Representation Learning for High-Dimensional Single-Cell Genomics with Variational Autoencoders

Using Associations Between Latent Factors and SNPs to Discover new eQTLs

David van der Ham

Student:	David van der Ham
Supervisors:	Kirti Biharie, Inez den Hond
Responsible Professor:	Marcel Reinders
Committee:	Marcel Reinders, Kirti Biharie, Inez den Hond, Christoph Lofi
Institution:	Delft University of Technology
Faculty:	EEMCS
Course:	CSE3000 Research Project
Date:	June 27, 2026

A Thesis Submitted to the EEMCS Faculty of Delft University of Technology
in Partial Fulfilment of the Requirements for the
Bachelor of Computer Science and Engineering

An electronic version of this thesis is available at

Abstract

Single-cell expression quantitative trait loci (eQTL) studies link genetic variants to changes in gene expression in that cell. This allows us to study the effect of genetics on diseases per cell instead of aggregated, since effects can differ per cell type. Traditional SNP to gene expression linking on the single-cell level suffers from the multiple testing burden, due to the great amount of SNPs and genes. To address this, a deep learning framework was developed recently to compress gene expression into low-dimensional encodings and reconstruct the gene expression linearly from these encodings, enabling direct interpretation of the latent space. This model is called Latent Interaction Variational Inference (LIVI). Here, we determine whether the latent factors of this model can serve as a quantitative trait for Single Nucleotide Polymorphisms (SNPs) that associate with Rheumatoid Arthritis (RA) on a dataset with RA patients. RA is a chronic disease characterized by progressive damage of the joints. In this study, we found 617 out of 700 latent factors correlating to at least one SNP, using a linear mixed model. We also found that genes that are associated with RA in a Genome Wide Association Study have a higher loading for associated SNP-Latent factor pairs then for none associated one. We also identified genes affected by GWAS-identified risk SNPs for which the original GWAS did not identify a functionally associated gene. We conclude that the latent factors of the LIVI model can be used as a quantitative trait for SNPs, and used these latent factors to discover trans-eQTLs.

1 Introduction

Gene expression is a measure of how often each gene in the DNA is transcribed into RNA. Gene expression can partially be altered by Single Nuceotide Polymorphisms (SNPs) [1], which are one-nucleotide variations in the DNA that occur in at least 1% of the population. When a SNP affects gene expression, it is referred to as an expression quantitative trait locus (eQTL). Depending on their genomic location relative to the affected gene, eQTLs are classified as either cis-eQTLs, which act on nearby genes within a defined window, or trans-eQTLs, which influence genes further away. The relationship between SNPs and gene expression has been researched before, as understanding how genetic variation alters gene expression can be used to understand the biological mechanism behind a disease. However, due to the large number of genes and SNPs that exist, the multiple testing burden grows rapidly, especially for trans-eQTL studies where the effects are often smaller and all the genes have to be tested against for each SNP.

To address this, a deep learning framework was developed recently, called Latent Interaction Variational Inference (LIVI). LIVI is build on the Variatonal Auto Encoder (VAE) framework, a generative model that differs from a normal auto encoder in that encodes each input not to a single point in latent space but as a probability distribution centered around a point. This enforces the latent space to be structured and continuous. [2]. Contrary to normal VAEs, LIVI uses a linear decoder, where each gene's reconstructed expression is a weighted sum of latent factors. This allows latent factors to be directly interpreted to gene reconstructions it affects. LIVI works with multiple latent spaces, also called embeddings. (Fig. 1) Each of these embeddings captures a different part of gene expression variation.

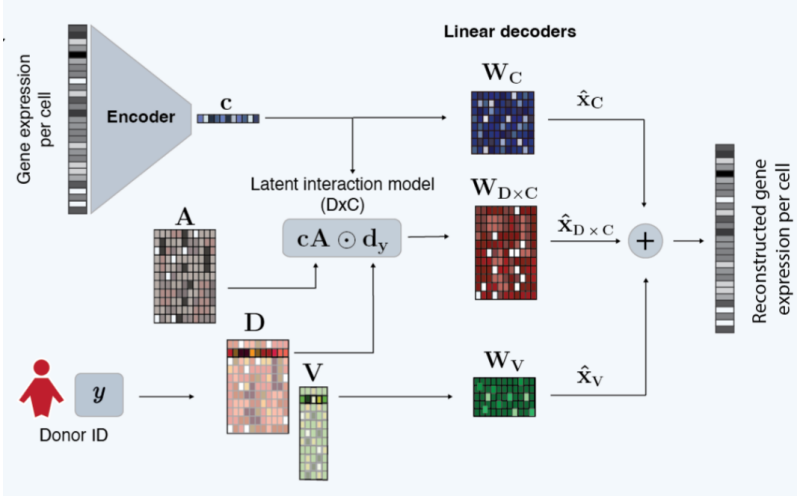

Figure 1: The architecture of LIVI. For each cell, it decomposes the gene expression into cell-specific effects (C) and donor-specific effects (D, V), where D captures cell-state-specific effects from the donor, and V captures global sources of inter-individual variation. C and V have their own linear decoder (with weights W_C, W_V), while D and C are combined according to a factor assignment matrix A. This combined matrix is then decoded with a linear decoder (using weights W_{DxC}). The results of all three decoders are added together to reconstruct the gene expression. *Figure from Vagiaki et al.* [3]

It is an open question whether the latent factors of the LIVI model can serve as quantitative traits for SNPs that increase an individual's risk of disease when trained on a cohort of patients of that disease. Answering this would not only be useful to test the biological relevance of LIVI but also relieve the multiple testing burden.
To investigate this, we applied and trained LIVI on a dataset containing Rheumatoid Arthritis patients. [4] Rheumatoid arthritis (RA) is a chronic autoimmune disease in which the immune system attacks the joints, leading to inflammation, pain, and potential joint damage over time. The SNPs we used are previously identified to affect risk of RA via a Genome Wide Association Study (GWAS), a large scale study amongst different ancestries [5]. (Fig. 7. The GWAS has two types of SNP-gene mappings: functional mappings, where there was a gene found that the SNP acts on, and positional mappings, where such a functional mapping was not found and the nearest gene was assigned instead.

2 Results

Application of LIVI on a single-cell dataset containing 82 Rheumatoid Arthritis patients
There are three sources of data used for our analysis. First, the SNPs we used are from a large scale multi ancestry GWAS [5]. Out of the 153 RA-associated SNPs that were found in this GWAS, 93 are also present in our genotype data. Each SNP also has a gene mapping: either a functional mapping, where there was a gene found that the SNP acts on, or a positional mapping, where such a functional mapping was not found and the nearest gene was assigned instead.

Secondly, LIVI was trained on a dataset containing single-cell gene expression data from 82 Rheumatoid Arthritis patients. It contains 314,000 cells in total and metadata such as sex, age, treatment received, disease severity and treatment site for each patient.
Third, the donor-level embedding (D) produced by this training was first tested for association to SNPs, and then used as a quantitative trait to discover trans-eQTLs for GWAS-identified risk SNPs with a positional mapping.

Direct correlation between SNPs and latent factors returns only one significant eQTL

We first wanted to know whether a direct linear association exists between the latent factors of the D-embedding and the GWAS-identified risk SNPs. To determine this, Pearson's test was calculated for each of the 65,100 latent factor SNP pairs. Only one association reached significance: Factor 291 with SNP rs56787183 (Fig. 2). Only 1 out of 65,100 results being significant indicates that SNPs are not directly correlating with the latent factors of the LIVI model, suggesting that a linear correlation is not enough to capture the association between SNPs and the latent factors at a statistically significant level.

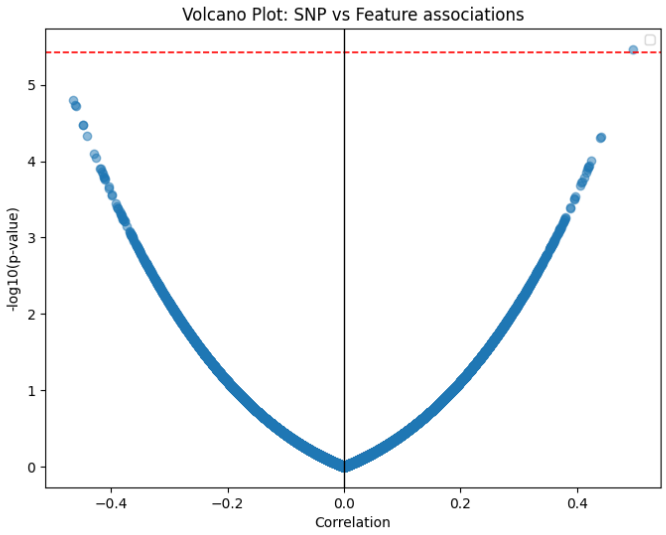

Figure 2: Volcano plot of Pearson correlations between LIVI latent factors and SNPs across all 65,100 pairs. The x-axis shows the sign and magnitude of the correlation, the y axis shows the uncorrected p-value. The dashed horizontal line shows the cutoff value at FDR = 0.25

A Linear Mixed Model identifies significant associations between latent factors and SNPs.

Given that a direct correlation ignores population structure, shared ancestry and other covariates, we suspected it might be too simple to model the complex biology. A Linear Mixed Model (LMM) was used to account for these factors. Using this technique, it was found that a subset of latent factors showed significant association with one or more SNPs, while others showed none. (Fig 3) This difference is expected: some latent factors might encode other

relevant information.

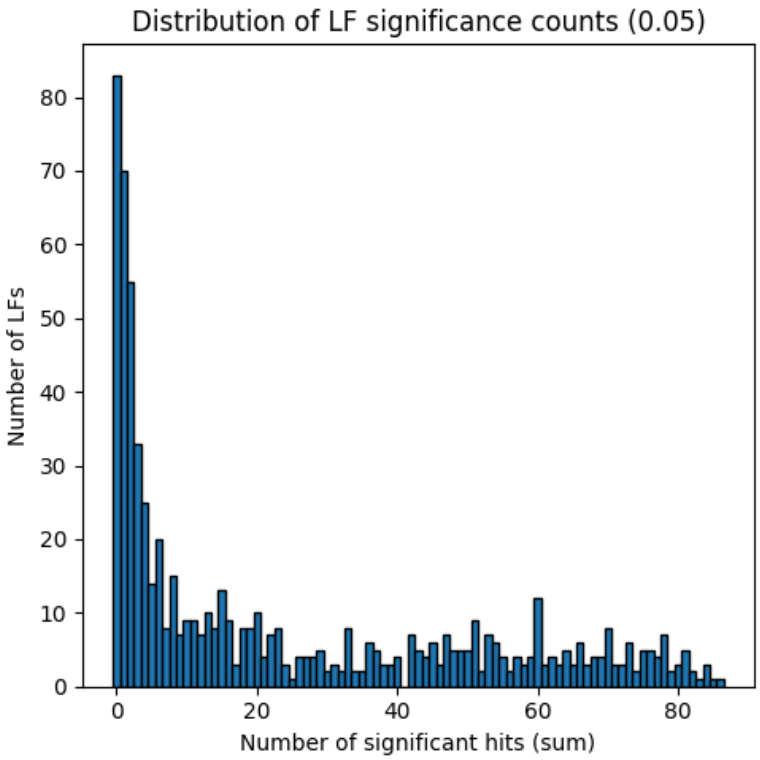

Figure 3: Distribution of the number of SNPs significantly associated with each latent factor when testing with an LMM. We see that most latent factors have few associations, some having zero.

We also wanted to know whether these significant results reflect biological signals. To assess this, the results of the LMM association were cross-referenced against a GWAS [5]. It provides for each SNP a gene that is affected by it. If the latent factors capture biological effect of the SNPs on the gene expression, the GWAS-identified risk genes should be more affected by the latent factors significantly associated to this SNP. We indeed found that the GWAS discovered gene-SNP effects had higher decoder values for significant results compared to non-significant results. (Fig. 4)

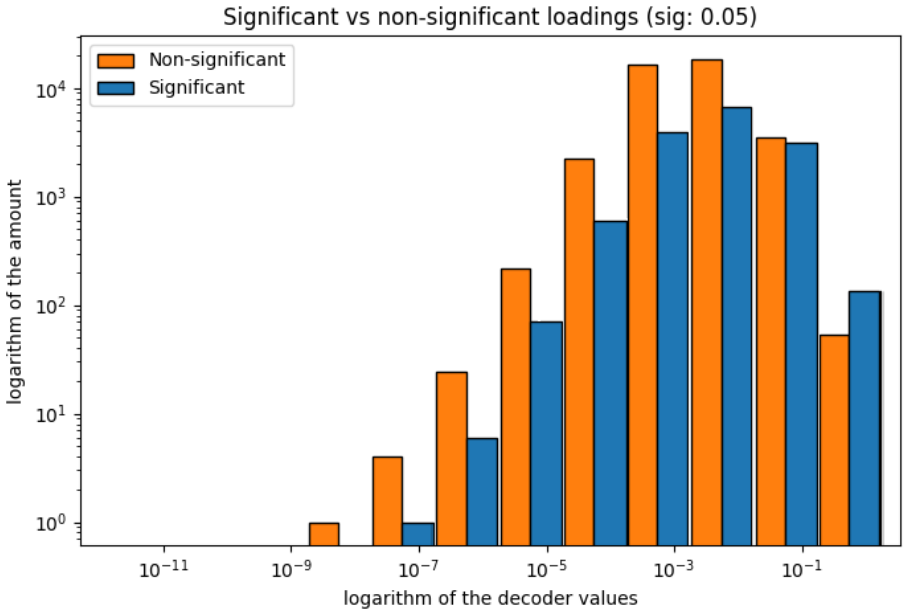

Figure 4: Comparing the distribution of decoder values the significant associations have on the GWAS-identified risk genes compared to the non-significant results. ($\alpha = 0.05$)

Furthermore, we also examined the rank of functionally mapped GWAS genes within the DxC decoder. These functionally mapped genes gave a significant difference in rank distribution between significant and non-significant associations. (Fig. 5, Table 1) indicating that the GWAS-identified target genes are stronger encoded in the latent factors that associate significantly with the corresponding GWAS-identified risk SNP. Positional mappings, where the GWAS did not detect a functional association, did not show such a difference. (Fig. 5, Table 1) This serves as a negative control. This indicates that significant associations actually capture the SNP-gene associations mapped in the GWAS.

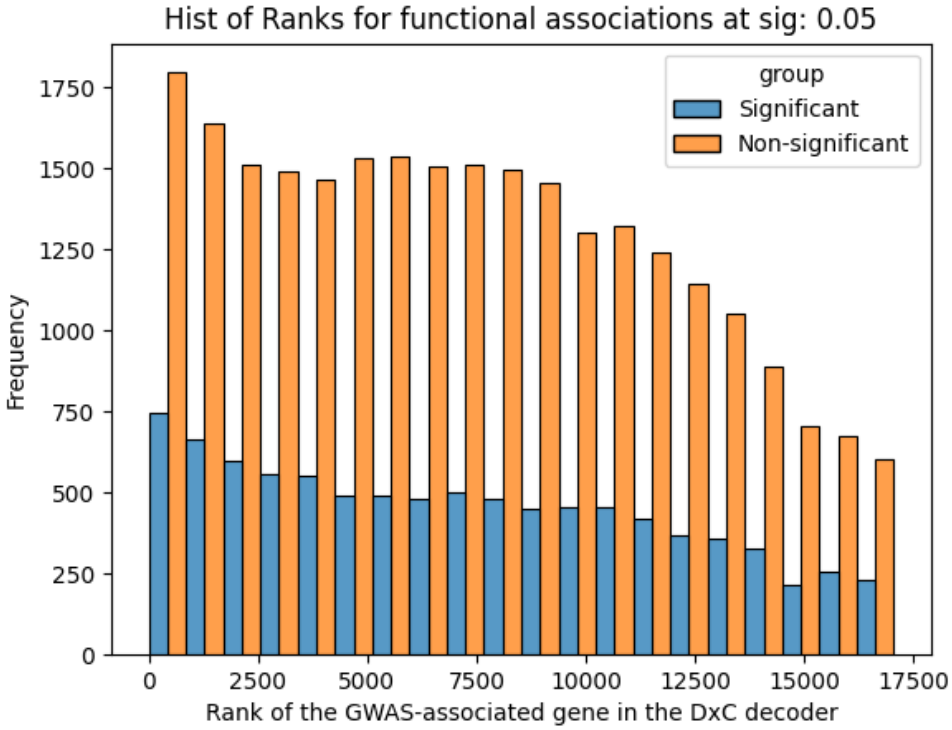

Figure 5: Comparison between the rank in the DxC embedding of functionally mapped SNPs for significant and non-significant results. A lower rank indicates a stronger loading.

GWAS type	**U statistic**	**p-value**	**Median rank difference)**
Positional	42818037.5	0.7269751	84
Functional	114326532.0	7.018e-06	-329.5

Table 1: Results of a one-sided Wilcoxon rank-sum test comparing DxC decoder rank distributions of GWAS-identified risk genes between significant and non-significant LMM associations, for functional SNP-to-gene mappings. A negative rank difference indicates higher decoder ranks in the significant group.

To get a better overview of the locations of the associations found with the LMM testing, we wanted to see the genomic position of the results. The SNPs that were significantly associated with a latent factor are spread over all chromosomes, consistent with the polygenic nature of RA [6]. This indicates that the LIVI model captures effects on the entire range of the DNA. (Fig. 6)

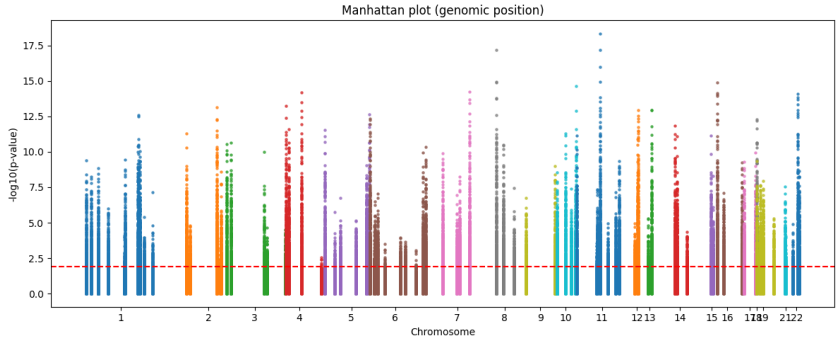

Figure 6: Manhattan Plot of the results of the LMM test between the 93 GWAS-identified risk SNPs and each of the 700 latent factors of the D-embedding. Each color represents a chromosome.

Genes affected by GWAS-identified risk SNPs for which no functionally associated target gene had been identified in the original GWAS were also identified. Out of the 93 GWAS-identified risk SNPs, 33 did not have a functional mapping assigned. For each of these SNPs, the gene most strongly encoded by the significantly associated latent factors was identified as a candidate target. All the effects we found (Table 2) using this method are trans-eQTLs. Certain genes (MRPL40, TRBC1) also show up multiple times. This supports the claim of the original LIVI paper ([3]) that LIVI is able to aggregate weak trans effects in a single latent factor.

SNP	Score Method			
	$w_{qval} = 0.01$	$w_{qval} = 0.05$	$w_{qval} = 0.1$	multiplication
1:198614892	ASPN (T)	IGKC (T)	IGKC (T)	IGKC (T)
14:75515513	TRBC1 (T)	TRBC1 (T)	SIRPA (T)	SIRPA (T)
7:100295525	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)
11:64340005	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)
9:34710263	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)
2:203745673	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)
6:14106966	SIRPA (T)	SIRPA (T)	SIRPA (T)	SIRPA (T)
14:104920174	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)	CEP170 (T)
4:10707742	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)
11:118810370	TRBC1 (T)	MRPL40 (T)	MRPL40 (T)	CEP170 (T)
6:426268	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)
11:95578258	TRBC1 (T)	TRBC1 (T)	TRBC1 (T)	MALAT1 (T)
1:157716547	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)
1:116738074	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)
8:101441122	MRPL40 (T)	MRPL40 (T)	CD37 (T)	MRPL40 (T)
6:23924793	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)	DMXL2 (T)
7:128936032	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)
6:15195451	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)
16:85982795	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)

Table 2 *(continued)*

SNP	Score Method			
	0.01	0.05	0.1	mult
1:38156596	CEP170 (T)	TPT1 (T)	IGLC2 (T)	IGLC2 (T)
5:134503843	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)
11:128318147	TRBC1 (T)	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)
10:62284216	MRPL40 (T)	MRPL40 (T)	CEP170 (T)	MRPL40 (T)
10:31097045	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)
16:30787368	TRBC1 (T)	TRBC1 (T)	IGKC (T)	IGKC (T)
6:44266884	MT-CO1 (T)	MT-CO1 (T)	MT-CO1 (T)	MALAT1 (T)
1:173337507	TRBC1 (T)	TRBC1 (T)	TRBC1 (T)	RPL10 (T)
5:40499188	HBEGF (T)	ANKRD28 (T)	ANKRD28 (T)	ANKRD28 (T)
14:68784174	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)
19:51514686	TRBC1 (T)	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)
6:137921300	MT-ATP6 (T)	MT-ATP6 (T)	MT-ATP6 (T)	B2M (T)
5:134085107	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)	MRPL40 (T)
3:189588861	TRBC1 (T)	TRBC1 (T)	TRBC1 (T)	TRBC1 (T)

Table 2: Top gene per score method for all GWAS-identified risk SNPs with a positional mapping. For each SNP, the associated latent factors were decoded and the highest-effect gene was determined using four score methods. The first three use: $S = w_q \cdot q_{\text{SNP}\to\text{LF}} + (1 - w_q) \cdot d_{\text{LF,Gene}}$, where $w_q \in \{0.01, 0.05, 0.1\}$. The fourth uses: $S = q_{\text{SNP}\to\text{LF}} \cdot DxC_{\text{LF,Gene}}$, where $q_{\text{SNP}\to\text{LF}}$ is the q-value of the SNP-to-LF association and $DxC_{\text{LF,Gene}}$ is the DxC decoder loading of the latent factor onto the gene.

3 Discussion and Conclusion

Understanding the genetic basis of complex diseases such as Rheumatoid Arthritis requires linking single nucleotide genetic variants to their effects on gene expression. Normal eQTL mapping faces a scalability problem, due to the large number of genes and SNPs. This study investigated whether the latent factors of LIVI can serve as quantitative traits for GWAS-identified Rheumatoid Arthritis risk SNPs.

A direct linear correlation is insufficient. Pearson's correlation between all 65,100 SNP latent factor pairs resulted in only one significant result. This is due to it not taking covariates and other confounders into account.

A Linear Mixed Model substantially increases results. By separating random and fixed effects, the Linear Mixed Model found 16,650 significant associations across 617 out of 700 latent factors in the D-embedding. Ablation testing confirmed the need for covariates: removing them resulted in no significant results.

GWAS-identified SNP-gene pairs show higher decoding effect for significant associations. The loadings of the pairs from the GWAS show higher loading in the decoding matrix DxC. Functionally mapped GWAS genes ranked significantly higher in the decoders of significant associations than non-significant ones. Positionally mapped genes showed no effect.

For the 33 risk SNPs not having a functional mapping in the GWAS, potential genes were identified via their associated latent factors. All of these are trans-eQTLs, suggesting that LIVI captures these effects that the original GWAS did not find.

The main limitation of this work is the small number of donors. After removing repeat samplings, only 72 donors remained. Another limitation is the difference in sample size between individuals, some samples having significantly more cells sampled than others.
We conclude that LIVI's latent factors can be used as quantitative traits for RA-associated SNPs. By reducing the testing space, the multiple testing problem can be reduced, and weak trans-eQTLs can be aggregated. More broadly, structured latent representations from deep generative models like LIVI are a promising approach for genetic association analysis across complex traits and diseases.

4 Responsible Research

This project was conducted with attention to responsible research practices, including privacy protection, research integrity, reproducibility, and broader societal impact. The data used in this project was handled with care in accordance with ARK Portal Controlled Access Data Use Certificate. Files were kept local or on the DAIC cluster. To minimize bias and maintain research integrity, the project critically evaluated the limitations of the datasets and methods used. Potential sources of bias, such as unbalanced data or overfitting, were considered during analysis. The dataset used was not evenly distributed, but was representative of the RA population. Results were reported honestly, including limitations and uncertainties, rather than selectively presenting favorable outcomes. Reproducibility was supported by open-sourcing the code on GitHub. Data preprocessing steps, model parameters, and analysis methods were clearly described so that others could reproduce the experiments. We also considered broader societal impact of this project. While the research may contribute positively to scientific understanding and future applications, there is also a risk of misinterpretation or misuse of the results. We believe the pros are bigger than the cons.

5 Code availability

The code used for the analysis is open source and available at `https://github.com/Dadievid/BEP`

6 Methods

Data used Genome Wide Association Study

Ishigaki et al. performed the GWAS on a large population across an ethnically diverse group [5] (Fig. 7) 93 of the 153 discovered SNPs in this GWAS are present in the genotype data we used in our analysis. Each SNP is also mapped to a gene. The GWAS has two types of SNP-gene mappings: functional mappings, where there was a gene found that the SNP acts on, and positional mappings, where such a functional mapping was not found and the nearest gene was assigned instead. This data source was used to validate whether LIVI does indeed encode biology.

Single-Cell gene expression

We trained the model on gene expression data containing 314,000 single-cell data points

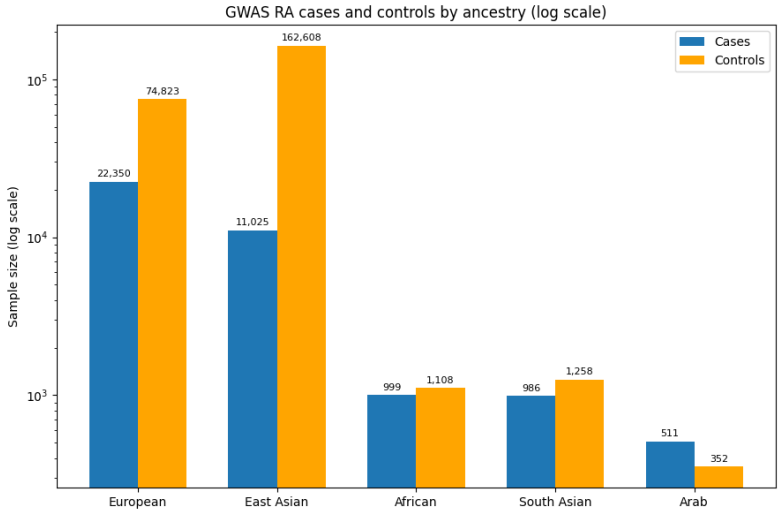

Figure 7: GWAS RA cases and controls by ancestry

from 82 different donors with Rheumatoid Arthritis (RA) [4]. It also contains metadata for each patient on the hospitalization, disease details and covariates. We excluded repeat samplings, as they might bias the model because the SNPs stay the same across samplings.

Embeddings from our training

Our inference of the LIVI model is largely the same as the original paper. We used 700 D factors and 5 V factors for each individual, and 15 C factors for each single-cell measurement. The batch size used for inference was set to 10,000. In contrast to the LIVI paper, we did not include cis-eQTLs during training. We decided on using the D-embedding for our analysis as the SNPs are the same in all cells of an individual.

Pearson's correlation between SNPs and latent factors

Pearson correlation was calculated between each pair of the 93 SNPs and 700 latent factors. This results in 65,100 tests. In this process, SNPs were encoded numerically as 0 (homozygous major allele), 1 (heterozygous), and 2 (homozygous minor allele). A B-H multiple testing correction [7] with an FDR of 0.25 was used.

Correlation between SNPs and latent factors while accounting for random effects

Linear mixed models (LMMs) are an approach for analysis of multiple traits. [8] It takes both fixed effects and random effects into account. Random effects capture correlation and shared ancestry between individuals by modeling these as being drawn from a normal distribution. The fixed effects are the associations from SNPs to latent factors that we try to detect, taking into account the random effects already present for each individual. Let $g \in \{0, 1, 2\}^{N_d}$ be a vector of SNP genotypes. For each latent factor k, we model

$$d_k = g\beta + M\alpha + u + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I), \tag{1}$$

where

$$u \sim \mathcal{N}(0, \sigma_u^2 K) \tag{2}$$

is a random effect parameterized by the kinship matrix K between individuals to account for genetic relatedness between individuals, $d_k \in \mathbf{R}^{N_d}$ is the vector of donor-level latent factor values for factor k across all $N_d = 72$ donors. The scalar $\beta \in \mathbf{R}$ is the fixed effect size to be estimated, representing the association between the SNP and the latent factor. $M \in \mathbf{R}^{N_d \times 10}$ is a matrix of confounding covariates, where the first 10 genotype principal components with $\alpha \in \mathbf{R}^{10}$ are the corresponding effect sizes. The residual noise is captured by ϵ The kinship matrix was calculated using PLINK with the KING option. [9] We used B-H multiple testing correction [7] at an FDR of 0.05 unless stated otherwise. LIMIX was used [8] to calculate the results for each of the 700 latent factors. In order to confirm the relevance of the covariates and the kinship matrix, we did ablation testing. After removing the covariates, no significant results showed up. Removing the kinship matrix resulted in roughly the same results as including both the covariates and the kinship matrix.

Using over-representation analysis to determine whether the latent factors associated to RA decode to RA pathway.

For each of the 93 SNPs, all the significantly associated latent factors at $\alpha = 0.01$ were identified. For each latent factor, the Kneedle algorithm [10] was applied to the corresponding column of the DxC decoder matrix to select the genes with the highest loadings. Over-representation analysis was then performed, using the KEGG_2021_Human as gene sets and all the genes in the DxC decoder as background genes.

Finding the top gene for each GWAS-identified risk SNP with a positional mapping.

To determine the eQTLs for the SNPs that did not have a functional mapping in the GWAS, all the significantly associated latent factors for this SNP were taken. For each SNP, the associated latent factors were decoded and the highest-effect gene was determined using four score methods. The first three use: $S = w_q \cdot q_{\text{SNP}\to\text{LF}} + (1 - w_q) \cdot d_{\text{LF,Gene}}$, where $w_q \in \{0.01, 0.05, 0.1\}$. The fourth uses: $S = q_{\text{SNP}\to\text{LF}} \cdot DxC_{\text{LF,Gene}}$, where $q_{\text{SNP}\to\text{LF}}$ is the q-value of the SNP-to-LF association and $DxC_{\text{LF,Gene}}$ is the DxC decoder loading of the latent factor onto the gene. To determine whether the eQTL is a cis- or trans-eQTL, a cutoff of 1,000,000 bases was used.

References

[1] Barkur S Shastry. SNPs: impact on gene function and phenotype. *Methods Mol. Biol.*, 578:3–22, 2009.

[2] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[3] Danai Vagiaki, Tobias Heinen, Manu Saraswat, Brian Clarke, and Oliver Stegle. Mapping *trans* -eQTLs at single-cell resolution using latent interaction variational inference. Preprint available on bioRxiv, February 2026.

[4] Fan Zhang, Anna Helena Jonsson, Aparna Nathan, Nghia Millard, Michelle Curtis, Qian Xiao, Maria Gutierrez-Arcelus, William Apruzzese, Gerald F M Watts, Dana Weisenfeld, Saba Nayar, Javier Rangel-Moreno, Nida Meednu, Kathryne E Marks, Ian Mantel, Joyce B Kang, Laurie Rumker, Joseph Mears, Kamil Slowikowski, Kathryn Weinand, Dana E Orange, Laura Geraldino-Pardilla, Kevin D Deane, Darren Tabechian, Arnoldas Ceponis, Gary S Firestein, Mark Maybury, Ilfita Sahbudin, Ami Ben-Artzi, Arthur M Mandelin, 2nd, Alessandra Nerviani, Myles J Lewis, Felice Rivellese, Costantino Pitzalis, Laura B Hughes, Diane Horowitz, Edward DiCarlo, Ellen M

Gravallese, Brendan F Boyce, Accelerating Medicines Partnership: RA/SLE Network, Larry W Moreland, Susan M Goodman, Harris Perlman, V Michael Holers, Katherine P Liao, Andrew Filer, Vivian P Bykerk, Kevin Wei, Deepak A Rao, Laura T Donlin, Jennifer H Anolik, Michael B Brenner, and Soumya Raychaudhuri. Deconstruction of rheumatoid arthritis synovium defines inflammatory subtypes. *Nature*, 623(7987):616–624, November 2023.

[5] Kazuyoshi Ishigaki, Saori Sakaue, Chikashi Terao, Yang Luo, Kyuto Sonehara, Kensuke Yamaguchi, Tiffany Amariuta, Chun Lai Too, Vincent A Laufer, Ian C Scott, Sebastien Viatte, Meiko Takahashi, Koichiro Ohmura, Akira Murasawa, Motomu Hashimoto, Hiromu Ito, Mohammed Hammoudeh, Samar Al Emadi, Basel K Masri, Hussein Halabi, Humeira Badsha, Imad W Uthman, Xin Wu, Li Lin, Ting Li, Darren Plant, Anne Barton, Gisela Orozco, Suzanne M M Verstappen, John Bowes, Alexander J MacGregor, Suguru Honda, Masaru Koido, Kohei Tomizuka, Yoichiro Kamatani, Hiroaki Tanaka, Eiichi Tanaka, Akari Suzuki, Yuichi Maeda, Kenichi Yamamoto, Satoru Miyawaki, Gang Xie, Jinyi Zhang, Christopher I Amos, Edward Keystone, Gertjan Wolbink, Irene van der Horst-Bruinsma, Jing Cui, Katherine P Liao, Robert J Carroll, Hye-Soon Lee, So-Young Bang, Katherine A Siminovitch, Niek de Vries, Lars Alfredsson, Solbritt Rantapää-Dahlqvist, Elizabeth W Karlson, Sang-Cheol Bae, Robert P Kimberly, Jeffrey C Edberg, Xavier Mariette, Tom Huizinga, Philippe Dieudé, Matthias Schneider, Martin Kerick, Joshua C Denny, BioBank Japan Project, Koichi Matsuda, Keitaro Matsuo, Tsuneyo Mimori, Fumihiko Matsuda, Keishi Fujio, Yoshiya Tanaka, Atsushi Kumanogoh, Matthew Traylor, Cathryn M Lewis, Stephen Eyre, Huji Xu, Richa Saxena, Thurayya Arayssi, Yuta Kochi, Katsunori Ikari, Masayoshi Harigai, Peter K Gregersen, Kazuhiko Yamamoto, S Louis Bridges, Jr, Leonid Padyukov, Javier Martin, Lars Klareskog, Yukinori Okada, and Soumya Raychaudhuri. Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis. *Nat. Genet.*, 54(11):1640–1651, November 2022.

[6] Cornelia M Weyand. Rheumatoid arthritis: A polygenic disease with multiple phenotypes. *Arthritis Res.*, 1(Suppl 1):S03, 2000.

[7] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, 57(1):289–300, January 1995.

[8] Christoph Lippert, Franceso Paolo Casale, Barbara Rakitsch, and Oliver Stegle. LIMIX: genetic analysis of multiple traits. Preprint available on bioRxiv, 2014.

[9] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, September 2007.

[10] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*. IEEE, June 2011.