# Graph embedded topic modeling to mine UK Biobank phenotypes and medications with a case study on pain phenotypes

Yuening Wang

**Master Thesis**

School of Computer Science


A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science, Computer Science

# Acknowledgement

I would like to thank and acknowledge all the support from many people and institutions without which I would not have been able to complete this thesis.

First, I would like to give my special thanks to Prof. Yue Li and Prof. Audrey Grant, through supervision I was introduced to bioinformatics. Prof. Yue Li has guided me on the machine learning model development, statistical analysis and important computational skills development. Prof. Audrey Grant has offered me great help to do data analysis and interpret research results from a biomedical perspective. Their everlasting encouragement and guidance have laid the foundation of my confidence to conducting research and equipping myself with important skills such as presenting research work and writing scientific reports. Thank you very much for supporting me through the hard pandemic period and inspiring me at those tough moments.

I would also like to thank my lab mates in the Center of Bioinformatics at McGill University for their genuine friendship and those interesting chats.

I also cannot thank enough to Prof. Luda Diatchenko for those professional and brilliant ideas as well as her knowledge on biological concepts. Many thanks also to Dr. Nikolay Dimitrov, Vivek Verma and other researchers from McGill Genome Center for sharing and processing quality medical data for me.

In the end, I would like to thank my parents and my friends for their consistent care and support.

# Abstract

The UK Biobank, thanks to extensive clinical data and a large sample size, is an unprecedented resource to understand pain and the development of chronic pain as well as impacts of other conditions on its development. However, in efforts to efficiently extract information from UK Biobank and to analyze many phenotypes simultaneously, it presented several challenges including high sparsity and heterogeneous data types. Therefore, to address the challenges, we develop the graph embedded topic model (GETM) as a novel multi-modal topic model which utilizes graph data by incorporating Node2Vec for pre-trained embedding. We successfully learned interpretable topics, which also captured relationships between conditions and medications. We achieved superior performance in experiments using individual-level information of past conditions predicting the status of chronic musculoskeletal pain of baseline. The model was also able to gain promising performance in missing record imputation as well as in medication recommendations. Lastly, our model led to interesting insights on how specific combinations of conditions and medications might affect musculoskeletal pain.

# Résumé

La UK Biobank, grâce à de nombreuses données cliniques et à une grande taille d'échantillon, est une ressource sans précédent pour comprendre la douleur et le développement de la douleur chronique ainsi que les impacts d'autres conditions sur son développement. Cependant, dans les efforts visant â extraire efficacement des informations de UK Biobank et â analyser de nombreux phénotypes simultanément, elle a présenté plusieurs défis, notamment des types de données très clairsemés et hétérogènes. Par conséquent, pour relever les défis, nous développons le modèle de sujet intégré au graphique (GETM) en tant que nouveau modèle de sujet multimodal qui utilise les données du graphique en incorporant Node2Vec pour l'intégration pré-entraînée. Nous avons appris avec succès des sujets interprétables, qui ont également capturé les relations entre les conditions et les médicaments. Nous avons obtenu des performances supérieures dans des expériences utilisant des informations au niveau individuel sur les conditions passées prédisant l'état de la douleur musculo-squelettique chronique de référence. Le modèle a également été en mesure d'obtenir des performances prometteuses dans l'imputation des enregistrements manquants ainsi que dans les recommandations de médicaments. Enfin, notre modèle a conduit â des informations intéressantes sur la façon dont des combinaisons spécifiques de conditions et de médicaments pourraient affecter la douleur musculo-squelettique.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Background

## 1.1 Chronic pain

Chronic pain is the result of dysfunction of the nociceptive circuitry leading to continued perceptions of pain, and was recently recognized by the World Health Organization (WHO) as a disease in its own right, resulting in revisions to the latest (11th) version of the International Classification of Diseases (ICD-11) [28]. High prevalence of chronic pain conditions was observed especially in aging people affecting 50% of older adults (>65y) [36]. It decreases mental and emotional health of people who are suffering chronic pain [34].

Chronic neuropathic pain may be initiated through many different pathologies. Though pain becomes related to the nervous system by perhaps only a few mechanisms, those mechanisms are not yet fully elucidated. It is therefore urgent to better understand chronic pain. One important aspect of the reasons why chronic neuropathic pain is universally recognized as one of the most difficult pain syndromes to treat is that the interrelationship of factors impacting outcomes of chronic pain is complex. [51]. Though several studies have identified the strong association between the presence of chronic pain and mental health conditions, such as depression [19, 20, 62] etc., we need to understand comorbidities of chronic pain to better categorize patients [31]. Besides, the uncertain etiology of chronic

pain often poses a challenge to health care providers who might give escalating doses of medications, which potentially expose patients to unnecessary treatments and associated side effects [61]. Therefore, uncovering the causes of chronic pain is necessary for determining better medication use.

## 1.2 UK Biobank

The UK Biobank [3] is a powerful data resource for understanding the determinants of common life-threatening and disabling diseases. The UK Biobank is a cohort study of 500,000 individuals from across the United Kingdom, aged between 40 and 69 at recruitment [27]. The recruitment started in 2007 and was declared complete in 2010 [3]. Extensive information was collected through questionnaires and physical measurements, as well as by storing biological samples that allow many different types of assays (e.g., genetic, proteomic, metabonomic, or biochemical) [30]. This allows investigations on combined effects of lifestyle, environment, genes, and a wide range of exposures on health outcomes. With the aid of effective computational methods, the UK Biobank promises to yield novel findings made up of multi-factor interactions. Besides, a benefit of a large-scale population allows for models the detection of finer resolution effects [63].

The establishment of the UK Biobank as a resource has led to numerous publications based on statistical analyses of the clinical data [3]. For example, in the pain field, prevalence and associative factors of facial pain were examined by standardized prevalence on UK Biobank estimates and Cox regression with results expressed as relative risk [72]. However, previous methods are still constrained by computing power, and do not take full advantage of diverse information and distill meaningful interactions and combinations from UK Biobank. There are several remaining challenges: the data is very sparse which makes it hard to extract information and correlate different data types. Besides, the size of UK Biobank data makes it difficult for simple statistical methods to incorporate multiple phenotypes simultaneously.

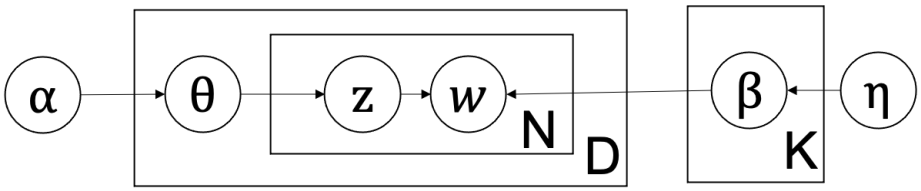


Figure 1.1: **Graphical model for Latent Dirichlet Allocation (LDA).**

Therefore, our goal is to bring novel graph embedded topic model (GETM) to the problem of using medical conditions and medication usage to predict the development of chronic pain.

## 1.3 Topic model

Topic models were originally developed to aid in understanding large collections of text documents. The main importance of topic models is to discover patterns of word use and to connect documents that share similar patterns. This is particularly useful to give keyword labels to documents which enable faster search without reading through the whole corpus [23]. Latent Dirichlet Allocation (LDA) is a classical and original topic model which relies on the commonly used bag-of-words assumption, which ignores the ordering of the words in the document [25]. The basic idea is that documents are represented as mixtures over topics, where each topic is characterized by a distribution over words. It assumes the following generative process for each document in a corpus D (Fig. 1.1):

1. Choose $\boldsymbol{\theta}$ from a Dirichlet distribution parameterized by $\boldsymbol{\alpha}$.

2. For each word $w_{dn}$ of the N words:

   (a) Draw a topic $z_{dn} \sim Multinomial(\boldsymbol{\theta_d})$.

   (b) Choose a word $w_{dn}$ from $p(w_{dn}|z_{dn}, \boldsymbol{\beta})$, a multinomial probability conditioned on the topic $z_{dn}$. $\boldsymbol{\beta}$ is a $K \times V$ matrix and it is topics' distribution over V words.

Joint probability distribution over the words, latent topics, topic proportions and topic assignment defines the generative process of creating a corpus as in mathematical equation

below [41]:

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\alpha}, \eta) = (\prod_{k=1}^{K} p(\boldsymbol{\beta_k}|\eta))(\prod_{d=1}^{D} p(\boldsymbol{\theta_d}|\boldsymbol{\alpha})(\prod_{n=1}^{N} p(z_{dn}|\boldsymbol{\theta_d})p(w_{dn}|z_{dn}, \boldsymbol{\beta}))) \tag{1.1}$$

where D is the number of documents, N is the number of words, K is the number of topics, $\boldsymbol{\theta_d}$ is the document topic mixture, $z_{dn}$ is the topic assignment for the nth word in the dth document, $w_{dn}$ is the nth observed word in the dth document, $\boldsymbol{\beta_k}$ is the kth topic word distribution, $\boldsymbol{\alpha}$ are the Dirichlet hyper-parameters and $\eta$ is the scalar topic hyper-parameter. Integrating over $\theta$ and summing over $z$, we obtain the marginal distribution of a document:

$$p(\boldsymbol{w_d}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \int p(\boldsymbol{\theta}|\boldsymbol{\alpha})(\prod_{n=1}^{N} \sum_{z_{dn}} p(z_{dn}|\boldsymbol{\theta})p(w_{dn}|z_{dn}, \boldsymbol{\beta}))d\boldsymbol{\theta} \tag{1.2}$$

Because the number of possible hidden topic structures is exponentially large, the marginal probability in equation 1.2 is intractable to compute [23, 25]. Hence, it is required to use approximate posterior inference algorithms such as Gibbs sampling [54], variational Bayesian inference [37], maximum a *posteriori* estimation [52], etc. Among all these approximation methods, variational Bayesian inference is suitable for large datasets and can easily be accommodated with deep neural networks. It is a good alternative to some classic methods such as Markov Chain Monte Carlo (MCMC), since it is scalable to large datasets and faster in different applications. In variational inference, we specify a family $\mathcal{Q}$. Each $q(z) \in \mathcal{Q}$ is a candidate approximation to the exact conditional. Inference now amounts to solving the following optimization problem:

$$q^*(z) = argmin_{q(\boldsymbol{z}) \in \mathcal{Q}} D_{KL}[(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{w})] \tag{1.3}$$

where $D_{KL}$ represents Kullback–Leibler (KL) divergence defined as $D_{KL}((\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{w})) = \mathbb{E}[logq(\boldsymbol{z})] - \mathbb{E}[logp(\boldsymbol{z}, \boldsymbol{w})] + logp(\boldsymbol{w})$. Due to its dependency on the evidence $logp(\boldsymbol{w})$, 1.3 is hard to compute. Thus, the idea is to determine the evidence lower bound (ELBO) of the

log probability of the observations, which could be derived using Jensen's inequality:

$$\begin{aligned} logp(\boldsymbol{w}) &= log \int p(\boldsymbol{w}, \boldsymbol{z}) d\boldsymbol{z} \\ &= log \int p(\boldsymbol{w}, \boldsymbol{z}) \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} d\boldsymbol{z} \\ &= log \mathbb{E}_q[\frac{p(\boldsymbol{w}, \boldsymbol{z})}{q(\boldsymbol{z})}] \\ &\geq \mathbb{E}_q[log \frac{p(\boldsymbol{w}, \boldsymbol{z})}{q(\boldsymbol{z})}] \\ &= \mathbb{E}_q[logp(\boldsymbol{w}, \boldsymbol{z})] - \mathbb{E}_q[logq(\boldsymbol{z})] \end{aligned} \tag{1.4}$$

Since the lower bound holds for any q, we could choose the proper $q(z)$ based on different assumptions to ensure that it is easily computable [64]. One popular approximation is mean field variational inference [69]. It breaks coupling between $\boldsymbol{\theta}$ and $\boldsymbol{z}$ by introducing free variational parameters $\boldsymbol{\gamma}$ over $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ over $\boldsymbol{z}$. As a result, the optimization problem becomes optimizing $q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})$ to best approximate $p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. The ELBO could be written as:

$$L(\boldsymbol{\gamma}, \boldsymbol{\phi}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = D_{KL}[q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})||p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})] - logp(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})) \tag{1.5}$$

Though LDA has the closed form coordinate descent equations of mean field variational inference optimization, it is impractical or even impossible for some new models to find closed form solutions. Due to its limitation in flexibility, the autoencoding variational inference (AEVB) method was proposed [40, 57, 69]. In AEVB, the equation 1.5 could be rewritten:

$$L(\boldsymbol{\gamma}, \boldsymbol{\phi}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = -D_{KL}[q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})||p(\boldsymbol{\theta}, \boldsymbol{z}|, \boldsymbol{\alpha})] - \mathbb{E}_{q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})} logp(\boldsymbol{w}|\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \tag{1.6}$$

The second term is the reconstruction term, which is used to ensure that the variational posterior favors values of the latent variables that are good at explaining the data [69]. The variational parameters are computed using a neural network that takes the observed data as

input. For example, if we assume the model prior for $\boldsymbol{\theta}$ as the logistic normal distribution, the inference network would be defined as a feed-forward neural network $(\mu(\boldsymbol{w}), \sigma(\boldsymbol{w})) = f(\boldsymbol{w}, \boldsymbol{\gamma})$ where $\boldsymbol{\gamma}$ is the network parameters, which gives a Gaussian variational distribution $q_\gamma(\boldsymbol{\theta})) = \mathcal{N}(\boldsymbol{\theta}; \mu(\boldsymbol{w}), \sigma(\boldsymbol{w}))$.

Following LDA which is the foundation for numerous latent factors discovery algorithms, there are several extensions that have been developed which are known collectively as topic models. For example, Hierarchical Latent Dirichlet Allocation (HLDA) [17, 24] introduced built topic hierarchies to model the tree of topics which automatically identified syntactic and lexical patterns. Author topic model (ATM) [59, 60] proposed an expansion which used metadata for extracting the topic distribution with respect to authors. Each author has a distribution over topics. This type of model adding authorship could be applied to pull out model hidden information such as similarities between authors, etc. Dynamic topic model (DTM) [22, 53] incorporates evolution of time to LDA which is capable of tracking pattern changes over time series. In addition, to adjust models to have better scalability and expedite the training, neural networks were involved. For instance, embedded topic model (ETM [33]) and dynamic embedded topic model (DETM) [32] extended LDA and DTM by decomposing topic word mixture and allowed the models to be more adaptable in deep learning systems.

As the topic models have become more sophisticated and able to solve more challenging problems, they have been widely used in biological or health-related problems to uncover underlying semantic associations among biomedical concepts. For instance, Arnold et al. [18] applied LDA to identify clinically significant topics by learning patients' case-specific notes. Zhang et al. [77] combined LDA and networking analysis to discover latent disease mechanisms by dissecting disease-gene associations from over 25 million PubMed indexed articles. Wu et al. [76] proposed an LDA-based model to rank gene-drug associations in biomedical literature for drug re-purposing. Rider et al. [58] used an ensembled topic model to facilitate effective transfer learning between distinct healthcare data and constructing a network for interpretations used by domain experts and the discovery of disease relationships.

Their work improved the estimation of patient disease risk by extracting hidden shared patterns. Elibol et al. [35] characterized trajectories of developmental disorders' change over time using a dynamic topic model, which is important for early treatments. The topic model they developed gave promising results with incomplete medical records and social media posts.

The topic models' applications were not limited to text data. Madhaven et al. [46] developed an LDA-based topic model to group lncRNAs from a collection of transcriptome sequences and successfully addressed the problem that lncRNAs are less conserved at their sequence level. Li et al. [44] treated patients as documents and developed an LDA-based multi-modal topic model to learn interpretable patient topic mixture which was used to classify target diseases and predict mortality of patients. Song et al. [68] developed a specialist-specific supervised topic model to predict disease diagnoses and treatments while accounting for multimodality among specialist-dependent topic distributions.

## 1.4 Node2Vec

Graph neural network (GNN) is an approach emerged to model with graph data. Graph data is non-Euclidean and thus hard to model using other deep learning methods, since the graph can be irregular, meaning a graph may have variable sizes of unordered nodes and varied numbers of neighbors to a node [47]. As much biomedical data is naturally represented as graphs, GNN has been used extensively to model health-related data and successfully captured topological information in biomedical systems. There are mainly two types of biomedical graphs: molecular-level graphs such as chemical molecules and network graphs such as the drug-drug interaction graph. For instance, Choi et al. [29] proposed a graph-based attention model to learn meaningful representations from the medical oncology graph and achieved better performance in sequential diagnoses prediction tasks and heart failure prediction tasks. You et al. [78] developed a multi-species graph neural network-based

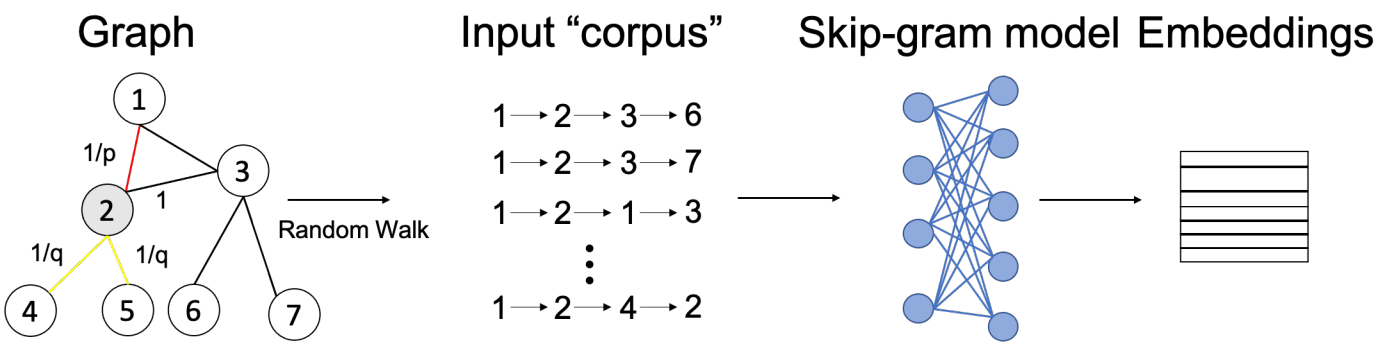


Figure 1.2: **Node2Vec Overview**. Node2Vec uses biased random walks to sample neighbors from graph. The biased factor $\alpha$ is determined by current state and potential next state. If the node 2 is current node and the previous node is node 1. Then the $\alpha$ for red line is 1/p. The yellow lines shows the case that the walk is going outward to the nodes which are not connected to the previous node, where the $\alpha$ is 1/q. For the last case, the walk goes to the node not identical to the previous node but connected to it, where $\alpha$ is 1. After generating the neighbor set for the resource nodes, the numerical embedding of the nodes is learned by a Skip-gram model.

method which made the most of both protein sequences and the high-order protein network information to do automated function prediction with a large-scale dataset. Nguyen et al. [50] compared multiple graph neural network methods in prediction of drug-target bindings and drug re-purposing on the bipartite graph. As the graph neural network approaches develop rapidly, longitudinal healthcare data could be modeled taking the temporal order of data into account. As an example, Lee et al. [43] has combined a unified graph representation learning framework with a long short term memory (LSTM) network to model heterogeneous medical entities and significantly improved the performance in subsequent code predication tasks.

Among all different graph neural network models, we have applied Node2Vec in our system. Node2Vec [39] is an embedding method which transforms graphs into numerical representations. The learned representation preserves the structure of the original network.

Given a graph $G = (V, E)$ where $V$ represents the set of vertices (nodes) and $E$ represents the set of edges, Node2Vec aims to find the optimized mapping function $f : V \rightarrow \mathbb{R}^d$ from nodes to numerical representations. For every source node $u$ in the graph, a set of

neighborhoods $N_S(u) \subset V$ is sampled. The model extends the Skip-gram architecture in natural language processing (NLP) systems, which uses the given target word to predict context words [49]. It optimizes the following objective function:

$$max_f \sum_{u \in V} logP(N_S(u)|f(u)) \tag{1.7}$$

There are two assumptions that make the objective tractable: conditional independence, which assumes observing a neighborhood node is independent from any other neighborhood node, and symmetry in feature space, which assumes a source node and a neighbor node have symmetric effects on each other. Thus, the objective function becomes:

$$\begin{aligned} \mathcal{L} &= max_f \sum_{u \in V} log( \prod_{n_i \in N_S(u)} \frac{exp(f(n_i), f(u))}{\sum_{v \in V} exp(f(v), f(u))}) \\ &= max_f \sum_{u \in V} (-log(\sum_{v \in V} exp(f(v), f(u))) + \sum_{n_i \in N_S(u)} (f(n_i), f(u))) \end{aligned} \tag{1.8}$$

For large graphs with millions of nodes, since the model has a tremendous number of weights to optimize which would be computationally expensive and time-consuming, negative sampling is applied in the training process. For each source node, the model samples both positive neighbors as well as false neighbors, which are referred as negative samples. The negative samples are generated following certain distributions depending on the settings and datasets. The model is trained to favor the probability of true neighbors to be large. Then the objective of the model is to minimize:

$$\mathcal{L} = \sum_{u \in V} \sum_{n_i \in N_S(u)} (-(f(n_i), f(u)) - \sum_{v' \in N'_S(u)} (f(v'), f(u))) \tag{1.9}$$

where v' is a negative sample form the negative neighborhood $N'_S(u)$.

Node2Vec generates positive neighbors via biased random walk. The sampling strategy includes four parameters: the number of walk, which is the number of random walks to be

generated for each node, walk length, which is the number of nodes in each walk, p, which is the return parameter, and q, which is the in-out parameter. p and q are used to determine $\alpha_{pq}(t,x)$ which is a bias factor to determine the unnormalized transition probability $\pi_{vx} = \alpha_{pq}$ (t,x)$\cdot w_{vx}$ given the current node v, the previous node t and the possible next node x. $w_{vx}$ are the static edge weights for edges (v,x). Consider a random walk that just traversed edge (1,2) and now resides at node 2 (Fig. 1.2). For the current state, if the next state is a node of previous state (1 in our example), $\alpha_{pq}(t,x)$ is 1/p. If the next state is a node which is not connected to the previous node, $\alpha_{pq}(t,x)$ is 1/q. For the rest of nodes, $\alpha_{pq}(t,x)$ is set to 1. A small q increases the probability of the walk moving outward from the localized neighborhood. In other words, if $q < 1$, the random walk is more inclined to perform depth-first search (DFS), by which the neighborhood consists of nodes sequentially sampled at increasing distances from the source node. In contrast, if $q > 1$, the random walk is more reflective to breadth-first search (BFS), by which the neighborhood is restricted to nodes which are immediate neighbors of the source node. So the nodes $c_i$ are generated by the following distribution:

$$P(c_i = x | c_{i-1}) = \begin{cases} \frac{\pi_{vx}}{Z} & \text{if } (v,x) \in E \\ 0 & \text{otherwise} \end{cases} \tag{1.10}$$

where Z is a normalizing constant.

## 1.5 Related work

There have been attempts to apply matrix factorization methods to UK Biobank data, in order to extract more hidden patterns or learn associations between different types of data. Bayesian non-negative matrix factorization (bNMF) clustering has been applied to genome-wide association study (GWAS) results for 94 independent Type 2 diabetes (T2D) genetic variants and 47 diabetes-related traits. The study found out individuals with T2D in the top genetic risk score decile for each cluster reproducibly exhibited the

predicted cluster-associated phenotypes, with approximately 30% of all individuals assigned to just one cluster top decile [73]. In the study using topic modeling via non-negative matrix factorization (NMF) for identifying associations between disease phenotypes and genetic variants, they identified a positive correlations of topics enriched for cardiovascular diseases and hyperlipidemia with rs10455872 in Lipoprotein(a) (LPA) and a negative correlations between LPA and a topic enriched for lung cancer [79]. Tanigawa et al. have applied truncated singular value decomposition (DeGAs) to matrices of summary statistics derived from genome-wide association analyses across 2,138 phenotypes measured in 337,199 White British individuals in the UK Biobank study to identify key components of genetic associations and the contributions of variants, genes, and phenotypes to each component [71]. A variant of Principal component analysis (PCA) has been used in the UK Biobank and the 1000 Genomes project datasets, which help make recommendations for best practices and provide efficient and user-friendly implementations of the proposed solutions in R packages bigsnpr and bigutilsr [55]. These methods, though have demonstrated advantages of learning associations, they ignored the phenotypic networks. We assumed those intra-relationships can also provide useful insights. Hence, we incorporate graph modelling to learn embeddings of phenotypes using relational graphs.

## 1.6 Uniform manifold approximation and projection

Uniform manifold approximation and projection (UMAP) is a novel manifold learning technique for dimension reduction [48]. We have applied this method to trained condition embedding $\rho_c$ concatenated with trained condition-defined topic embedding $\alpha_c$ and to trained medication embedding $\rho_m$ concatenated with trained medication-defined topic embedding $\alpha_m$. Visualizing 2-dimensional embedding obtained from UMAP, we were able to observe whether the topics enriched with certain diseases and medications are assigned to common disease groups.

The UMAP algorithm consists of two steps: construction of a graph in high dimensions and an optimization step to find the most similar graph in lower dimensions. It constructs a weighted graph from high dimensional data, and then projects this graph down to a lower dimensionality. In the graph, the edge strength represents how "close" a given point is to another. For each data point, UMAP extends some radius r and connects points that overlap, so to construct sets of 1-, 2-, and higher-dimensional simplices. To solve the "curse of dimensionality", UMAP uses a flexible radius determined for each point based on the distance to its kth nearest neighbor. Once this weighted graph is constructed, UMAP projects the data into lower dimensions essentially via a force-directed graph layout algorithm [48]. The algorithm proceeds by iteratively applying attractive and repulsive forces at each edge or vertex. The attractive force between two vertices i and j at coordinates $y_i$ and $y_j$ is given by:

$$\frac{-2ab||\boldsymbol{y_i} - \boldsymbol{y_j}||_2^{2(b-1)}}{1 + ||\boldsymbol{y_i} - \boldsymbol{y_j}||_2^2} w((x_i - x_j))(\boldsymbol{y_i} - \boldsymbol{y_j}) \tag{1.11}$$

where $x_i$ and $x_j$ are inputs, while a and b are hyper-parameters.

Repulsive forces are computed via sampling due to computational constraints.

$$\frac{2b}{(\epsilon + ||\boldsymbol{y_i} - \boldsymbol{y_j}||_2^2)(1 + a||\boldsymbol{y_i} - \boldsymbol{y_j}||_2^{2b})} (1 - w((x_i - x_j)))(\boldsymbol{y_i} - \boldsymbol{y_j}) \tag{1.12}$$

where $\epsilon$ is a small number to avoid zero division.

# Chapter 2

# Methods

## 2.1 Graph embedded topic model generative process

We present graph embedded topic model (GETM) and our model is inspired by ETM [33]. We modeled the medication and condition data using a generative model (Fig. 2.1). In our framework, each individual was treated as a document and each feature (a medication or a condition from two different vocabulary sets), was treated as a word. We assumed each individual could be represented as a mixture of latent topics, which captured hidden information of medications and conditions. In contrast to LDA, which defines the topic distribution over terms by K independent Dirichlet priors $\beta_k \sim Dirichlet(\tau_\beta)$, we decomposed the topic distribution over medications $\beta_m$ to medication-defined topic embedding $\alpha_m \in \mathbb{R}^{K \times L_1}$, and medication embedding $\rho_m \in \mathbb{R}^{L_1 \times M}$, where $L_1$ denotes the medication embedding dimension and $M$ denotes the number of unique medications. Similarly, the topic distribution over condition $\beta_c$ is proportional to the inner product of condition-defined topic embedding $\alpha_c \in \mathbb{R}^{K \times L_2}$, and condition embedding $\rho_c \in \mathbb{R}^{L_2 \times C}$, where $L_2$ denotes the condition embedding dimension and $C$ denotes the number of unique conditions. For an individual d, the generative process started by drawing $\theta_d$ from logistic

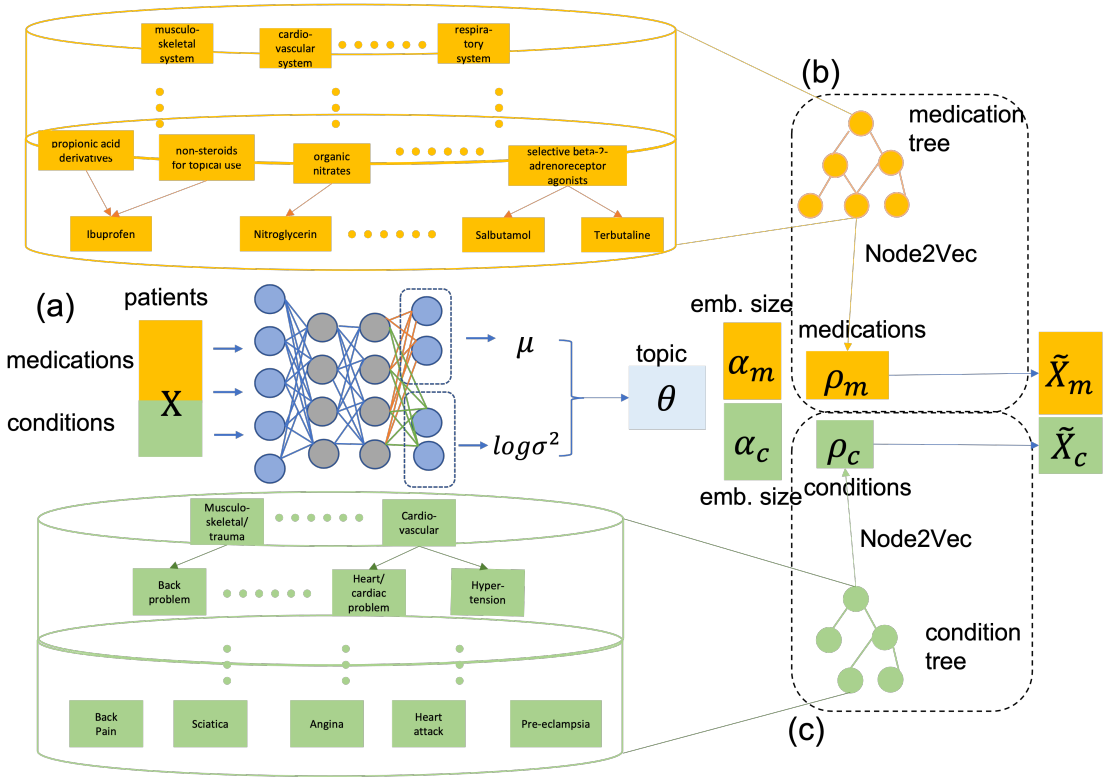


Figure 2.1: **Graph embedded topic model overview**. **(a)** GETM training. A variational autoencoder (VAE) takes individuals' condition and medication information as input and produces latent topic mixture $\theta$, which could be used in multiple tasks such as pain prediction, data imputation, etc. There are two linear decoders. One learns medication-defined topic embedding $\alpha_m$ and medication embedding $\rho_m$. The other learns condition-defined topic embedding $\alpha_c$ and $\rho_c$. **(b-c)**. Graph learning. The embedded topic model could take $\rho_m$ and $\rho_c$ generated from Node2Vec, which leverages structural information of medications or conditions. Medication and condition hierarchical trees were treated as graphs and trained separately.

normal $\theta_d \sim \mathcal{LN}(0, \mathbf{I})$ as :

$$\eta_d \sim \mathcal{N}(0, \mathbf{I}); \quad \theta_d = softmax(\eta_d) = \frac{e^{\eta_{d,k}}}{\sum_{k=1}^{K} e^{\eta_{d,k}}} \tag{2.1}$$

For each feature n, the topic assignment from categorical distribution with the per-individual variable $\theta_d$ as: $z_{dn} \sim Cat(\theta_d)$ is drawn. Next an observed medication using $m_{dn} = Cat(softmax(\rho_m^T \alpha_{m(z_{dn})}))$ or an observed condition using $c_{dn} = Cat(softmax(\rho_c^T \alpha_{c(z_{dn})}))$ is drawn. Notably, to model the sparsity of the data (an individual usually takes a very small fraction of the medication and did not have more than five conditions simultaneously), the softmax function was used to normalize the likelihood of each drawn feature.

## 2.2 Model inference and estimation

Given D individuals, M medications and C conditions, to fit GETM, we want to maximize the marginal likelihood of the individuals with respect to $\alpha_m, \rho_m, \alpha_c, \rho_c$:

$$\mathcal{L}(\alpha_m, \rho_m, \alpha_c, \rho_c) = \sum_{d=1}^{D} logp(x_d | \alpha_m, \rho_m, \alpha_c, \rho_c) \tag{2.2}$$

where $x_d$ is the bag of words of medications and conditions for individual d. The problem is that this marginal likelihood is intractable to compute since it involves the following difficult integral:

$$\begin{aligned} p(x_d | \alpha_m, \rho_m, \alpha_c, \rho_c) &= \int p(\eta_d) \prod_{n=1}^{M} p(m_{dn} | \eta_d, \alpha_m, \rho_m) \prod_{n=M+1}^{M+C} p(c_{dn} | \eta_d, \alpha_c, \rho_c) d\eta_d \\ &= \int p(\eta_d) \sum_{k=1}^{K} (\prod_{n=1}^{M} \theta_{dk} \beta_{dm_{dn}} \prod_{n=M+1}^{M+C} \theta_{dk} \beta_{dc_{dn}}) d\eta_d \end{aligned} \tag{2.3}$$

To sidestep this intractable integral, we took a variational inference approach to optimize a sum of of per-individual bounds on the log of the marginal likelihood of Eq. 2.3.

To begin, we proposed distribution $q(\eta_d | x_d, W_\theta)$ to approximate true posterior $p(\eta_d | x_d)$.

Specifically,$q(\eta_d|x_d)$ is Gaussian with mean and variance come from a neural network parameterized by the shared variational parameter $W_\theta$ :

$$q(\eta_d|x_d) = \mu_d + diag(\sigma_d)\mathcal{N}(0, \mathbf{I}); \quad [\mu_d, log\sigma_d^2] = NNET(x_d; W_\theta) \tag{2.4}$$

The evidence lower bound (ELBO) of the log-likelihood was optimized to learn the the model parameters and variational parameters:

$$\begin{aligned}\mathcal{L}(\alpha_m, \rho_m, \alpha_c, \rho_c, W_\theta) = \sum_{d=1}^{D}(\sum_{n=1}^{M}\mathbb{E}_q[logp(m_{dn}|\eta_d, \alpha_m, \rho_m)] + \sum_{n=M+1}^{M+C}\mathbb{E}_q[logp(c_{dn}|\eta_d, \alpha_c, \rho_c)] \\ - KL(q(\eta_d|x_d, W_\theta)||p(\eta_d)))\end{aligned} \tag{2.5}$$

## 2.3 Baseline Models

We compared our method to different baselines (Table 2.1): 1). We applied ETM to only condition data or only medication data without using Node2Vec pre-trained embedding. 2). We applied ETM to only condition data or medication data using Node2Vec pre-trained embedding. 3). We applied GETM without either condition embedding or medication embedding from Node2Vec. 4). We applied GETM without one of condition embedding or medication embedding from Node2Vec. 5). We treated conditions and medications as the same features and modeled the resulting data using ETM. With the comparison of results from above baseline models in different tasks, we were able to gain a better understanding of how each component of our method contributing to the overall improvements in performance.

| | |
|---|---|
| ETM | ETM without Node2Vec pre-trained embedding |
| ETM + emb. | ETM with Node2Vec pre-trained embedding |
| GETM | GETM without either Node2Vec condition embedding or Node2Vec medication embedding |
| GETM + emb._cond | GETM with Node2Vec condition embedding but without Node2Vec medication embedding |
| GETM + emb._med | GETM without Node2Vec condition embedding but with Node2Vec medication embedding |
| GETM + emb._cond+med | GETM with both Node2Vec condition embedding and Node2Vec medication embedding |
| Flattened | ETM takes both conditions and medications which were treated as same feature |

Table 2.1: **Description of abbreviation of algorithm names**. ETM only takes one type of feature and GETM takes both conditions and medications as different features.

## 2.4 UK Biobank data processing

For condition data, we have datafield 20002, which contains information on individuals' self-reported non-cancer diseases. This was collected by questionnaire during participant interviews. The participants were asked whether or not they have been diagnosed with certain conditions as well as when that condition was first diagnosed by doctor [4]. For medication usage data, we used datafield 20003 which contains treatment/medication codes [5]. Besides, we have also referred to individuals' demographic information such as ethnicity, etc. and a pain-related questionnaire which confirmed whether the individual had pain on seven body sites and how long they experienced the pain if any. Pain-related information is included in datafield 6159 [14]: pain type(s) experienced in last month, datafield 3799 [12]: headaches for 3+ months, datafield 4067 [13]: facial pains for 3+ months, datafield 3404 [7]: neck/shoulder pain for 3+ months, datafield 3571 [9]: back pain for 3+ months, datafield 3741 [10]: stomach/abdominal pain for 3+ months, datafield 3414 [8]: hip pains for 3+ months, datafield 3773 [11]: knee pains for 3+ months, and datafield 2956 [6]: general pain for 3+ months. More specifically, to collect data in datafield 6159, participants were asked "In the last month have you experienced any of the following that interfered with your usual activities? (You can select more than one answer)." If they said "yes" to any pain, for example, back pain, they were further asked "Have you had back pains for more than 3 months?" in datafield 3571. We kept 457461 individuals of European descent individuals to reduce confounding caused by different ethic groups. 802 active ingredients were kept as medications and 443 conditions were extracted. Here we encoded the medications and

diseases as binary variables. Without a temporal component in the model, only the data from the first visit of individuals was included.

As mentioned before, the medication embedding $\rho_m$ and condition embedding $\rho_c$ could either be randomly initialized and then learned, or we could use trained embedding from pre-trained models. This allows us to incorporate previous knowledge of medications and conditions which could enrich the information going through the topic model. The additional information is helpful especially when the data is sparse. To leverage the structural information and internal relational information among medications or conditions. We applied Node2Vec separately on hierarchical trees of medications and conditions (Fig. 2.1). The condition tree graph was formed using coding tree designed in datafield 20002. The tree describes the topology of the conditions with 473 nodes and 4 levels. The medication graph was formed based on Anatomical Therapeutic Chemical (ATC) classification system [15]. The entire tree is composed of 5 levels. We first kept the top 4 levels of ATC of which the first level contains main anatomical or pharmacological groups; the second level includes pharmacological or therapeutic subgroups; and in the third and fourth levels are chemical, pharmacological or therapeutic subgroups. Then we mapped the names of active ingredients from UKBiobank datafield 20003 to the fifth level codes of ATC which are chemical substances. We replaced matched substances with UKBiobank medications. In particular, for some medications in UKBiobank, they could be mapped to multiple ATC fifth level codes, because they could belong to different subgroups with respect to different usages. For those medications, we replaced all mapped ATC fifth level nodes with one UKBiobank medication active ingredients node. As a result, the final medication graph contains 2561 nodes in total. The trees are treated as undirected graphs (Fig. 2.1).

## 2.5 Topic quality evaluation

We aimed to identify topics that may be interpreted. Besides, it is expected that different topics could be associated with different medications or conditions. Therefore, we measured the topic quality with two metrics: topic coherence and topic diversity. To gain better interpretations, the top features of one topic are expected to come from the same category. Therefore, for medication, the topic coherence was calculated as:

$$TC_{med} = \frac{1}{K}\sum_{k=1}^{K}\frac{m}{n} \tag{2.6}$$

where n is the number of top medications and m is the maximum number of medications that are from the same category. To avoid overestimation of the topic quality, the categories we used to evaluate the topic coherence were not processed from the ATC graph since it was involved in the pre-trained model. Instead, we employed 59 categories which are physician-curated and pain-focused. However, we do not have other classification approaches for conditions. As a result, we decided to calculate the topic coherence for conditions using average pointwise mutual information of two conditions drawn from the same individual as follows:

$$TC_{cond} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n} f(c_i^k, c_j^k) \tag{2.7}$$

where $c_i^k$ is the $i^{th}$ top most likely condition in topic k. and $f(\cdot,\cdot)$ is normalized pointwise mutual information:

$$f(c_i, c_j) = \frac{log\frac{P(c_i,c_j)}{P(c_i)P(c_j)}}{-logP(c_i,c_j)} \tag{2.8}$$

where $P(c_i, c_j)$ is the probability of condition i and condition j co-occurring in one individual and $P(c_i)$ is the marginal probability of condition i. The probabilities were approximated by empirical counts. Topic diversity was defined as the percentage of unique features of certain number of top features among different topics. We chose 50 for our evaluation. The closer the topic diversity to 1, the more varied the topic is.

## 2.6 Study of medication and condition relations

In GETM, since the encoder takes both medication and condition information as input (Fig. 2.1), the top medications and conditions from the same topic (i.e. the same index) are related. To quantitatively measure the ability of the model to capture condition-medication relations, we got all combinations from top 3 conditions and top 3 medications for each topic and then counted the number of condition-medication pairs which are known to be related. The reference of known pairs was extracted from Comparative Toxicogenomics Database (CTD) [1] and DrugBank [2]. We eventually mapped 222 conditions and 529 medications from UKBiobank to these two databases. Then we obtained 2444 positive pairs of which the medication has treatment effects on the condition and 3231 negative pairs of which the condition belongs to the adverse effects of the medication.

## 2.7 Data imputation

Since our method can impute data for new individuals, two useful applications are imputing incomplete records and recommending medications. Therefore, two experiments were performed to evaluate how well the model could complete these two tasks. To simulate missing records, we randomly masked 50% of medications and conditions for test individuals. Then we calculated reconstruction error, which is the negative log-likelihood of the reconstructed matrix $\tilde{X}$ (Fig. 2.1):

$$NLL = \frac{1}{D}\sum_{d=1}^{D} -log(\theta\beta) \odot X \tag{2.9}$$

where $\beta = softmax(\alpha\rho)$.

The treatment of chronic illnesses commonly includes the long-term use of pharmacotherapy. Poor adherence to medication leads to increased morbidity and death and is estimated to incur costs of approximately $100 billion per year [26]. Thus, it is crucial for

individuals to take all medications that could improve their conditions. The other experiment was masking the entire medication data of the test individuals and then reconstructing the medication matrix. This experiment mimics the scenario that we recommend relevant medications based only on patients' conditions, which namely is similar to the process diagnosis. In addition to reconstruction error, we also calculated recall@5 and precision@5 as evaluation metrics. We sorted the probability of medication to be recommended and chose top five medications, of which we calculated recall and precision with respect to medications that the individual is taking as true labels. The recall and precision are calculated as:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{2.10}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{2.11}$$

## 2.8 Chronic musculoskeletal pain prediction

### 2.8.1 Fisher's exact test

To assess marginal association between each condition and medication with chronic musculoskeletal pain, we performed Fisher's exact test. Fisher's exact test is typically the first step for analyzing marginal association because of its simplicity and interpretability. For each feature (condition or medication), we formed a contingency table (Table. 2.2). Then p-values and odds ratios were calculated as follows:

$$\begin{aligned} p &= \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \\ &= \frac{(a+b)!(a+c)!(b+d)!(c+d)!}{a!b!c!d!n!} \end{aligned} \tag{2.12}$$

| Feature \ Pain | Yes | No |
|---|---|---|
| Yes | a | b |
| No | c | d |

Table 2.2: **Contingency table used for each feature to do Fisher's exact test**

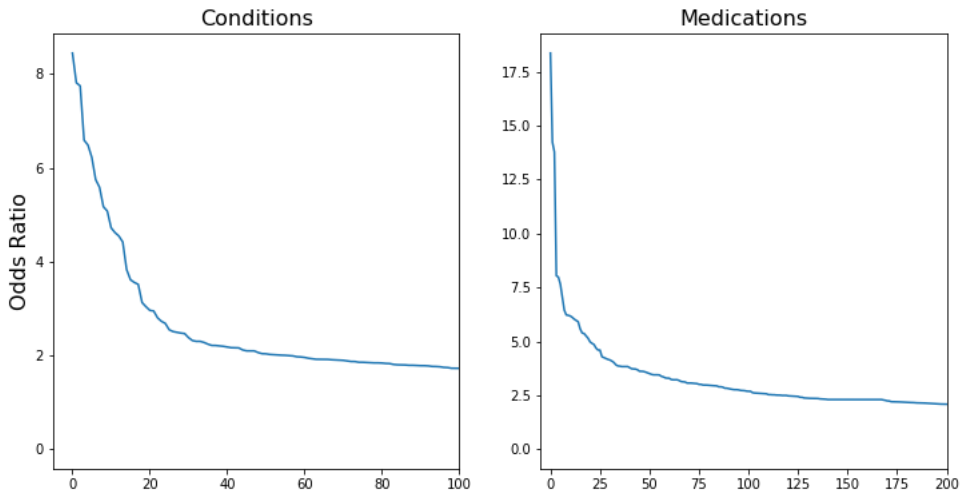


Figure 2.2: **Fisher's exact test for chronic musculoskeletal pain**. The significance of the association of conditions or medications with chronic musculoskeletal pain was tested. The analysis was used to determine the conditions and medications to be removed for certain datasets in the prediction task.

where $n = a + b + c + d$

$$OR = \frac{ad}{bc} \tag{2.13}$$

Fisher's exact test results (Fig. 2.2) were used for two purposes: 1). determining the signature conditions and medications to be removed in certain datasets for chronic musculoskeletal pain prediction. 2). a baseline when comparing with pain-related conditions and medication lists created by a physician based on professional knowledge.

### 2.8.2 Chronic musculoskeletal pain prediction

Chronic musculoskeletal pain was defined as musculoskeletal pain which has lasted more than three months, and musculoskeletal pain was pain from any of four body sites: knee, back, neck or shoulder and face. The label information was obtained from pain questionnaires as described in section 2.4.

We used individual topic mixture $\theta$ to predict whether a individual has chronic musculoskeletal pain at the time of the visit. GETM was first trained with training data to get $\theta_{train}$. Then we fit logistic regression on $\theta_{train}$. To evaluate the performance, we first obtained $\theta_{test}$ using trained GETM and then predicted chronic musculoskeletal pain status using the trained logistic regression model.

As mentioned in the previous section, we have done Fisher's exact test to test the significance of the associations that conditions and medications have with chronic musculoskeletal pain. Based on the results, we picked 50 conditions (C1) and 150 medications (M1) to remove (Fig. 2.2). Combining with 63 conditions (C2) and 122 medications (M2) which a physician declared to be related to general pain, we created three new condition sets: one removing C1, one removing C2 and one removing C1 $\bigcup$ C2, and two new medication sets: one removing M1 and one removing M1 $\bigcup$ M2. Those new condition sets and new medication sets form six datasets which are all combinations of created condition and medication sets. Together with the original dataset including 802 medications and 443 conditions, we have done analysis on seven datasets (details in section 2.4 and Fig. 3.3). In this way, we were able to observe how the pipeline worked when removing those associated conditions and medications. Removing those associated conditions and medications in prediction could help those individuals with no obvious symptoms prevent chronic musculoskeletal pain in advance.

### 2.8.3 Pain-related conditions and medications

After we got trained coefficients $\omega$ from logistic regression, we investigated the relevance of medications and conditions to chronic musculoskeletal pain by calculating relevance vectors:

$$V = \omega^T \alpha \rho \tag{2.14}$$

The top N medications and conditions were then selected from $V_m$ and $V_c$, after which we calculated the proportions of these N medications or conditions overlapping with the pain-related lists created by a physician. This analysis enables us to examine the ability of our model to extract associative information.

# Chapter 3

# Results

## 3.1 Topic quality evaluation

For condition-defined topics, the highest topic coherence was achieved by ETM with Node2Vec pre-trained embedding (0.0253) and topic number 15. The GETM with only condition embedding from Node2Vec and topic number of 50 was the second highest (0.0196)(Table 3.1). For medication-defined topic, the highest coherence was obtained using GETM with only medication embedding from Node2Vec (0.7860) and topic number of 100. Therefore, applying embedding learned from Node2Vec improved topic coherence. It tells us that pre-trained embedding captured the inner associations of medications or conditions and thus enriched the information provided to the topic model. Another notable result is that the values of topic coherence for conditions are generally low, though GETM with both medication and condition embedding from Node2Vec actually learned highly interpretable topics of which the top conditions or medications were from same category (Fig. 3.1). For example, The top 5 conditions from topic 8 are all from musculoskeletal/trauma category while the top 5 medications from topic 8 are all from dermatological category. This result suggests that we have chosen an inappropriate metric, which was average pointwise mutual information of two conditions drawn from the same patient. A patient rarely gained multiple

| Algorithm \ Topic # | 15 | 50 | 75 | 100 | 15 | 50 | 75 | 100 | 15 | 50 | 75 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Topic Coherence | | | | Topic Diversity | | | | Topic Quality | | | |
| ETM | 0.0035 | 0.003 | 0.0019 | 0.0076 | 0.92 | 0.32 | 0.2 | 0.16 | 0.0032 | 0.001 | 0.0004 | 0.0012 |
| ETM + emb. | **0.0253** | 0.0153 | 0.0172 | 0.017 | 0.92 | 0.48 | 0.24 | 0.16 | **0.0233** | 0.0074 | 0.0041 | 0.0027 |
| GETM | 0.0125 | 0.0105 | 0.0044 | 0.0046 | 0.92 | 0.32 | 0.24 | 0.16 | 0.0115 | 0.0034 | 0.0011 | 0.0007 |
| GETM + emb._cond | 0.0213 | **0.0196** | 0.0193 | 0.0169 | 0.88 | **0.56** | 0.28 | 0.12 | 0.0188 | 0.011 | 0.0054 | 0.002 |
| GETM + emb._med | 0.0144 | 0.007 | 0.0087 | 0.0087 | 0.92 | 0.52 | 0.24 | **0.28** | 0.0132 | 0.0036 | 0.0021 | 0.0024 |
| GETM + emb._cond+med | 0.0206 | 0.0188 | **0.0186** | **0.0186** | **0.96** | 0.64 | **0.32** | 0.2 | 0.0198 | **0.0121** | **0.0057** | **0.0037** |

Table 3.1: **Condition-defined topic quality**. We have calculated topic coherence, topic diversity and topic quality in terms of conditions for 6 algorithms using different topic numbers. The description of algorithm name is in table 2.1. The feature ETM accepted as input was the condition.

| Algorithm \ Topic # | 15 | 50 | 75 | 100 | 15 | 50 | 75 | 100 | 15 | 50 | 75 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Topic Coherence | | | | Topic Diversity | | | | Topic Quality | | | |
| ETM | 0.2800 | 0.3320 | 0.3173 | 0.3460 | 1.00 | 0.64 | 0.56 | 0.24 | 0.2800 | 0.2125 | 0.1777 | 0.0830 |
| ETM + emb. | 0.4933 | **0.7160** | 0.7040 | 0.7800 | 1.00 | 0.84 | 0.44 | 0.24 | 0.4933 | 0.6014 | 0.3098 | 0.1872 |
| GETM | 0.3200 | 0.3000 | 0.3494 | 0.3460 | 1.00 | **0.96** | 0.92 | 0.96 | 0.3200 | 0.2880 | 0.3214 | 0.3322 |
| GETM + emb._cond | 0.3733 | 0.4520 | 0.3307 | 0.4060 | 0.96 | 0.92 | 0.96 | 1.00 | 0.3584 | 0.4158 | 0.3174 | 0.4060 |
| GETM + emb._med | 0.4933 | 0.7000 | **0.7707** | **0.7860** | 1.00 | 1.00 | 1.00 | 0.96 | 0.4933 | **0.7000** | **0.7707** | **0.7546** |
| GETM + emb._cond+med | **0.5200** | 0.7040 | 0.7200 | 0.7220 | 0.96 | 0.92 | 1.00 | 0.96 | **0.4992** | 0.6477 | 0.7200 | 0.6931 |

Table 3.2: **Medication-defined topic quality**. We have calculated topic coherence, topic diversity and topic quality in terms of medications for 6 algorithms using different topic number. The description of algorithm name is in table 2.1. The feature ETM accepted as input was medication. The results of last four columns are from correspondingly same models in table 3.1

conditions from same category simultaneously, which led to low probability of co-occurrence of top conditions from a specific topic.

## 3.2 Study of medication and condition relations

We have compared the total number of unique known pairs between medications and conditions that could be generated by five algorithms (Table 3.3). GETM with medication and condition embedding from Node2Vec can extract most pairs of correlated conditions and medications through various number of topic using topic number of 50 (161 pairs), 75 (175 pairs), and 100 (203 pairs). The examples could also be visualized in panel (b) and panel (d) in Fig. 3.1. For instance, bisoprolol in topic 32 is known to be used to treat heart failure in topic 32. Salmeterol is prescribed to treat asthma and chronic obstructive airways (COPD) [2]. They are both enriched in topic 60. The results indicate that GETM combined with

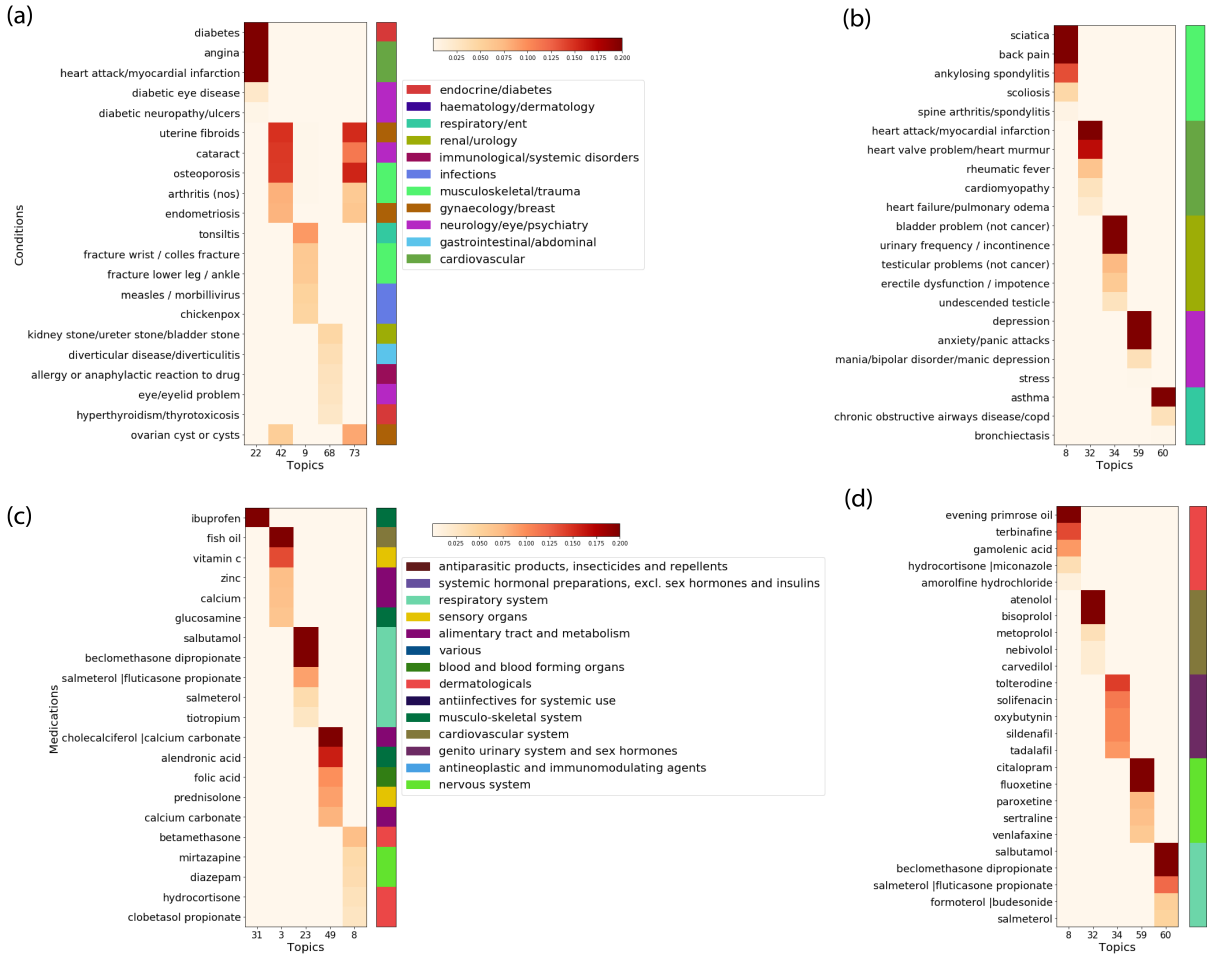


Figure 3.1: **Topic quality visualization**. For all four panels, we visualized 5 topics out of 75 topics. **(a)**. Top conditions selected from topics using ETM on patient data condition data information without Node2Vec condition embedding **(b)**. Top conditions selected from topics using GETM on patient data with both Node2Vec conditions embedding and Node2Vec medications embedding. **(c)**. Top medications selected from topics using ETM on patient medication data without Node2Vec embedding. **(d)**. Top medications selected from topics using same model as in panel (b). Comparing panel (a) with panel (b) and panel (c) with panel (d), we could observe that GETM with pre-trained embedding learns much more coherent and interpretable topics. If looking close into panel (b) and panel (d). it could be told that the conditions and medications from same topic are correlated.

| Algorithm \ Topic # | 15 | 50 | 75 | 100 |
|---|---|---|---|---|
| | Number of matched pairs | | | |
| Flattened | 53 | 84 | 119 | 132 |
| GETM | 63 | 86 | 105 | 126 |
| GETM + emb._cond | 62 | 118 | 159 | 162 |
| GETM + emb._med | **65** | 135 | 171 | 178 |
| GETM + emb._cond+med | 61 | **161** | **175** | **203** |

Table 3.3: **Number of known pairs between conditions and medications**. We have mapped our medications and conditions to CTD and DrugBank databases to get reference for links between conditions and medications. Then we got all combinations from top 3 conditions and top 3 medications for each topic and then summarized number of condition-medication pairs that exists in those known links. We compared performance of five algorithms with different topic number.The description of algorithm name is in table 2.1.

Node2Vec could capture correlations between different types of features within the same topic. It is also worthwhile to explore potential relations of those pairs that are not present in current databases. It is reasonable to assume they are related for the reason that they are learned the same way as those known pairs. For example, though solifenacin in topic 59 are not associated with depression according to current database, there is a recent research showing that solifenacin and mirabegron act mainly via peripheral pathways in overactive bladder (OAB), whereas the central pathways are responsible for the effects of duloxetine, 72h after discontinuation of which, positive changes in the corticosterone-induced depression, detrusor overactivity, and inflammation were observed [75].

## 3.3 Data imputation

### 3.3.1 Missing record imputation

For all three datasets: data with only conditions, data with only medications and data with both conditions and medications, the minimum reconstruction errors were obtained after applying Node2Vec embedding for all features in that dataset (6.008, 8.5732 and 25.2462 respectively). This indicates that the structural information was preserved during

| Algorithm \ Topic # | 15 | 50 | 75 | 100 |
|---|---|---|---|---|
| | Reconstruction Error | | | |
| ETM(cond) | 6.2267 | 6.2189 | 6.2905 | 6.2176 |
| ETM(cond) + emb. | **6.0008** | **6.2056** | **6.1305** | **6.1472** |
| ETM(med) | 8.7118 | 9.0881 | 8.9889 | 9.1980 |
| ETM(med) + emb. | **8.5732** | **8.7869** | **8.6748** | **8.8807** |
| GETM | 26.3611 | 27.0904 | 27.4865 | 27.3511 |
| GETM + emb._cond | 25.7661 | 25.9377 | 26.7536 | 26.5983 |
| GETM + emb._med | 26.1501 | 25.9594 | 26.1133 | 26.1765 |
| GETM + emb._cond+med | **25.2462** | **25.4741** | **25.9872** | **25.5122** |

Table 3.4: **Reconstruction error of masked features**. The 50% of test data was randomly masked. Then we reconstructed the matrix with learned $\theta, \alpha$ and $\rho$. The reconstruction error (i.e. negative log-likelihood) was calculated for the held-out data. It could be observed that applying embedding pre-trained by Node2Vec could enhance the reconstructing ability for both ETM and GETM. Description of algorithm names are in table 2.1.

the training process of the topic model (Table 3.4).

### 3.3.2 Medication recommendation

For the medication recommendation task, we observed that GETM with condition and medication embedding both from Node2Vec outperformed all other models in terms of the reconstruction error (14.6125), the precision@5 (0.2612) and the recall@5 (0.5787). It gives the closest performance to upper bounds which are acquired from unmasked test data ( 11.4936, 0.4226 and 0.8027, respectively) (Table 3.5, 3.6). Besides, we found out interestingly that on average around 44% of medications from the top 10 medications recommended by our model which were however not taken by patients indeed have a treatment effect based on condition-medication association information extracted from CTD and DrugBank (as shown in panel (b) of Fig. 3.2). This finding implies that our model not only recommended most specific medications, but also predicted medications that were ignored but actually help improve the health condition of the individuals.

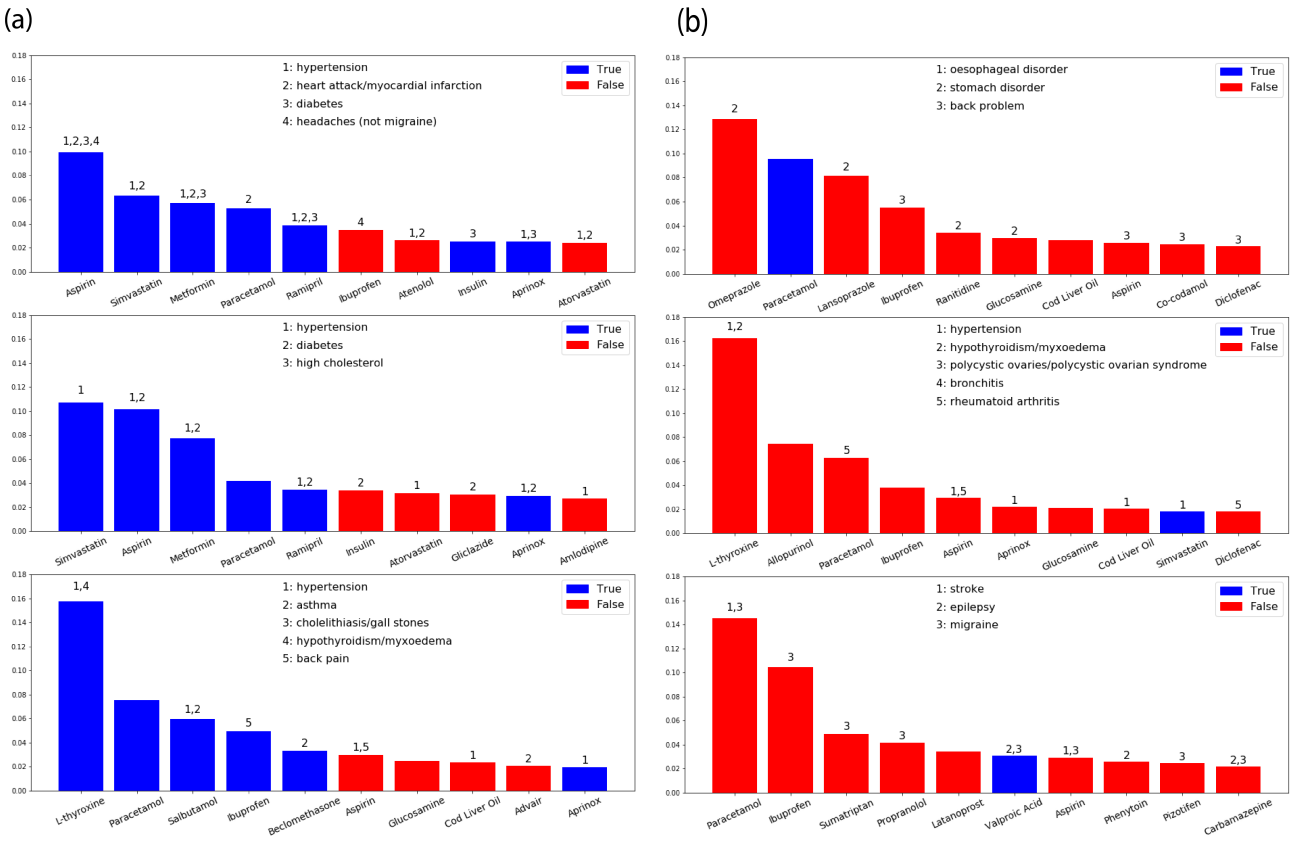


Figure 3.2: **Examples of patients with most and least matched recommended medications compared to medications truly taken**. Y axis is normalized probability of medication to be recommended to the patient. **(a)**. Three patients who got most overlapped medications between true medications they are taking and top 10 medications recommended. **(b)**. Three patients who got least overlapped medications between true medications they are taking and top 10 medications recommended. The numbers on the bar are conditions that could be treated with the medication represented by the bar. It is noteworthy that in panel (b), for those medications that patients do not take, most of them have treatment effects with the patients' current conditions. This suggests that our model have recommended relevant medications.

| Algorithm \ Topic # | 15 | 50 | 75 | 100 |
|---|---|---|---|---|
| | Medication Recovery Error | | | |
| Upper bound | 12.2535 | 11.2364 | 11.4986 | 11.5028 |
| Flattened | 18.5142 | 19.8497 | 19.8070 | 20.0636 |
| GETM | 14.9877 | 14.8770 | 14.9392 | 15.0474 |
| GETM + emb._cond | 15.2422 | 14.9481 | 14.9500 | 14.9556 |
| GETM + emb._med | 15.1829 | 14.8816 | 14.9012 | 14.8819 |
| GETM + emb._cond+med | **14.8178** | **14.6533** | **14.6125** | **14.6525** |

Table 3.5: **Reconstruction error of medication data**. The medication data was masked for test patients. Then we reconstructed the medication data with learned $\theta, \alpha$ and $\rho$. The reconstruction error (i.e. negative log-likelihood) was calculated. The upper bounds were obtained using reconstruction errors calculated from unmasked test data using GETM with both condition and medication embedding from Node2Vec. Flattened refers to results obtained using ETM for which condition and medication are treated as same features. Description of algorithm names are in table 2.1.

| Algorithm \ Topic # | 15 | 50 | 75 | 100 | 15 | 50 | 75 | 100 |
|---|---|---|---|---|---|---|---|---|
| | recall@5 | | | | precision@5 | | | |
| Upper bound | 0.6672 | 0.7943 | 0.8027 | 7862 | 0.3342 | 0.4084 | 0.4226 | 0.4049 |
| Flattened | 0.4397 | 0.4486 | 0.4667 | 0.4614 | 0.1901 | 0.2109 | 0.2111 | 0.2162 |
| GETM | 0.5543 | 0.5639 | 0.5568 | 0.5388 | 0.2479 | 0.2516 | 0.2504 | 0.2417 |
| GETM + emb._cond | 0.5479 | 0.5664 | 0.5670 | 0.5647 | 0.2373 | 0.2524 | 0.2493 | 0.2486 |
| GETM + emb._med | 0.5519 | 0.5722 | 0.5668 | 0.5716 | 0.2440 | 0.2533 | 0.2521 | 0.2532 |
| GETM + emb._cond+med | **0.5692** | **0.5787** | **0.5732** | **0.5753** | **0.2504** | **0.2612** | **0.2606** | **0.2578** |

Table 3.6: **Performance of medication recommendation**. The medication data was masked for test patient before we reconstructed the medication data with learned $\theta, \alpha$ and $\rho$. Then we chose top 5 medications for each patient and calculated the precision and recall to evaluate the ability of our model to recover medication information. The upper bounds were obtained using reconstruction error calculated from unmasked test data using GETM with both condition and medication embedding from Node2Vec. Description of algorithm names are in table 2.1.

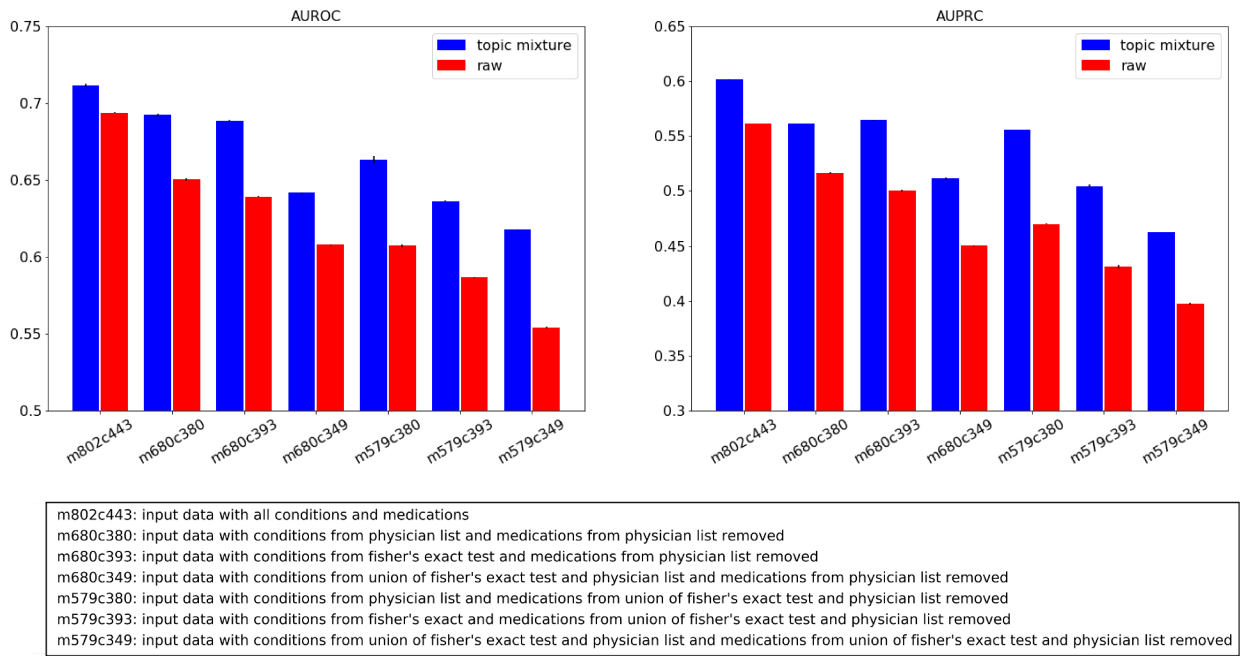


Figure 3.3: **Performance of logistic regression for chronic musculoskeletal pain**. Logistic regression was performed on seven datasets using patient topic mixture $\theta$ with 128 topics as input to predict musculoskeletal pain. The baseline was using raw condition and medication data $X$ (Fig. 1) as input. Here shows the AUROC and AUPRC of two input and all datasets.

## 3.4 Chronic musculoskeletal pain prediction

### 3.4.1 Prediction results

We explored the ability of our model to predict chronic musculoskeletal pain using logistic regression (Fig. 3.3). Specifically, we compared the performance of using the patient topic mixture with the performance directly using condition and medication information on seven datasets which have been described in the **Methods** section. We observed that 1). The topic mixture achieved larger area under the receiver operating characteristic (AUROC) and larger area under the precision-recall curve (AUPRC) for all datasets. 2). As we removed more signature conditions and medications which are related to pain, the performance of using raw features dropped faster than that using the patient topic mixture. The differences between the performance of using patient topic mixture and using raw data got larger without those indicative conditions and medications. The difference increased 0.046 for AUROC and

0.026 for AUPRC if comparing the result using all data (m802c443) with the result using least conditions and medications (m579c379)(Fig. 3.3). Therefore, GETM demonstrated the advantages of using heterogeneous data and incorporating pre-trained embedding to make up for information loss after removing informative conditions and medications to some extent.

### 3.4.2 Pain-related conditions and medications

We investigated the most pain-related conditions and medications based on logistic regression coefficients and calculated overlapping proportions with lists provided by the physician (Fig. 3.4). In comparison with the overlapping proportions from ETM and Fisher's exact test, GETM identified the most conditions (36.7% from top 10 conditions, 33.3% from top 30 conditions and 30.0% from top 50 conditions) and medications (60.0% from top 10 medications, 33.3% from top 30 medications and 32.0% from top 50 medications) in the provided lists. This suggests that GETM improves the ability to extract otherwise hidden associations, which could identify pain-related comorbidities.

### 3.4.3 Topic clustering

We visualized the clustering using UMAP (Fig. 3.5). We randomly chose 5 topics each from condition clusters and medication clusters and then confirmed the topics were assigned to correct clusters which are consistent to the categories of their top features of the topics. Thus, GETM allows identifying feature groups of heterogeneous data in a data-driven manner. We also had a close look at three most positively associated topics and three most negatively associated topics to chronic musculoskeletal pain based on learned $\omega$. It showed that topic 56, 34, and 51 are strongly positively associated with chronic musculoskeletal pain and topic 73, 68, 89 are strongly negatively associated with chronic musculoskeletal pain, respectively. For each topic, we reveal their meaning by top conditions and top medications.

Particularly, topic 56 and 34 contains conditions and medications from musculoskeletal system, which makes clinical sense of that they are highly related to chronic musculoskeletal

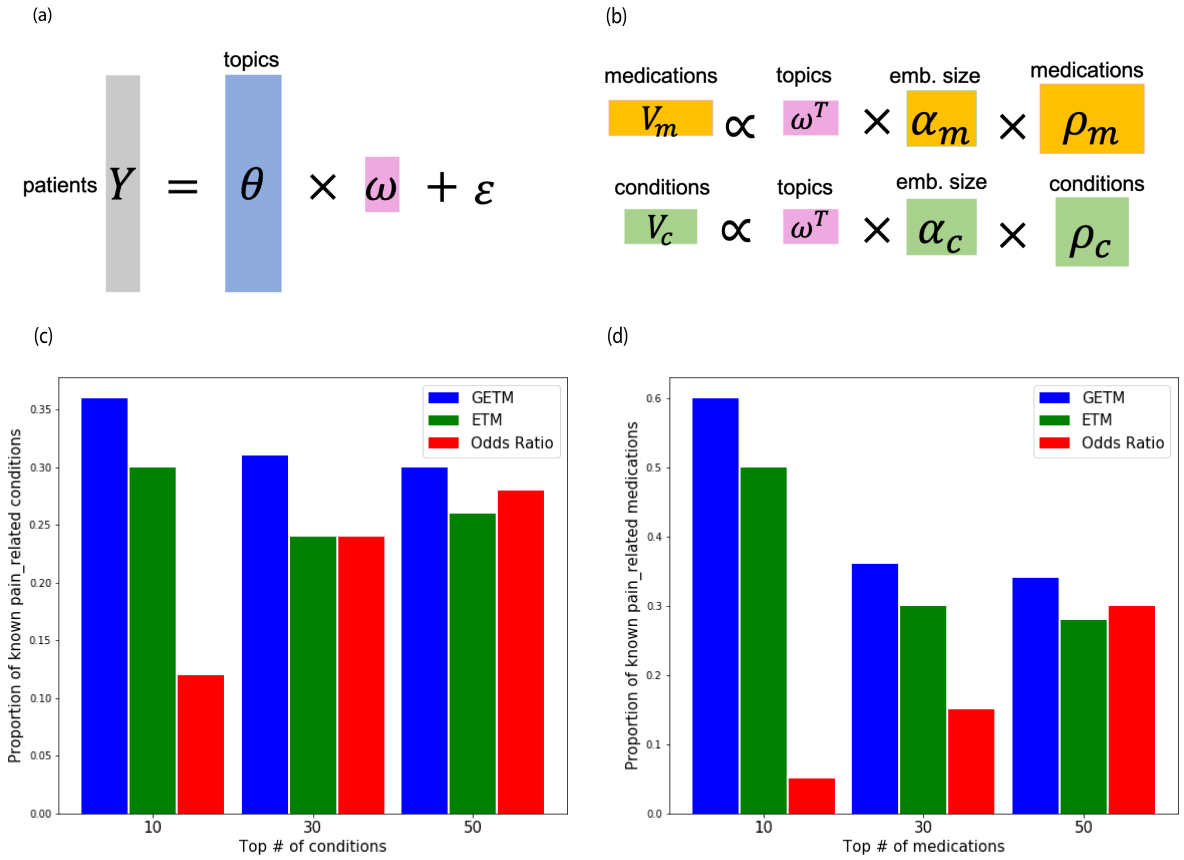


Figure 3.4: **Analysis of chronic pain-related conditions and medications. a**. Logistic regression using patient topic mixture $\theta \in \mathbb{R}^{D\times K}$ of D patients and K topics. **(b)**. Sorting $V_m \in \mathbb{R}^{1\times M}$ obtained from matrix factorization of coefficients $\omega \in \mathbb{R}^{K\times 1}$ in panel (a) , $\alpha_m \in \mathbb{R}^{K\times L_1}$ and $\rho_m \in \mathbb{R}^{L_1\times M}$ from GETM, we could get top relevant medications to chronic musculoskeletal pain among M medications. Similarly, we could get top relevant conditions to chronic musculoskeletal pain among C conditions by selecting top conditions from $V_c \in \mathbb{R}^{1\times C}$ calculated by matrix factorization of coefficients $\omega$ in panel (a) , $\alpha_c \in \mathbb{R}^{K\times L_2}$ and $\rho_c \in \mathbb{R}^{L_2\times C}$. **(c)**. We chose different number of top conditions, and then calculated the percentage of those conditions that are in physician list of pain-related conditions. **(d)**. Similar analysis as in panel (c) for medications. We compared with two baselines: (1). Using ETM which treated conditions and medications as same features and then selecting top medications and conditions all from $V \in \mathbb{R}^{1\times(M+C)}$ (2). Implementing fisher's exact test and picking top conditions and medications as in Fig. 2.2.

pain. Specifically, this combination in topic 34 provides a proof of concept as prolapsed disc or slipped disc as a condition is painful and ibuprofen is an analgesic in the NSAID (non-steroidal anti-inflammatory drug) class. Topic 51 in the cardiovascular category, acetylsalicylic acid is used in prevention of stroke and heart attacks; it acts as a "blood thinner". Dipyridamole inhibits blood clot formation and therefore prevents potential consequences of blood clotting. Atherosclerosis is a process of deposition of fatty material in the walls of arteries, and this thickening leads to an increase in stroke and heart attack risk. A common cause of atherosclerosis is high cholesterol levels. Thus, although the two medications are not directly used as a cure for the condition high cholesterol, by way of atherosclerosis, high cholesterol leads to higher risks for other cardiovascular outcomes and the conditions prevent those outcomes [16].

Our findings also give insight to directions for further investigations. In topic 73, Ramipril, lisinopril and enalapril are ACE inhibitors, used to treat high blood pressure and may be used in response to heart failure and a heart attack. All of the conditions have an allergic component. This is a particularly interesting finding, suggesting that a particular subset of individuals suffering from allergic conditions who also are undergoing cardiovascular treatment are at lower risk for musculoskeletal pain. One of the etiological mechanisms that may underlie chronification of musculoskeletal pain is central sensitization [74]. Given the implication of immune cells in pain signalling [21], it may be that pharmacological intervention on hypertension may define a subcategory of individuals who are thereby protected from musculoskeletal pain. For topic 89, this topic is focused on women given that estrogen taking only applies to women. Hypertension and estrogen taking are associated with protection from musculoskeletal pain chronification. It is possible that taking of estrogens has as a secondary effect the prevention of pain chronification. It is known that menopause with the associated decrease in estrogens is a risk factor for musculoskeletal pain chronification.

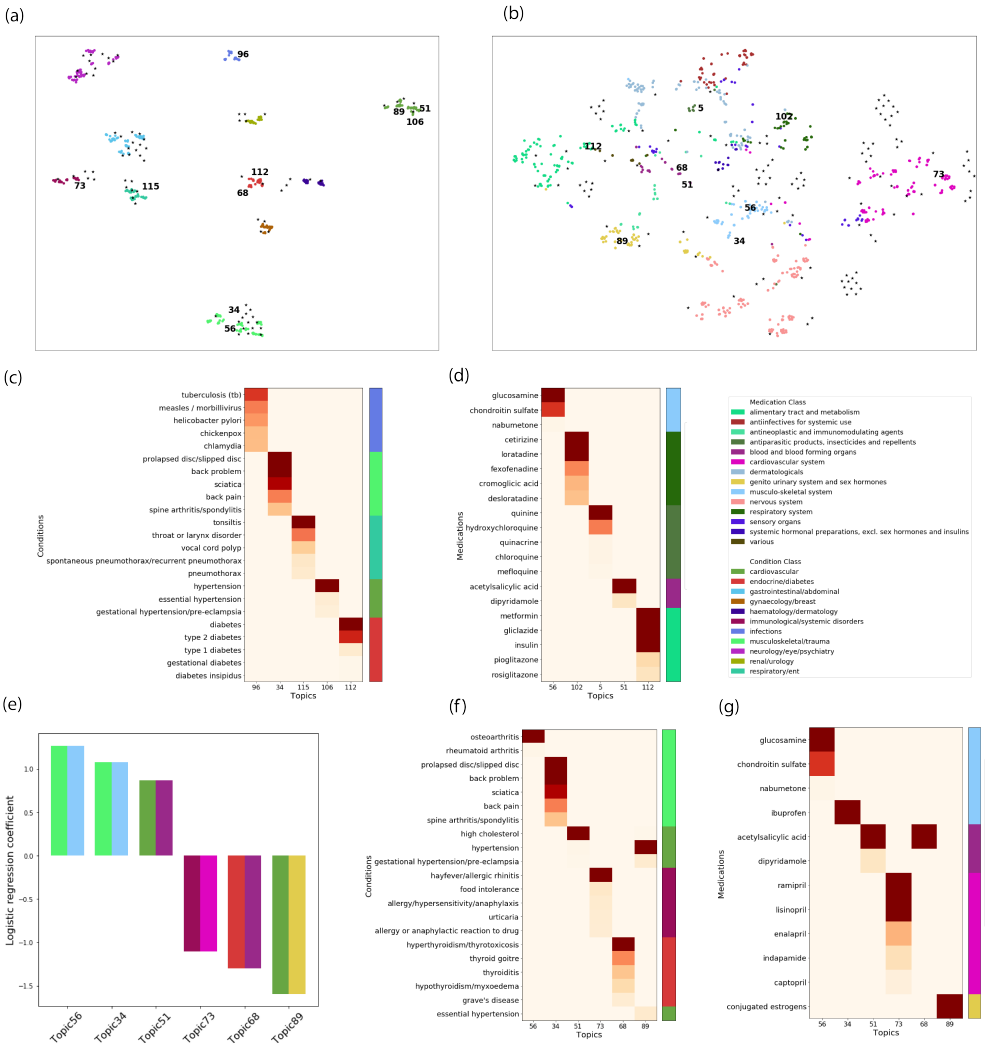


Figure 3.5: **Topic clusters visualization and analysis for chronic musculoskeletal pain**. The analysis was based on results from GETM model with both condition and medication embedding from Node2Vec, which used all condition and medication data and topic number of 128. **(a)**. Condition and topic embedding clusters. **(b)**. Medication and topic embedding clusters. (c) A heatmap of 5 randomly picked topic from panel (a). It is apparent that the category of top conditions for each topic matches its position on panel (a). **(d)**. A heatmap of 5 randomly picked topic from panel (b). The category of top conditions for each topic is consistent with its position on the panel (b). **(e)**. We chose three topics with highest coefficients from logistic regression of musculoskeletal pain prediction and another three with lowest coefficients. Each bar is consisted of two columns colored by condition category on the left and medication category on the right. **(f)**. Top condition visualization of topics in panel (e). **(g)**. Top medication visualization of topics in panel (e).

# Chapter 4

# Discussion

While the data from the UK Biobank is an unparalleled resource to understand chronic pain, its sparsity, heterogeneity and the large size of its data pose great challenges to classical statistical models to get the full benefit from UK Biobank. To address those challenges, we developed GETM and presented its promising performance on different tasks. GETM demonstrates excellent capabilities of extracting hidden information.

## 4.1 Topic quality

By introducing the knowledge graph and simultaneously training the different types of features (i.e. conditions and medications), GETM was able to infer more coherent topics comparing to topics learned without using Node2Vec embedding or without a different type of feature, in a sense that the top medications or conditions of a specific topic are from the same categories. This allows us to interpret topics and any finding related to certain topics with a clear clinical ground. This also proves that our model is a very good tool to identify comorbidities, which is very important in clinical researches. Current researches in comorbidities identification mainly focus on one disease or one group of diseases. [38, 67, 65], while our model could search on many diseases and find various disease groups at the same time. Another thing that worthy mentioning is that though there have been researches

utilizing a multimodal topic model to find meaningful latent topics, the data they have used was health insurance claim records with many contextual information and many more tokens compared to our 802 medications and 443 conditions [45]. In comparison, we pulled out meaningful topics with medication usage and condition history of individuals. This suggests broader application of our model to datasets without text descriptions.

## 4.2 Study of medication and condition relations

As for link extraction between conditions and medications, many existing methods need to feed their models the drug-disease network information to predict relationships [42, 56, 70]. In contrast, we could extract meaningful pairs without feeding the model links between conditions and medications ahead of training. The findings also give insight on possible associations and pathways among conditions and medications.

## 4.3 Data imputation

In terms of data imputation, GETM achieved lower reconstruction error on 50% held-out data and also gave most precise medication recommendations given only condition information. Besides, it turned out that many medications GETM recommended which were not taken by participants actually have treatment effects on conditions they had. This is the advantage that our model introduced that the model could learn from large population and give better medication suggestions to certain individuals.

## 4.4 Chronic musculoskeletal pain prediction

We applied GETM to predict chronic musculoskeletal pain. GETM has made more accurate prediction across datasets compared to prediction results obtained using raw data. In addition, its predicting power is less sensitive to the removal of informative conditions and

medications, which offered a practical use case to predict chronic musculoskeletal pain for those individuals with no obvious pain symptoms. As we compared the top pain-related condition and medication lists to those created by the physician based on background knowledge, the lists created using GETM topic mixtures overlapped the most with physician lists. This result puts forward the potential for GETM to be applied in associative analysis to draw more hidden associations.

Finally, the clustering visualization by UMAP demonstrated that GETM could assign topics to condition or medication clusters that are most representative of the topics. This is beneficial to classifying conditions or medications without category information. Additionally, the combinations of conditions and medications of topics make clinical sense of why they are strongly associative with chronic musculoskeletal pain positively and negatively. This result implies that there might lie clinical grounds of any undiscovered combinations in those topics.

## 4.5 Future work

One limitation of our model is that it did not take the temporal information of data into account. The development of conditions as the participants' age increases could not be observed using our method. However, the UK Biobank provides valuable information such as the age of participants when conditions were first diagnosed, multi-visit records for same participants, etc. Therefore, one future direction is to incorporate time series as part of the model so that we will be able to gain a dynamic understanding of associations.

Besides, we will design an end-to-end training system to combine Node2Vec and GETM. Putting two models together might take better advantage of information sharing and further improves the topic quality and performance in downstream tasks. We will also choose a more appropriate approach to evaluate the condition-defined topic. One proposed metric is to calculate the co-occurence probability for conditions of the same topic in PubMed

literature.

Though Node2Vec improves the overall performance of our model, it ignores the hierarchy in our condition and medication trees. The representations learned by Node2Vec thus might not be able to capture that hierarchical information. It is worth experimenting using a hierarchical graph neural network [66] to learn both trees which might further benefits the learning process by adding more information.

Since our method could successfully find known condition-medication links, it is then worthwhile to examine whether there are any real relationships in those novel condition-medication pairs. Besides, we have also proved, using all conditions and medications, that we could find meaningful condition-medication combinations that are related to chronic musculoskeletal pain. There are two directions to extend this analysis: 1). Removing those signature conditions and medications to find out novel combinations that potentially have impact on chronic musculoskeletal pain; 2) Creating lists for different pain labels and conducting similar analysis. Investigating intersecting and non-intersecting combinations that have impact on different pain types will enable us to have more detailed ideas on how conditions and medications affect specific pain outcomes.

Lastly, although we used pain as a case study, our GETM can be used to characterize other phenotypes in UK Biobank or other data. We will explore more applications of GETM in our future work.

# Bibliography


[1] Comparative Toxicogenomics Database kernel description. http://ctdbase.org/. Accessed: 2021-07-31.

[2] DRUGBANK kernel description. https://go.drugbank.com/. Accessed: 2021-07-31.

[3] Uk biobank. https://www.ukbiobank.ac.uk/. Accessed: 2021-07-31.

[4] Uk biobank data-field 20002: non-cancer illness code, self-reported. https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=20002. Accessed: 2021-07-31.

[5] Uk biobank data-field 20003: treatment/medication code. https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=20003. Accessed: 2021-07-31.

[6] Uk biobank data-field 2956: general pain for 3+ months for 3+ months. https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=2956. Accessed: 2021-07-31.

[7] Uk biobank data-field 3404: neck/shoulder pain for 3+ months for 3+ months. https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=3404. Accessed: 2021-07-31.

[8] Uk biobank data-field 3414: hip pain for 3+ months for 3+ months. https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=3414. Accessed: 2021-07-31.

[9] Uk biobank data-field 3571: back pain for 3+ months. https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=3571. Accessed: 2021-07-31.

[10] Uk biobank data-field 3741: stomach/abdominal pain for 3+ months for 3+ months. https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=3741. Accessed: 2021-07-31.

[11] Uk biobank data-field 3773: knee pain for 3+ months for 3+ months. https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=3773. Accessed: 2021-07-31.

[12] Uk biobank data-field 3799: headaches for 3+ months. https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=3799. Accessed: 2021-07-31.

[13] Uk biobank data-field 4067: facialpain for 3+ months for 3+ months. https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=4067. Accessed: 2021-07-31.

[14] Uk biobank data-field 6159: pain type(s) experienced in last month. https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=6159. Accessed: 2021-07-31.

[15] WHO Collaborating Centre for Drug Statistics Methodology kernel description. https://www.whocc.no/. Accessed: 2021-07-31.

[16] Sameer Al-Ghamdi, Mamdouh M. Shubair, Ashraf El-Metwally, Majid Alsalamah, Saeed Mastour Alshahrani, Badr F Al-Khateeb, Salwa Bahkali, Sara M. Aloudah, Jamaan Al-Zahrani, Turky H. Almigbal, and Khaled K. Aldossari. The relationship between chronic pain, prehypertension, and hypertension. a population-based cross-sectional survey in al-kharj, saudi arabia. *Postgraduate Medicine*, 133(3):345–350, 2021. PMID: 33317375.

[17] Enrique Alfonseca, Katja Filippova, Jean-Yves Delort, and Guillermo Garrido. Pattern learning for relation extraction with a hierarchical topic model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, page 54–59, USA, 2012. Association for Computational Linguistics.

[18] C. Arnold, S. El-Saden, A. Bui, and R. Taira. Clinical case-based retrieval using latent topic analysis. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2010:26–30, 2010.

[19] B. Arnow, C. Blasey, Janelle Y Lee, B. Fireman, E. Hunkeler, R. Dea, R. Robinson, and C. Hayward. Relationships among depression, chronic pain, chronic disabling pain, and medical costs. *Psychiatric Services*, 60:344–350, 2009.

[20] Matthew J. Bair, Rebecca L. Robinson, Wayne Katon, and Kurt Kroenke. Depression and pain comorbidity: A literature review. *Archives of Internal Medicine*, 163(20):2433–2445, 11 2003.

[21] Pankaj Baral, Swalpa Udit, and I. Chiu. Pain and immunity: implications for host defence. *Nature Reviews Immunology*, pages 1–15, 2019.

[22] Arnab Bhadury, Jianfei Chen, Jun Zhu, and Shixia Liu. Scaling up dynamic topic models. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 381–390, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.

[23] David Blei, Lawrence Carin, and David Dunson. Probabilistic topic models: A focus on graphical model design and applications to document and image analysis. *IEEE signal processing magazine*, 27(6):55–65, 2010. PMID: 4122269.

[24] David M. Blei, Michael I. Jordan, Thomas L. Griffiths, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, page 17–24, Cambridge, MA, USA, 2003. MIT Press.

[25] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, March 2003.

[26] Marie T Brown and Jennifer K Bussell. Medication adherence: Who cares? *Mayo Clinic proceedings*, 86(4):304–314, 2011.

[27] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 10 2018.

[28] Fiona Campbell, Maria Hudspith, Melissa Anderson, Chiniére Manon, Hani El-Gabalawy, Jacques Laliberté, Jaris Swidrovich, and Linda Wilhelm. Chronic pain in canada: Laying a foundation for action. Technical report, Health Canada, 2019.

[29] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. GRAM: graph-based attention model for healthcare representation learning. *CoRR*, abs/1611.07012, 2016.

[30] R. Collins. What makes uk biobank special? *The Lancet*, 379:1173–1174, 2012.

[31] Jessica A Davis, Rebecca L Robinson, Trong Kim Le, and Jin Xie. Incidence and impact of pain conditions and comorbid illnesses. *Journal of pain research*, 4:331–345, 2011.

[32] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. The dynamic embedded topic model. *CoRR*, abs/1907.05545, 2019.

[33] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic modeling in embedding spaces. *CoRR*, abs/1907.04907, 2019.

[34] M. Dueñas, B. Ojeda, A. Salazar, J. Micó, and I. Failde. A review of chronic pain impact on patients, their social environment and the health care system. *Journal of Pain Research*, 9:457 – 467, 2016.

[35] Huseyin Melih Elibol, Vincent Nguyen, Scott Linderman, Matthew Johnson, Amna Hashmi, and Finale Doshi-Velez. Cross-corpora unsupervised learning of trajectories in autism spectrum disorders. *J. Mach. Learn. Res.*, 17(1):4597–4634, jan 2016.

[36] A Fayaz, P Croft, R M Langford, L J Donaldson, and G T Jones. Prevalence of chronic pain in the uk: a systematic review and meta-analysis of population studies. *BMJ Open*, 6(6), 2016.

[37] James Foulds, Levi Boyles, Christopher DuBois, Padhraic Smyth, and Max Welling. Stochastic collapsed variational bayesian inference for latent dirichlet allocation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 446–454, New York, NY, USA, 2013. Association for Computing Machinery.

[38] R. Groen, O. Ryan, J. Wigman, H. Riese, B. Penninx, E. Giltay, M. Wichers, and C. Hartman. Comorbidity between depression and anxiety: assessing the role of bridge mental states in dynamic psychological networks. *BMC Medicine*, 18, 2020.

[39] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. *CoRR*, abs/1607.00653, 2016.

[40] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.

[41] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[42] Sumit Kumar, Raj Ratn Pranesh, and Ambesh Shekhar. Biomedical network link prediction using neural network graph embedding. In *8th ACM IKDD CODS and 26th COMAD*, CODS COMAD 2021, page 412, New York, NY, USA, 2021. Association for Computing Machinery.

[43] Dongha Lee, Xiaoqian Jiang, and Hwanjo Yu. Harmonized representation learning on dynamic ehr graphs. *Journal of Biomedical Informatics*, 106:103426, 2020.

[44] Y. Li, Pratheeksha Nair, Xing Han Lu, Zhi Wen, Yuening Wang, Amir Ardalan Kalantari Dehaghi, Yanchun Miao, Weiqi Liu, T. Ordog, J. Biernacka, E. Ryu, J. Olson, M. Frye, Aihua Liu, Liming Guo, A. Marelli, Y. Ahuja, J. Davila-Velderrain, and Manolis Kellis. Inferring multimodal latent topics from electronic health records. *Nature Communications*, 11, 2020.

[45] Hsin-Min Lu, Chih-Ping Wei, and Fei-Yuan Hsiao. Modeling healthcare data using multiple-channel latent dirichlet allocation. *Journal of Biomedical Informatics*, 60:210–223, 2016.

[46] Manu Madhavan and Gopakumar G. A tf-idf based topic model for identifying lncrnas from genomic background. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, SAC '18, page 40–46, New York, NY, USA, 2018. Association for Computing Machinery.

[47] Jonathan Masci, Emanuele Rodolà, Davide Boscaini, Michael M. Bronstein, and Hao Li. Geometric deep learning. In *SIGGRAPH ASIA 2016 Courses*, SA '16, New York, NY, USA, 2016. Association for Computing Machinery.

[48] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[49] Tomas Mikolov, Kai Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

[50] Thin Nguyen, Hang Le, and S. Venkatesh. Graphdta: prediction of drug–target binding affinity using graph convolutional networks. 2019.

[51] Bruce Nicholson and Sunil Verma. Comorbidities in chronic neuropathic pain. *Pain Medicine*, 5(suppl_1):S9–S27, 02 2004.

[52] Zhenxing Niu, Gang Hua, Xinbo Gao, and Qi Tian. Context aware topic model for scene recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2743–2750, 2012.

[53] Sungrae Park, Wonsung Lee, and Il-Chul Moon. Supervised dynamic topic models for associative topic extraction with a numerical time series. In *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*, TM '15, page 49–54, New York, NY, USA, 2015. Association for Computing Machinery.

[54] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 569–577, New York, NY, USA, 2008. Association for Computing Machinery.

[55] Florian Privé, Keurcien Luu, Michael G B Blum, John J McGrath, and Bjarni J Vilhjálmsson. Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics*, 36(16):4449–4457, 05 2020.

[56] Farshid Rayhan, Sajid Ahmed, Zaynab Mousavian, Dewan Md Farid, and Swakkhar Shatabda. Frnet-dti: Deep convolutional neural network for drug-target interaction prediction. *Heliyon*, 6(3):e03444, 2020.

[57] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models, 2014.

[58] Andrew K. Rider and Nitesh V. Chawla. An ensemble topic model for sharing healthcare data and predicting disease risk. In *Proceedings of the International Conference on*

*Bioinformatics, Computational Biology and Biomedical Informatics*, BCB'13, page 333–340, New York, NY, USA, 2013. Association for Computing Machinery.

[59] Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. Learning author-topic models from text corpora. *ACM Trans. Inf. Syst.*, 28(1), January 2010.

[60] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, page 487–494, Arlington, Virginia, USA, 2004. AUAI Press.

[61] S. Salduker, E. Allers, S. Bechan, R. Hodgson, F. Meyer, H. Meyer, J. Smuts, E. Vuong, and D. Webb. Practical approach to a patient with chronic pain of uncertain etiology in primary care. *Journal of Pain Research*, 12:2651 – 2662, 2019.

[62] J. Sareen, B. Cox, I. Clara, and G. Asmundson. The relationship between anxiety disorders and physical disorders in the u.s. national comorbidity survey. *Depression and Anxiety*, 21, 2005.

[63] Marc-Andre Schulz, B. Yeo, J. Vogelstein, Janaina Mourao-Miranada, J. Kather, Konrad Paul Kording, Blake A. Richards, and D. Bzdok. Different scaling of linear models and deep learning in ukbiobank brain images versus machine-learning datasets. *Nature Communications*, 11, 2020.

[64] Rui Shu, Hung H. Bui, Shengjia Zhao, Mykel J. Kochenderfer, and Stefano Ermon. Amortized inference regularization, 2019.

[65] Umesh Singh, Victoria Wangia-Anderson, and J. Bernstein. Chronic rhinitis is a high-risk comorbidity for 30-day hospital readmission of patients with asthma and chronic obstructive pulmonary disease. *The journal of allergy and clinical immunology. In practice*, 7 1:279–285.e6, 2019.

[66] Stanislav Sobolevsky. Hierarchical graph neural networks. *CoRR*, abs/2105.03388, 2021.

[67] Jia Song, M. Zeng, Hai Wang, C. Qin, H. Hou, Zi-Yong Sun, San-Peng Xu, Guo ping Wang, Cuilian Guo, Yi ke Deng, Zhi chao Wang, J. Ma, L. Pan, B. Liao, Zhi-Hui Du, Q. Feng, Y. Liu, Jun-Gang Xie, and Z. Liu. Distinct effects of asthma and copd comorbidity on disease expression and outcome in patients with covid-19. *Allergy*, 76:483 – 496, 2020.

[68] Ziyang Song, Xavier Sumba Toral, Yixin Xu, Aihua Liu, Liming Guo, Guido Powell, Aman Verma, David Buckeridge, Ariane Marelli, and Yue Li. Supervised multi-specialist topic model with applications on large-scale electronic health record data. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '21, New York, NY, USA, 2021. Association for Computing Machinery.

[69] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models, 2017.

[70] Mehmet Tan, Ozan Fırat Özgül, Batuhan Bardak, Işıksu Ekşioğlu, and Suna Sabuncuoğlu. Drug response prediction by ensemble learning and drug-induced gene expression signatures. *Genomics*, 111(5):1078–1088, 2019.

[71] Yosuke Tanigawa, Jiehan Li, Johanne Marie Justesen, Heiko Horn, Matthew Aguirre, Christopher M. DeBoever, Chris Chang, Balasubramanian Narasimhan, Kasper Lage, Trevor J. Hastie, Chong Y. Park, Gill Bejerano, Erik Ingelsson, and Manuel A. Rivas. Components of genetic associations across 2,138 phenotypes in the uk biobank highlight adipocyte biology. *Nature Communications*, 10, 2019.

[72] Macfarlane TV, Beasley M, and Macfarlane GJ. Self-reported facial pain in uk biobank study: Prevalence and associated factors. *J Oral Maxillofac Res*, 5, 2014.

[73] Miriam S Udler, Jaegil Kim, Marcin von Grotthuss, Bonás-Guarch, and et. al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLOS Medicine*, 15(9):e1002654, 2018.

[74] C. Woolf. Central sensitization: Implications for the diagnosis and treatment of pain. *PAIN*, 152:S2–S15, 2011.

[75] Andrzej Wróbel, Anna Serefko, Andrzej Woźniak, Jacek Kociszewski, Aleksandra Szopa, Radosław Wiśniewski, and Ewa Poleszak. Duloxetine reverses the symptoms of overactive bladder co-existing with depression via the central pathways. *Pharmacology Biochemistry and Behavior*, 189:172842, 2020.

[76] Yonghui Wu, M. Liu, W. J. Zheng, Zhongming Zhao, and Hua Xu. Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 422–33, 2012.

[77] Zhang Y, Shen F, Mojarad MR, Li, Liu S, Tao C, Yu Y, and Liu H. Systematic identification of latent disease-gene associations from pubmed articles. *PLoS One*, 2018.

[78] Ronghui You, Shuwei Yao, Hiroshi Mamitsuka, and Shanfeng Zhu. DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction. *Bioinformatics*, 37(Supplement_1):i262–i271, 07 2021.

[79] Juan Zhao, QiPing Feng, Patrick Wu, Jeremy L. Warner, Joshua C. Denny, and Wei-Qi Wei. Using topic modeling via non-negative matrix factorization to identify relationships between genetic variants and disease phenotypes: A case study of lipoprotein(a) (lpa). *bioRxiv*, 2018.