



McGill

Disentangling chronic pain genetic heterogeneity from UK Biobank data using Graph-Embedded Topic Model

Hsuan Megan Tsao

Department of Human Genetics,

Faculty of Medicine and Health Sciences,

McGill University, Montreal, Quebec, Canada

April 2025

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

© Hsuan Megan Tsao, 2025

*This work is dedicated to my family,
my mother Ms. Yin-Ling Lee, and my father Mr. Hui-Tung Tsao,
for their endless love, support, and encouragement throughout my life.*

Abstract

Chronic pain, defined as pain persisting for three or more months, is a multifactorial problem with a heterogeneous genetic architecture and diverse phenotypic expressions, including chronic musculoskeletal pain (CMSKP). Individuals with chronic pain often present with comorbidities, most commonly psychiatric disorders, cardiovascular conditions, and autoimmune diseases. Standard genome-wide association (GWA) studies typically focus on single phenotypes, potentially overlooking shared genetic and comorbidity patterns. To address this limitation, we explored the application of Graph-Embedded Topic Modeling (GETM) (Wang et al., iScience 2022) to UK Biobank self-reported non-cancer conditions, medications, and hospital-derived ICD-10 codes to identify latent comorbidity patterns. Using data from 401,013 individuals of European ancestry from UK Biobank, we derived the probability of individuals by topics (θ) and assessed the feasibility of using θ as the phenotypic input for the GWA scans to investigate genetic associations underlying CMSKP. We used REGENIE to run GWA scans adjusting for age and sex and Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) for post-GWA scan analysis. While our study demonstrated that GETM effectively reduces high-dimensional clinical data into structured topics, topics lacked biologically meaningful features related to CMSKP, leading to the failure of identifying novel genetic associations. Nevertheless, this study introduces a novel framework to use GETM-derived GWA scan to study CMSKP, highlighting the potential and challenges of leveraging multimorbidity patterns to study the genetic architecture of complex traits.

Résumé

La douleur chronique, définie comme une douleur persistant pendant trois mois ou plus, est un problème multifactoriel dont l'architecture génétique est hétérogène et les expressions phénotypiques diverses, y compris la douleur musculo-squelettique chronique (CMSKP). Les personnes souffrant de douleur chronique présentent souvent des comorbidités, le plus souvent des troubles psychiatriques, des affections cardiovasculaires et des maladies auto-immunes. Les études standard d'association à l'échelle du génome (GWA) se concentrent généralement sur des phénotypes uniques, négligeant potentiellement les schémas génétiques et de comorbidité partagés. Pour remédier à cette limitation, nous avons exploré l'application de Graph-Embedded Topic Modeling (GETM) (Wang et al., iScience 2022) aux affections non cancéreuses autodéclarées de la UK Biobank, aux médicaments et aux codes CIM-10 dérivés des hôpitaux, afin d'identifier des schémas de comorbidité latents. En utilisant les données de 401 013 individus d'ascendance européenne de la UK Biobank, nous avons dérivé les individus par thèmes (θ) et évalué la faisabilité de l'intégration de θ avec les analyses GWA pour étudier les associations génétiques sous-jacentes à la CMSKP. Nous avons utilisé REGENIE pour effectuer des analyses GWA ajustées en fonction de l'âge et du sexe et Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) pour l'analyse post-GWA. Bien que notre étude ait démontré que le GETM réduit efficacement les données cliniques à haute dimension en sujets structurés, les sujets manquaient de caractéristiques biologiquement significatives liées au CMSKP, ce qui a conduit à l'échec de l'identification de nouvelles associations génétiques. Néanmoins, cette étude introduit un nouveau cadre d'utilisation de l'analyse GWA dérivée du GETM pour étudier le CMSKP, soulignant le potentiel et les difficultés de l'exploitation des schémas de multimorbidité pour étudier l'architecture génétique de traits complexes.

Table of Contents

<i>Abstract</i>	3
<i>Résumé</i>	4
<i>Table of Contents</i>	5
<i>List of Abbreviations</i>	8
<i>List of Figures</i>	10
<i>List of Tables</i>	12
<i>Acknowledgments</i>	13
<i>Contribution of Authors</i>	14
<i>Chapter 1 – Introduction</i>	15
1.1 Pain	15
1.2 Chronic Pain	15
1.2.1 Chronic Musculoskeletal Pain	16
1.3 Biobank Era	17
1.3.1 Emergence of Large-Scale Biobanks	17
1.3.2 Leveraging Biobank Data in Pain Research.....	17
1.4 Medical Ontologies	18
1.4.1 Self-Reported Non-Cancer Conditions	18
1.4.2 Treatment/Medication Code.....	19
1.4.3 Hospital-derived ICD-10 Codes.....	20
1.5 Chronic Pain Genetics	21

1.5.1	Genetic Basis of Chronic Pain	21
1.5.2	Genome-wide Association Studies (GWAS) in Chronic Pain	22
1.6	Topic Modeling	24
1.6.1	Common Topic Models	24
1.6.2	Topic Modeling in the Biomedical Context	26
1.7	Objectives and Hypothesis.....	26
Chapter 2 – Materials and Methods.....		28
2.1	Data Availability	28
2.2	Graph-Embedded Topic Modeling	29
2.3	Topic Selection for GWA Scans	33
2.4	Comparison of GWA Scans	33
2.5	Genetic Analysis.....	34
Chapter 3 – Results		35
3.1	Summary Statistics.....	35
3.2	Topic Selection for GWA Scans.....	38
3.2.1	Evaluation of Topic Quality	38
3.2.2	Identifying the Most Discriminating Topic for CMSKP.....	42
3.3	GWAS of CMSKP Using Binary Case-Control Design.....	44
3.4	Refined GWAS of CMSKP with Stricter Case and Control Definitions.....	48
3.5	GWAS of CMSKP individuals Using Topic 9 Theta Values as a Continuous Trait.....	49
3.6	GWAS of CMSKP Case-only Using Topic 9 Theta Values as a Continuous Trait.....	50
Chapter 4 – Discussion		52

4.1	CMSKP Findings Before Incorporating GETM	52
4.2	Insights into GETM Development and Application.....	54
4.2.1	Challenges in Including Medication Data for Phenotyping	54
4.2.2	Embedding Approaches in GETM: Strengths and Limitations	56
4.2.3	Evaluating Topic Quality: Methodological Insights	57
4.3	Application of GETM-GWAS to Pain Genomics	58
4.3.1	Relevance of GETM-derived GWA Scans in Studying CMSKP.....	58
4.4	Overall Strengths and Limitations of this Study	59
<i>Chapter 5 – Conclusions and Future Directions.....</i>		<i>61</i>
<i>Chapter 6 – References.....</i>		<i>63</i>

List of Abbreviations

ARI – Adjusted Rand Index

ATC – Anatomical Therapeutic Chemical Classification

CMSKP – Chronic Musculoskeletal Pain

EHR – Electronic Health Record

ETM – Embedded Topic Model

FUMA – Functional Mapping and Annotation of Genome-Wide Association Studies

GETM – Graph-Embedded Topic Modeling

GWA – Genome-wide Association

HES – Hospital Episode Statistics

HLA – Human Leukocyte Antigen

HRC – Haplotype Reference Consortium

IASP – International Association for the Study of Pain

MHC – Major Histocompatibility Complex

NNI – Normalized Mutual Information

LDA – Latent Dirichlet Allocation

MAF – Minor Allele Frequency

MAGMA – Multi-marker Analysis of GenoMic Annotation

MCP – Multisite Chronic Pain

NHS – National Health Service

SNP – Single Nucleotide Polymorphisms

treeLFA – Tree-structured Logistic Factor Analysis

UMAP – Uniform Manifold Approximation and Projection

WHO – World Health Organization

List of Figures

Figure 1. An example of the ATC code, adapted from the WHO website on the ATC classification system	20
Figure 2. An example of the ICD-10 code, adapted from Figure 1 of Sammani et al. (2021)	21
Figure 3. Overview of GETM	30
Figure 4. Mathematical representation of GETM	32
Figure 5. Overall study design	36
Figure 6. (A) Histogram and table of condition counts. (B) Histogram and table of medication counts. (C) Histogram and table of ICD-10 code counts	37
Figure 7. UMAP for condition embedding and topic embedding	39
Figure 8. UMAP for ICD-10 embedding and topic embedding	39
Figure 9. UMAP for medication embedding and topic embedding	40
Figure 10. Heatmap displaying the top 5 conditions for a random selection of 5 out of the 10 topics, with the color intensity representing the probability of each condition belonging to its respective topic	42
Figure 11. (A) Manhattan plot of GWA scan summary statistics between CMSKP cases and controls. Each point represents a SNP. The red horizontal line indicates the genome-wide significance threshold ($P = 5 \times 10^{-8}$); (B) QQ plot of GWA scan summary statistics between CMSKP cases and controls	44

Figure 12. (A) Manhattan plot; (B) QQ plot of GWA scan gene-based test between CMSKP cases and controls.....	45
Figure 13. MAGMA Tissue Expression Analysis – GTEx v8 53 tissue types	46
Figure 14. Differentially Expressed Genes – GTEx v8 53 tissue types	47
Figure 15. Gene Expression Heatmap of the top 5 genes	47
Figure 16. (A) Manhattan plot; (B) QQ plot of GWA scan summary statistics between stricter CMSKP cases and controls.....	48
Figure 17. (A) Manhattan plot; (B) QQ plot of GWA scan gene-based test between stricter CMSKP cases and controls	49
Figure 18. (A) Manhattan plot; (B) QQ plot of GWA scan summary statistics for CMSKP individuals with theta values.....	50
Figure 19. Figure 19. (A) Manhattan plot; (B) QQ plot of GWA scan gene-based test for CMSKP individuals with theta values	50
Figure 20. (A) Manhattan plot; (B) QQ plot of GWA scan summary statistics for CMSKP cases with Topic 9 theta values.....	51
Figure 21. (A) Manhattan plot; (B) QQ plot of GWA scan gene-based test for CMSKP cases with Topic 9 theta values.....	51

List of Tables

Table 1. Summary of chronic musculoskeletal pain by body site and as a unified phenotype between European individuals and European individuals with at least one conditions, medications, or ICD-10 codes	38
Table 2. ICD-10-specific and condition-specific topic quality evaluation across 10, 20, 30, 40, and 50 topics; SD: Standard Deviation; coh: coherence; div: diversity	41
Table 3. T-test results for identifying the most predictive topic for GWA scans in CMSKP	43

Acknowledgments

First, I would like to express my sincere gratitude to my supervisor, Dr. Audrey V. Grant. Her encouragement to persevere through challenges and her belief in my ability to overcome obstacles have been invaluable. Dr. Grant inspired me to view difficulties not as setbacks but as opportunities for growth and learning. Her continued trust and unwavering support have been crucial in shaping my academic development and resilience throughout this journey. Second, I am also thankful to my Supervisory Committee members, Dr. Luda Diatchenko and Dr. Marc O. Martel. Their constructive suggestions and thoughtful reviews have played a pivotal role in improving the quality of this research. Their diverse expertise and perspectives have greatly enhanced my understanding of pain. Third, I am grateful to Dr. Yue Li from the Department of Computer Science. His guidance in computational methods and data analysis has been crucial in advancing the technical aspects of this study. His patience in explaining complex concepts and his willingness to assist with problem-solving have been indispensable. Fourth, I would like to thank my colleagues, Dr. Goodarz Koli Farhood, Marc Parisien, Rachael Osagie, Bangli Cao, and Amandeep Kaur. Goodarz's expertise in hands-on genetic analysis has greatly supported my work. I also want to thank Rachael Osagie for her unwavering mental support. Her encouragement and thoughtful conversations motivated me throughout this journey. A heartfelt thank you goes to all the Alan Edwards Centre for Research on Pain members. Their collective expertise and the interdisciplinary resources they provided greatly flourished my research journey here. Last but not least, I am proud to be a student in the Department of Human Genetics at McGill University. The academic environment, resources, faculty, and staff, particularly Ross MacKay and Rimi Joshi, have made my research experience both stimulating and fulfilling.

Contribution of Authors

All chapters in this thesis were entirely written by me, Hsuan Megan Tsao. All research was conducted independently, under the supervision of Dr. Audrey Grant.

Chapter 1 – Introduction

1.1 Pain

Pain is a complex health issue that often acts as an initial signal of underlying medical conditions, accounting for approximately 78% of emergency department visits (Todd et al., 2007). It arises from diverse biological mechanisms and presents various forms and intensities across individuals. Pain can be categorized into distinct types based on its origin: nociceptive pain, resulting from tissue damage (Gold & Gebhart, 2010); neuropathic pain, caused by lesions in the somatosensory system (Colloca et al., 2017); and inflammatory pain, driven by the activation of immune cells (Baral et al., 2019). These classifications highlight the complexity of pain. In addition to its physiological effects, pain significantly diminishes people's quality of life, leading to significant socioeconomic and public health burdens (Henschke et al., 2015). As a result, advancing the understanding of pain is an urgent and essential priority for public health.

1.2 Chronic Pain

Chronic pain is defined as pain that persists or recurs for three or more months. Individuals experiencing chronic pain frequently present with comorbid conditions, particularly psychiatric disorders, cardiovascular diseases, and autoimmune conditions (Foley et al., 2021). In clinical observations, depression, anxiety, and substance use disorders are highly prevalent among chronic pain patients, with a bidirectional relationship potentially driven by shared neural mechanisms and overlapping genetic risk factors (Hooten, 2016; Johnston & Huckins, 2023; Weihua Meng et al., 2020). Chronic pain is also associated with an increased risk of coronary artery disease, myocardial infarction, heart failure, and stroke (Lin et al., 2023; Rönnegård et al., 2022). Additionally, reciprocal interactions between neural and immune systems contribute to associations between

chronic pain and autoimmune disorders, such as multiple sclerosis and rheumatoid arthritis (Tang et al., 2023). These comorbidities present significant challenges for studying and managing chronic pain.

The International Association for the Study of Pain (IASP) Task Force categorizes chronic pain in three ways: (1) by etiology, such as cancer-related pain; (2) by pathophysiological mechanisms, such as neuropathic pain; and (3) by body site, such as back pain (Treede et al., 2015). In this thesis, we studied pain by body site, as it is the most common assessment recorded in biobanks.

Chronic pain can affect various body sites. In a large population-based study of ~500,000 individuals, the UK Biobank, which is the primary data resource for this study, chronic pain was assessed at seven body sites, including headache, facial pain, neck or shoulder pain, back pain, stomach or abdominal pain, hip pain, knee pain, and pain all over the body. The data was collected by asking the participants: “Have you had *** pain for more than 3 months?”

1.2.1 Chronic Musculoskeletal Pain

Chronic musculoskeletal pain (CMSKP), defined as persistent pain arising from joints, bones, and muscles, is a leading cause of disability worldwide, affecting about 1.71 billion people (El-Tallawy et al., 2021). The definition and classification of CMSKP vary across studies and literature, reflecting the complexities and inconsistencies in how this condition is characterized (El-Tallawy et al., 2021; Hodges et al., 2023). For this study, we defined CMSKP based on common occurrences reported in the literature, focusing on chronic pain localized to one or more of four body sites: knee, back, neck/shoulder, or hip (Pan et al., 2019). Given the shared characteristics of these sites, such as the involvement of musculoskeletal tissues and similar

pathological mechanisms, we analyzed them as a unified phenotype. This approach enhances the statistical power of the analysis by maximizing the sample size and enables a binary classification of chronic pain cases.

1.3 Biobank Era

1.3.1 Emergence of Large-Scale Biobanks

Large-scale biobanks have transformed biomedical research by enabling investigations of genetic and environmental factors influencing health and disease. National biobanks in the UK, Canada, Finland, Japan, Taiwan, and the United States systematically collect biological samples, demographic data, and longitudinal health records, supporting studies on disease risk, progression, and outcomes (Coppola et al., 2019; De Souza & Greenspan, 2013). Pioneering biobanks such as the UK Biobank, which includes data from approximately 500,000 participants, have featured in over 14,000 findings (Free, 2024). Other major biobanks, including the All of Us Research Program in the US and the FinnGen project in Finland, continue to expand the diversity of participant populations, enhancing cross-biobank research opportunities (Bick et al., 2024). By integrating genetic, imaging, and health record data, these biobanks advance the study of multifactorial diseases, including chronic pain, facilitating biomarker discovery and novel therapeutic targets.

1.3.2 Leveraging Biobank Data in Pain Research

The application of biobank data in pain research has provided an integrative approach to understanding the multifactorial etiology of chronic pain (Khoury et al., 2022). Notably, population-based biobanks facilitate large-scale genome-wide association studies (GWAS), enabling the identification of genetic loci associated with pain susceptibility and persistence

(Mocci et al., 2023). These loci often reside in or near genes that influence gene expression, epigenomic regulation, and downstream biological pathway. By integrating GWAS findings with functional genomics approaches, researchers have uncovered genetic correlations between chronic pain and a wide range of comorbid conditions, highlighting shared biological pathways and risk factors. Beyond genetics, biobank data allow researchers to incorporate lifestyle, demographic, and environmental risk factors, providing a more comprehensive understanding of pain risk and progression (Tanguay-Sabourin et al., 2023).

Variability in the definition, measurement, and quantification of pain further complicates research efforts, limiting the reproducibility and comparability of findings (Robinson-Papp et al., 2015). Addressing these challenges by expanding the inclusion of pain-related health data may enhance the accuracy and depth of future pain research.

1.4 Medical Ontologies

Medical ontologies provide standardized frameworks for classifying medical diagnoses (Haendel et al., 2018). In population-based biobank studies, medical ontologies are often used. The UK Biobank includes multiple medical ontologies. In this study, we focused on three key ontologies: (1) self-reported non-cancer conditions, which capture participants' medical histories, (2) treatment/medication code, recorded through structured questionnaires, and (3) hospital-derived ICD-10 codes, which provide clinically validated disease classifications.

1.4.1 Self-Reported Non-Cancer Conditions

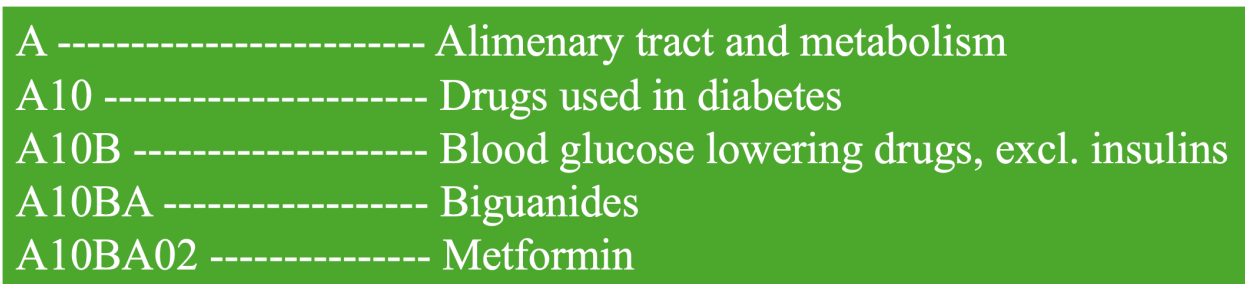
Self-reported non-cancer conditions in the UK Biobank were captured through structured questionnaires, stored in data field 20002. Participants were asked to report lifetime medical conditions from a predefined list of 474 non-cancer illnesses (Mutz et al., 2021). These conditions

are organized into 10 top-level parent categories, such as cardiovascular, musculoskeletal/trauma, and immunological/systemic disorders. These responses provide insights into chronic and long-term health conditions that may not always be reflected in hospital records. The self-reported data is particularly valuable for tracking conditions that do not require hospitalization or specialist diagnoses, offering a broader perspective on participants' health histories. However, self-reported medical histories are subject to certain limitations. Recall bias may lead to inaccuracies in reporting, as participants might forget or misclassify past diagnoses. Additionally, the absence of clinical validation means that some conditions may be under- or over-reported (Papez et al., 2022). For example, autoimmune disorders such as rheumatoid arthritis (RA) are often over-reported, as indicated by the discrepancy between self-reported RA and clinical validated cases based on MRI findings (Stanciu et al., 2022).

1.4.2 Treatment/Medication Code

Medication use in the UK Biobank is recorded in data field 20003, where participants self-reported their regular medication and health supplements at the time of recruitment (Wu et al., 2019). This dataset provides insight into commonly used drugs among the cohort but lacks longitudinal tracking, meaning changes in medication use over time are not captured. Unlike self-reported non-cancer conditions, the UK Biobank does not incorporate a predefined medical ontology for medications. To standardize the reported medication data, we manually mapped the entries to the Anatomical Therapeutic Chemical (ATC) Classification system. The ATC classification categorizes drugs based on their therapeutic use and chemical composition, providing a hierarchical structure that facilitates pharmacological analysis (Chen et al., 2012). Figure 1 illustrates the layout of ATC codes, which consist of five levels: the first level represents the anatomical group, the second level represents the therapeutic subgroup, the third level

represents the pharmacological subgroup, the fourth level represents the chemical subgroup, and the fifth level represents the chemical substance (WHO, n.d.). At the first level, the ATC system comprises 14 main anatomical groups.



A	-----	Alimentary tract and metabolism
A10	-----	Drugs used in diabetes
A10B	-----	Blood glucose lowering drugs, excl. insulins
A10BA	-----	Biguanides
A10BA02	-----	Metformin

Figure 1. An example of the ATC code, adapted from the WHO website on the ATC classification system

1.4.3 Hospital-derived ICD-10 Codes

The International Classification of Diseases (ICD) is a standardized diagnostic coding system developed by the World Health Organization (WHO) to classify diseases and health conditions. Originally designed for mortality statistics in the 19th century, the ICD has evolved into a comprehensive framework for tracking diseases, supporting clinical care, and facilitating research (Hirsch et al., 2016). ICD-10, the tenth revision, introduced increased granularity and a hierarchical structure based on etiology, anatomic location, severity, and other relevant clinical characteristics to capture disease complexity (Hirsch et al., 2016). In the UK Biobank, hospital-derived ICD-10 codes are recorded in data field 41270, capturing primary and secondary diagnoses from the National Health Service (NHS) Hospital Episode Statistics (HES) (Davis et al., 2018). Primary diagnoses represent the main reason for hospitalization, and secondary diagnoses include comorbid conditions or contributing factors that provide additional clinical context. The hierarchical structure of ICD-10 enables the precise classification of diseases. Each code consists

of an alphanumeric sequence, as shown in Figure 2 (Sammani et al., 2021). The first character is a letter indicating a general disease category, with total 22 categories, and the next two characters specify a disease group. Additional characters add specificity regarding laterality, severity, or etiology. We included ICD-10 data as a source of phenotypic information in our study.

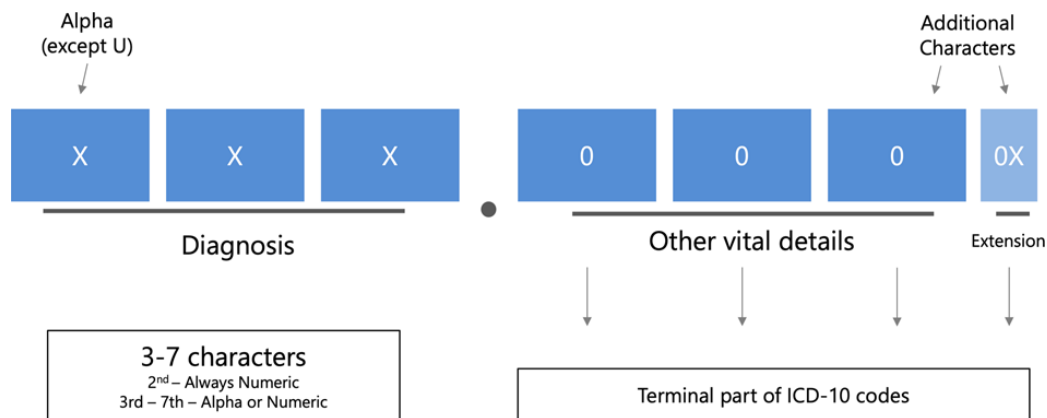


Figure 2. An example of the ICD-10 code, adapted from Figure 1 of Sammani et al. (2021)

1.5 Chronic Pain Genetics

Chronic pain results from the complex interplay of genetic, environmental, and psychosocial factors. Studying the genetic architecture of chronic pain can improve our understanding of underlying mechanisms of chronic pain development.

1.5.1 Genetic Basis of Chronic Pain

As introduced in Section 1.1 and 1.2, individuals experience different types of chronic pain, with substantial variability in pain susceptibility and sensitivity. Genetic factors contribute significantly to these individual differences, as evidenced by heritability estimates ranging from ~10–50%, depending on chronic pain subtype and study design (Diatchenko et al., 2005; Zorina-Lichtenwalter et al., 2016). For CMSKP, heritability estimates vary from 10.2% - 54%, depending on its definition (Johnston et al., 2019; Rahman et al., 2021).

A major challenge in studying chronic pain genetics is its substantial genetic heterogeneity, meaning the genetic factors contributing to the development and experience of chronic pain vary across individuals and populations (Tsepilov et al., 2020). This heterogeneity complicates the identification of specific genetic markers and the development of targeted treatments, as different genetic variants may underlie similar pain conditions in different individuals. Several genes, such as *ADRB2*, HLA/MHC region, and *SLC6A4*, have been implicated in CMSKP, but the underlying mechanism of CMSKP remains unclear (Diatchenko et al., 2013). This genetic complexity highlights the polygenic nature of chronic pain, wherein numerous genetic variants each exert small effects. Chronic pain development results from the cumulative influence of both common and rare variants, further modulated by environmental and lifestyle factors.

Another challenge in studying chronic pain genetic architecture is pleiotropy, where genetic variants associated with chronic pain also influence other complex traits. This shared genetic architecture explains the high comorbidity between chronic pain and other medical conditions. For instance, a recent study identified several colocalized genes between chronic back pain and inflammation-related musculoskeletal conditions, such as rheumatoid arthritis and osteoporosis, suggesting that chronic pain and its comorbidities may arise from common inflammatory mechanisms (Kasher et al., 2023). To disentangle these complex genetic relationships and identify key loci and genes associated with chronic pain, genome-wide association studies (GWAS) have been an important tool.

1.5.2 Genome-wide Association Studies (GWAS) in Chronic Pain

GWAS provides a hypothesis-free approach to identifying genetic variants associated with complex traits, including chronic pain development. By examining genetic variation across the

genome, GWAS can pinpoint susceptibility loci linked to different pain conditions, offering insights into potential biological pathways and mechanisms (Mills et al., 2020).

Several GWAS have leveraged the UK Biobank to identify risk loci for chronic pain by body site. For knee pain, a GWAS found significant associations with *GDF5* and *COL27A1* genes, involved in skeletal development and collagen formation, respectively (Meng et al., 2019). For neck or shoulder pain, a GWAS identified 3 genetic loci, and the most significant locus was in an intergenic region on chromosome 17, with the lead SNP rs12453010 ($P = 1.66 \times 10^{-11}$; $\beta = 0.0095$) (W. Meng et al., 2020). A review summarized GWAS for different chronic pain conditions by body site: 4 studies for chronic back pain with total 7 loci been identified, 1 study for chronic knee pain with 2 loci been identified (as mentioned above), and 1 study for chronic neck or shoulder pain with 3 loci been identified (as mentioned above) (Li et al., 2023). Up to date, the most recent GWAS meta-analysis on chronic low back pain, involving 325,078 participants from the UK Biobank and HUNT population studies from Norway, identified 18 genetic loci (Martinsen et al., 2025). Overall, while these studies have provided insights into the genetic architecture of chronic pain, the number of loci identified for each chronic pain condition by body site remains relatively small.

Given the limited number of loci identified for chronic pain at individual body sites, studies have combined multiple body sites to explore chronic pain under varying definitions of cases and controls. Johnston et al. (Johnston et al., 2019) studied multisite chronic pain (MCP), a measure of the number of chronic pain sites in individuals, and they found 76 independent lead SNPs at 39 risk loci within the same UK Biobank population. Khoury et al. (Khoury et al., 2022) identified distinct genetic signals between single-site chronic pain and chronic overlapping pain, with the latter exhibiting significantly stronger genetic associations. These findings suggest that evaluating

chronic pain as a quantitative trait may uncover a broader and more comprehensive genetic architecture of chronic pain.

Considering the complex nature of chronic pain, traditional GWAS that focus on a single chronic pain body site or MCP phenotypes are not sufficient to capture the genetic complexity of chronic pain, especially when excluding comorbid conditions. A more integrative approach that accounts for chronic pain across multiple body sites and incorporates comorbid conditions could potentially cover the complex genetic heterogeneity by identifying shared genetic factors and common biological pathways that underlie chronic pain across different medical conditions. Therefore, we need a method that can effectively study chronic pain development by incorporating genetic heterogeneity and accounting for comorbidities.

1.6 Topic Modeling

Topic modeling is an unsupervised machine learning technique developed to discover the underlying themes, or topics, among the words within a collection of documents. In this context, a topic is a set of words that frequently co-occur in the same documents, representing a distinct semantic theme. Unlike clustering, which assigns each word to a single cluster, topic modeling allows words to belong to several different topics with different probabilities. There are many different types of topic modeling, which are briefly discussed below.

1.6.1 Common Topic Models

Several topic modeling algorithms have been developed, including Latent Dirichlet Allocation (LDA), Tree-structured Logistic Factor Analysis (treeLFA), and Embedded Topic Model (ETM).

LDA is a classical probabilistic topic modeling technique that extracts topics from a collection of documents based on the bag-of-words assumption, which treats all words independently without considering their orders in the documents (Blei et al., 2003). LDA models the joint probability distribution over words, latent topics, topic proportions, and topic assignments. Specifically, it assumes that each document is generated from a mixture of latent topics, where each topic is characterized by a distinct distribution over words. The Dirichlet distribution is a multivariate probability distribution that serves as a prior for both the topic distribution within documents and the word distribution within topics, ensuring that the assigned probabilities sum to one and enabling the modeling of proportions in applications of topic modeling. By leveraging this probabilistic framework, LDA can effectively infer the latent topic structure from the batches of documents.

treeLFA is a topic model designed for analyzing patterns of disease co-occurrence in binary healthcare data (Zhang et al., 2023). Based on the presence or absence of disease codes for individuals, treeLFA factors data into two matrices: one representing topics (disease clusters) and the other representing individual topic weights.

Embedded Topic Modeling (ETM) is a probabilistic framework that combines neural network embeddings and traditional topic models to infer latent topics (Dieng et al., 2019). Unlike LDA and treeLFA, ETM represents words in a continuous latent space using word embeddings, which capture the semantic relationships between words. An embedding is a mathematical representation that maps discrete entities, such as words, into a continuous vector space, where similar entities are positioned closer together. By leveraging embeddings, ETM can capture nuanced relationships in language, making it particularly effective for analyzing high-dimensional datasets.

1.6.2 Topic Modeling in the Biomedical Context

Topic modeling has been increasingly applied in biomedical research to uncover latent structures within complex health data. In this context, each study subject is treated as a document, and their features, such as medical conditions or medications, are treated as a mixture of words. Topics then represent the clusters and categories of related features that are grouped based on shared patterns or associations in the data. Previous research has demonstrated the use of topic modeling to derive phenotypic categories from electronic health record (EHR) data. These topic weights, calculated at the individual level, have subsequently been utilized as derived phenotypes in genome-wide association (GWA) scans. By incorporating these novel clusters/topics into GWA scans, researchers may identify different or novel genetic loci associated with these broader phenotypic categories rather than using single-condition labels (Jiang et al., 2023). For instance, Zhang et al. (Zhang et al., 2023) had a topic about metabolic and heart diseases with 50 associated loci, including 5 unique ones, suggesting that topic-derived GWAS enhanced the discovery of novel loci. This approach improved risk prediction for some disorders by uncovering genetic variants in complex traits. However, prior uses of topic modeling in a GWA study context did not allow for the incorporation of prior knowledge of relationships among medical condition variables. Furthermore, topic modeling has not been applied to chronic pain.

1.7 Objectives and Hypothesis

This thesis explores the application of Graph-Embedded Topic Modeling (GETM) (Wang et al., 2022) to high-dimensional UK Biobank data. GETM has the potential to uncover multimorbidity patterns because it integrates biomedical hierarchies into pre-trained graph embeddings, which are structured representations that capture the relationships between features in a continuous vector space. This study aims to test whether GETM can effectively generate

informative topics from UK Biobank medical data and, subsequently, whether these topics can serve as refined phenotypic inputs for GWA scans, improving the identification of genetic variants associated with CMSKP.

Objective 1: To apply GETM to reduce the dimensionality of medical conditions, medication codes, and ICD-10 codes in the UK Biobank. By structuring these high-dimensional data into a smaller set of topics, we aimed to capture meaningful disease patterns and generate topics that represent multimorbidity.

Objective 2: To use GETM-derived outputs as phenotypic inputs for GWA scans to identify genetic variants associated with CMSKP. For our investigation of the genetic heterogeneity of CMSKP and verification of whether GETM can improve GWA scan outputs, we compared GWA scan results across three different schemes. This evaluation aims to determine whether GETM-derived phenotypes improve GWA scan resolution by refining trait definitions and enhancing the detection of pain-related genetic loci.

Hypothesis: The hypothesis is that phenotypes derived from GETM topics, used to run GWA studies, may be used to identify genetic variants underlying CMSKP with less genetic heterogeneity than a single-trait CMSKP GWA scan.

Chapter 2 – Materials and Methods

2.1 Data Availability

In this study, we retrieved all of our data from the UK Biobank. The UK Biobank is one of the largest biobanks available to researchers worldwide, which includes extensive phenotypic and genotypic data of ~500,000 individuals aged between 40 to 69 years recruited between 2007 and 2010 (Sudlow et al., 2015). Participants' blood samples were collected at the baseline visit, and whole genome sequencing of every participant was conducted to study the genetic determinants across various medically related traits. We drew on the more restricted UK Biobank genetic dataset, which included genotypes for 488,377 participants, 49,950 genotyped using the UK BiLEVE Axiom Array (807,411 markers), and the remaining 438,427 genotyped using the UK Biobank Axiom Array (825,927 markers) (Bycroft et al., 2018). Genotypes were further imputed using a combination of reference panels, including the UK10K, 1000 Genomes Project, and HRC (Haplotype Reference Consortium), to increase variant density and accuracy (Bycroft et al., 2018).

To avoid spurious association due to population stratification, we restricted both the topic modeling and GWA scan analyses to individuals of White British, maximizing the number of available participants. We conducted a series of quality control steps on these participants, including filtering SNPs and individuals with a high level of missingness, checking for the inconsistencies between reported sex and genetic data, keeping autosomal SNPs, deleting SNPs with minor allele frequency (MAF) lower than 0.05, and filtering SNPs which are not in Hardy-Weinberg equilibrium (HWE), resulting in 577,232 SNPs that were used for modeling polygenic effects to account for relatedness in running the GWA scans.

For the phenotypic data, we extracted 443 self-reported non-cancer conditions, 529 medications, and 1,922 top-3-digit ICD-10 codes from the UK Biobank.

2.2 Graph-Embedded Topic Modeling

In this study, we used Graph-Embedded Topic Modeling (GETM) (Wang et al., 2022), an advanced topic modeling approach that integrates medical ontologies into pre-trained graph embeddings to uncover latent topics within a collection of health-related data. Unlike traditional ETM, GETM incorporates graph structures, which consist of nodes (features) and edges that capture hierarchical relationships among nodes.

GETM incorporates two existing algorithms, node2vec and Embedded Topic Modeling (ETM). Node2vec is a graph-based algorithm that captures the relationships between words by constructing a graph and generating embeddings, which are vectors that encode the relationships between features with their biomedical hierarchies (Grover & Leskovec, 2016). ETM then takes these graph-based embeddings as input and uses them to generate topics that reveal patterns of co-occurrence and meaning across large document collections (Dieng et al., 2019). By integrating graph-structured data into topic modeling, GETM preserves hierarchical connections between features, unlike traditional models that treat them as independent entities.

The overview of the GETM design is shown in Figure 3. We first constructed graphs for medical conditions and ICD-10 codes using biomedical hierarchies from the UK Biobank and for medications using the ATC classification system. These graphs were then processed using node2vec, which learned their structure and generated vector embeddings for each feature. Next, these embeddings were incorporated into ETM’s encoder network, along with individual-level data containing corresponding features. The encoder generated latent variables capturing relationships

between embeddings and topics, and the decoder network reconstructed the original data and generated three matrices, individuals-by-topics (θ), topics-by-embedding (α), and embedding-by-conditions/ICD-10 codes (ρ). These three matrices allow the exploration of topics, individuals, and relationships among features in a highly interpretable way.

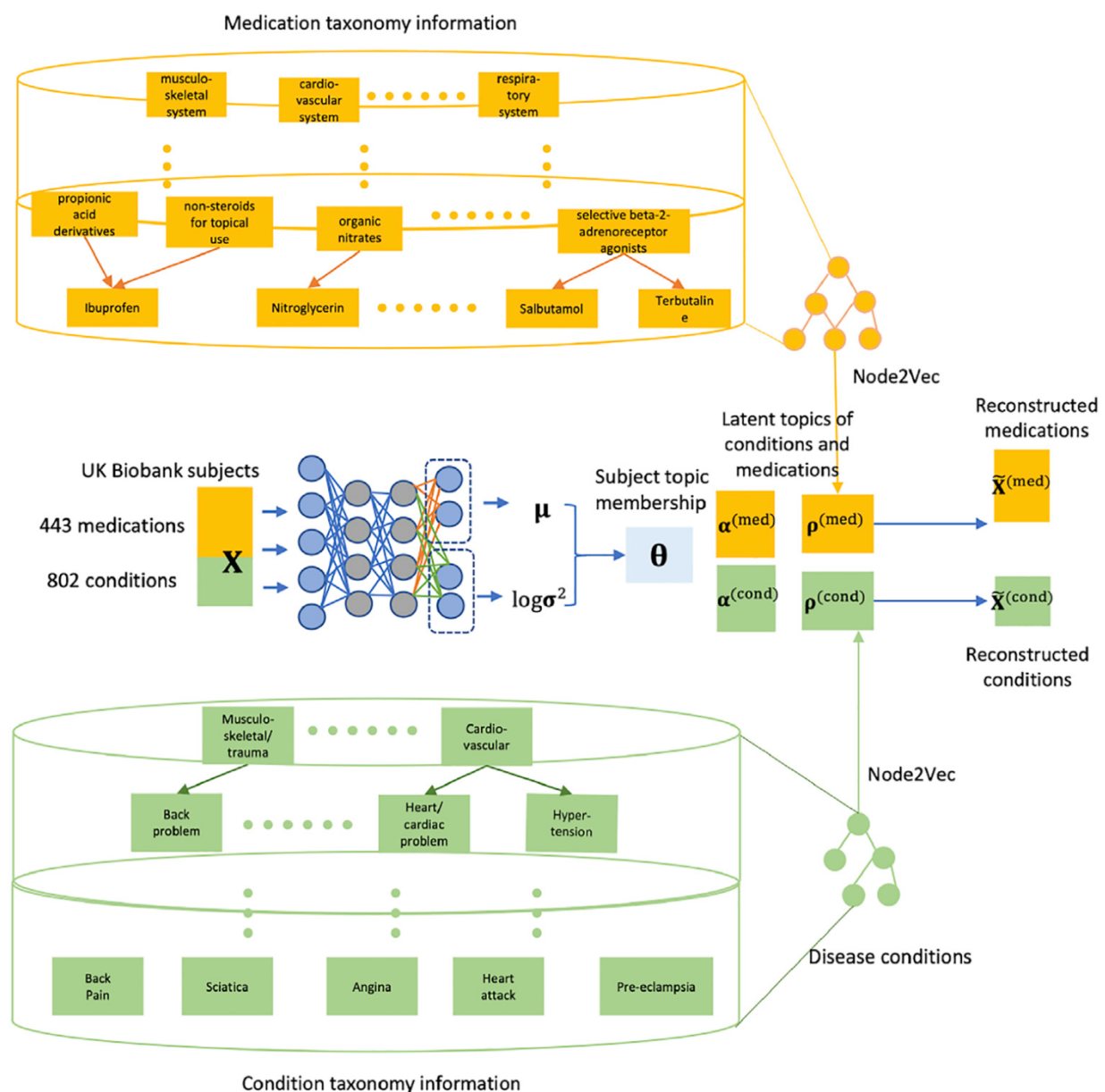


Figure 3. Overview of GETM

For an alternative explanation, Figure 4 presents the mathematical representation of GETM. Given input features X_i for individual i , a neural network (NN) encoder parameterized by W_θ estimated the variational parameters μ_i and $\log \sigma_i^2$, which defined a Gaussian latent space:

$$[\mu_i, \log \sigma_i^2] = \text{NN}(x_i; W_\theta)$$

The latent variable δ_i was then sampled as:

$$q(\delta_i) = \mu_i + \text{diag}(\sigma_i) \mathcal{N}(0, I)$$

To obtain a topic mixture θ_i , we applied a softmax transformation:

$$q(\theta_i | x_i) = \text{softmax}(\delta_i)$$

Where each topic proportion $\theta_{i,k}$ was computed as:

$$\theta_{i,k} = \frac{\exp(\delta_{i,k})}{\sum_{k'} \exp(\delta_{i,k'})}$$

The decoder reconstructed the observed features using a linear model. Specifically, feature representation $r_{i,f}$ was determined by a weighted combination of topic embeddings α and feature embeddings ρ , with batch effect correction λ :

$$\hat{r}_{i,f} = \theta_i \alpha \rho_f + \lambda s(i), f$$

The final probability distribution over features was obtained using a softmax transformation:

$$r_{i,f} = \frac{\exp(\hat{r}_{i,f})}{\sum f' \exp(\hat{r}_{i,f'})}$$

This formulation enabled the model to capture latent topic structures underlying UKB features while accounting for batch effects, improving the interpretability of feature associations.

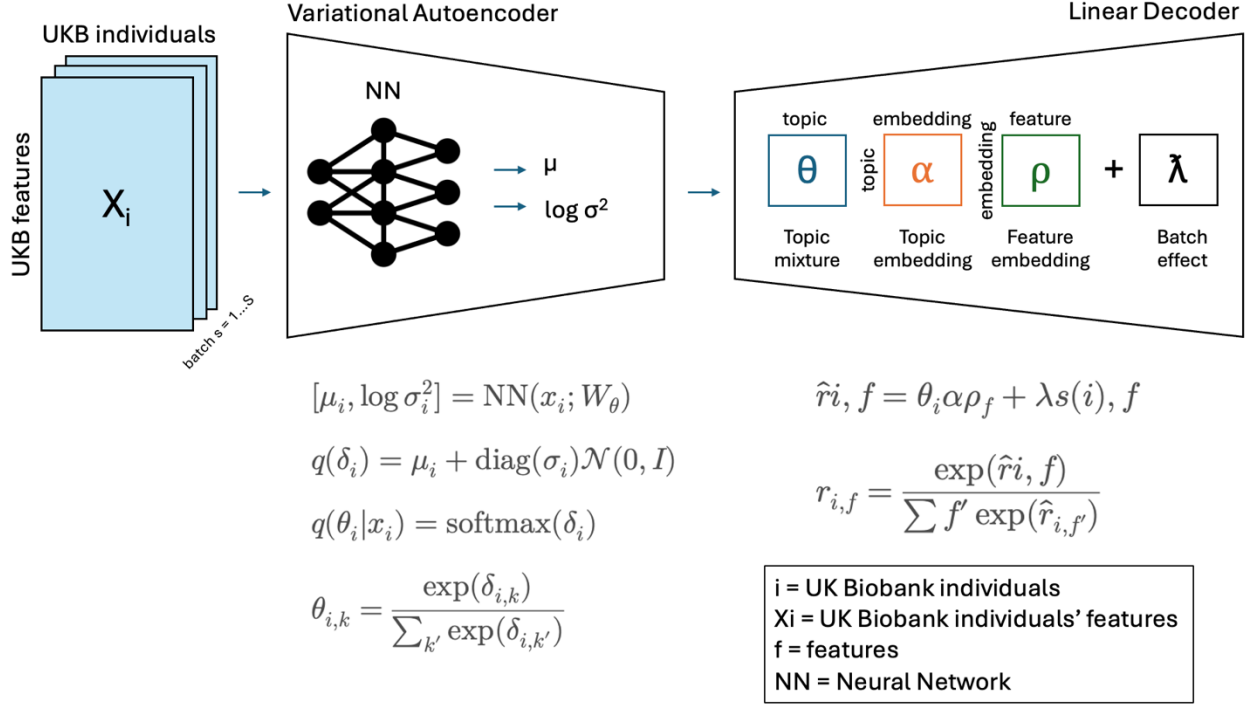


Figure 4. Mathematical representation Of GETM

Due to the design of GETM, only two features are allowed as the inputs for topic generation. To assess embedding performance, we used Uniform Manifold Approximation and Projection (UMAP) to visualize the generated embeddings after topic modeling (McInnes et al., 2018). Based on these evaluations, we selected two out of three feature classes (self-reported conditions, medications, and hospital-derived ICD-10 codes) to ensure optimal interpretability and downstream genetic analysis.

2.3 Topic Selection for GWA Scans

To identify the most discriminative topics for CMSKP cases and controls in downstream genetic analysis, we implemented a three-step evaluation process. First, we assessed average topic quality, which combines topic coherence (how closely related the top 5 features are within a topic) and topic diversity (the proportion of distinct features among the top 5 across all topics) (Wang et al., 2023). We generated topics five times using different node2vec embeddings under a predefined number of topics and averaged the quality scores. The second step is to test for the difference in θ value means between CMSKP cases and controls using the t-test p-value. The selected M topics were used as the phenotypic inputs for the GWA scans.

2.4 Comparison of GWA Scans

We compared four GWA scan approaches to evaluate how different methods of integrating topics with CMSKP to generate derived phenotypes impact the identification of genetic variants associated with CMSKP. Approaches 1 and 2 use binary traits as phenotype inputs, while Approaches 3 and 4 use quantitative traits. In Approach 1, individuals are classified as either cases or controls: cases are individuals who have at least one pain site at any of the four body sites of CMSKP, and controls are individuals without acute or chronic pain at any body site. Approach 2 derives categorical phenotypes across the M topics by taking the top 50% θ values from CMSKP cases (new cases), contrasted with the bottom 50% from CMSKP controls (new controls). Approach 3 uses the topic weights (θ) of all individuals as quantitative input for GWA scans, while Approach 4 interprets the topic weights (θ) of CMSKP cases only.

2.5 Genetic Analysis

We used REGENIE (Mbatchou et al., 2021) to run GWA scans, adjusting for age, sex, and the top 40 principal components to account for population stratification among subjects. For post-GWA scan analysis, we used Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) (Watanabe et al., 2017) to conduct functional annotations of the genetic loci and interactive visualizations. Utilizing the 1000 Genomes Project Phase 3 (Auton et al., 2015) as the reference panel, we identified independent significant SNPs with a genome-wide significance threshold ($P < 5 \times 10^{-8}$) that were independent at $r^2 < 0.6$. From these, lead SNPs were defined as those further independent at $r^2 < 0.1$. Genomic risk loci were established by merging lead SNPs within a 250 kb window, incorporating all SNPs in linkage disequilibrium ($r^2 \geq 0.6$) with the independent significant SNPs. For gene definition, SNPs were mapped to 19,111 protein-coding genes using a ± 10 kilobase window around each gene to include potential regulatory regions, and the genome-wide significance threshold for gene-based test was $P < 2.616 \times 10^{-6}$ (Watanabe et al., 2017). Additionally, we conducted Multi-marker Analysis of GenoMic Annotation (MAGMA) Tissue Expression Analysis utilizing GTEx v8 data across 53 tissue types to identify tissue-specific gene expression patterns associated with the traits of interest (de Leeuw et al., 2015).

Chapter 3 – Results

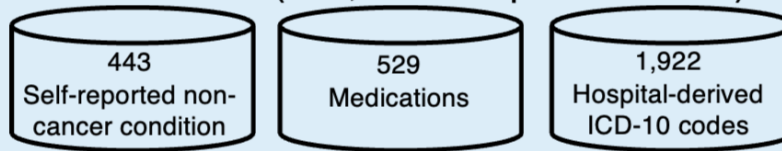
3.1 Summary Statistics

Figure 5 shows the overall design of the study. A total of 401,013 people of European descent have either condition, medication, or ICD-10 data in the UK Biobank. Fig. 6A, 6B, and 6C show the distributions of the condition, medication, and ICD-10 code counts in European individuals, respectively. On average, each individual has 2.16 conditions, 2.70 medications, and 12.55 ICD-10 codes. The distributions indicate that most of the individuals have more than one medical conditions.

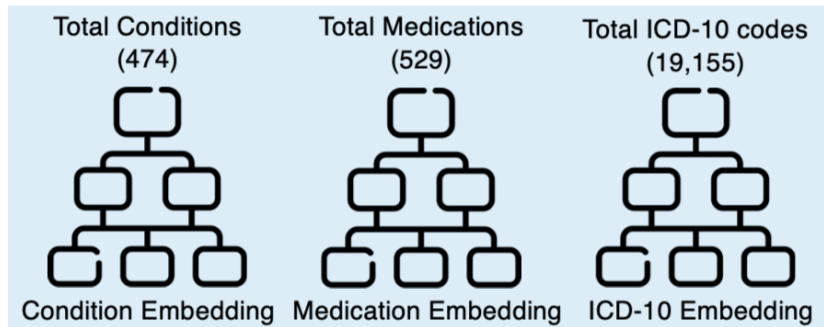
Study Design

1. Data Preparation

UK Biobank (401,013 European descent)



2. Graph Embedding



3. Topic Generation

Individuals by topics
(θ)

Topic Embedding
(α)

Feature Embedding
(p_{cond} , p_{med} , p_{ICD10})

4. Topic Selection for GWA Scans

A. Topic Quality (coherence x diversity)

B. T-test difference in θ means between case and control

- Case (N = 156,235): individuals who have at least one chronic pain site at any of the four body sites: knee, back, neck/shoulder, or hip
- Control (N= 154,045): individuals who do NOT have acute or chronic pain at any body sites

5. GWA Scans Approaches

A. Binary Trait

- Case vs. Control
- Top 50% θ values from CMSKP Cases vs. Bottom 50% θ values from CMSKP Controls

B. Quantitative Trait

- θ values for the selected topic (All individuals)
- θ values for the selected topic (Case only)

Figure 5. Overall study design

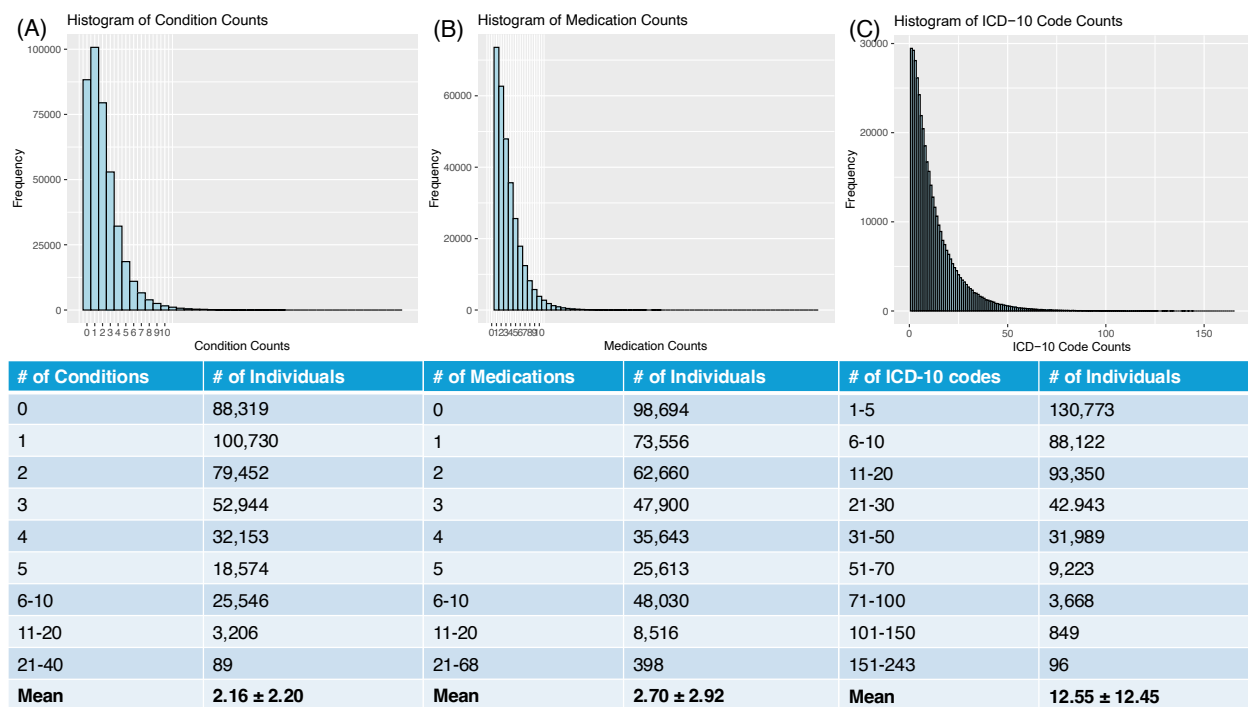


Figure 6. (A) Histogram and table of condition counts; (B) Histogram and table of medication counts; (C) Histogram and table of ICD-10 code counts

Table 1 summarizes the number of individuals with CMSKP by body site and as a unified phenotype, comparing all European individuals to those with at least one conditions, medications, or ICD-10 codes in the UK Biobank. European individuals with at least one conditions or ICD-10 codes have higher percentages of chronic neck or shoulder pain (16.30% vs. 15.46%), chronic back pain (18.07% vs. 17.12%), chronic hip pain (9.23% vs. 8.56%), chronic knee pain (17.47% vs. 16.42%), and CMSKP (38.96% vs. 37.17%). All differences are statistically significant based on two-sample t-tests, with CMSKP showing the most significant difference (P-value = 3.13×10^{-65}).

	European individuals (N = 457,461)	European individuals with at least 1 or more condition/medication/ICD-10 code (N = 401,013)	P-value
Chronic neck or shoulder pain	70,723 (15.46%)	65,385 (16.30%)	1.11e-26
Chronic back pain	78,298 (17.12%)	72,445 (18.07%)	8.68e-31
Chronic hip pain	39,144 (8.56%)	37,016 (9.23%)	6.55e-28
Chronic knee pain	75,109 (16.42%)	70,043 (17.47%)	3.47e-38
Chronic musculoskeletal pain	170,033 (37.17%)	156,235 (38.96%)	3.13e-65

Table 1. Summary of chronic musculoskeletal pain by body site and as a unified phenotype between European individuals and European individuals with at least one conditions, medications, or ICD-10 codes

3.2 Topic Selection for GWA Scans

3.2.1 Evaluation of Topic Quality

To determine which two of the three feature classes (self-reported conditions, medications, and hospital-derived ICD-10 codes) had the best embedding performance, we visualized both feature embeddings and topic embeddings using UMAP (Figure 7-9). Each dot represents a feature, color-coded by category, while black stars indicate topic embeddings.

In the UMAP plots for condition and ICD-10 embeddings (Figures 6 and 7), features from the same category are well-clustered, and black stars are randomly distributed, indicating strong alignment between feature embeddings and topic embeddings. However, in the UMAP plot for medication embeddings (Figure 8), although features of the same category tend to cluster together, there is substantial overlap between groups, making it difficult to distinguish distinct categories.

This lack of clear separation reduces interpretability, making medications a less suitable feature class for downstream genetic analysis. Based on these observations, we selected self-reported conditions and ICD-10 codes for subsequent analyses.

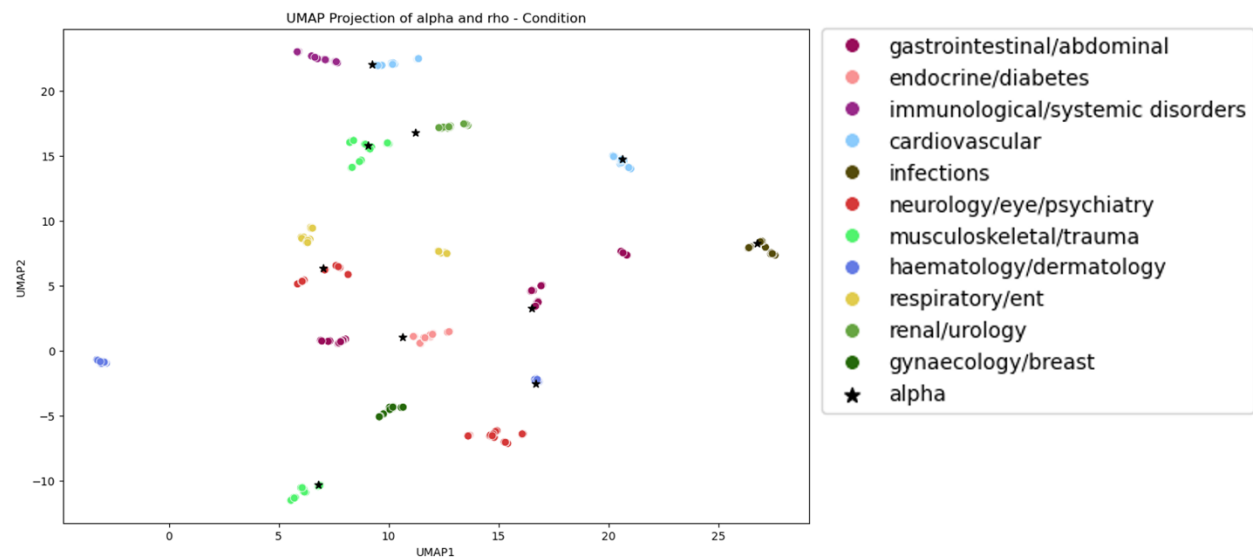


Figure 7. UMAP for condition embedding and topic embedding

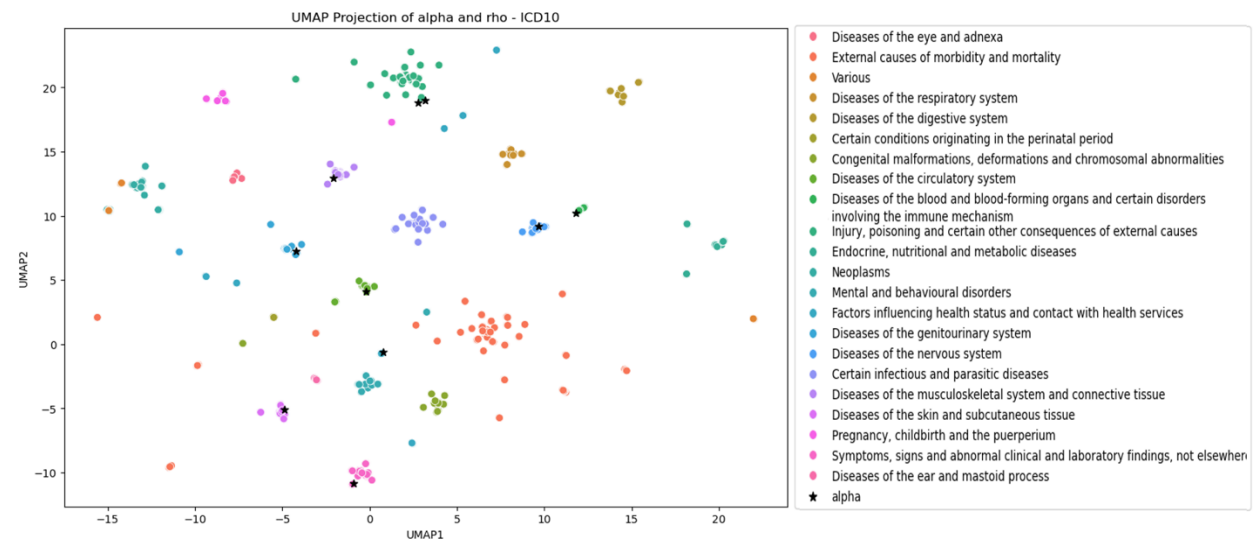


Figure 8. UMAP for ICD-10 embedding and topic embedding

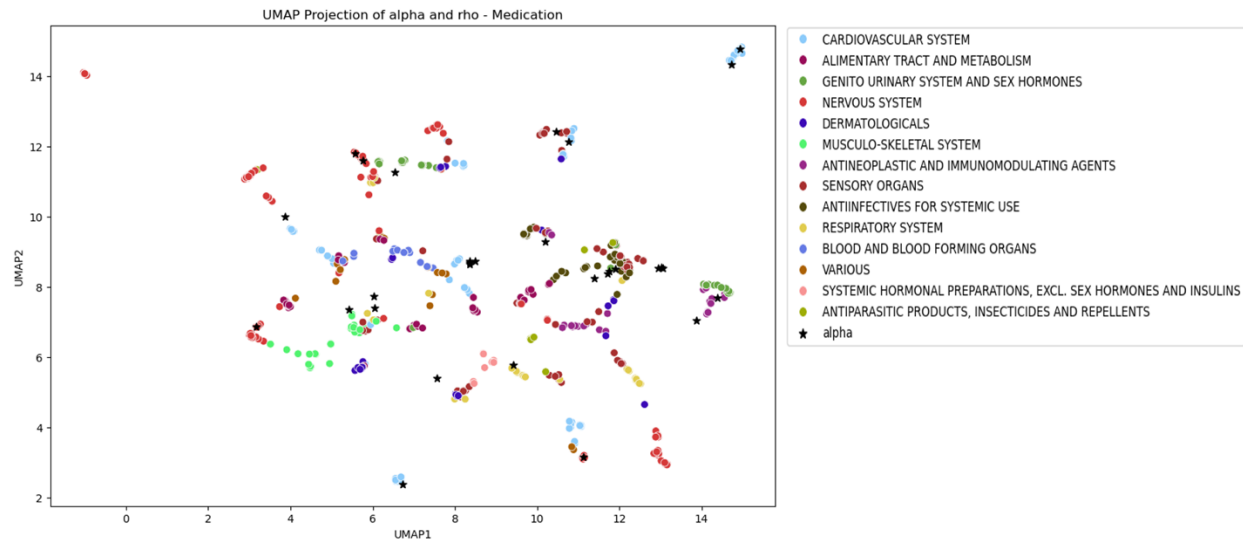


Figure 9. UMAP for medication embedding and topic embedding

After selecting the two input features for topic generation, we determined the optimal number of topics for GETM by testing predefined values of 10, 20, 30, 40, and 50. These values were selected based on previous literature and a balance between interpretability and granularity. A smaller number of topics may lead to noisy or overly broad groupings, while a larger number may introduce feature redundancy and reduce interpretability. Table 2 presents the average topic coherence, topic diversity, and overall topic quality for ICD-10 codes and conditions across five different embeddings. Notably, these five embeddings represent the results from different runs of node2vec, reflecting the inconsistency in the version of embeddings generated, which contributed to variability in topic coherence and diversity across analyses. Although setting the topic number at 10 yielded the worst ICD-10 topic quality, the differences across the five scenarios were minimal. Since conditions achieved the highest topic quality at 10 topics, we selected this as the best out of the tested options for downstream analysis.

	Mean \pm SD					
Topic #	ICD-10 Coherence	ICD-10 Diversity	Condition Coherence	Condition Diversity	ICD-10 Quality	Condition Quality
10	0.550 \pm 0.035	0.940 \pm 0.055	0.550 \pm 0.035	1.000 \pm 0.000	0.517 \pm 0.044	0.550 \pm 0.035
20	0.535 \pm 0.029	0.970 \pm 0.027	0.535 \pm 0.014	0.970 \pm 0.045	0.519 \pm 0.032	0.519 \pm 0.022
30	0.523 \pm 0.007	0.960 \pm 0.015	0.560 \pm 0.019	0.900 \pm 0.053	0.509 \pm 0.009	0.505 \pm 0.045
40	0.518 \pm 0.026	0.940 \pm 0.049	0.545 \pm 0.034	0.850 \pm 0.035	0.486 \pm 0.033	0.463 \pm 0.022
50	0.544 \pm 0.011	0.920 \pm 0.032	0.556 \pm 0.021	0.796 \pm 0.026	0.500 \pm 0.020	0.443 \pm 0.021

Table 2. ICD-10-specific and condition-specific topic quality evaluation across 10, 20, 30, 40, and 50 topics; SD: Standard Deviation; coh: coherence; div: diversity

To further investigate the identified topics, we visualized the top five features for each topic using a heatmap, with conditions as an illustrative example (Figure 10). The heatmap illustrates the probability of each feature belonging to its respective topic. However, we observed that the top five conditions are not from the same category. For example, top 3 features in Topic 3 were from three different categories: “peripheral nerve injury” (neurology/eye/psychiatry), “fracture metatarsal” (musculoskeletal/trauma), and “glomerulonephritis” (renal/urology). This discrepancy highlights the need for further refinement in topic modeling to ensure that the identified topics are both coherent and meaningful. As this thesis aims to explore the potential and feasibility of using topic-derived result as the phenotypic input for GWA scans, we continued our analysis to identify the most discriminating topic between CMSKP cases and controls.

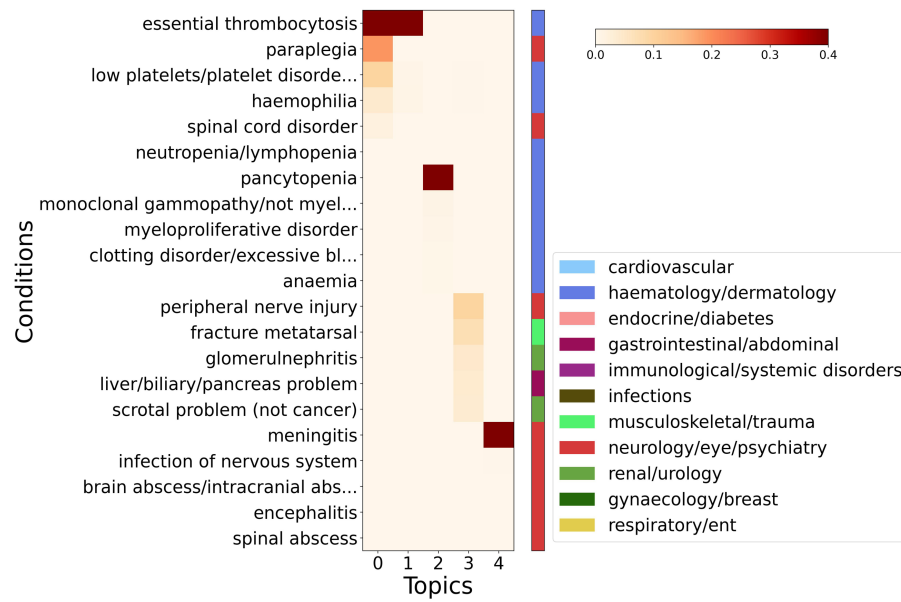


Figure 10. Heatmap displaying the top 5 conditions for a random selection of 5 out of the 10 topics, with the color intensity representing the probability of each condition belonging to its respective topic

3.2.2 Identifying the Most Discriminating Topic for CMSKP

To identify the most predictive topic for GWA Scans of CMSKP, we performed a t-test (Table 3) and found that Topic 9 exhibited a statistically significant difference between CMSKP cases and controls (P-value = 0.004). We further examined the top 5 features in Topic 9 based on their probabilities. Given that the total sample size was 401,013 individuals, the top features in Topic 9 were predominantly rare features. Top 5 conditions of Topic 9 with its corresponding individuals with this feature were 'fracture nose' (668 individuals; 0.17%), 'nasal polyps' (3,017 individuals; 0.75%), 'pulmonary fibrosis' (0 individual; 0%), 'thyroiditis' (1,989 individuals; 0.50%), and 'bronchiectasis' (1,096 individuals; 0.27%). Top 5 ICD-10 codes of Topic 9 with its corresponding individuals with this feature were 'I88 Nonspecific lymphadenitis' (164 individuals; 0.04%), 'P14 Birth injury to peripheral nervous system' (8 individuals; 0.002%), 'G08 Intracranial

and intraspinal phlebitis and thrombophlebitis' (97 individuals; 0.02%), 'I80 Phlebitis and thrombophlebitis' (7,306 individuals; 1.8%), and 'S76 Injury of muscle and tendon at hip and thigh level' (632 individuals; 0.16%). This finding was in contrast the GETM paper (Wang et al., 2022), which reported that topics were primarily composed of common features. The reason for this discrepancy remains unclear.

Since this thesis aims to explore the feasibility of implementing GETM to refine phenotypes for GWA scans, we proceeded with applying Topic 9 to GWA scans, despite the small mean differences between cases and controls and the large standard deviations.

	Case		Control		P-value
	Mean	SD	Mean	SD	
Topic 1	0.571	2.410	0.573	2.417	0.783
Topic 2	0.777	1.875	0.777	1.882	0.999
Topic 3	-0.626	3.048	-0.618	3.045	0.458
Topic 4	0.002	2.066	0.007	2.068	0.486
Topic 5	-0.056	2.762	-0.045	2.776	0.299
Topic 6	1.005	1.954	1.008	1.962	0.642
Topic 7	0.251	2.194	0.255	2.194	0.533
Topic 8	-0.093	2.373	-0.092	2.365	0.964
Topic 9	-1.564	4.242	-1.609	4.301	0.004
Topic 10	-0.272	3.131	-0.262	3.154	0.392

Table 3. T-test results for identifying the most predictive topic for GWA scans in CMSKP;

Statistical significance was determined using a threshold of $P < 0.05$.

3.3 GWAS of CMSKP Using Binary Case-Control Design

A total of 310,300 individuals were included in three different approaches to GWA scans. In the first approach, CMSKP cases ($N = 156,245$) were defined as individuals with at least one chronic pain site at any of the following four body sites: neck or shoulder, back, hip, and knee, and CMSKP controls ($N = 154,055$) were defined as individuals who do not have acute or chronic pain at any body site. Figure 11 shows the Manhattan and QQ plots for the GWA scan summary statistics, identifying 27 genomic risk loci, 29 lead SNPs, and 58 independent significant SNPs. Figure 12 shows the Manhattan plot and the QQ plot for the GWA scan gene-based tests. The gene-based test revealed 95 genes, and based on the ranking of the p-value significance, the top 5 genes are *PABPC4*, *TCTA*, *DCC*, *MACF1*, and *BSN*. Notably, the QQ plot for the gene-based test shows a lambda value of 2.36, indicating substantial inflation. This level of inflation is common in chronic pain GWAS and reflects the polygenic nature of the trait.

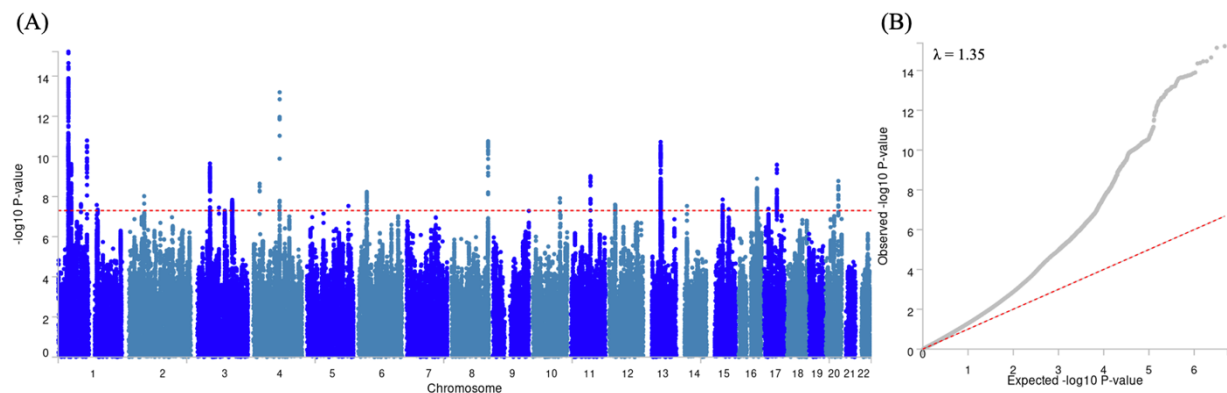


Figure 11. (A) Manhattan plot of GWA scan summary statistics between CMSKP cases and controls. Each point represents a SNP. The red horizontal line indicates the genome-wide significance threshold ($P = 5 \times 10^{-8}$); (B) QQ plot of GWA scan summary statistics between CMSKP cases and controls

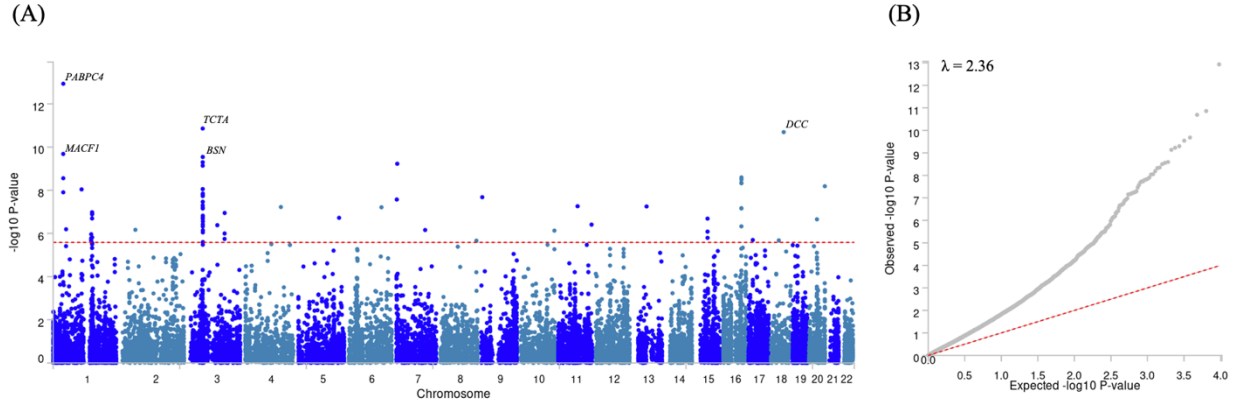


Figure 12. (A) Manhattan plot; (B) QQ plot of GWA scan gene-based test between CMSKP cases and controls

In addition to the gene-level information, Multi-marker Analysis of GenoMic Annotation (MAGMA) tissue expression analysis also integrates the tissue-specific expression to annotate the functional information, and the result shows that the genetic signals associated with CMSKP are significantly expressed in cerebellum, cortex, and basal ganglia (Figure 13). Complementary to MAGMA result, the differentially expressed genes (DEG) results show that the genes are significantly downregulated in basal ganglia (Figure 14). To visualize the expression of each of the top 5 gene across different tissues, Figure 15 shows the gene expression heatmap of the top 5 genes. *PABP4* shows a high expression across various tissues, suggesting its involvement in multiple CMSKP pathways. In contrast, *DCC* shows a low expression across all tissues.

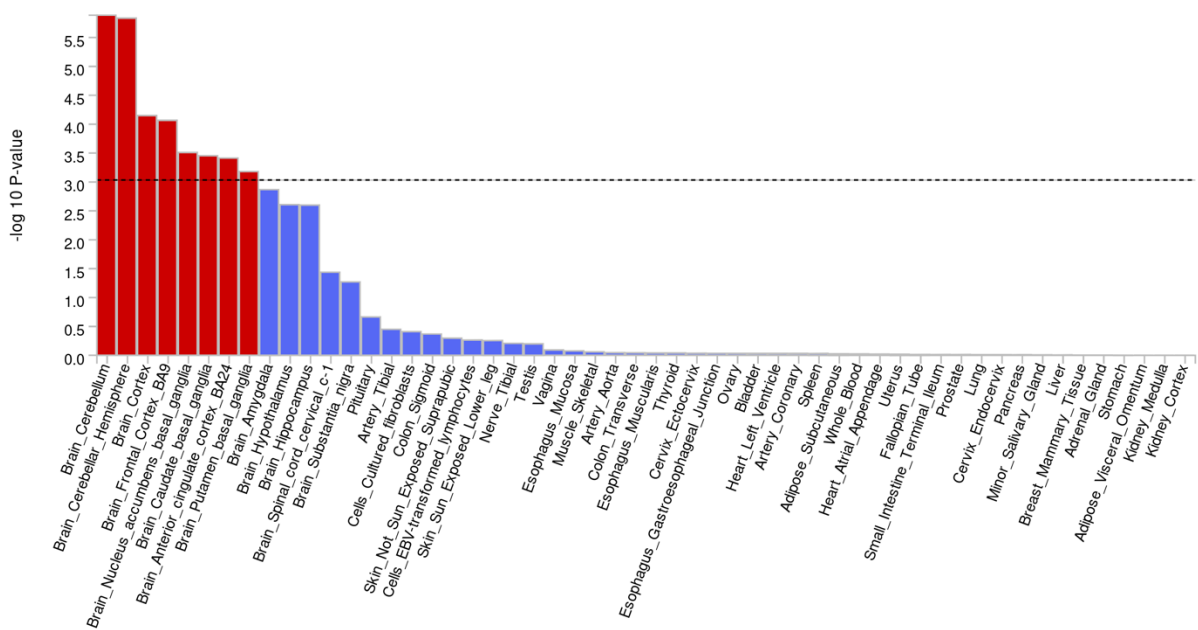


Figure 13: MAGMA Tissue Expression Analysis – GTEx v8 53 tissue types

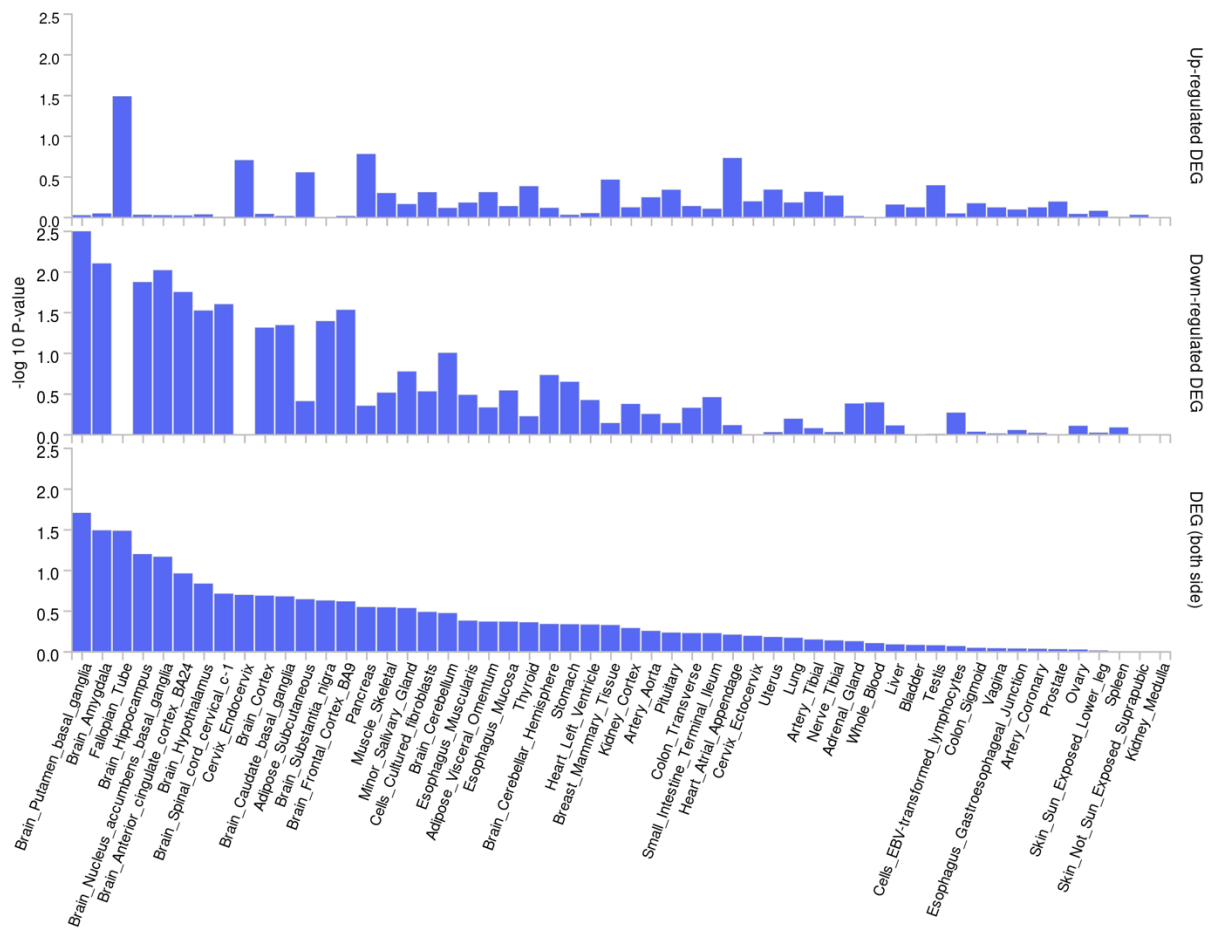


Figure 14. Differentially Expressed Genes – GTEx v8 53 tissue types

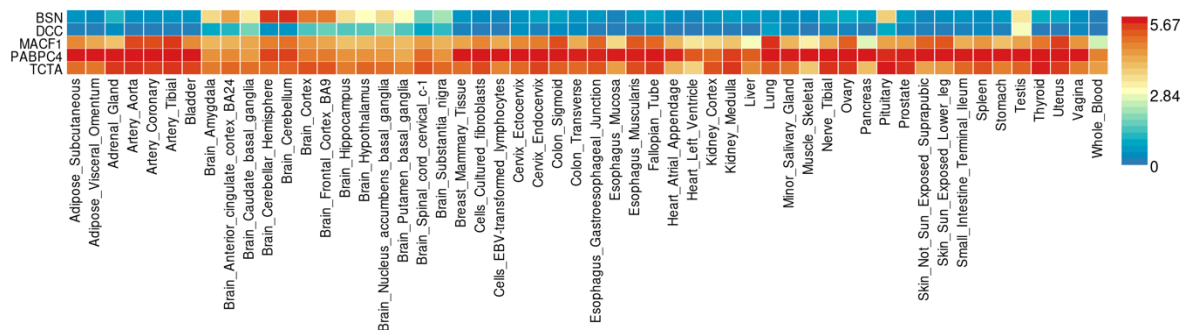


Figure 15. Gene Expression Heatmap of the top 5 genes

3.4 Refined GWAS of CMSKP with Stricter Case and Control

Definitions

The second approach, which selected the top 50% θ values of Topic 9 from CMSKP cases (new cases) contrasted with the bottom 50% of Topic 9 from CMSKP controls (new controls), included 78,117 CMSKP cases and 77,022 CMSKP controls. Figure 16 shows the Manhattan and QQ plots for the GWA scan summary statistics, identifying 9 genomic risk loci, 9 lead SNPs, and 16 independent significant SNPs. Figure 17 shows the Manhattan and QQ plots for the GWA scan gene-base tests. The top 5 genes based on the ranking of the p-value significance were *TCTA*, *BSN*, *AMIGO3*, *GMPPB*, and *RNF123*. Although Approach 2 had lost 50% of individuals compared to Approach 1, it replicated the finding of *TCTA* and *BSN*, along with new genes.

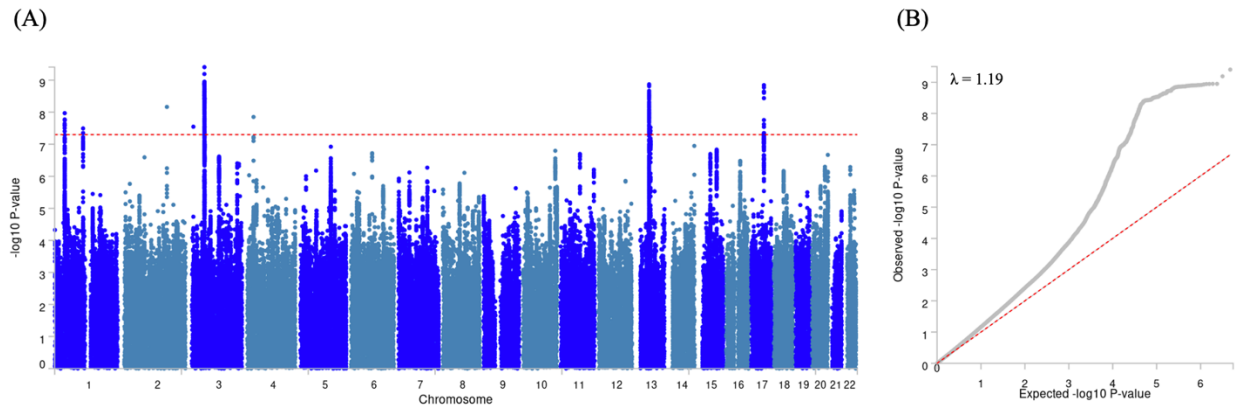


Figure 16. (A) Manhattan plot; (B) QQ plot of GWA scan summary statistics between stricter CMSKP cases and controls

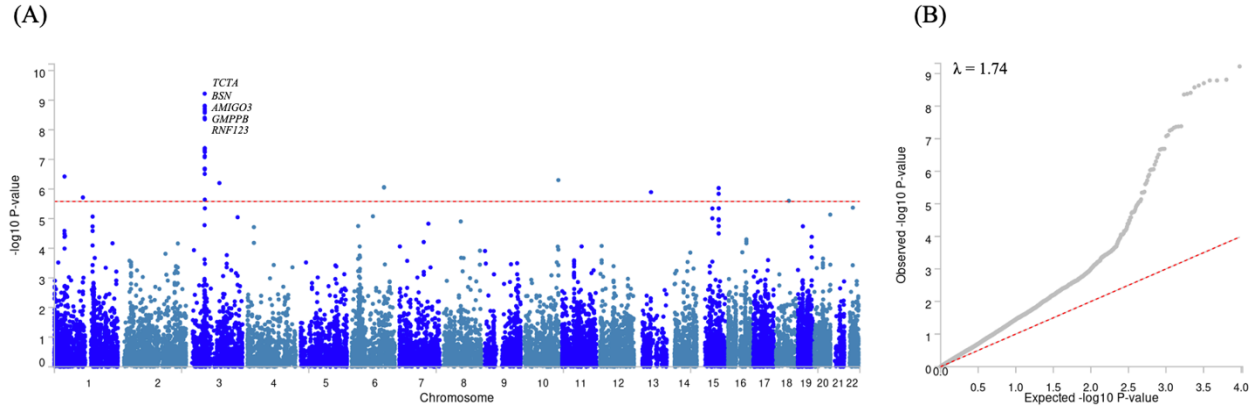


Figure 17. (A) Manhattan plot; (B) QQ plot of GWA scan gene-based test between stricter CMSKP cases and controls

3.5 GWAS of CMSKP individuals Using Topic 9 Theta Values as a Continuous Trait

The third approach included a total of 310,280 CMSKP individuals, using their Topic 9 theta values as a continuous input trait for GWA scans. No SNPs passed the genome-wide significance threshold of $P < 5 \times 10^{-8}$ (Figure 18), and no genes passed the genome-wide significance threshold of $P < 2.616 \times 10^{-6}$ (Figure 19).

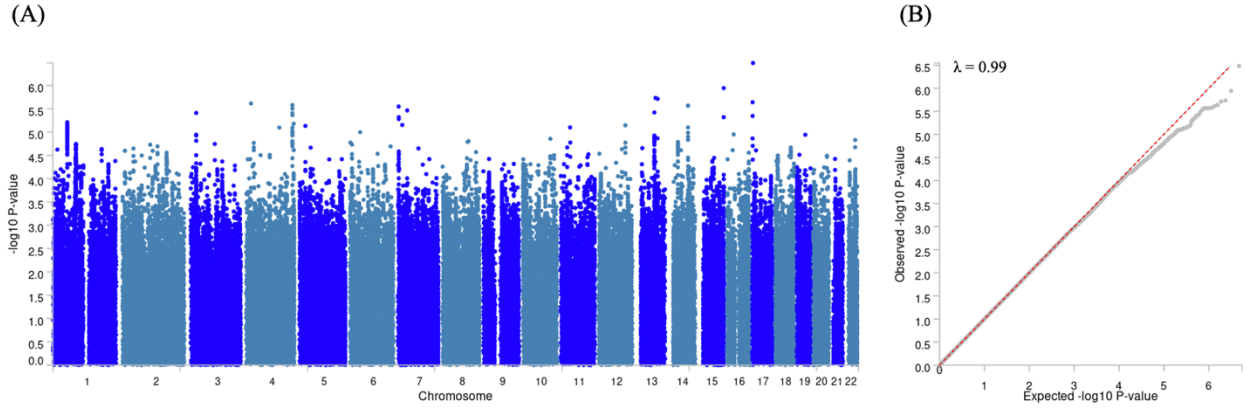


Figure 18. (A) Manhattan plot; (B) QQ plot of GWA scan summary statistics for CMSKP individuals with theta values

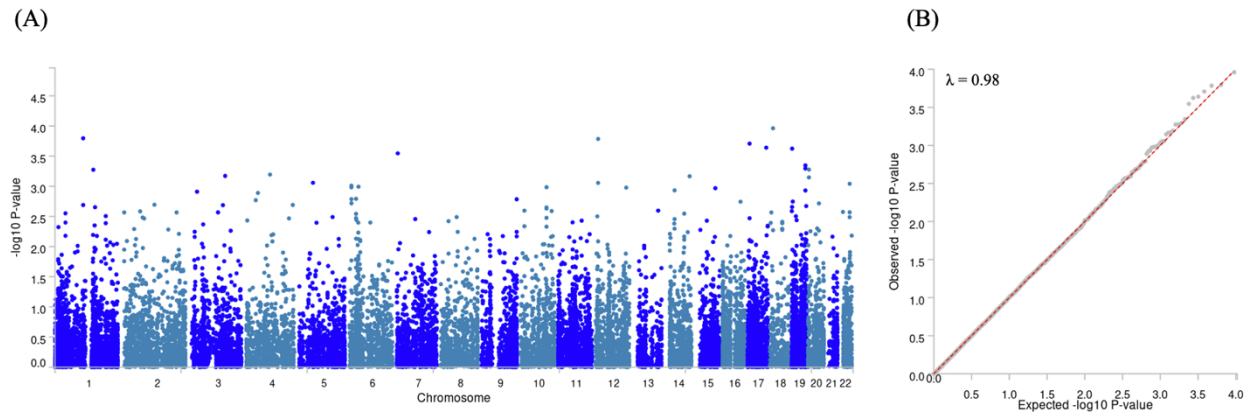


Figure 19. (A) Manhattan plot; (B) QQ plot of GWA scan gene-based test for CMSKP individuals with theta values

3.6 GWAS of CMSKP Case-only Using Topic 9 Theta Values as a Continuous Trait

The fourth approach focused on only CMSKP cases, a total of 156,235 individuals. No SNPs passed the genome-wide significance threshold of $P < 5 \times 10^{-8}$ (Figure 20), and no genes passed the genome-wide significance threshold of $P < 2.616 \times 10^{-6}$ (Figure 21).

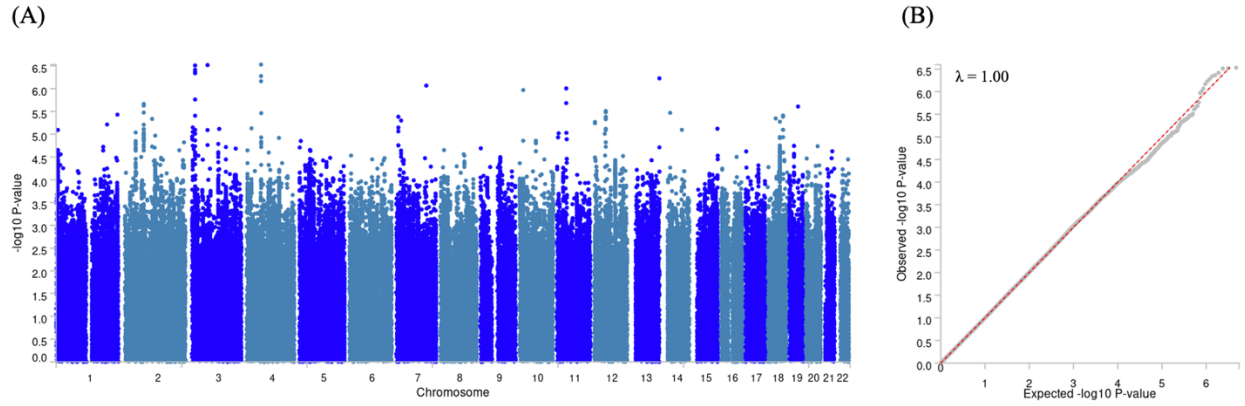


Figure 20. (A) Manhattan plot; (B) QQ plot of GWA scan summary statistics for CMSKP cases with Topic 9 theta values

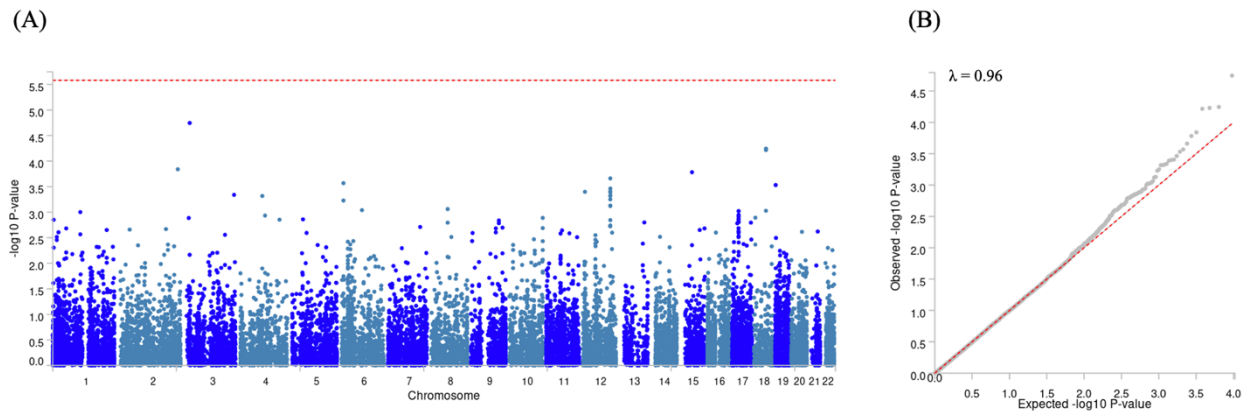


Figure 21. (A) Manhattan plot; (B) QQ plot of GWA scan gene-based test for CMSKP cases with Topic 9 theta values

Chapter 4 – Discussion

This study applied GETM to three UK Biobank features, self-reported non-cancer conditions, medications, and ICD-10 codes, to investigate its utility in structuring biomedical data into biologically meaningful topics. By integrating biomedical hierarchies into pre-trained graph embeddings, GETM aimed to capture multimorbidity patterns and refine phenotypic representations for CMSKP GWA scans. The results demonstrated that GETM effectively reduced about 2,000 features into 10 topics, uncovering latent patterns of disease co-occurrence. However, it failed to generate coherent and biologically relevant topics for CMSKP. To assess the feasibility of topic-derived phenotypes for GWA scans, four GWA scan approaches with different phenotype definitions were designed. Only the traditional binary case-control GWA scan yielded meaningful results for CMSKP, while all GETM-related approaches failed to confidently identify relevant genetic loci. Despite this limitation, this study introduces an innovative framework for integrating multi-phenotype data with genetic analyses, contributing to exploration of the genetic architecture of complex traits such as CMSKP.

4.1 CMSKP Findings Before Incorporating GETM

Since Section 3.3 (GWAS of CMSKP using a binary case-control design) was the only approach to yield biologically meaningful results for CMSKP, it serves as a baseline for evaluating the effectiveness of implementing topic-derived phenotypes in GWA scan. Examining these findings provides a foundation for assessing the impact of incorporating GETM-derived topics on the identification of genetic variants associated with CMSKP.

In 29 lead SNPs and 58 independent significant SNPs, three SNPs, rs3811474, rs13135092, and rs11599236, have been previously reported in the literature. rs3811474 has been associated

with an increased risk of MCP (Johnston et al., 2019), and has also been implicated in the genetic architecture of osteoarthritis (Aubourg et al., 2022). rs13135092 has been linked to a decreased risk of MCP (Johnston et al., 2019), and has shown a significant association with MCP in females, with a consistent protective direction of effect (Johnston et al., 2021). rs11599236 has been reported to increase the risk of MCP (Johnston et al., 2019), and has also been identified in a multi-trait GWA study of atherosclerosis, suggesting potential pleiotropic effects (Bellomo et al., 2021). These findings highlight the shared genetic mechanisms between CMSKP and MCP, while also revealing 26 novel lead SNPs associated with CMSKP. Notably, these

In the gene-based test, the top 5 genes were *PABPC4*, *TCTA*, *DCC*, *MACF1*, and *BSN*, each with varying degrees of evidence linking them to pain mechanisms.

PABPC4 (Poly(A) Binding Protein Cytoplasmic 4) encodes an RNA-binding protein that regulates gene expression by stabilizing mRNA and enhancing translation (Jiao et al., 2021). The inhibition of *PABPC4* has been shown to reduce pain sensitization in mice, and its expression in spinal cord, dorsal root ganglia (DRG), and sciatic nerve tissues overlaps with pain-related markers, such as *TRPV1*, *CGRP*, and IB4-positive fibers, suggesting a potential role in nociception and pain regulation (Barragán-Iglesias et al., 2018). Notably, our study is the first to identify *PABPC4* in relation to pain phenotypes in humans.

TCTA (T Cell Leukemia Translocation Altered) was originally identified in chromosomal translocations in T-cell leukemia and involves in osteoclast differentiation and fusion, with potential implications for bone-related conditions (Kotake et al., 2009). However, its relevance to CMSKP or other pain phenotypes remains unexplored.

DCC (Deleted in Colorectal Cancer) encodes a netrin 1 receptor involved in neuronal guidance and has been reported to be associated with pain. *DCC* is the only gene that is significantly associated with MCP in both genders (Johnston et al., 2021).

MACF1 (Microtubule-Actin Crosslinking Factor 1) encodes a cytoskeletal crosslinking protein involved in Wnt signaling and neural processes (Moffat et al., 2017). Despite its role in neuronal function, no direct evidence currently links *MACF1* to pain mechanisms.

BSN (Bassoon Presynaptic Cytomatrix) encodes Bassoon, a large scaffolding protein essential for synaptic transmission. A genomic structural equation modeling (SEM) study on 24 distinct pain conditions in the UK Biobank identified *BSN* among the top 31 genes associated with a general pain factor (Zorina-Lichtenwalter et al., 2023).

4.2 Insights into GETM Development and Application

4.2.1 Challenges in Including Medication Data for Phenotyping

When we incorporated medication data into GETM, the UMAP of medication embedding and topic embedding showed that medications from the same anatomical category were poorly grouped together. The inadequate clustering of medications influenced the quality and interpretability of the embeddings and topic generation, ultimately leading us to exclude medication data from the further analyses.

This discrepancy may arise from the inherent complexity of medication categorization. For example, ibuprofen, a common pain-relieving medication, can be classified under multiple categories due to its diverse applications, such as an anti-inflammatory drug and an analgesic (Guo et al., 2024). While categorizing each medication based on its first level in the ATC classification system ensures a manageable number of categories and maintains comparability with other

features, such as condition classifications, it also overlooks the medication’s therapeutic versatility. This simplification can introduce noise into the embeddings and reduce the reliability of the topics generated by GETM. Additionally, the ATC classification provides detailed information, including detailed chemical properties of medications, which may not work well in helping with bringing comorbidities together. Medications often have multiple indications and off-label uses, further complicating their integration into topic modeling (Rusz et al., 2021). For example, a medication prescribed for pain relief may also be used for non-pain-related conditions, introducing ambiguity in the relationships between medications and disease phenotypes. This ambiguity can hinder the interpretability of topics and reduce the precision of embeddings.

In addition, medication records may reflect socio-economic and behavioral factors. Although the UK healthcare system aims to provide equal access to care, individuals with higher socio-economic status may be more proactive in seeking preventive care or receive earlier diagnoses, which could influence the likelihood and type of medications prescribed. These differences may introduce biases into the medication data, potentially affecting how GETM infers topics.

In summary, while medication data may offer valuable insights into chronic pain management and treatment, its inclusion in GETM presented several challenges, including inadequate clustering, misclassification, and excessive granularity. These challenges highlight the need for careful consideration when integrating medication data into phenotyping models. For this study, excluding medication data allowed us to focus more effectively on identifying meaningful pain-related phenotypes in CMSKP.

4.2.2 Embedding Approaches in GETM: Strengths and Limitations

The uniqueness of GETM lies in its graph-embedding approach, which captures the biomedical hierarchy of features and encodes them into a low-dimensional matrix. This strength makes GETM particularly valuable for studying multifactorial phenotypes, such as chronic pain (Wang et al., 2022). We aimed to evaluate whether GETM’s ability to account for these relationships improves the discovery of underlying patterns in clinical data.

By incorporating node2vec to generate graph-based embeddings as input to the ETM algorithm, GETM presents several strengths. First, it effectively captures and preserves relationships among medical conditions and ICD-10 codes, such as parent-child or comorbid associations, in its data representation. Second, its flexibility of input data allows it to take in different health-related data depending on the phenotype and associated features being studied. Additionally, unlike traditional bag-of-words approaches, graph embeddings enhance the interpretability of the resulting topics, enabling GETM to investigate the probability and contribution of features within each topic.

While the embedding function of GETM offers several advantages, it also has limitations when applied to the study of chronic pain. First, the quality of node2vec embeddings highly depends on the input graph structure. The confusing relationships within features, such as ibuprofen that can belong to anti-inflammatory or analgesic category, can propagate through the embedding process, resulting in less distinct topic generation. Second, the graph embedding may bias the model toward relationships explicitly encoded in the graph, such as the over-detailed representation of condition in ICD-10 codes, potentially overshadowing less structured but important associations in the text data. These limitations reduce the model’s capacity to capture

unexpected or novel connections and contribute to the failure of using topic-derived phenotypes for GWA scans.

4.2.3 Evaluating Topic Quality: Methodological Insights

Evaluating the quality of topics generated by the GETM is crucial for two main reasons. First, topic quality reflects whether the topics effectively capture the patterns in the input data. Second, it informs the selection of the optimal number of topics, which is essential for conducting GWA scans. The optimal number balances the features' coherence within the topics and the features' distinction across topics, enabling accurate associations between genetic variants and clinical phenotypes. A topic count that is too high may lead to overfitting, producing overly specific or fragmented topics with limited generalizability. On the other hand, a count that is too low may result in underfitting, capturing excess noise and masking important relationships. Striking the right balance enhances both interpretability and precision in subsequent analyses, such as GWA scans.

In this study, we adopted the topic quality evaluation method from the GETM publication, where topic quality is determined by the product of topic coherence and topic diversity (Wang et al., 2022). Our results indicated that the best number of topics out of tested options for this study was 10, which aligns with findings from similar studies that also used 11 topics for subsequent GWA scans (Zhang et al., 2023). As expected, we observed a trend where increasing the number of topics led to an increase in coherence, accompanied by a decrease in diversity. This trend is reasonable, as smaller topics tend to be more focused and cohesive, while larger topics may capture broader, less specific patterns. However, despite this expected trend, the values for both topic coherence and diversity did not show substantial variation across different numbers of topics. This suggests that the current method may not be optimal for evaluating topic qualities. For future work,

we plan to implement other methods, such as using K-nearest neighbors to evaluate the accuracy at various values of K to identify the optimal number of topics (Zhang, 2016).

4.3 Application of GETM-GWAS to Pain Genomics

4.3.1 Relevance of GETM-derived GWA Scans in Studying CMSKP

The GETM-derived GWAS offers a novel framework for studying chronic pain development. By using the topic as the latent representation of co-occurrence patterns of medical conditions, GETM-derived GWA scans have the potential to address genetic heterogeneity and enhance the detection of genetic signals. This is particularly important for complex conditions like CMSKP. Our goal was to observe whether GETM-derived GWA scans could provide new insights into CMSKP by identifying novel loci that may be overlooked by traditional GWA scans. However, due to the lack of biologically meaningful topics for CMSKP, we could not confidently assess its ability to uncover genetic variants that traditional methods might miss.

Several challenges remain in applying GETM-derived GWAS to study CMSKP. Topic interpretability emerged as a notable limitation, with features within the predictive topic lacking direct relevance to pain phenotypes. Incorporating more precise data, such as medication use with careful annotations regarding dosages, treatment durations, and off-label uses, could potentially improve the clinical relevance and accuracy of derived topics. Also, optimizing the GETM algorithm by focusing on one feature input could be another approach. Despite these challenges, this study provides a novel design for incorporating multi-phenotype data into pain genetics and proposes a possible approach for future research on the genomics of complex traits.

4.4 Overall Strengths and Limitations of this Study

This study presents several strengths that highlight its contributions to the field of chronic pain development. First, it represents a novel application of GETM-derived GWA scans to investigate CMSKP, marking the first time topic modeling has been used in a pain context. Second, by leveraging GETM for dimensionality reduction, the study successfully distilled complex medical data into latent topics, capturing meaningful patterns of disease co-occurrence. Third, the integration of these topic-derived phenotypes into GWA scans provides a new approach for studying CMSKP. Unlike traditional single-condition GWA scans, topic-derived GWA scans have been shown to identify novel or distinct genetic loci associated with broader phenotypic categories (Zhang et al., 2023). This was also evident in our analysis, where comparisons between the binary case-control design and the refined GWA of CMSKP with different case definitions, using their respective theta values, highlighted the same and additional genetic loci. Fourth, the use of the UK Biobank, a large and well-characterized dataset with pain-related data, provided high statistical power and enhanced the validity of the findings. Finally, this study demonstrates scalability in terms of its ability to be adapted for use with diverse datasets and complex traits. The framework is not limited to CMSKP but can be applied to other pain-related conditions by incorporating relevant data features.

This study also has several limitations that should be considered when interpreting the findings. First, the lack of a consistent version of embeddings generated by node2vec and the absence of consistent topic generation across analyses introduced variability in the topic modeling process, which may affect the reproducibility and robustness of results. Second, GETM was designed to incorporate only two classes of features. While the self-reported medical conditions and hospital-derived ICD-10 codes used in this study were informative for chronic pain

development, other features, such as functional assessments (e.g., pain severity scales, psychological questionnaires, or quality of life measures), could provide additionally meaningful insights. Third, the interpretability of topics was impossible. Fourth, the definition of CMSKP itself posed challenges. Although CMSKP in the UK Biobank is typically defined as chronic pain localized to one or more out of four body sites (knee, back, neck/shoulder, or hip), there is no consistent or absolute definition of this condition. This variability in the definition could limit the generalizability of findings across different study designs or populations. Fifth, we only kept autosomal SNPs from our analyses. Finally, the limitations of the GETM framework, including its reliance on unsupervised learning, may have hindered the ability to uncover pain-specific biological mechanisms. These limitations highlight the need for further methodological refinement in future research.

Chapter 5 – Conclusions and Future Directions

This study introduces a novel application of GETM-derived GWA scans to explore the genetic architecture of CMSKP within the UK Biobank dataset. By structuring self-derived medical conditions and hospitalized ICD-10 codes into topics, this approach captured disease co-occurrence patterns that highlighted the phenotypic complexity. These topic-derived phenotypes, when integrated into GWA scans, enabled the identification of novel genetic loci and reinforced findings from traditional design of binary CMSKP cases and controls. Furthermore, this work emphasizes the value of incorporating comorbid conditions into genetic studies, offering insights into the shared biological pathways between CMSKP and related disorders. Collectively, these findings underscore the promise of combining topic modeling with genetic analyses as a transformative framework for unraveling the complexities of chronic pain and similar multifactorial traits.

Building on the findings of this study, several future directions could enhance the utility and applicability of GETM in pain genomics. First, refining topic modeling methodologies is essential to improve the interpretability and clinical relevance of the derived topics. Strategies to achieve this include Adjusted Rand Index (ARI) (Chacón & Rastrojo, 2020) or Normalized Mutual Information (NMI) (McDaid et al., 2011) to check embedding performance and enhance the relevance of topics. Additionally, incorporating domain knowledge, such as clinical expert input or phenotypic annotations, could help ensure that topics reflect meaningful clinical phenotypes while avoiding the inclusion of irrelevant features. Second, integrating additional data types could improve the clinical relevance and validity of the derived topics. These additional layers of information would offer a more comprehensive view of the phenotypic diversity in CMSKP and refine our understanding of how genetic variation interacts with environmental factors. Third,

increasing the sample size could increase the statistical power for identifying the genetic loci associated with CMSKP, such as leveraging multi-biobank meta-analyses. Fourth, refining CMSKP phenotype definitions could also lead to more precise and reproducible genetic associations. This could involve adopting more granular phenotypic classifications, such as considering specific subtypes of chronic pain (e.g., neuropathic versus nociceptive pain) or incorporating biomarkers to enhance phenotypic stratification. Collectively, these future directions have the potential to bridge the gap between complex, data-driven methodologies and their clinical and translational applications, thereby advancing the study of chronic pain.

Chapter 6 – References

- Aubourg, G., Rice, S. J., Bruce-Wootton, P., & Loughlin, J. (2022). Genetics of osteoarthritis. *Osteoarthritis Cartilage*, 30(5), 636-649. <https://doi.org/10.1016/j.joca.2021.03.002>
- Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., . . . National Eye Institute, N. I. H. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68-74. <https://doi.org/10.1038/nature15393>
- Baral, P., Udit, S., & Chiu, I. M. (2019). Pain and immunity: implications for host defence. *Nature Reviews Immunology*, 19(7), 433-447. <https://doi.org/10.1038/s41577-019-0147-2>
- Barragán-Iglesias, P., Lou, T.-F., Bhat, V. D., Megat, S., Burton, M. D., Price, T. J., & Campbell, Z. T. (2018). Inhibition of Poly(A)-binding protein with a synthetic RNA mimic reduces pain sensitization in mice. *Nature Communications*, 9(1), 10. <https://doi.org/10.1038/s41467-017-02449-5>
- Bellomo, T. R., Bone, W. P., Chen, B. Y., Gawronski, K. A. B., Zhang, D., Park, J., Levin, M., Tsao, N., Klarin, D., Lynch, J., Assimes, T. L., Gaziano, J. M., Wilson, P. W., Cho, K., Vujkovic, M., O'Donnell, C. J., Chang, K. M., Tsao, P. S., Rader, D. J., . . . Voight, B. F. (2021). Multi-Trait Genome-Wide Association Study of Atherosclerosis Detects Novel Pleiotropic Loci. *Front Genet*, 12, 787545. <https://doi.org/10.3389/fgene.2021.787545>
- Bick, A. G., Metcalf, G. A., Mayo, K. R., Lichtenstein, L., Rura, S., Carroll, R. J., Musick, A., Linder, J. E., Jordan, I. K., Nagar, S. D., Sharma, S., Meller, R., Basford, M., Boerwinkle, E., Cicek, M. S., Doheny, K. F., Eichler, E. E., Gabriel, S., Gibbs, R. A., . . . Staff, N. I. H.

- A. o. U. R. P. (2024). Genomic data in the All of Us Research Program. *Nature*, 627(8003), 340-346. <https://doi.org/10.1038/s41586-023-06957-x>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726), 203-209. <https://doi.org/10.1038/s41586-018-0579-z>
- Chacón, J. E., & Rastrojo, A. I. (2020). Minimum adjusted Rand index for two clusterings of a given size. *arXiv preprint arXiv:2002.03677*.
- Chen, L., Zeng, W. M., Cai, Y. D., Feng, K. Y., & Chou, K. C. (2012). Predicting Anatomical Therapeutic Chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. *PLoS One*, 7(4), e35254. <https://doi.org/10.1371/journal.pone.0035254>
- Colloca, L., Ludman, T., Bouhassira, D., Baron, R., Dickenson, A. H., Yarnitsky, D., Freeman, R., Truini, A., Attal, N., Finnerup, N. B., Eccleston, C., Kalso, E., Bennett, D. L., Dworkin, R. H., & Raja, S. N. (2017). Neuropathic pain. *Nature Reviews Disease Primers*, 3(1), 17002. <https://doi.org/10.1038/nrdp.2017.2>
- Coppola, L., Cianflone, A., Grimaldi, A. M., Incoronato, M., Bevilacqua, P., Messina, F., Baselice, S., Soricelli, A., Mirabelli, P., & Salvatore, M. (2019). Biobanking in health care: evolution and future directions. *Journal of Translational Medicine*, 17(1), 172. <https://doi.org/10.1186/s12967-019-1922-3>

- Davis, K. A. S., Bashford, O., Jewell, A., Shetty, H., Stewart, R. J., Sudlow, C. L. M., & Hotopf, M. H. (2018). Using data linkage to electronic patient records to assess the validity of selected mental health diagnoses in English Hospital Episode Statistics (HES). *PLoS One*, 13(3), e0195002. <https://doi.org/10.1371/journal.pone.0195002>
- de Leeuw, C. A., Mooij, J. M., Heskes, T., & Posthuma, D. (2015). MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology*, 11(4), e1004219. <https://doi.org/10.1371/journal.pcbi.1004219>
- De Souza, Y. G., & Greenspan, J. S. (2013). Biobanking past, present and future: responsibilities and benefits. *Aids*, 27(3), 303-312. <https://doi.org/10.1097/QAD.0b013e32835c1244>
- Diatchenko, L., Fillingim, R. B., Smith, S. B., & Maixner, W. (2013). The phenotypic and genetic signatures of common musculoskeletal pain conditions. *Nature Reviews Rheumatology*, 9(6), 340-350. <https://doi.org/10.1038/nrrheum.2013.43>
- Diatchenko, L., Slade, G. D., Nackley, A. G., Bhalang, K., Sigurdsson, A., Belfer, I., Goldman, D., Xu, K., Shabalina, S. A., Shagin, D., Max, M. B., Makarov, S. S., & Maixner, W. (2005). Genetic basis for individual variations in pain perception and the development of a chronic pain condition. *Human Molecular Genetics*, 14(1), 135-143. <https://doi.org/10.1093/hmg/ddi013>
- Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2019). Topic Modeling in Embedding Spaces. In: arXiv.
- El-Tallawy, S. N., Nalamasu, R., Salem, G. I., LeQuang, J. A. K., Pergolizzi, J. V., & Christo, P. J. (2021). Management of Musculoskeletal Pain: An Update with Emphasis on Chronic Musculoskeletal Pain. *Pain and Therapy*, 10(1), 181-209. <https://doi.org/10.1007/s40122-021-00235-2>

- Foley, H. E., Knight, J. C., Ploughman, M., Asghari, S., & Audas, R. (2021). Association of chronic pain with comorbidities and health care utilization: a retrospective cohort study using health administrative data. *Pain*, 162(11), 2737-2749.
<https://doi.org/10.1097/j.pain.0000000000002264>
- Free, T. (2024). UK Biobank: what can it do, how you can use it and how is it being used? *BioTechniques*, 76(12), 553-557. <https://doi.org/10.1080/07366205.2024.2441639>
- Gold, M. S., & Gebhart, G. F. (2010). Nociceptor sensitization in pain pathogenesis. *Nature Medicine*, 16(11), 1248-1257. <https://doi.org/10.1038/nm.2235>
- Grover, A., & Leskovec, J. (2016, 2016//08/13). node2vec: Scalable Feature Learning for Networks. *KDD '16*
- Guo, H., Ma, R., Zhang, Y., Yin, K., Du, G., Yin, F., Li, H., Wang, Z., & Yin, D. (2024). Ibuprofen inhibits anaplastic thyroid cells in vivo and in vitro by triggering NLRP3-ASC-GSDMD-dependent pyroptosis. *Inflammopharmacology*, 32(1), 733-745.
<https://doi.org/10.1007/s10787-023-01379-7>
- Haendel, M. A., Chute, C. G., & Robinson, P. N. (2018). Classification, Ontology, and Precision Medicine. *N Engl J Med*, 379(15), 1452-1462. <https://doi.org/10.1056/NEJMr1615014>
- Henschke, N., Kamper, S. J., & Maher, C. G. (2015). The Epidemiology and Economic Consequences of Pain. *Mayo Clinic Proceedings*, 90(1), 139-147.
<https://doi.org/10.1016/j.mayocp.2014.09.010>
- Hirsch, J. A., Nicola, G., McGinty, G., Liu, R. W., Barr, R. M., Chittle, M. D., & Manchikanti, L. (2016). ICD-10: History and Context. *American Journal of Neuroradiology*, 37(4), 596-599. <https://doi.org/10.3174/ajnr.A4696>

- Hodges, S., Guler, S., Sacca, V., Vangel, M., Orr, S., Pace-Schott, E., Wen, Y., Ge, T., & Kong, J. (2023). Associations among acute and chronic musculoskeletal pain, sleep duration, and C-reactive protein (CRP): A cross-sectional study of the UK biobank dataset. *Sleep Med*, 101, 393-400. <https://doi.org/10.1016/j.sleep.2022.11.013>
- Hooten, W. M. (2016). Chronic Pain and Mental Health Disorders: Shared Neural Mechanisms, Epidemiology, and Treatment. *Mayo Clinic Proceedings*, 91(7), 955-970. <https://doi.org/https://doi.org/10.1016/j.mayocp.2016.04.029>
- Jiang, X., Zhang, M. J., Zhang, Y., Durvasula, A., Inouye, M., Holmes, C., Price, A. L., & McVean, G. (2023). Age-dependent topic modeling of comorbidities in UK Biobank identifies disease subtypes with differential genetic risk. *Nature Genetics*, 55(11), 1854-1865. <https://doi.org/10.1038/s41588-023-01522-8>
- Jiao, Y., Kong, N., Wang, H., Sun, D., Dong, S., Chen, X., Zheng, H., Tong, W., Yu, H., Yu, L., Huang, Y., Wang, H., Sui, B., Zhao, L., Liao, Y., Zhang, W., Tong, G., & Shan, T. (2021). PABPC4 Broadly Inhibits Coronavirus Replication by Degrading Nucleocapsid Protein through Selective Autophagy. *Microbiol Spectr*, 9(2), e0090821. <https://doi.org/10.1128/Spectrum.00908-21>
- Johnston, K. J. A., Adams, M. J., Nicholl, B. I., Ward, J., Strawbridge, R. J., Ferguson, A., McIntosh, A. M., Bailey, M. E. S., & Smith, D. J. (2019). Genome-wide association study of multisite chronic pain in UK Biobank. *PLOS Genetics*, 15(6), e1008164. <https://doi.org/10.1371/journal.pgen.1008164>
- Johnston, K. J. A., & Huckins, L. M. (2023). Chronic Pain and Psychiatric Conditions. *Complex Psychiatry*, 9(1-4), 24-43. <https://doi.org/10.1159/000527041>

- Johnston, K. J. A., Ward, J., Ray, P. R., Adams, M. J., McIntosh, A. M., Smith, B. H., Strawbridge, R. J., Price, T. J., Smith, D. J., Nicholl, B. I., & Bailey, M. E. S. (2021). Sex-stratified genome-wide association study of multisite chronic pain in UK Biobank. *PLoS Genet*, 17(4), e1009428. <https://doi.org/10.1371/journal.pgen.1009428>
- Kasher, M., Williams, F. M., Freidin, M. B., Cherny, S. S., Malkin, I., Livshits, G., & Group, C. I. W. (2023). Insights into the pleiotropic relationships between chronic back pain and inflammation-related musculoskeletal conditions: rheumatoid arthritis and osteoporotic abnormalities. *Pain*, 164(3), e122-e134.
- Khoury, S., Parisien, M., Thompson, S. J., Vachon-Preseu, E., Roy, M., Martinsen, A. E., Winsvold, B. S., Mundal, I. P., Zwart, J. A., Kania, A., Mogil, J. S., & Diatchenko, L. (2022). Genome-wide analysis identifies impaired axonogenesis in chronic overlapping pain conditions. *Brain*, 145(3), 1111-1123. <https://doi.org/10.1093/brain/awab359>
- Kotake, S., Nanke, Y., Kawamoto, M., Yago, T., Udagawa, N., Ichikawa, N., Kobashigawa, T., Saito, S., Momohara, S., Kamatani, N., & Yamanaka, H. (2009). T-cell leukemia translocation-associated gene (TCTA) protein is required for human osteoclastogenesis. *Bone*, 45(4), 627-639. <https://doi.org/https://doi.org/10.1016/j.bone.2009.06.019>
- Li, S., Brimmers, A., van Boekel, R. L. M., Vissers, K. C. P., & Coenen, M. J. H. (2023). A systematic review of genome-wide association studies for pain, nociception, neuropathy, and pain treatment responses. *Pain*, 164(9), 1891-1911. <https://doi.org/10.1097/j.pain.0000000000002910>
- Lin, L., Lin, J., Qiu, J., Liufu, N., Lin, S., Wei, F., Liu, Q., Zeng, J., Zhang, M., & Cao, M. (2023). Genetic liability to multi-site chronic pain increases the risk of cardiovascular

- disease. *British Journal of Anaesthesia*, 131(2), 373-384.
- <https://doi.org/https://doi.org/10.1016/j.bja.2023.04.020>
- Martinsen, A. E., Børte, S., Spildrejorde, M., Brumpton, B. M., Heuch, I., Zwart, J. A., & Winsvold, B. S. (2025). Insights into Chronic Low Back Pain Etiology: Population-based genome-wide Association Study Identifies 18 Risk Loci. *Spine (Phila Pa 1976)*.
- <https://doi.org/10.1097/brs.00000000000005254>
- Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., Habegger, L., Ferreira, M., Baras, A., Reid, J., Abecasis, G., Maxwell, E., & Marchini, J. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, 53(7), 1097-1103.
- <https://doi.org/10.1038/s41588-021-00870-7>
- McDaid, A. F., Greene, D., & Hurley, N. (2011). Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*.
- McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- Meng, W., Adams, M. J., Palmer, C. N. A., Agee, M., Alipanahi, B., Bell, R. K., Bryc, K., Elson, S. L., Fontanillas, P., Furlotte, N. A., Hicks, B., Hinds, D. A., Huber, K. E., Jewett, E. M., Jiang, Y., Kleinman, A., Lin, K.-H., Litterman, N. K., McCreight, J. C., . . . The 23andMe Research, T. (2019). Genome-wide association study of knee pain identifies associations with GDF5 and COL27A1 in UK Biobank. *Communications Biology*, 2(1), 321.
- <https://doi.org/10.1038/s42003-019-0568-2>
- Meng, W., Adams, M. J., Reel, P., Rajendrakumar, A., Huang, Y., Deary, I. J., Palmer, C. N. A., McIntosh, A. M., & Smith, B. H. (2020). Genetic correlations between pain phenotypes

- and depression and neuroticism. *European Journal of Human Genetics*, 28(3), 358-366.
<https://doi.org/10.1038/s41431-019-0530-2>
- Meng, W., Chan, B. W., Harris, C., Freidin, M. B., Hebert, H. L., Adams, M. J., Campbell, A., Hayward, C., Zheng, H., Zhang, X., Colvin, L. A., Hales, T. G., Palmer, C. N. A., Williams, F. M. K., McIntosh, A., & Smith, B. H. (2020). A genome-wide association study finds genetic variants associated with neck or shoulder pain in UK Biobank. *Hum Mol Genet*, 29(8), 1396-1404. <https://doi.org/10.1093/hmg/ddaa058>
- Mills, M., Barban, N., & Tropf, F. C. (2020). *An introduction to statistical genetic data analysis / Melinda C. Mills, Nicola Barban, and Felix C. Tropf*. The MIT Press.
- Mocci, E., Ward, K., Perry, J. A., Starkweather, A., Stone, L. S., Schabrun, S. M., Renn, C., Dorsey, S. G., & Ament, S. A. (2023). Genome wide association joint analysis reveals 99 risk loci for pain susceptibility and pleiotropic relationships with psychiatric, metabolic, and immunological traits. *PLoS Genet*, 19(10), e1010977.
<https://doi.org/10.1371/journal.pgen.1010977>
- Moffat, J. J., Ka, M., Jung, E.-M., Smith, A. L., & Kim, W.-Y. (2017). The role of MACF1 in nervous system development and maintenance. *Seminars in Cell & Developmental Biology*, 69, 9-17. <https://doi.org/https://doi.org/10.1016/j.semcdb.2017.05.020>
- Mutz, J., Roscoe, C. J., & Lewis, C. M. (2021). Exploring health in the UK Biobank: associations with sociodemographic characteristics, psychosocial factors, lifestyle and environmental exposures. *BMC Med*, 19(1), 240. <https://doi.org/10.1186/s12916-021-02097-z>
- Pan, F., Byrne, K. S., Ramakrishnan, R., Ferreira, M., Dwyer, T., & Jones, G. (2019). Association between musculoskeletal pain at multiple sites and objectively measured physical activity

- and work capacity: Results from UK Biobank study. *Journal of Science and Medicine in Sport*, 22(4), 444-449. <https://doi.org/https://doi.org/10.1016/j.jsams.2018.10.008>
- Papez, V., Moinat, M., Voss, E. A., Bazakou, S., Van Winzum, A., Peviani, A., Payralbe, S., Kallfelz, M., Asselbergs, F. W., Prieto-Alhambra, D., Dobson, R. J. B., & Denaxas, S. (2022). Transforming and evaluating the UK Biobank to the OMOP Common Data Model for COVID-19 research and beyond. *J Am Med Inform Assoc*, 30(1), 103-111. <https://doi.org/10.1093/jamia/ocac203>
- Rahman, M. S., Winsvold, B. S., Chavez Chavez, S. O., Børte, S., Tsepilov, Y. A., Sharapov, S. Z., Aulchenko, Y. S., Hagen, K., Fors, E. A., Hveem, K., Zwart, J. A., van Meurs, J. B., Freidin, M. B., & Williams, F. M. (2021). Genome-wide association study identifies RNF123 locus as associated with chronic widespread musculoskeletal pain. *Ann Rheum Dis*, 80(9), 1227-1235. <https://doi.org/10.1136/annrheumdis-2020-219624>
- Robinson-Papp, J., George, M. C., Dorfman, D., & Simpson, D. M. (2015). Barriers to Chronic Pain Measurement: A Qualitative Study of Patient Perspectives. *Pain Med*, 16(7), 1256-1264. <https://doi.org/10.1111/pme.12717>
- Rönnegård, A.-S., Nowak, C., Äng, B., & Ärnlöv, J. (2022). The association between short-term, chronic localized and chronic widespread pain and risk for cardiovascular disease in the UK Biobank. *European Journal of Preventive Cardiology*, 29(15), 1994-2002. <https://doi.org/10.1093/eurjpc/zwac127>
- Rusz, C.-M., Ősz, B.-E., Jîtcă, G., Miklos, A., Bătrînu, M.-G., & Imre, S. (2021). Off-Label Medication: From a Simple Concept to Complex Practical Aspects. *International Journal of Environmental Research and Public Health*, 18(19), 10447. <https://www.mdpi.com/1660-4601/18/19/10447>

- Sammani, A., Bagheri, A., van der Heijden, P. G. M., te Riele, A. S. J. M., Baas, A. F., Oosters, C. A. J., Oberski, D., & Asselbergs, F. W. (2021). Automatic multilabel detection of ICD10 codes in Dutch cardiology discharge letters using neural networks. *npj Digital Medicine*, 4(1), 37. <https://doi.org/10.1038/s41746-021-00404-9>
- Stanciu, I., Anderson, J., Siebert, S., Mackay, D., & Lyall, D. M. (2022). Associations of rheumatoid arthritis and rheumatoid factor with mental health, sleep and cognition characteristics in the UK Biobank. *Sci Rep*, 12(1), 19844. <https://doi.org/10.1038/s41598-022-22021-6>
- Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3), e1001779. <https://doi.org/10.1371/journal.pmed.1001779>
- Tang, Y., Liu, W., Kong, W., Zhang, S., & Zhu, T. (2023). Multisite chronic pain and the risk of autoimmune diseases: A Mendelian randomization study. *Front Immunol*, 14, 1077088. <https://doi.org/10.3389/fimmu.2023.1077088>
- Tanguay-Sabourin, C., Fillingim, M., Guglietti, G. V., Zare, A., Parisien, M., Norman, J., Sweatman, H., Da-ano, R., Heikkala, E., Perez, J., Karppinen, J., Villeneuve, S., Thompson, S. J., Martel, M. O., Roy, M., Diatchenko, L., & Vachon-Presseau, E. (2023). A prognostic risk score for development and spread of chronic pain. *Nature Medicine*, 29(7), 1821-1831. <https://doi.org/10.1038/s41591-023-02430-4>
- Todd, K. H., Ducharme, J., Choiniere, M., Crandall, C. S., Fosnocht, D. E., Homel, P., & Tanabe, P. (2007). Pain in the Emergency Department: Results of the Pain and Emergency

- Medicine Initiative (PEMI) Multicenter Study. *The Journal of Pain*, 8(6), 460-466.
<https://doi.org/10.1016/j.jpain.2006.12.005>
- Treede, R. D., Rief, W., Barke, A., Aziz, Q., Bennett, M. I., Benoliel, R., Cohen, M., Evers, S., Finnerup, N. B., First, M. B., Giamberardino, M. A., Kaasa, S., Kosek, E., Lavand'homme, P., Nicholas, M., Perrot, S., Scholz, J., Schug, S., Smith, B. H., . . . Wang, S. J. (2015). A classification of chronic pain for ICD-11. *Pain*, 156(6), 1003-1007.
<https://doi.org/10.1097/j.pain.0000000000000160>
- Tsepilov, Y. A., Freidin, M. B., Shadrina, A. S., Sharapov, S. Z., Elgaeva, E. E., Zundert, J. v., Karssen, L. C., Suri, P., Williams, F. M. K., & Aulchenko, Y. S. (2020). Analysis of genetically independent phenotypes identifies shared genetic factors associated with chronic musculoskeletal pain conditions. *Communications Biology*, 3(1), 329.
<https://doi.org/10.1038/s42003-020-1051-9>
- Wang, Y., Benavides, R., Diatchenko, L., Grant, A. V., & Li, Y. (2022). A graph-embedded topic model enables characterization of diverse pain phenotypes among UK biobank individuals. *iScience*, 25(6), 104390. <https://doi.org/10.1016/j.isci.2022.104390>
- Wang, Y., Grant, A. V., & Li, Y. (2023). Implementation of a graph-embedded topic model for analysis of population-level electronic health records. *STAR protocols*, 4(1), 101966.
<https://doi.org/10.1016/j.xpro.2022.101966>
- Watanabe, K., Taskesen, E., van Bochoven, A., & Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*, 8(1), 1826.
<https://doi.org/10.1038/s41467-017-01261-5>
- Wu, Y., Byrne, E. M., Zheng, Z., Kemper, K. E., Yengo, L., Mallett, A. J., Yang, J., Visscher, P. M., & Wray, N. R. (2019). Genome-wide association study of medication-use and

associated disease in the UK Biobank. *Nature Communications*, 10(1), 1891.

<https://doi.org/10.1038/s41467-019-09572-5>

Zhang, Y., Jiang, X., Mentzer, A. J., McVean, G., & Lunter, G. (2023). Topic modeling identifies novel genetic loci associated with multimorbidities in UK Biobank. *Cell Genomics*, 3(8), 100371. <https://doi.org/10.1016/j.xgen.2023.100371>

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Ann Transl Med*, 4(11), 218. <https://doi.org/10.21037/atm.2016.03.37>

Zorina-Lichtenwalter, K., Bango, C. I., Van Oudenhove, L., Čeko, M., Lindquist, M. A., Grotzinger, A. D., Keller, M. C., Friedman, N. P., & Wager, T. D. (2023). Genetic risk shared across 24 chronic pain conditions: identification and characterization with genomic structural equation modeling. *Pain*, 164(10), 2239-2252.

<https://doi.org/10.1097/j.pain.0000000000002922>

Zorina-Lichtenwalter, K., Meloto, C. B., Khoury, S., & Diatchenko, L. (2016). Genetic predictors of human chronic pain conditions. *Neuroscience*, 338, 36-62.

<https://doi.org/10.1016/j.neuroscience.2016.04.041> (Nociception, Pain, and Analgesia)