COMP 565 Fall 2023 Research-based Final Project*

1 "Red pill" or "blue pill"

You have been given an option of taking the "blue pill" by doing an individual survey as a project for the equal worth of **30% of the course grade**. See the separate instruction for the survey. That being said, the research-based projects below are designed for anyone who is interested in "going deeper into the rabbit hole" (by taking the "red pill") or you just can't get enough from the Matrix.

2 General guideline for the final project report

Choose one of the 3 projects below. You may work on your own or with another classmate. For the latter, the two students will be assigned the same grade for the project. Therefore, it will be your responsibility to make sure the work is divided equally between you and your teammate. Regardless of your choice of the project, write your report in maximum 10 pages excluding figures and references. Use 0.75 inch margin all around, minimum 11 font size, and single-space (which are the same as this instruction). You are encouraged to discuss the project with your instructor and the TA. In addition, each project has a dedicated mentor PhD student, who has done research in the area. Their contact information are listed under each project below. I will add them to Ed so you may discuss with your mentor under specific project thread in there as well. **Strictly follow the anti-plagiarism policy posted on MyCourses Content**.

1. Background
    - Opportunities
    - Computational challenges
    - Related methods/studies
    - One paragraph summarizing the proposed approach and results
2. Materials and Methods
    - Overview of the proposed model
    - Details of the learning algorithm
    - Data processing
    - Simulation (if applicable)
    - Evaluation metric
3. Results Discuss your results centering at each figure:
    - Fig. 1 method/study overview
    - Fig. 2 quantitive/qualitative results 1 (e.g., method comparison)
    - Fig. 3 quantitive/qualitative results 2
    - Fig. 4 quantitive/qualitative results 3
4. Discussion
    - Summarize what have been accomplished in this study
    - Limitations of the presented method
    - Future work

*Latest version: `https://www.overleaf.com/read/xtbdswbgmxrj`

3 Deliverable

Submit the final report as instructed above. If you are a team of two, one of you can submit the report on behalf of the other. Also, submit your scripts in either zip file or a single file (e.g., Python notebook or R markdown).

4 Project 1: Variational inference of polygenic risk score using empirical prior

This project is an extension of Assignment 3 and covers the population genetics axis of the course.

Project mentor Shadi Zabad (`shadi.zabad@mail.mcgill.ca`) (PhD candidate in Li lab, research in population genetics and the first author of the VIPRS paper [1]).

4.1 VIPRS with empirical prior

In Lecture 9 & 10, we learned an efficient Bayesian inference algorithm VIPRS [1] for high-dimensional multiple regression applied in the context of polygenic risk score (PRS) problem. We also gained an appreciation of the reference genomic annotations in earlier lectures including Lec 3 on Stratified LDSC and Lec 7 on inferring causal SNPs based on the empirical prior.

In this project, we will explore the usefulness of empirical prior in PRS. For P SNPs and and N individuals, let $\mathbf{X}$ denotes $P \times N$ genotype matrix and $\mathbf{y} \in \mathbb{R}^{N \times 1}$ the phenotype vector. Let $a_{j,k} \in \{0, 1\}$ be the binary indicator for whether SNP j is in annotation k, i.e., $j \in \mathcal{C}_k$, where $\mathcal{C}_k$ is the set of SNPs that are in annotation k. Therefore, for K annotations, we have a $P \times K$ binary annotation matrix $\mathbf{A}$.

Let $\mathbf{w} \in \mathbb{R}^{K \times 1}$ be the unknown annotation weights, σ_k^2 the variance explained by SNP j if it is in annotation k, and $\boldsymbol{\beta}$ and $\mathbf{s}$ be the effect size and binary indicator variables of causal SNPs, respectively. Together, we have the following data generative process.
Prior:

$$\pi_j = \frac{1}{1 + \exp(-\sum_k a_{j,k} w_k)} = \sigma(\mathbf{a}_j \mathbf{w})$$
$$p(s_j | \pi_j) = Bernoulli(s_j | \pi_j) = \pi_j^{s_j} (1 - \pi_j)^{1 - s_j}$$
$$p(\beta_j | s_j = 1, a_{j,k} = 1) = \mathcal{N}(\beta_j; 0, \sigma_k^2)$$

Likelihood:

$$p(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}) = \mathcal{N}(\mathbf{X}(\mathbf{s} \circ \boldsymbol{\beta}), \sigma_\epsilon^2 I)$$

Mean-field variational distribution to approximate the true posterior for $p(\boldsymbol{\beta}, \mathbf{s} | X, y)$:

$$q(\boldsymbol{\beta}, \mathbf{s}) = \prod_j q(\beta_j | s_j) q(s_j)$$

4.2 Objective

Modify the code in A3 or the actual VIPRS code in `https://github.com/shz9/viprs` by implementing the updates of point estimates for the annotation weights $\mathbf{w} \in \mathbb{R}^{K \times 1}$ and the point estimates for the variance explained per annotation k: $\sigma_k^2 = \tau_k^{-1}$ for $k \in \{1, \ldots, K\}$. To achieve that, you will follow the Expectation-Maximization (EM) algorithm (aka Empirical Bayes) outlined in Algorithm 1. The algorithm is similar to the original VIPRS you implemented in A3 but with the changes highlighted in red to incorporate the functional priors. Equations (6) and (7) were derived by taking the partial derivative of the ELBO w.r.t. τ_k and $\mathbf{w}$, respectively. Make sure you know how to derive them. Because of the logistic function, there is no closed-form update for $\mathbf{w}$. Instead, we perform gradient ascent in (7) with some fixed learning rate η. Equation (5) is optional. You may fix it to 1 if you have trouble of convergence.

Algorithm 1 The EM algorithm of Bayesian PRS with spike-and-slab prior

E-step: For each SNP j given the estimates for all other SNPs $i \neq j$, do:

$$q^*(\beta_j|s_j = 1) = \mathcal{N}(\mu^*_{\beta_j}, 1/\tau^*_{\beta_j}) \tag{1}$$

$$\tau^*_{\beta_j} = \mathbf{x}'_j\mathbf{x}_j\tau_\epsilon + \sum_{k \text{ where } a_{j,k}=1} \tau_k \tag{2}$$

$$\mu^*_{\beta_j} = N\frac{\tau_\epsilon}{\tau^*_{\beta_j}}\left(\mathbf{y}'\mathbf{x}_j/N - \sum_{i\neq j}\gamma^*_i\mu^*_{\beta_i}r_{ij}\right) \tag{3}$$

$$q^*(s_j = 1) = \frac{1}{1+\exp(-u_j)} \equiv \gamma^*_j, \quad u_j = \ln\frac{\pi_j}{1-\pi_j} + \frac{1}{2}\ln\frac{\tau_\beta}{\tau^*_{\beta_j}} + \frac{\tau^*_{\beta_j}}{2}{\mu^*_{\beta_j}}^2 \tag{4}$$

M-step: update hyperparameters:

$$\tau_\epsilon = \frac{1}{N}\Big(\mathbf{y}'\mathbf{y} - 2(\gamma^* \circ \mu^*)'\mathbf{X}'\mathbf{y} + \sum_j \gamma^*_j({\mu^*_{\beta_j}}^2 + 1/\tau^*_{\beta_j})\mathbf{x}'_j\mathbf{x}_j + 2\sum_{j=1}^{M}\sum_{k=j+1}^{M}\gamma^*_j\mu^*_{\beta_j}\gamma^*_k\mu^*_{\beta_k}\mathbf{x}'_k\mathbf{x}_j\Big) \tag{5}$$

$$\tau_k^{-1} = \sum_{j\in\mathcal{C}_k}\gamma^*_j({\mu^*_j}^2 + {\tau^*_{\beta_j}}^{-1})/\sum_{j\in\mathcal{C}_k}\gamma^*_j \tag{6}$$

$$\mathbf{w} = \mathbf{w} - \eta\nabla\mathbf{w}, \quad \text{where} \quad \nabla\mathbf{w} = (\boldsymbol{\gamma}^* - \boldsymbol{\pi})\mathbf{A} \tag{7}$$

4.3 Data

Linkage-Disequilibrium matrices The LD matrices can be computed from a reference panel, such as the 1000 Genome dataset (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`) or you can download pre-computed LD matrices from the UK Biobank dataset from Zenodo: `https://zenodo.org/record/7036625#.ZBeDiOzMJQI`.

Annotations The 'annotations' folder stores the annotations for SNPs on chromosome 21 and 22. In total, there are 101 columns and the last 96 columns are the annotations. The above algorithm works only for binary annotations. Some annotations are continuous (e.g., GERP.NS, MAF_Adj_LLD_AFR, Recomb_Rate_10kb, Nucleotide_Diversity_10kb, etc). You may ignore those or choose a reasonable threshold to binarize them.

Summary statistics data The training and testing summary statistics data computed for human standing height on chromosome 21 and 22 from 337,205 White British individuals from the UK Biobank are stored in here `https://drive.google.com/drive/folders/1qbaGULJ3IFSW3qpOWh354EyCRoSPG05b?usp=sharing`. The data are divided into 5 folds to facilitate evaluation described below.

4.4 Evaluation metric

You will need to use R-squared to evaluate how your trained model performs on the corresponding testing fold. Assuming standardized phenotype, the R-squared on testing fold can be computed by summary

statistics:

$$
\begin{aligned}
R^2 &= 1 - \frac{SS_{res}}{SS_{tot}} \\
SS_{res} &= \frac{1}{N_{test}}(\mathbf{y}_{test} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y}_{test} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= \frac{1}{N_{test}}\mathbf{y}'_{test}\mathbf{y}_{test} - 2\frac{\mathbf{y}'_{test}\mathbf{X}_{test}}{N_{test}}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\frac{1}{N_{test}}\mathbf{X}'_{test}\mathbf{X}_{test}\hat{\boldsymbol{\beta}} \\
&= 1 - 2\tilde{\boldsymbol{\beta}}'_{test}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{R}\hat{\boldsymbol{\beta}} \\
SS_{tot} &= \frac{1}{N_{test}}(\mathbf{y}_{test} - \bar{\mathbf{y}}_{test})'(\mathbf{y}_{test} - \bar{\mathbf{y}}_{test}) = 1
\end{aligned}
$$

As an essential baseline, evaluate on the same data using the VIPRS without annotation (i.e., the model in your A3) or the VIPRS in here `https://github.com/shz9/viprs`. You can perform 5-fold cross-validation by training and evaluating your model on the training and testing fold with the same index and repeat your experiments 5 times to compute standard error of the R-squared estimates for each method.

5 Project 2: Graph-embedded topic model for diagonal integration of single-cell data

This project is an extension of Assignment 4 and covers the single-cell genomics axis of the course.

Project mentor Liam Hodgson (`liam.hodgson@mail.mcgill.ca`) (research in single-cell genomics; also your TA for the course) and Yuesong Zou (`yuesong.zou@mail.mcgill.ca`) (CS master candidate in Li lab, research in graph representational learning, and first author on GAT-ETM [2]).

5.1 Diagonal integration of scRNA-seq and scATAC-seq data

Single-cell RNA sequencing (scRNA-seq) measures gene expression output at each cell providing a detailed view of the cell-type-specific transcriptome. scATAC-seq measures the open chromatin regions (i.e., "peaks") at the DNA level, providing detailed view of the cis-regulatory elements (CRE) that dictate the gene expression regulation. Linking CREs with gene expression in cell-type-specific context is an ongoing and important research challenge as it can help explain GWAS SNPs, which are mostly harboured in the regulatory regions of the genome [3].

In Lecture 15-18, we discussed three types of data integration, namely horizontal, vertical, and diagonal integration. The last is the most challenging and yet promising direction. As the atlas-level single-cell data such Tabula Sapiens [4] (scRNA-seq) and Human Enhancer Atlas [5] (scATAC-seq) have become available, there is a need to learn from the vast amount of data compendia, each profiling over one million of cells spanning for all primary tissues in human or mouse. In this project, we will develop a model that utilizes static biological networks information to perform diagonal integration of scRNA-seq and scATAC-seq data that are measured in different cells.

5.2 Objective

Modify scETM [6] by incorporating a graph neural network (GNN), which produces the embedding of genes and peaks simultaneously. As illustrated in Figure 1, one type of GNN is called Graph Convolutional Network (GCN) [7], which convolves the adjacency matrix by the convolutional filters in a way analogous to the convolutional neural network. Also, refer to Graph VAE [8]. Another popular GNN is called Graph Attention Network (GAT) [9]. We have previously developed a related framework called GAT-ETM [2] (`https://github.com/li-lab-mcgill/GAT-ETM`) to model EHR data.

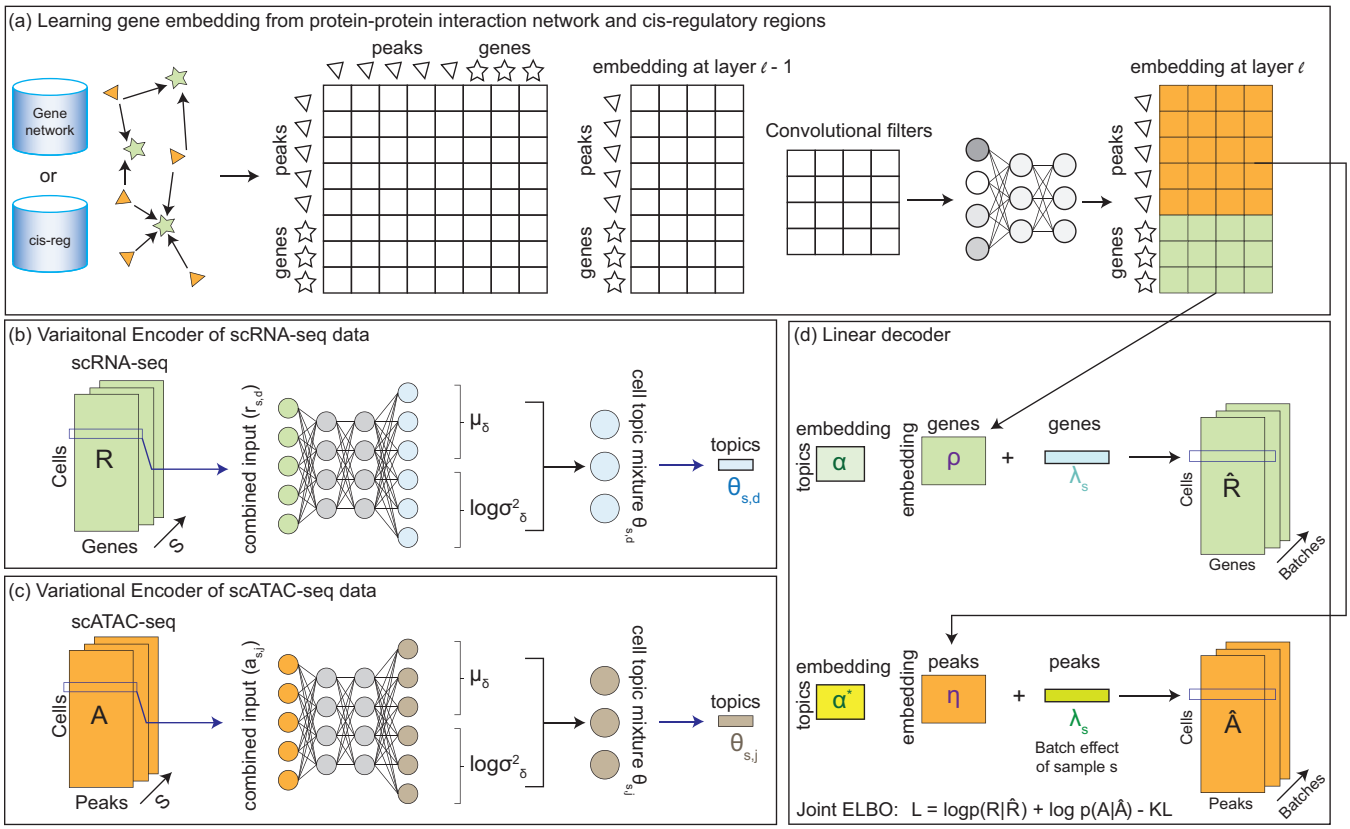

Figure 1: Schematic for the Graph-embedded embedded topic model for diagonal integration of scRNA-seq and scATAC data.

5.3 Data

For the purpose of this project, to diagonally integrate the two omics, you can focus on one tissue (e.g., pancreas) from each of the omics rather than modelling all of them altogether.

scRNA-seq data The Tabula Sapiens [4] are available at `https://figshare.com/articles/dataset/Tabula_Sapiens_release_1_0/14267219`.

scATAC-seq data The Human Enhancer Atlas [5] data are available at `http://catlas.org/humanenhancer/#!/`. You can either use `Cell_by_cCRE` or `Cell_by_gene`. See `Cell_ontology.tsv` for a summary of the cell types and `Cell_metadata.tsv.gz` for specific cell-type labels for each cell.

Biological network data The minimum network to be used is the peaks-by-genes network. You may derive such network based on the proximity of the peaks and the genes. A common way is to set the peak-gene pair to 1 if the peak is within the promoter region of the gene, which is -200 to +200 of the transcription start site (TSS) of the gene. To pair peaks with genes, you can obtain the genomic coordinates of the cCREs `cCRE_hg38.tsv.gz` from (`http://yed.ucsd.edu:8787/`) with the gene locations in gene annotation from Gencode `https://www.gencodegenes.org/human/`. There are other biological networks you can use. For example, you can consider incorporating gene-gene interactions in terms of protein-protein interaction from the String database (`https://string-db.org/`); you can consider peaks-peaks network via chromatin interaction measured by Hi-C `https://www.encodeproject.org/search/?type=Experiment&assay_title=Hi-C&status=released&assay_title=intact+Hi-C`. These are completely optional.

5.4 Evaluation Metrics

You will need to evaluate your model performance by Adjusted Rand Index (ARI) on cell clustering by algorithm such as K-means, Louvain or Leiden on the embedding. You will need to compare ablated models such as scETM trained on single-omic data to demonstrate the benefits of the diagonal integration. You will also need to show qualitative results including UMAP or t-SNE and top-gene per topic heatmaps similar to A4 and those presented in the course Lectures.

6 Project 3: Hierarchical guided-topic model for electronic health record data

This project is an extension of Assignment 5 and covers the EHR application axis of the course.

Project mentor Ziyang Song (`ziyang.song@mail.mcgill.ca`) (PhD candidate in Li lab; research in topic models in EHR, first authors for two MixEHR extensions including MixEHR-supervised [10] and MixEHR-seed [11]).

6.1 A guided hierarchical topic model

In MixEHR-guided [12], when modelling cross-sectional EHR data such as MIMIC-III, we anchor each topic to a specific pre-defined phenotype concept known as PheCode by setting $\alpha_{d,k}$ to 1 if individual d has one of the ICD-9 codes that define PheCode k otherwise 0, where $\boldsymbol{\alpha}_d$ is a $1 \times K$ vector for patient d. The corresponding data generative process for patient d and each ICD code j is:

$$\theta_d \sim Dir(\boldsymbol{\alpha}_d) \tag{8}$$

$$z_{d,j} \sim Cat(\theta_d) \tag{9}$$

$$x_{d,j} \sim Cat(\boldsymbol{\Phi}_{.,z_{d,j}}) \tag{10}$$

where θ_d is the phenotype topic mixture membership that follows a K-dimensional Dirichlet distribution with the hyperparameter fixed to $\boldsymbol{\alpha}_d$, $z_{d,j}$ is the topic index for ICD token j, and $x_{d,j}$ is the ICD index for ICD token j; $\boldsymbol{\Phi}_{.,z_{d,j}} \sim Dir(\boldsymbol{\eta})$ is a V-dimensional Dirichlet variable for topic indexed by $z_{d,j}$, where all V η_v's are fixed to a constant value (say 0.1). Note that for each PheCode, we dedicated exactly one topic distribution.

To extend it, we can have a $K \times M$ matrix for $\boldsymbol{\alpha}_d$ for K PheCode-guided topics and M sub-topics per topic. If the patient has the PheCode k, we will set the k^{th} row of $\boldsymbol{\alpha}_d$ to 1 and otherwise 0. For example, suppose we have 5 PheCode-guided topics and 3 sub-topics per topic. If the patient d has ICD-9 code that belongs to the definition of PheCode 2 but nothing else, then the topic prior hyperparameter matrix $\boldsymbol{\alpha}_d$ is set to:

$$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

The data generative process is identical to the above MixEHR-guided except having $K \times M$ topics instead of K topics. In particular, to sample θ_d as in (8), we flatten the $K \times M$ matrix for $\boldsymbol{\alpha}_d$ to have a row vector of $1 \times (K \times M)$ so that the θ_d Dirichlet variable will only have non-zero values for the M sub-topics where the corresponding PheCode is observed for patient d.

The resulting model, namely *MixEHR-nest*, allows us to learn M sub-phenotype topic distributions under each of the K main PheCode-guided topics. For instance, we may learn M sub-phenotype topics for different disease stages of the same chronic disease (e.g., COPD). Note that MixEHR-guide is a special case of MixEHR-nest when $M = 1$.

6.2 Objective

For the default option, the inference is the same as in MixEHR-guided or just the regular LDA. Once you have your code working, focus on the application on the MIMIC-III data. The main contribution in this project is the experimentation on the MIMIC-III data and interpretation of the sub-phenotypes for a select number of diseases.

6.3 Data

MIMIC-III data The full diagnostic ICD code data of all subjects along with the meta data describing the ICD codes are available at:
`https://drive.google.com/drive/folders/1rsIuwX_SoMjlS8wPIIlbEc8UNNCqkVMJ?usp=sharing`.

PheCode PheCode concepts are available at `https://phewascatalog.org/phecodes`.

6.4 Evaluation metric

You will be evaluated by the correctness of your derivation so you will need to clearly show your steps to arrive at your update equations. For quantitative evaluation, you can compute perplexity score, which is the negative log likelihood on the held-out subjects while fixing the inferred topic distributions: $-\log p(\mathbf{x}|\hat{\mathbf{\Phi}})$, which is similar to the reconstruction loss. Compare your MixEHR-nest with the baseline MixEHR-guided or standard LDA.

6.5 (5% class Bonus): a more general 2-level hierarchical topic model

Besides the above "simple hack" on the prior, we can design a hierarchical topic model with the following data generative process:

$$\begin{aligned}
\boldsymbol{\theta}_d &\sim Dir(\boldsymbol{\alpha}_d) \\
z_{d,j}^{\{1\}} &\sim Cat(\boldsymbol{\theta}_d) \\
\boldsymbol{\gamma}_{d,j} &\sim Dir(\boldsymbol{\beta}) \\
z_{d,j}^{\{2\}} &\sim Cat(\boldsymbol{\gamma}_{d,j}) \\
x_{d,j} &\sim Cat(\mathbf{\Phi}_{.,z_{d,j}^{\{1\}},z_{d,j}^{\{2\}}})
\end{aligned}$$

Here for each of the K PheCode-guided topics we dedicate $M \geq 1$ topics. All K β_k's are fixed to a constant say 0.1 and do not need to be estimated. $\mathbf{\Phi}$ is a $V \times K \times M$ tensor, where $\mathbf{\Phi}_{.,z_{d,j}^{\{1\}},z_{d,j}^{\{2\}}} \sim Dir(\boldsymbol{\delta})$ follows a V-dimensional Dirichlet with all V δ_v's set to a fixed values (say 0.1). You will need to derive collapsed Gibbs sampling or mean-field variational inference update of the topic assignments for $z_{d,j}^{\{1\}}$ and $z_{d,j}^{\{2\}}$. The recommended steps to take is to first integrate out θ, $\boldsymbol{\gamma}$, and $\mathbf{\Phi}$ and then derive closed-form conditional probabilities for the two types of topic assignments. The derivation is similar to what we cover in class. The main contribution in this project is the derivation of the inference. If you manage to complete it, you may implement and experiment your model on the MIMIC-III data based on the above evaluation metric. To receive the 5% class bonus, you will need to come up with a correct derivation for the inference of $z_{d,j}^{\{1\}}$ and $z_{d,j}^{\{2\}}$.

References

[1] Shadi Zabad, Simon Gravel, and Yue Li. Fast and accurate bayesian polygenic risk modeling with variational inference. *bioRxiv*, 2022.

[2] Yuesong Zou, Ahmad Pesaranghader, Ziyang Song, Aman Verma, David L Buckeridge, and Yue Li. Modeling electronic health record data using an end-to-end knowledge-graph-informed topic model. *Scientific Reports*, 12(1):17868, 2022.

[3] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2019.

[4] Tabula Sapiens Consortium*, Robert C Jones, Jim Karkanias, Mark A Krasnow, Angela Oliveira Pisco, Stephen R Quake, Julia Salzman, Nir Yosef, Bryan Bulthaup, Phillip Brown, et al. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022.

[5] Kai Zhang, James D Hocker, Michael Miller, Xiaomeng Hou, Joshua Chiou, Olivier B Poirion, Yunjiang Qiu, Yang E Li, Kyle J Gaulton, Allen Wang, et al. A single-cell atlas of chromatin accessibility in the human genome. *Cell*, 184(24):5985–6001, 2021.

[6] Yifan Zhao, Huiyu Cai, Zuobai Zhang, Jian Tang, and Yue Li. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nature communications*, 12(1):5261, 2021.

[7] Thomas N Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv.org*, September 2016.

[8] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

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

[10] Ziyang Song, Xavier Sumba Toral, Yixin Xu, Aihua Liu, Liming Guo, Guido Powell, Aman Verma, David Buckeridge, Ariane Marelli, and Yue Li. Supervised multi-specialist topic model with applications on large-scale electronic health record data. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–26, 2021.

[11] Ziyang Song, Yuanyi Hu, Aman Verma, David L Buckeridge, and Yue Li. Automatic phenotyping by a seed-guided topic model. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4713–4723, 2022.

[12] Yuri Ahuja, Yuesong Zou, Aman Verma, David Buckeridge, and Yue Li. Mixehr-guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using the electronic health record. *Journal of biomedical informatics*, 134:104190, 2022.