA
ran.xu@emory.edu

Yuzhang Xie
Emory University
Atlanta, Georgia, USA
yuzhang.xie@emory.edu

Joyce C. Ho
Emory University
Atlanta, Georgia, USA
joyce.c.ho@emory.edu

Carl Yang*
Emory University
Atlanta, Georgia, USA
j.carlyang@emory.edu

ABSTRACT

The growing availability of well-organized Electronic Health Records (EHR) data has enabled the development of various machine learning models towards disease risk prediction. However, existing risk prediction methods overlook the heterogeneity of complex diseases, failing to model the potential disease subtypes regarding their corresponding patient visits and clinical concept subgroups. In this work, we introduce **TACCO**, a novel framework that jointly discovers clusters of clinical concepts and patient visits based on a hypergraph modeling of EHR data. Specifically, we develop a novel self-supervised co-clustering framework that can be guided by the risk prediction task of specific diseases. Furthermore, we enhance the hypergraph model of EHR data with textual embeddings and enforce the alignment between the clusters of clinical concepts and patient visits through a contrastive objective. Comprehensive experiments conducted on the public MIMIC-III dataset and Emory internal CRADLE dataset over the downstream clinical tasks of phenotype classification and cardiovascular risk prediction demonstrate an average 31.25% performance improvement compared to traditional ML baselines and a 5.26% improvement on top of the vanilla hypergraph model without our co-clustering mechanism. In-depth model analysis, clustering results analysis, and clinical case studies further validate the improved utilities and insightful interpretations delivered by **TACCO**. Code is available at https://github.com/PericlesHat/TACCO.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Computing methodologies** → **Cluster analysis**; **Neural networks**.

KEYWORDS

Self-supervised; Clustering; Hypergraph; Electronic Health Records

*Carl Yang is the corresponding author.

KDD '24, August 25–29, 2024, Barcelona, Spain.

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671594

ACM Reference Format:
Ziyang Zhang, Hejie Cui, Ran Xu, Yuzhang Xie, Joyce C. Ho, and Carl Yang. 2024. TACCO: Task-guided Co-clustering of Clinical Concepts and Patient Visits for Disease Subtyping based on EHR Data. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3637528.3671594

1 INTRODUCTION

Electronic Health Records (EHR) is a significant advancement in medical data management. They store patient data such as medical history, treatments, and lab results. EHRs have played a crucial role in advancing healthcare applications, such as treatment decision support [63] and preventive care [35, 57]. In recent years, machine learning (ML) has been employed to extract valuable insights from EHRs and is widely studied in healthcare informatics. This leads to innovations in various applications such as suicide risk prediction [6, 61], diagnosis prediction [26, 42, 52], phenotypes classification [19, 71, 72], and drug recommendation [75].

Among various ML models for EHR data, graph-based models have shown promise with their capabilities of modeling complex structures within EHRs [9, 11, 38, 58]. Given the lack of generic ways of constructing reliable graph structures from EHR, hypergraphs have recently been used as a flexible data structure that directly models the interactions between clinical concepts (*i.e.*, names of medical codes in EHR) and patient visits [71, 72]. These higher-order modeling approaches are robust with sparse incomplete features and can generate interpretable predictions on the important clinical concepts for analyzed diseases.

Understanding disease subtypes is crucial for studying the mechanisms of complex diseases and establishing personalized treatments. Disease subtypes refer to specific variations within a broader disease category, differentiated by unique characteristics, symptoms, or treatment responses. In EHR, disease subtypes can be defined as subgroups of patients who exhibit similar patterns related to clinical concepts such as diagnoses, medications, and procedures received in their medical visits [44, 68]. In this work, beyond the commonly defined disease types such as in Current Procedural Terminology (CPT) and International Classification of Diseases (ICD), we are interested in further discovering fine-grained disease subtypes such as regarding subgroups of diabetic patients with high risks of stroke, retinopathy, neuropathy, or nephropathy [41].

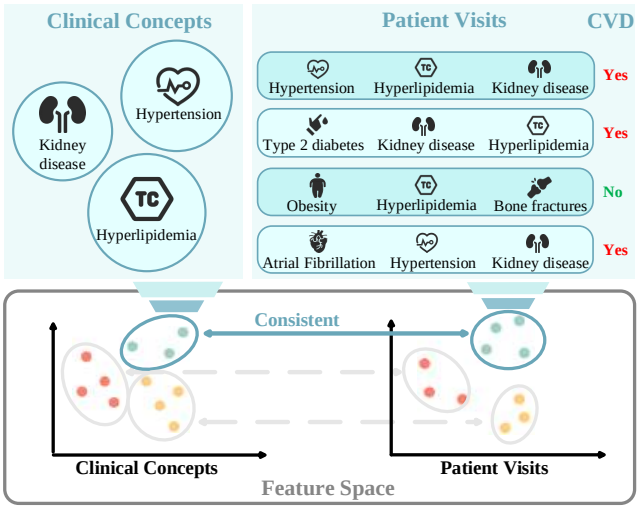


Figure 1: Co-clustering analysis of clinical concepts and patient visits for cardiovascular disease (CVD) identification for diabetic patients. TACCO performs co-clustering over a hypergraph to yield consistent clinical concept and patient visit subgroups. A node cluster of 3 clinical concepts *(blue circle in the left panel)* and a hyperedge cluster of 4 patient visits *(blue circle in the right panel)* suggest potential disease subtypes in correlation with CVD outcomes.

On the other hand, modeling disease subtypes and the complex interactions between clinical concepts and patient visits can also assist in the risk prediction of specific diseases. For example, in Figure 1, clinical concepts such as hypertension, kidney disease, and hyperlipidemia can indicate a group of patients at high risk of CVD [71], who may require targeted therapies to mitigate the risk of further diabetes-related complications [1]. These connections can help medical professionals gain fine-grained understandings of the risks and design precise effective interventions.

To the best of our knowledge, jointly analyzing the subgroups of clinical concepts and patient visits is rarely studied in healthcare informatics. Previous models have used simple algorithms like K-means [1, 59, 60], single variable analysis [22], and matrix factorization [65] to identify disease subtypes. However, these statistical methods lack guidance from risk prediction tasks and thus need manually pre-defined sets of clinical concepts for specific diseases [13, 53]. In recent years, more advanced clustering techniques such as Deep Embedded Clustering [70] and other self-supervised approaches [23, 32, 77] have been studied for EHR-based clinical predictions. However, these methods can only cluster one type of entity and do not consider the higher-order interactions among clinical concepts and patient visits [28, 79, 80].

We aim to develop a model to jointly identify clusters of clinical concepts and patient visits for disease subtyping on EHR data. Task-guided co-clustering is used to analyze clinical concepts and patient visits, providing meaningful interpretations for predictive tasks. However, some challenges need to be addressed: (1) *Inadequate Graph Representation.* Conventional GNNs struggle to efficiently represent medical concepts and patient visits due to their focus on pairwise relationships. EHR data contains multiple medical codes that can be repeated across various visits, requiring a higher-order graph modeling technique for accurate representation. Existing work [40, 43, 51, 71] only focus on geometric structures and ignore clinical natural language descriptions from medical coding systems, limiting the ability of GNNs to learn a comprehensive representation and negatively impacting downstream tasks. (2) *Lack of Supervision for Interpretation.* Recent ML research in healthcare has increasingly shifted towards not only showcasing model performance but also providing interpretability. Approaches such as factual counterfactual reasoning [21, 72] and time-aware mechanisms [62, 78] can extract interpretable subsets for EHR data, but they require supervised learning. In our case, we aim to uncover patterns through co-clustering without supervision and find consistency between clinical concepts and patient visits for interpretations.

In this work, we introduce **TACCO** (**Ta**sk-guided **Co**-**C**lustering), the first framework that clusters clinical concepts and patient visits on EHR networks using self-supervised co-clustering and a contrastive alignment module. Our framework TACCO "homogenizes" clusters to reveal deeper insights into the connections between subgroups of clinical concepts and patient visits. The contributions of our work are summarized as follows:

- We identify a novel task for disease subtyping in EHR analysis, where clusters of clinical concepts and patient visits are jointly studied and contribute to understanding complex diseases.
- We develop TACCO, a task-guided self-supervised framework that uncovers patterns through co-clustering without supervision. The model is built based on a text-enhanced hypergraph transformer with a dual application of deep clustering on both nodes and hyperedges. The model further aligns clinical concepts and patient visit clusters through contrastive learning for identifying consistent disease subtypes.
- Extensive quantitative experiments and clustering analysis are conducted on two clinical EHR datasets, the publicly available MIMIC-III [29] and the private CRADLE. TACCO outperforms the previous state-of-the-art model [72] and demonstrates a notable 5.26% improvement across four metrics compared to hypergraph model backbone [71]. Case studies further show that TACCO is capable of grouping consistent clinical concepts and patient visits that reveal disease subtypes related to a specific disease (e.g., CVD). As validated by a domain expert, the captured disease subtypes could have different levels of relationships (e.g., positive, weak, or negative) with specific diseases for a fine-grained understanding in practice.

2 RELATED WORK

Machine Learning in Healthcare. Medical research has utilized various model architectures to analyze healthcare data. Earlier research efforts mainly employed fundamental model architectures. For instance, Liu et al. [36] used auto-encoders to diagnose Alzheimer's disease. Choi et al. [12] utilized word2vec [47] to learn the representations of medical concepts. More recent models, such as Convolutional Neural Networks and Recurrent Neural Networks have also been widely applied in various applications [8, 10, 24, 49, 55].

To capture the intrinsic structures of healthcare data, there is a growing interest in graph-based methods. For example, event

sequences are modeled as weight graphs for heart failure prediction [33], robust relations among medical codes are learned [11], and medical knowledge graphs are incorporated for downstream reasoning [9, 43]. To consider the higher-level relations in structured data, Xu et al. [71, 72] proposed to model patient visits as hyperedges in a hypergraph transformer, which overcomes the limitations of pairwise relations and provides interpretable insights.

Clustering Methods. Traditional clustering methods are mostly algorithmic and heuristic-based, e.g., K-means [39], hierarchical clustering [30], and density-based spatial clustering of applications with noise (DBSCAN) [16]. While effective for linearly separable data, these algorithms are not learning-based and cannot generalize to unseen data. To address this limitation, deep clustering techniques like DEC [70] introduced self-supervised techniques for complex cluster representation. Subsequently, several self-supervised clustering approaches [23, 27, 32, 76, 77] were proposed to enhance the robustness of clustering and were generalized to various settings.

Clustering techniques have been employed to identify meaningful subsets of medical concepts in healthcare. For example, iCluster [60] jointly estimated the latent tumor subtypes by refining K-means and Gaussian latent variable models. SilHAC [50] tackled cluster number estimation and identification in cancer data based on the Silhouette Index. NCIS [37] identified cancer subtypes based on gene expression with network-assisted co-clustering. MODEC [79] leveraged DEC for cancer subtype identification and clinical feature analysis. However, none of these studies has explored the co-clustering of clinical concepts and patient visits to improve downstream prediction tasks and provide interpretations.

3 METHOD

3.1 Problem Definition

We define a hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is a set of vertices, each representing a medical code, and $\mathcal{E}$ is a set of hyperedges, each representing a patient visit that includes a subset of medical codes from $\mathcal{V}$. The goals of this study include: (1) Given a patient's clinical record, predict the clinical outcome y of that patient. (2) Analyze clusters of clinical concepts and patient visits for disease subtyping based on the hypergraph $\mathcal{G}$ constructed from EHR data.

3.2 Text-enhanced Hypergraph Transformer

To effectively capture the intricate relationships in EHR, we adopt a hypergraph transformer [7]. As shown in the middle part of Figure 2, the node embedding of each medical code in the hypergraph are initialized with information from two aspects: structures and semantics. The structural part $\boldsymbol{X}_{\text{structure}}$ is obtained from DeepWalk [54], where we apply random walks on $\mathcal{G}$ to train a Skip-gram model. For the semantical part, we process the medical code descriptions of nodes with SapBERT [34], a transformer-based model pre-trained on extensive biomedical literature, to generate text embeddings $\boldsymbol{X}_{\text{text}}$. We directly concatenate these two vectors as the initial node embeddings:

$$\boldsymbol{X} = [\boldsymbol{X}_{\text{structure}}; \boldsymbol{X}_{\text{text}}]. \tag{1}$$

For the l-th layer of the hypergraph, node and hyperedge embeddings are denoted by $\boldsymbol{X}^{(l)} \in \mathbb{R}^{|\mathcal{V}| \times d}$ and $\boldsymbol{E}^{(l)} \in \mathbb{R}^{|\mathcal{E}| \times d'}$, where d and d' are dimensionality parameters of the node and hyperedge feature spaces, respectively. The embeddings are updated through a two-step message-passing mechanism:

$$\boldsymbol{E}_e^{(l)} = f_{\mathcal{V} \to \mathcal{E}}(\mathcal{V}_{e,\boldsymbol{X}^{(l-1)}}), \quad \boldsymbol{X}_v^{(l)} = f_{\mathcal{E} \to \mathcal{V}}(\mathcal{E}_{v,\boldsymbol{E}^{(l)}}), \tag{2}$$

where $\mathcal{V}_{e,\boldsymbol{X}} = \{\boldsymbol{X}_{u,:} : u \in e\}$ is the representation of nodes contained in the hyperedge e, and $\mathcal{E}_{v,\boldsymbol{E}} = \{\boldsymbol{E}_{e,:} : v \in e\}$ denote the representations of hyperedges that contain the node v. For the two functions $f(\cdot)$, we leverage a self-attention mechanism [66] that allows the model to focus on the most informative parts:

$$\text{Self-Att}(\boldsymbol{S}) = \text{LayerNorm}(\boldsymbol{Y} + \text{FFN}(\boldsymbol{Y})), \tag{3}$$

where $\boldsymbol{Y}$ is the output from the multi-head self-attention block:

$$\boldsymbol{Y} = \text{LayerNorm}(\boldsymbol{S} + \Big\|_{i=1}^{h} \text{SA}_i(\boldsymbol{S})). \tag{4}$$

$\text{SA}_i(\boldsymbol{S})$ denotes the scaled dot-product attention mechanism:

$$\text{SA}_i(\boldsymbol{S}) = \text{softmax}\left(\frac{\boldsymbol{W}_i^Q (\boldsymbol{S}\boldsymbol{W}_i^K)^\top}{\sqrt{\lfloor d/h \rfloor}}\right)\boldsymbol{S}\boldsymbol{W}_i^V, \tag{5}$$

where $\boldsymbol{W}_i^Q$, $\boldsymbol{W}_i^K$, and $\boldsymbol{W}_i^V$ are parameter matrices for the i-th head corresponding to queries, keys, and values, respectively. The input sequence $\boldsymbol{S}$ will be projected into different h heads. The output of each head is then concatenated (denoted by $\|$) to form the multi-head attention output. The input dimensionality d is evenly split across the heads in $\lfloor d/h \rfloor$ dimensions. The multi-head attention output is combined with a feed-forward neural network (FFN), which is composed of a 2-layer Multilayer Perceptron (MLP) with a ReLU activation. In our task, we do not include the position encoding technique in the standard Transformer due to the lack of such information in our datasets. A 2-layer MLP is utilized for the disease risk prediction task, along with a sigmoid activation function, denoted by σ:

$$\hat{y} = \sigma\left(\text{MLP}\left(\Big\|_{l=1}^{L} \hat{\boldsymbol{E}}^{(l)}\right)\right). \tag{6}$$

The learning objective is a binary cross-entropy loss, where y represents the truth label and $\hat{y}$ is the predicted probability:

$$\mathcal{L}_{\text{cls}} = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}). \tag{7}$$

3.3 Deep Self-Supervised Co-clustering

The main challenge in our problem is the lack of supervision for generating clusters of clinical concepts and patient visits for disease subtyping. Since there are no labels available, traditional supervised methods such as classification cannot be applied directly. Some previous works use simple K-means to extract specific disease subtypes, but these methods are not data-driven and cannot be integrated into deep models. Inspired by Xie et al. [70], we employ a deep clustering method that iteratively learns the cluster assignments in a self-supervised manner. This deep clustering technique has been proven to be effective in graphs and can be jointly optimized with embedding propagation [74].

In TACCO, we propose a dual application of Deep Embedded Clustering (DEC) on both clinical concepts and patient visits, as shown in the left and right parts of Figure 2. Specifically, we seek to jointly learn soft clustering assignments $\boldsymbol{Q}$ for both nodes and hyperedges, denoted as $\boldsymbol{Q}_{\mathcal{V}}$ and $\boldsymbol{Q}_{\mathcal{E}}$, respectively. For any given

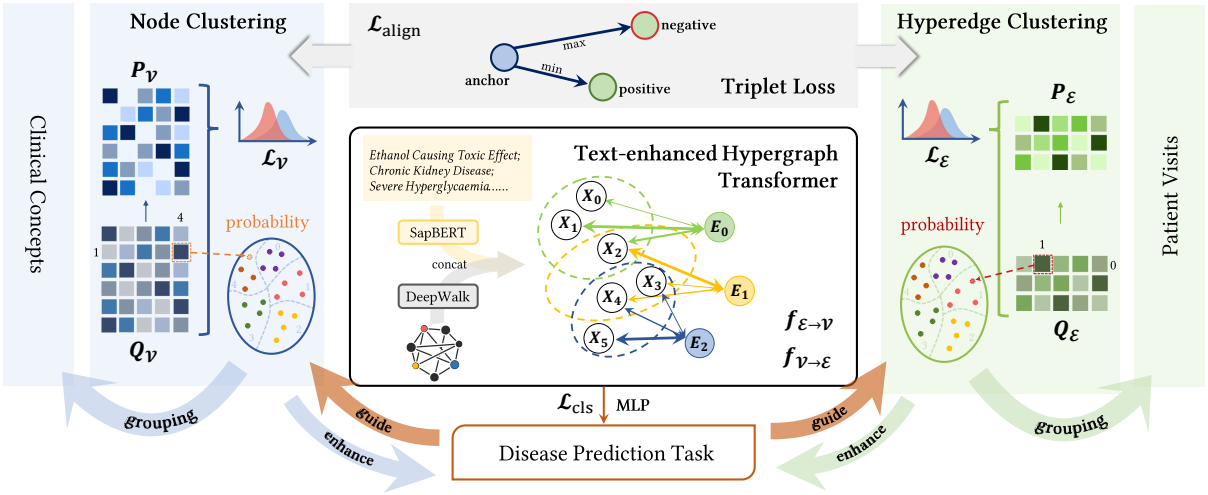


Figure 2: Pipeline of TACCO. A hypergraph transformer *(middle)* is used as a backbone to model node and hyperedge interactions. Node clustering *(left)* and hyperedge clustering *(right)* are jointly optimized to produce clusters of clinical concepts and patient visits. A triplet loss function *(top)* is applied for a consistent cluster alignment across two domains.

node (or hyperedge) i and cluster k, the soft assignment q_{ik} is calculated based on the similarity between the node's (or hyperedge's) embedding and the cluster centroid, formalized as:

$$q_{ik} = \frac{(1 + \|\mathbf{x}_i - \mathbf{u}_k\|^2)^{-1}}{\sum_j (1 + \|\mathbf{x}_i - \mathbf{u}_j\|^2)^{-1}}, \tag{8}$$

where $\mathbf{x}_i$ is the embedding of node (or hyperedge) i and $\mathbf{u}_k$ is the centroid of the k-th cluster. Similarly to DEC, we perform standard K-means to initialize K centroids $\{\mathbf{u}_j\}_{j=1}^K$.

Subsequent to the determination of $\boldsymbol{Q}$, we construct a refined target distribution $\boldsymbol{P}$, which aims to enhance cluster purity by emphasizing confident assignments. The components of $\boldsymbol{P}$ are computed by squaring the elements of $\boldsymbol{Q}$ and normalizing them across each cluster, as follows:

$$p_{ik} = \frac{q_{ik}^2 / f_k}{\sum_j q_{ij}^2 / f_j}, \tag{9}$$

with $f_k = \sum_i q_{ik}$ representing the sum of the soft assignments to the k-th cluster. We then minimize the Kullback-Leibler divergence between $\boldsymbol{Q}$ and $\boldsymbol{P}$, which serves as the self-training clustering loss:

$$\mathcal{L}_\mathcal{V} = \mathcal{L}_\mathcal{E} = KL(\boldsymbol{P}||\boldsymbol{Q}) = \sum_i \sum_k p_{ik} \log \frac{p_{ik}}{q_{ik}}. \tag{10}$$

3.4 Cluster Contrastive Alignment

We further align these clusters across two domains in a shared feature space, which helps to generate consistent clusters on clinical concepts and patient visits for interpretations. The cluster centroids are computed by incorporating the soft assignment probabilities from the matrix $\boldsymbol{Q}$. Specifically, for each cluster k, the centroid $\mathbf{c}^k$ is determined by the weighted average of the node (or hyperedge) embeddings, with the weights given by the soft assignments q_{ik} for node (or hyperedge) i. This is formally expressed as:

$$\mathbf{c}^k = \frac{\sum_i q_{ik} \mathbf{x}_i}{\sum_i q_{ik}}, \tag{11}$$

where the denominator $\sum_i q_{ik}$ is the sum of the soft assignments to cluster k. Unlike the rigid nature of hard clustering that strictly assigns nodes/hyperedges based on maximal probability, soft assignment reflects the intrinsic nature of disease subtypes, where clinical concepts (e.g., *obesity*) may serve as potential causes for multiple diseases. With soft assignments, TACCO preserves such natural but often overlooked overlaps present in EHR data.

To enable an unsupervised cluster alignment across two domains, existing strategies [15, 69] propose to match the first-order moments of the k-th cluster from the source domain and target domain. Their loss function is designed to minimize the distance $\mathcal{D}$ between the node clusters embedding $\boldsymbol{C}_e \in \mathbb{R}^{|\mathcal{E}|\times K}$ and hyperedge clusters embedding $\boldsymbol{C}_v \in \mathbb{R}^{|\mathcal{V}|\times K}$ of the corresponding clusters across two domains. However, these strategies presuppose that the indices of clusters between the two domains are pre-aligned. Our task centers on discovering these correspondences in situations where clusters are randomly generated. Thus, we cannot rely on alignment methods that assume index matching between two domains.

Using the principle of contrastive learning, we align the consistent clusters by minimizing the distance between each node cluster centroid and its nearest hyperedge cluster centroid while maximizing the separation from less similar centroids. This process is illustrated on the top of Figure 2. In this design, the node and hyperedge embeddings from Eq. (2) are independently processed by a projection MLP head [20], which contains 2 linear layers with batch normalization and a ReLU activation:

$$\boldsymbol{Z}_v = \text{MLP}(\boldsymbol{C}_v), \boldsymbol{Z}_e = \text{MLP}(\boldsymbol{C}_e). \tag{12}$$

We use the triplet loss function to align cross-domain embedding without explicit label correspondences, which is calculated as:

$$\mathcal{L}_{\text{align}} = \sum_{i=1}^{|\mathcal{V}|} \max(0, \mathcal{D}(\mathbf{z}_v^i, \mathbf{z}_e^{i+}) - \mathcal{D}(\mathbf{z}_v^i, \mathbf{z}_e^{i-}) + m), \tag{13}$$

where $\mathbf{z}_v^i$ (*i.e.*, anchor) is the i-th embedding from Z_v, $\mathbf{z}_e^{i+}$ is the positive sample for $\mathbf{z}_v^i$ within Z_e, and $\mathbf{z}_e^{i-}$ represents other negative samples, with $\mathcal{D}$ measuring distance and m setting the minimum desired difference between positive and negative samples. In this case, the distance $\mathcal{D}$ is measured by negative cosine similarity, which has been proven to be effective in contrastive learning [5]:

$$\mathcal{D}(\mathbf{z}_v, \mathbf{z}_e) = -\frac{\mathbf{z}_v \cdot \mathbf{z}_e}{\|\mathbf{z}_v\|\|\mathbf{z}_e\|}, \tag{14}$$

where $\cdot$ denotes dot product, and $\|\cdot\|$ is the l_2-norm.

3.5 Learning Objective

The final learning objective of TACCO is the sum of three parts:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \alpha(\mathcal{L}_{\mathcal{V}} + \mathcal{L}_{\mathcal{E}}) + \beta\mathcal{L}_{\text{align}}, \tag{15}$$

where α and β are hyperparameters for weighting different losses.

4 EXPERIMENTS

In this section, we evaluate TACCO on two EHR datasets, in terms of the performance of downstream clinical tasks, in-depth model analysis, clustering analysis, and case studies.

4.1 Experiment Settings

Datasets and Tasks. We adopt the public **MIMIC-III** dataset [29] for a phenotype classification task. MIMIC-III comprises de-identified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. We follow the setting of [24] to identify 25 phenotypes, including 12 acute conditions (e.g., pneumonia), 8 chronic conditions (e.g., chronic kidney disease), and 5 mixed conditions (e.g., conduction disorders). In terms of experiments, we choose patients with more than one visit and utilize the records from a preceding visit to predict the diagnostic phenotypes of the subsequent visit. These visits are represented as hyperedges within our hypergraph modeling framework, each annotated with a 25-category multihot label.

We also utilize the **CRADLE** (Emory Clinical Research Analytics Data Lake Environment) dataset for a CVD risk prediction task. Project CRADLE contains close to 48 thousand de-identified patient records with type 2 diabetes seen at Emory Healthcare System between 2013 and 2017. Following [71], our study aims to predict the onset of CVD within one year following the initial diagnosis of type 2 diabetes, utilizing ICD-9 and ICD-10 codes to identify CVD events such as coronary heart disease (CHD), congestive heart failure (CHF), myocardial infarction (MI), or stroke. The patients are considered positive if they develop a CVD complication within a year and negative otherwise.

The dataset statistics are summarized in Table 1. Our tasks also include a clustering analysis, where task-guided clusters for clinical concepts and patient visits are generated to discover disease subtypes. We also visualize the distribution of disease subtypes and patient subgroups using t-SNE. Their significant relationships are captured and analyzed through case studies in Sec. 4.5.

Metrics. To deal with the imbalanced labels within the EHR data, we follow [71] to adopt Accuracy, AUROC, AUPR, and Macro F1 score as the metrics of downstream clinical tasks. To measure the quality of clusters TACCO generates, we utilize Silhouette Coefficient [56] for an unsupervised evaluation.

Table 1: Statistics of MIMIC-III and CRADLE datasets. For MIMIC-III, there are only 12535 are processed with labels.

Stats	MIMIC-III	CRADLE
# of diagnosis	846	7915
# of procedure	2032	4321
# of service	20	—
# of prescription	4525	489
# of nodes	7423	12725
# of hyperedges	36875/12353	36611

Baselines. In terms of the two disease prediction tasks, we compare our TACCO with the following baselines:

(1) Traditional ML. The traditional ML baselines include models that process EHR data without utilizing graph structures. Specifically, we consider:

- **LR [46]**: Logistic Regression, a linear model for classification that estimates probabilities using a logistic function.
- **SVM [14]**: Support Vector Machine, an algorithm that finds the hyperplane that best separates different classes.
- **MLP [48]**: Multilayer Perceptron, an artificial neural network that consists of at least three layers of nodes.
- **XGBoost [4]**: an implementation of gradient-boosted decision trees designed for speed and performance.

(2) Graph-based Models. We also compare advanced GNN models for further analysis of EHR data. In this category, we evaluate:

- **GCT [11]**: Graph Convolutional Transformer, an improved GNN that combines mechanisms of convolution with attention.
- **GAT [67]**: Graph Attention Network, an improved GNN that uses attentions mechanism to weight the significance of nodes.

(3) Hypergraph-based Models. Hypergraph modeling has become the state-of-the-art approach in EHR analysis. We select several representative methods including:

- **HGNN [17]**: Hypergraph Neural Network, a hypergraph model that learns the hidden representation via high-order structures.
- **HyperGCN [73]**: Hypergraph Convolutional Network, a model that uses convolution for semi-supervised learning based on higher-order graph modeling.
- **HCHA [2]**: Hypergraph Convolution and Hypergraph Attention, a hypergraph model that integrates both convolution and attention mechanisms.
- **HypEHR [71]**: A hypergraph transformer based on AllSetTransformer [7] that predicts disease risks on EHR data.

We also compare several clustering methods with the default DEC in TACCO in terms of the clustering quality:

- **HDBSCAN [45]**: Hierarchical Density-Based Spatial Clustering of Applications with Noise, a clustering algorithm that identifies clusters of varying densities by building a hierarchy of clusters using a density-based approach.
- **IDEC [23]**: Improved Deep Embedded Clustering, which improves DEC by introducing an additional autoencoder for embedding reconstructions.
- **DCC [77]**: Deep Constraint Clustering, which explores different constraints that benefit the clustering performance. In our implementation, we choose a global size constraint that assumes each cluster should be approximately the same size.

Table 2: Performance of clinical outcome predictions on MIMIC-III and CRADLE compared with different baselines. The presented results are averages of the best metrics from 10 individual runs of the models. **Bold** numbers indicate the best results, and underlined numbers indicate the second-best results in each category. TACCO uses DEC as the default clustering module; we also discuss the performance of DCC and IDEC in our framework. We use * to indicate statistically significant results ($p < 0.05$).

Model	MIMIC-III				CRADLE			
	Accuracy	AUROC	AUPR	Marcro-F1	Accuracy	AUROC	AUPR	Marcro-F1
LR	68.66 ± 0.24	64.62 ± 0.25	45.63 ± 0.32	13.74 ± 0.40	76.22 ± 0.30	57.22 ± 0.28	25.99 ± 0.28	42.18 ± 0.35
SVM	72.02 ± 0.12	55.10 ± 0.14	34.19 ± 0.17	32.35 ± 0.21	68.57 ± 0.13	53.57 ± 0.11	23.50 ± 0.15	52.34 ± 0.22
MLP	70.73 ± 0.24	71.20 ± 0.22	52.14 ± 0.23	16.39 ± 0.30	77.02 ± 0.17	63.89 ± 0.18	33.28 ± 0.23	45.16 ± 0.26
XGBoost	76.40 ± 0.42	67.68 ± 0.35	47.26 ± 0.34	36.14 ± 0.59	79.28 ± 0.26	68.65 ± 0.58	39.12 ± 0.39	56.57 ± 0.65
GCT	76.58 ± 0.23	78.62 ± 0.21	63.99 ± 0.27	35.48 ± 0.34	77.26 ± 0.22	67.08 ± 0.19	35.90 ± 0.20	56.66 ± 0.25
GAT	76.75 ± 0.26	78.89 ± 0.12	66.22 ± 0.29	34.88 ± 0.33	77.82 ± 0.20	66.55 ± 0.27	36.06 ± 0.18	56.43 ± 0.26
HGNN	77.93 ± 0.41	80.12 ± 0.30	68.38 ± 0.24	40.04 ± 0.35	76.77 ± 0.24	67.21 ± 0.25	37.93 ± 0.18	58.05 ± 0.23
HyperGCN	78.01 ± 0.23	80.34 ± 0.15	67.68 ± 0.16	39.29 ± 0.20	78.18 ± 0.11	67.83 ± 0.18	38.28 ± 0.19	60.24 ± 0.21
HCHA	78.07 ± 0.28	80.42 ± 0.17	68.56 ± 0.15	37.78 ± 0.22	78.60 ± 0.15	68.05 ± 0.17	39.23 ± 0.13	59.26 ± 0.21
HypEHR	79.07 ± 0.31	82.19 ± 0.13	71.08 ± 0.17	41.51 ± 0.25	79.76 ± 0.18	70.07 ± 0.13	40.92 ± 0.12	61.23 ± 0.18
TACCO								
w/ DCC	79.56 ± 0.25*	82.47 ± 0.15*	71.37 ± 0.29*	40.45 ± 0.62*	80.24 ± 0.30*	72.67 ± 0.23*	45.48 ± 0.44*	61.17 ± 0.54*
w/ IDEC	80.75 ± 0.09*	84.08 ± 0.22*	73.63 ± 0.28*	**45.59 ± 0.67***	80.06 ± 0.23*	73.48 ± 0.26*	48.09 ± 0.45*	64.55 ± 0.15*
w/ DEC	**81.02 ± 0.26***	**84.31 ± 0.15***	**73.67 ± 0.25***	45.53 ± 0.17*	**81.00 ± 0.32***	**74.23 ± 0.36***	**49.08 ± 0.43***	**64.64 ± 0.57***

Table 3: Ablation studies on MIMIC-III and CRADLE. The presented results are averages of the best metrics from 10 individual runs of the models. All models are based on the same backbone model hypergraph transformer. *text* refers to the use of textual information, *node* means clustering on nodes, *edge* means clustering on hyperedges, and *align* represents cluster alignment.

hypergraph w/				MIMIC-III				CRADLE			
text	node	edge	align	Accuracy	AUROC	AUPR	Macro-F1	Accuracy	AUROC	AUPR	Macro-F1
				79.07 ± 0.31	82.19 ± 0.13	71.08 ± 0.17	41.51 ± 0.25	79.76 ± 0.18	70.07 ± 0.13	40.92 ± 0.12	61.23 ± 0.18
✓				80.69 ± 0.28	83.71 ± 0.38	72.96 ± 0.30	45.50 ± 0.41	80.30 ± 0.41	73.47 ± 0.22	47.66 ± 0.28	64.39 ± 0.59
✓	✓			80.66 ± 0.06	83.97 ± 0.08	72.96 ± 0.12	45.20 ± 0.54	80.55 ± 0.22	73.60 ± 0.19	47.67 ± 0.49	63.94 ± 0.69
✓		✓		80.73 ± 0.05	84.04 ± 0.08	73.10 ± 0.12	45.01 ± 0.33	80.55 ± 0.20	73.82 ± 0.19	47.75 ± 0.32	64.45 ± 0.59
✓	✓	✓		80.77 ± 0.08	84.10 ± 0.06	73.52 ± 0.16	45.21 ± 0.47	80.73 ± 0.26	73.73 ± 0.25	48.19 ± 0.58	64.27 ± 0.68
✓	✓	✓	✓	**81.02 ± 0.26**	**84.31 ± 0.15**	**73.67 ± 0.25**	**45.53 ± 0.17**	**81.00 ± 0.32**	**74.23 ± 0.36**	**49.08 ± 0.43**	**64.64 ± 0.57**

Implementation Details. All the experiments are run on one NVIDIA H100 Tensor Core GPU. We implement most of our experiments using PyTorch[1]. For traditional ML baselines, we implement the code with scikit-learn[2]. For the other baselines, we follow the original settings suggested by the authors to train them. For our TACCO, we use Adam optimizers for all modules and tune the learning rates in $\{5e-4, 1e-3, 5e-3, 1e-2\}$. The frozen SapBERT encoder is implemented from Hugging Face[3]. To ensure a fair comparison, we strictly adhere to the hyperparameter settings of the backbone hypergraph [71], with $L = 3$ layers, $h = 4$ heads, and $d = 48$ as the hidden feature dimension in each hypergraph layer. For our best model in Table 2, we set $\alpha = 10$, $\beta = 0.1$, $K = 5$, and $m = 1$. The datasets are split into train/validation/test sets by 7:1:2. For the training, we warm up the hypergraph transformer alone with 100 epochs. For further insights into our hyperparameter selection, please refer to Sec. 4.3.

[1]https://pytorch.org/
[2]https://scikit-learn.org/
[3]https://huggingface.co/cambridgeltl/SapBERT-from-PubMedBERT-fulltext

4.2 Overall Performance Comparison

We perform two downstream tasks, *i.e.*, the phenotype classification on MIMIC-III and the CVD risk prediction on CRADLE, to evaluate the predictive performance of our proposed TACCO. The comparison with other baselines we discuss in Sec. 4.1 is presented in Table 2. It is evident that TACCO consistently outperforms all baselines on four metrics across both datasets, with DEC performing the best as the clustering module. We can observe a significant improvement of 31.25% on top of the traditional ML models, which often suffer from the sparse nature of large-scale EHR networks. TACCO also gains an improvement of 12.36% over the two graph-based models. These notable improvements highlight the effectiveness of modeling higher-order relations within complex EHR data.

Hypergraph-based models such as HCHA demonstrate better results compared to the traditional ones as they model EHR data beyond pairwise relations and learn robust representations. Compared to these advanced approaches, TACCO still maintains its lead by an average improvement of 7.89% across two datasets. Compared with the suboptimal HypEHR, our model raises the overall performance by 5.26%. These results further validate the effectiveness of our task-guided co-clustering in terms of improving downstream predicting.

4.3 In-depth Model Analysis

Model Ablation. We conduct detailed ablation studies to better understand the efficacy of different components in TACCO. As shown in Table 3, we observe a significant improvement in all four metrics when semantic information is involved compared to the backbone model. This suggests that clinical concept semantics are vital in providing more information for structured modeling. Additionally, both node and hyperedge clustering techniques improve the model's predictive power, with hyperedge clustering bringing a slightly better gain. The model's performance is further improved by equipping it with cross-domain alignment loss, which minimizes the distance between similar cluster centroids to generate more consistent results. Optimal performance is achieved when all the clustering and aligning components are integrated. This highlights the collective contribution of all four proposed components towards the enhanced model performance, with better interpretability as an additional benefit as shown in Sec. 4.5.

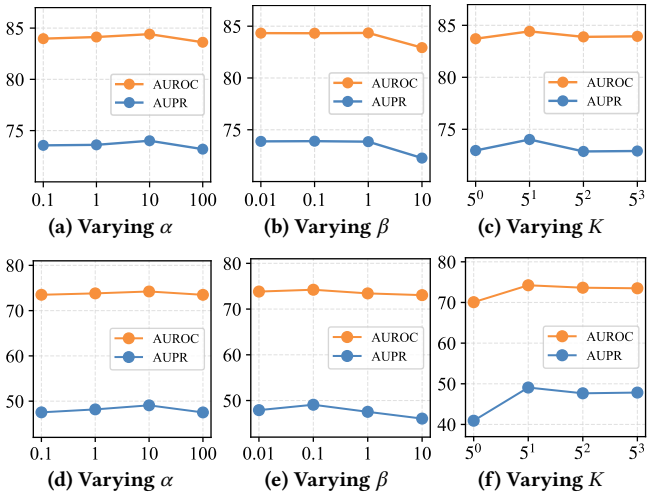


Figure 3: Effect of hyperparameters of TACCO. (a), (b), and (c) are on MIMIC-III dataset. (d), (e), and (f) are on CRADLE dataset.

Hyperparameter Study. We analyze the impact of important hyperparameters in our TACCO model, which includes the loss weight parameters α and β in Eq. (15), and the number of clusters K. We vary the contribution of different terms in Eq. (15) by adjusting their respective loss weights, as their numerical values are at different scales. The results are displayed in Figure 3. Our findings indicate that the best performance is attained when α is set at 10 and β is set at 0.1. We also change the number of clusters K and select the value that yielded the best performance for the model.

Efficiency Study. Efficiency experiments on our downstream tasks reveal that TACCO achieves the best trade-off of efficiency and performance. As shown in Figure 4, while TACCO's training time is slightly longer than HypEHR's by less than a minute, it achieves the highest AUROC scores in both MIMIC-III and CRADLE datasets. TACCO also easily outperforms other graph-based and hypergraph-based models. Especially for GCT and GAT, their extended training times can be attributed to their lack of hypergraph architecture, which results in significant computational overhead due to the need to flatten all hyperedges.

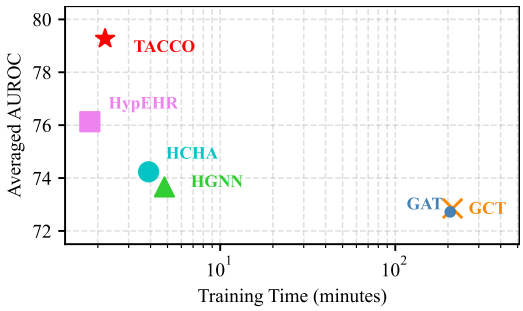


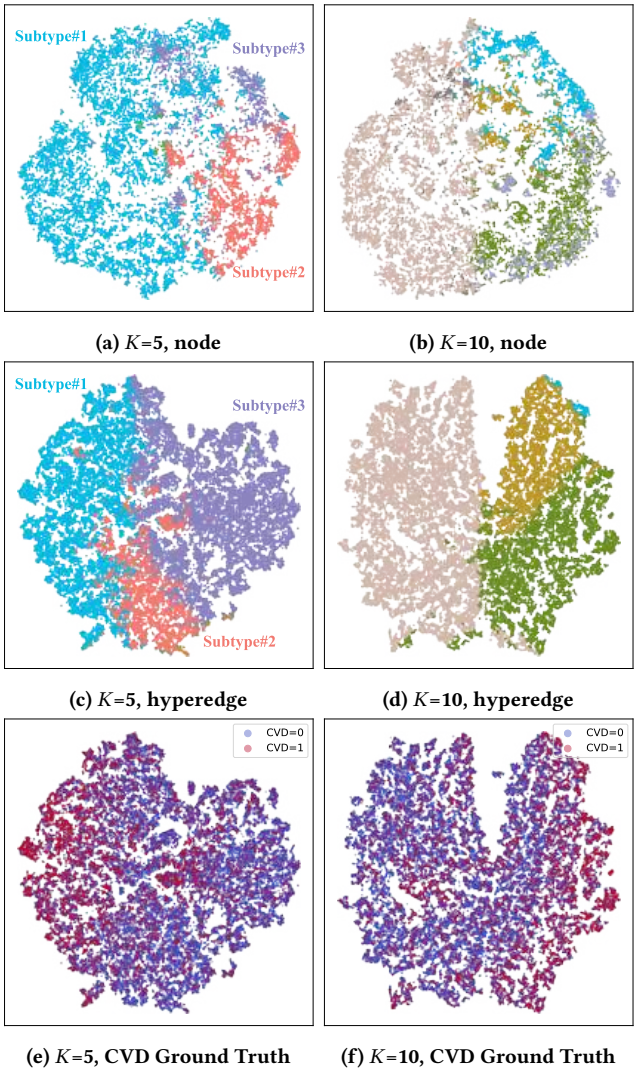
Figure 4: Efficiency and performance trade-off. The y-axis represents the model performance, measured by the averaged AUROC scores from the MIMIC-III and CRADLE datasets. The x-axis shows the training time using a logarithmic scale, also averaged from the two datasets. Some traditional ML methods are not included due to the scikit-learn package's inability to utilize GPU acceleration.

Table 4: Clustering quality comparison. SC denotes the Silhouette Coefficient. $\mathcal{V}$ and $\mathcal{E}$ denote the metrics that are calculated in node clustering and hyperedge clustering, respectively. Each method is executed with random seeds from 1 to 5 to ensure the stability of learned embeddings. The results are averaged.

Model	MIMIC-III		CRADLE	
	$SC_{\mathcal{V}}$	$SC_{\mathcal{E}}$	$SC_{\mathcal{V}}$	$SC_{\mathcal{E}}$
HDBSCAN	0.1270	0.4186	0.0399	0.0461
K-means ($K = 5$)	0.4131	0.8811	0.1715	0.2313
K-means ($K = 10$)	0.2852	0.6350	0.1136	0.1531
TACCO ($K = 5$)				
w/ DCC	0.0959	0.3476	0.1390	0.1441
w/ IDEC	0.4993	0.7639	0.2197	0.2226
w/ DEC	0.4999	**0.8881**	**0.3033**	**0.5083**
TACCO ($K = 10$)				
w/ DCC	0.1209	0.2834	0.1717	0.0879
w/ IDEC	0.4877	0.7648	0.2044	0.2204
w/ DEC	**0.5849**	0.7888	0.2242	0.3727

4.4 Clustering Analysis

To investigate the quality of our co-clustering assignment, we run the model on the CRADLE dataset with $K = 5$ and $K = 10$, respectively. The high-dimensional embeddings of nodes and hyperedges are then projected into a shared 2D space via t-SNE [64]. The visualizations are presented in Figure 5, with (a), (c), and (e) from the first model with $K = 5$, as well as (b), (d), and (f) from the second model with $K = 10$. The clustering outcomes reveal that, regardless of whether 5 or 10 clusters are targeted, the results consistently form 3-4 major distinct clusters. This consistency highlights the model's stable ability to capture the major patterns underlying the interactions of clinical concepts and patient visits within EHR data. Notably, the clusters on hyperedges and ground-truth CVD labels on hyperedges demonstrate a certain level of concordance. When $K = 5$, the cyan cluster in panel (c) (denoted as Subtype#1) largely coincides with the population diagnosed with CVD. This indicates that our self-supervised clustering on hyperedges is significantly guided by signals from the specific downstream disease prediction task. Such guiding signals are further propagated to the node



(a) K**=5, node** **(b)** K**=10, node**

(c) K**=5, hyperedge** **(d)** K**=10, hyperedge**

(e) K**=5, CVD Ground Truth** **(f)** K**=10, CVD Ground Truth**

Figure 5: Visualization of output clustering distribution and CVD ground truth on the CRADLE dataset via t-SNE. (a) and (c) show clusters distribution on nodes and hyperedges, (e) labels CVD on hyperedges embeddings, all from a TACCO with $K = 5$. (b), (d), and (f) are in the same order from a model with $K = 10$.

clusters through our contrastive alignment objective in Eq. (13), as evidenced by the similarity in color distributions observed between panels (a) and (c), as well as (b) and (d).

From a more quantitative perspective, we provide a clustering comparison on MIMIC-III and CRADLE datasets in Table 4 with HDBSCAN [45] and K-means applied post-training at the final epoch of the hypergraph model. We also discuss the how different deep clustering methods (IDEC [23] and DCC [77]) perform in our framework. We take the Silhouette Coefficient as our metric, which offers a robust measure of cluster purity and separation. The Silhouette Coefficient is defined as $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$, where $a(i)$ is the average intra-cluster distance, *i.e.*, cohesion, and $b(i)$ is the average distance to the nearest cluster that i is not part of, *i.e.*, separation. We can observe that TACCO using the default DEC (Sec. 3.3) generates clusters with the highest quality in both nodes and hyperedges. The other deep clustering methods show suboptimal results potentially due to additional learning targets they introduce. The Silhouette Coefficient of TACCO averaged over 5 runs is 0.5213, which is 77.93% higher than the average of HDBSCAN and K-means. These quantitative results further demonstrate our method's superior capability in discerning and grouping clinical concepts and patient visits within the EHR datasets.

4.5 Case Studies

To demonstrate the aligned clusters of clinical concepts and patient visits generated by TACCO, as well as their practical values in clinical settings, we illustrate actual clinical concepts and patient visits from our clusters of nodes and hyperedges. Specifically, we select 3 clusters from each of the two domains corresponding to the cyan, pink, and purple colors in Figure 5 (a) and (c). From the node clusters, we select the top 15 clinical concept candidates based on the clustering assignment probability $\boldsymbol{Q}_{\mathcal{V}}$ from Eq. (8). Those are presented in the left panel of Table 5. Notably, we mark the subtype-indicative clinical concepts with color after consulting a model-blinded clinical expert. Similarly, from the hyperedge clusters, we select 3 patient visits based on the clustering assignment probability $\boldsymbol{Q}_{\mathcal{E}}$ from Eq. (8), and illustrate each of them with the top 5 clinical concepts determined by the highest attention weights within the hypergraph transformer. They are shown in the right panel of Table 5. We also color the subtype-indicative clinical concepts as suggested by the clinical expert. As shown in Table 5, we can observe a notable overlap in clinical concepts between the clinical concept and patient visit clusters across all three subtypes, which validates the effectiveness of our cluster alignment in producing consistent clusters.

For **Subtype#1**, after consulting with the clinical expert, we interpret this subtype as indicative of a heightened risk for CVD. In general, the presence of medications like *Imidazoline Derivatives* and conditions such as *Severe Hyperglycaemia* and *End-stage Renal Disease* signals a significant cardiovascular risk [18, 25, 31]. Diagnostic practices, including *Preoperative Cardiovascular Evaluations* and *Electrocardiograms*, further enhance the probability of cardiovascular disease presence in this subtype [3]. We identify that 34 out of the top 50 patients were confirmed to have CVD a year after their visits, which is consistent with the distribution in Figure 5 where dots of Subtype#1 in Figure 5 (c) largely align with dots with **CVD labels** in Figure 5 (e). The records of Patient#681 and #716, who are diagnosed with CVD, also align closely with the representative clinical concepts within Subtype#1.

Subtype#2 emphasizes metabolic and thyroid disorders. It delineates conditions closely associated with *Type 2 Diabetes Mellitus*. Within the top 50 patients in the cluster aligning with Subtype#2, 20 individuals are diagnosed with CVD, which indicates a comparatively moderate association with cardiovascular risk. Unlike Subtype#1, Subtype#2 is not marked by the use of potent pharmacological interventions and the presence of severe disease states. This demonstrates that TACCO is capable of capturing nuanced disease subtypes, thereby facilitating more targeted monitoring and intervention strategies in clinical practice for these subgroups. This

Table 5: Case studies of disease subtypes. The same colors are used to indicate the correspondence with clusters in Figure 5 ($K = 5$). The colored clinical concepts represent the subtype-indicative ones as suggested by a clinical expert. Best viewed in color.

Clinical Concept Clusters	Patient Visit Clusters
Subtype#1: High CVD risk from potent medications, severe conditions, and essential diagnostics. **{** Imidazoline Derivatives; Opioid Intoxication; Sympathomimetics; Antihidrotics; End-stage Renal Disease; Electrocardiogram; Vancomycin; Preoperative Cardiovascular Examination; Myelodysplastic Disease; Ethanol Causing Toxic Effect; Chronic Kidney Disease Stage 5; Severe Hyperglycaemia; Carbon Disulfide Causing Toxic Effect; Adrenergic and Dopaminergic Agents; Influenza A Virus Subtype H5N1 **}**	**CVD rate@50: 68.00%** **Patient#681 {** Adrenergic and Dopaminergic Agents; Sympathomimetics; End-stage Renal Disease; Antihidrotics; Phenylpiperidine Derivatives **} CVD = 1** **Patient#716 {** Electrocardiogram; Preoperative Cardiovascular Examination; End-stage Renal Disease; Atherosclerosis Renal Artery; Cardiovascular Stress Test**} CVD = 1** **Patient#2336 {** ACE Inhibitors and Calcium Channel Blockers; Type II Diabetes Mellitus w/o Complication; Antihidrotics; Anesthetics for Topical Use; Long-term Drug Therapy **} CVD = 0**
Subtype#2: Type 2 diabetes without complications, including metabolic and thyroid disorders. **{** Disorder of Carbohydrate Metabolism; Blood Tests; Thyrotoxicosis; Hypothyroidism; Diabetes Mellitus; Hemoglobin Glycosylated (A1c); Benign Tumor of Descending; Screening for Osteoporosis; Bacterial Disease Screening; Type II Diabetes Mellitus w/o Complication; Thyroid Stimulating Hormone; Tuberculosis screening; Disorder of Thyroid Gland; Urinary System Symptoms; Mononeuropathy due to Type 2 Diabetes Mellitus **}**	**CVD rate@50: 40.00%** **Patient#13821 {** Hypothyroidism; Diabetes Mellitus; Blood Tests; Thyroxine (Free); Hemoglobin Glycosylated (A1c) **} CVD = 1** **Patient#34892 {**Diabetes Mellitus; Hyperlipidemia; Long-term Drug Therapy; Benign Essential Hypertension; Disorder of Transplanted Kidney**} CVD = 0** **Patient#2767 {** Hyperlipidemia; Type II Diabetes Mellitus w/o Complication; Hemoglobin Glycosylated (A1c); Blood Tests; Screening for Osteoporosis **} CVD = 0**
Subtype#3: Orthopedic and neurologic injuries affecting multiple body regions. **{** Arthropathy of Multiple Joints; Office or Other Outpatient Visit; Tear of Lateral Meniscus of Knee; Sprain of Knee; Postoperative Follow-Up Visit; Brachial Plexus Injury; Corticosteroids; Pre-surgery Evaluation; Cast; Closed Fracture Lumbar Vertebra; Traumatic Arthropathy of Shoulder; Closed Traumatic Dislocation of Elbow Joint; Fracture of First Lumbar Vertebra; Flatback Syndrome; Cancer (Mesothelioma) **}**	**CVD rate@50: 18.00%** **Patient#16379 {** Pain in Right Knee; Tear of Lateral Meniscus of Knee; Pre-surgery Evaluation; Postoperative Follow-Up Visit; Corticosteroids **} CVD = 0** **Patient#7276 {** Corticosteroids; Knee Pain; Pre-surgery Evaluation; Meniscectomy; Tear of Lateral Meniscus of Knee **} CVD = 0** **Patient#12488 {** Shoulder Joint Pain; Office Or Other Outpatient Visit; Accidental Physical Contact; Musculoskeletal Symptom; Pain in Limb **} CVD = 0**

advantage could potentially mitigate patients' risk of progressing to cardiovascular diseases.

Subtype#3 focuses on orthopedic and neurologic injuries affecting multiple body regions, such as *Sprain of Knee* and *Closed Fracture Lumbar Vertebra*, with no direct ties to CVD, as evidenced by the ground truth distribution in Figure 5(e). We also cannot observe obvious patients with CVD-related patterns. Instead, most of them share similar clinical records in musculoskeletal disorders, which are likely negatively correlated with CVD risk. This highlights that our deep co-clustering benefits from the guidance of specific disease predictions, and thereby efficiently identifies subtypes that are positively, weakly, or negatively correlated with a particular disease, such as CVD. These insights are advantageous for medical professionals in conducting precise clinical stratification and management.

5 CONCLUSION AND DISCUSSION

In this work, we introduce TACCO, a novel framework that jointly clusters clinical concepts and patient visits in EHR data. Specifically, we encode semantic information within the clinical concepts into a hypergraph transformer. We design a deep self-supervised co-clustering module that jointly learns a soft clustering assignment for both nodes and hyperedges. The learned clusters are then aligned through a contrastive learning objective for capturing the consistent patterns between clinical concepts and patient visits within the EHR data. Our comprehensive experiments demonstrate the superior performance of TACCO, which is 5.26% higher than the vanilla hypergraph backbone model and 31.25% higher than other ML baselines. Notably, TACCO is capable of discerning insightful disease subtypes related to specific diseases at different levels, enabling more targeted clinical interventions.

Currently, TACCO is in a stage of secondary data analysis. It has been tested on data from both academic benchmark MIMIC-III, and Project CRADLE, which is an actual application within the Emory Hospital Systems that provides substantial support to medical staff and researchers in the greater Atlanta and Georgia areas. Specifically, this work contributes to Project CRADLE's ongoing efforts in cardiovascular and diabetes disease management over 48 thousand patients. In a significant expansion, the method is also being adapted for use in the National Institutes of Health's All of Us[4] research program. This deployment aims to harness the diverse medical records of over 38 thousand participants across the United States. Looking forward, we aim to broaden TACCO's applicational scope by extending the framework to accommodate large-scale, heterogeneous datasets that contain more modalities.

ACKNOWLEDGMENTS

Research reported in this publication was mainly supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under Award Number K25DK135913. The research also receives partial support from the National Science Foundation under Award Number IIS-2145411, IIS-2312502, and IIS-2319449. Opinions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Government. The research has also benefited from the Microsoft Accelerating Foundation Models Research (AFMR) grant program.

[4]https://allofus.nih.gov/

REFERENCES

[1] Emma Ahlqvist, Petter Storm, Annemari Käräjämäki, Mats Martinell, Mozhgan Dorkhan, Annelie Carlsson, Petter Vikman, Rashmi B Prasad, Dina Mansour Aly, Peter Almgren, et al. 2018. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The lancet Diabetes & endocrinology* 6, 5 (2018), 361–369.

[2] Song Bai, Feihu Zhang, and Philip HS Torr. 2021. Hypergraph convolution and hypergraph attention. *Pattern Recognition* 110 (2021), 107637.

[3] R Sacha Bhatia and Paul Dorian. 2018. Screening for cardiovascular disease risk with electrocardiography. *JAMA Internal Medicine* 178, 9 (2018), 1163–1164.

[4] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[5] Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15750–15758.

[6] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. 2016. Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the 2016 SIAM international conference on data mining*. SIAM, 432–440.

[7] Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. 2021. You are allset: A multiset function framework for hypergraph neural networks. *arXiv preprint arXiv:2106.13264* (2021).

[8] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*. PMLR, 301–318.

[9] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. GRAM: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 787–795.

[10] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems* 29 (2016).

[11] Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. 2020. Learning the graphical structure of electronic health records with graph convolutional transformer. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 606–613.

[12] Youngduck Choi, Chill Yi-I Chiu, and David Sontag. 2016. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings* 2016 (2016), 41.

[13] Diana Hedevang Christensen, Sia K Nicolaisen, Emma Ahlqvist, Jacob V Stidsen, Jens Steen Nielsen, Kurt Hojlund, Michael H Olsen, Sonia García-Calzón, Charlotte Ling, Jørgen Rungby, et al. 2022. Type 2 diabetes classification: a data-driven cluster study of the Danish Centre for Strategic Research in Type 2 Diabetes (DD2) cohort. *BMJ Open Diabetes Research and Care* 10, 2 (2022), e002731.

[14] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20 (1995), 273–297.

[15] Zhijie Deng, Yucen Luo, and Jun Zhu. 2019. Cluster alignment with a teacher for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9944–9953.

[16] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, Vol. 96. 226–231.

[17] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 3558–3565.

[18] Robert N Foley, Patrick S Parfrey, and Mark J Sarnak. 1998. Clinical epidemiology of cardiovascular disease in chronic renal disease. *American Journal of Kidney Diseases* 32, 5 (1998), S112–S119.

[19] Tianfan Fu, Trong Nghia Hoang, Cao Xiao, and Jimeng Sun. 2019. Ddl: Deep dictionary learning for predictive phenotyping. In *IJCAI: proceedings of the conference*, Vol. 2019. NIH Public Access, 5857.

[20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems* 33 (2020), 21271–21284.

[21] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2019. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems* 34, 6 (2019), 14–23.

[22] Unjali P Gujral, R Pradeepa, Mary Beth Weber, KM Venkat Narayan, and Vishwanathan Mohan. 2013. Type 2 diabetes in South Asians: similarities and differences with white Caucasian and other populations. *Annals of the New York Academy of Sciences* 1281, 1 (2013), 51–63.

[23] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. 2017. Improved deep embedded clustering with local structure preservation.. In *Ijcai*, Vol. 17. 1753–1759.

[24] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data* 6, 1 (2019), 96.

[25] GA Head and DN Mayorov. 2006. Imidazoline receptors, novel agents and therapeutic potential. *Cardiovascular & Hematological Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Cardiovascular & Hematological Agents)* 4, 1 (2006), 17–32.

[26] Anahita Hosseini, Ting Chen, Wenjun Wu, Yizhou Sun, and Majid Sarrafzadeh. 2018. Heteromed: Heterogeneous information network for medical diagnosis. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 763–772.

[27] Peihao Huang, Yan Huang, Wei Wang, and Liang Wang. 2014. Deep embedding network for clustering. In *2014 22nd International conference on pattern recognition*. IEEE, 1532–1537.

[28] Yufang Huang, Yifan Liu, Peter AD Steel, Kelly M Axsom, John R Lee, Sri Lekha Tummalapalli, Fei Wang, Jyotishman Pathak, Lakshminarayanan Subramanian, and Yiye Zhang. 2021. Deep significance clustering: a novel approach for identifying risk-stratified and predictive patient subgroups. *Journal of the American Medical Informatics Association* 28, 12 (2021), 2641–2653.

[29] Alistair Johnson, Tom Pollard, and Roger Mark. 2016. MIMIC-III clinical database (version 1.4). *PhysioNet* 10, C2XW26 (2016), 2.

[30] Stephen C Johnson. 1967. Hierarchical clustering schemes. *Psychometrika* 32, 3 (1967), 241–254.

[31] Markku Laakso. 1999. Hyperglycemia and cardiovascular disease in type 2 diabetes. *Diabetes* 48, 5 (1999), 937–942.

[32] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. 2021. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 8547–8555.

[33] Chuanren Liu, Fei Wang, Jianying Hu, and Hui Xiong. 2015. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 705–714.

[34] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. 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*. 4228–4238.

[35] Jingshu Liu, Zachariah Zhang, and Narges Razavian. 2018. Deep ehr: Chronic disease prediction using medical notes. In *Machine Learning for Healthcare Conference*. PMLR, 440–464.

[36] Siqi Liu, Sidong Liu, Weidong Cai, Sonia Pujol, Ron Kikinis, and Dagan Feng. 2014. Early diagnosis of Alzheimer's disease with deep learning. In *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*. IEEE, 1015–1018.

[37] Yiyi Liu, Quanquan Gu, Jack P Hou, Jiawei Han, and Jian Ma. 2014. A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. *BMC bioinformatics* 15, 1 (2014), 1–11.

[38] Zheng Liu, Xiaohan Li, Hao Peng, Lifang He, and S Yu Philip. 2020. Heterogeneous similarity graph neural network on electronic health records. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 1196–1205.

[39] Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE transactions on information theory* 28, 2 (1982), 129–137.

[40] Haohui Lu and Shahadat Uddin. 2021. A weighted patient network-based framework for predicting chronic diseases using graph neural networks. *Scientific reports* 11, 1 (2021), 22607.

[41] Ronan L'heveder and Tim Nolan. 2013. International diabetes federation. *Diabetes research and clinical practice* 101, 3 (2013), 349–351.

[42] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 1903–1911.

[43] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 743–752.

[44] Enrico Maiorino and Joseph Loscalzo. 2023. Phenomics and Robust Multiomics Data for Cardiovascular Disease Subtyping. *Arteriosclerosis, Thrombosis, and Vascular Biology* 43, 7 (2023), 1111–1123.

[45] Leland McInnes and John Healy. 2017. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 33–42.

[46] Scott Menard. 2002. *Applied logistic regression analysis*. Number 106. Sage.

[47] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[48] Parisa Naraei, Abdolreza Abhari, and Alireza Sadeghian. 2016. Application of multilayer perceptron neural networks and support vector machines in classification of healthcare data. In *2016 Future technologies conference (FTC)*. IEEE, 848–852.

[49] Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. 2016. Deepr: a convolutional net for medical records. *IEEE journal of biomedical and health informatics* 21, 1 (2016), 22–30.

[50] N Nidheesh, KA Abdul Nazeer, and PM Ameer. 2020. A Hierarchical Clustering algorithm based on Silhouette Index for cancer subtype discovery from genomic data. *Neural Computing and Applications* 32 (2020), 11459–11476.

[51] Juan G Diaz Ochoa and Faizan E Mustafa. 2022. Graph neural network modelling as a potentially effective method for predicting and analyzing procedures based on patients' diagnoses. *Artificial Intelligence in Medicine* 131 (2022), 102359.

[52] Xueqin Pang, Christopher B Forrest, Félice Lê-Scherban, and Aaron J Masino. 2021. Prediction of early childhood obesity with machine learning and electronic health record data. *International journal of medical informatics* 150 (2021), 104454.

[53] Hemang M Parikh, Cassandra L Remedios, Christiane S Hampe, Ashok Balasubramanyam, Susan P Fisher-Hoch, Ye Ji Choi, Sanjeet Patel, Joseph B McCormick, Maria J Redondo, and Jeffrey P Krischer. 2023. Data Mining Framework for Discovering and Clustering Phenotypes of Atypical Diabetes. *The Journal of Clinical Endocrinology & Metabolism* 108, 4 (2023), 834–846.

[54] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 701–710.

[55] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2016. Deepcare: A deep dynamic memory model for predictive medicine. In *Advances in Knowledge Discovery and Data Mining: 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part II 20*. Springer, 30–41.

[56] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.

[57] Setu Shah and Xiao Luo. 2017. Extracting modifiable risk factors from narrative preventive healthcare guidelines for EHR integration. In *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 514–519.

[58] Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019. Pre-training of graph augmented transformers for medication recommendation. In *International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence.

[59] Jess J Shen, Phil Hyoun Lee, Jeanette JA Holden, and Hagit Shatkay. 2007. Using cluster ensemble and validation to identify subtypes of pervasive developmental disorders. In *AMIA Annual Symposium Proceedings*, Vol. 2007. American Medical Informatics Association, 666.

[60] Ronglai Shen, Adam B Olshen, and Marc Ladanyi. 2009. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25, 22 (2009), 2906–2912.

[61] Chang Su, Robert Aseltine, Riddhi Doshi, Kun Chen, Steven C Rogers, and Fei Wang. 2020. Machine learning for suicide risk prediction in children and adolescents with electronic health records. *Translational psychiatry* 10, 1 (2020), 413.

[62] Chenxi Sun, Hongna Dui, and Hongyan Li. 2021. Interpretable time-aware and co-occurrence-aware network for medical prediction. *BMC medical informatics and decision making* 21, 1 (2021), 1–12.

[63] Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine* 3, 1 (2020), 17.

[64] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[65] Jithin Sam Varghese, Joyce C Ho, Ranjit Mohan Anjana, Rajendra Pradeepa, Shivani A Patel, Saravanan Jebarani, Viswanathan Baskar, KM Venkat Narayan, and Viswanathan Mohan. 2021. Profiles of Intraday glucose in type 2 diabetes and their association with complications: an analysis of continuous glucose monitoring data. *Diabetes technology & therapeutics* 23, 8 (2021), 555–564.

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[67] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[68] Richard C Wang and Zhixiang Wang. 2023. Precision medicine: Disease subtyping and tailored treatment. *Cancers* 15, 15 (2023), 3837.

[69] Shuang Wang, Dong Zhao, Chi Zhang, Yuwei Guo, Qi Zang, Yu Gu, Yi Li, and Licheng Jiao. 2022. Cluster Alignment With Target Knowledge Mining for Unsupervised Domain Adaptation Semantic Segmentation. *IEEE Transactions on Image Processing* 31 (2022), 7403–7418.

[70] Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*. PMLR, 478–487.

[71] Ran Xu, Mohammed K Ali, Joyce C Ho, and Carl Yang. 2023. Hypergraph Transformers for EHR-based Clinical Predictions. *AMIA Summits on Translational Science Proceedings* 2023 (2023), 582.

[72] Ran Xu, Yue Yu, Chao Zhang, Mohammed K Ali, Joyce C Ho, and Carl Yang. 2022. Counterfactual and factual reasoning over hypergraphs for interpretable clinical predictions on ehr. In *Machine Learning for Health*. PMLR, 259–278.

[73] Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. 2019. Hypergcn: A new method for training graph convolutional networks on hypergraphs. *Advances in neural information processing systems* 32 (2019).

[74] Carl Yang, Liyuan Liu, Mengxiong Liu, Zongyi Wang, Chao Zhang, and Jiawei Han. 2020. Graph clustering with embedding propagation. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 858–867.

[75] Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. 2021. SafeDrug: Dual Molecular Graph Encoders for Recommending Effective and Safe Drug Combinations. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence, IJCAI 2021 (IJCAI International Joint Conference on Artificial Intelligence)*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence, 3735–3741.

[76] Jianwei Yang, Devi Parikh, and Dhruv Batra. 2016. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5147–5156.

[77] Hongjing Zhang, Sugato Basu, and Ian Davidson. 2020. A framework for deep constrained clustering-algorithms and advances. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I*. Springer, 57–72.

[78] Xianli Zhang, Buyue Qian, Yang Li, Shilei Cao, and Ian Davidson. 2021. Context-aware and time-aware attention-based model for disease risk prediction with interpretability. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[79] Yanting Zhang and Hisanori Kiryu. 2022. MODEC: an unsupervised clustering method integrating omics data for identifying cancer subtypes. *Briefings in Bioinformatics* 23, 6 (2022), bbac372.

[80] Jing Zhao, Bowen Zhao, Xiaotong Song, Chujun Lyu, Weizhi Chen, Yi Xiong, and Dong-Qing Wei. 2023. Subtype-DCC: decoupled contrastive clustering method for cancer subtype identification based on multi-omics data. *Briefings in Bioinformatics* 24, 2 (2023), bbad025.