GUIDE deconstructs genetic architectures using association studies

Daniel Lazarev[1,2,3,4,5], Grant Chau[2,3,4], Alex Bloemendal[2,3,4], Claire Churchhouse[2,3,4], and Benjamin M Neale[2,3,4,6]

[1]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[2]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
[3]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA
[4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
[5]Renaissance School of Medicine at Stony Brook University, Stony Brook, New York, USA
[6]Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

May 3, 2024

Abstract

Genome-wide association studies have revealed that the genetic architecture of most complex traits is characterized by a large number of distinct effects scattered across the genome. Functional enrichment analyses of these results suggest that the associations for any given complex trait are not purely random. Thus, we set out to leverage the genetic association results from many traits with a view to identifying the set of modules, or latent factors, that mediate these associations. The identification of such modules may aid in disease classification as well as the elucidation of complex disease mechanisms. We propose a method, Genetic Unmixing by Independent Decomposition (GUIDE), to estimate a set of statistically independent latent factors that best express the patterns of association across many traits. The resulting latent factors not only have desirable mathematical properties, such as sparsity and a higher variance explained (for both traits and variants), but are also able to single out and prioritize key biological features or pathophysiological mechanisms underlying a given trait or disease. Moreover, we show that these latent factors can index biological pathways as well as epidemiological and environmental influences that compose the genetic architecture of complex traits.

Mappings between the human genome and phenome—already complicated by the high dimensional structures involved—are further convoluted by polygenic and pleiotropic relationships. Deconvolving the genetic architecture underlying a given set of phenotypes may be achieved by learning a set of modules, or latent factors, which mediate between a set of genetic variants and a set of traits. These latent factors can be viewed as the components of the genetic architecture of the outcome traits and may relate to biochemical mechanisms (*e.g.*, LDL cholesterol and its impact on coronary artery disease) or genetic influences on exposures (*e.g.*, smoking as a risk factor for lung cancer). By using genetic association results across a wide range of traits it may be possible to identify different subsets of genetic variants that point to such biochemical or exposure related processes and in turn facilitate the interpretation of results from genome-wide association studies (GWAS). This is especially crucial for complex traits—those emerging from the interplay of several intertwined molecular and environmental mechanisms represented by distinct latent factors—for which the number of causal variants can be in the tens of thousands [1].

Existing methods that group subsets of genetic variants, such as genomic structural equation modeling (genomic SEM) [2], Mash [3], and pleiotropic decomposition regression (PDR) [4], are not necessarily

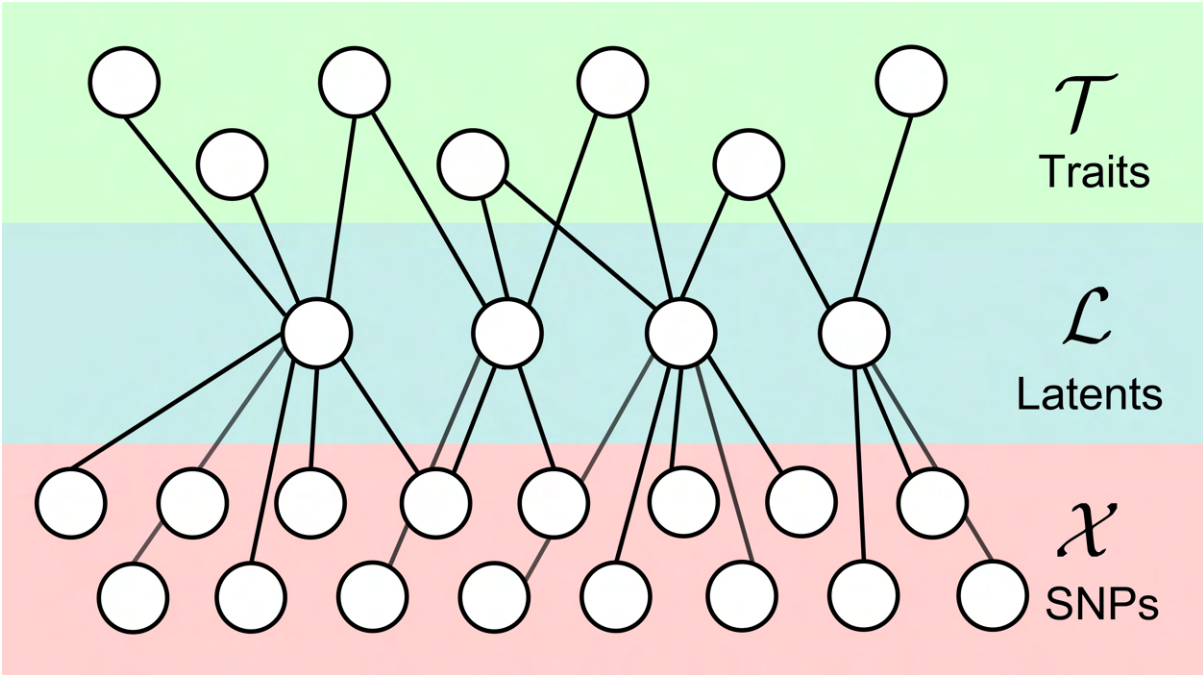


Figure 1: An example of a three-layer SNP-phenotype network. Here, $\mathcal{X}$ is the layer of SNPs, $\mathcal{L}$ the layer of latent factors, and $\mathcal{T}$ the layer of traits. We write, $M := |\mathcal{X}|$, $L := |\mathcal{L}|$, $T := |\mathcal{T}|$.

informed by the traits they affect, cannot be applied to a large number of traits, or require user-specified model parameters, architecture or latent factors. In contrast, a recently developed method, Decomposition of Genetic Associations (DeGAs) [5, 6], applies truncated singular value decomposition (TSVD) to GWAS summary statistics to derive a set of latent factors for this purpose. Applying singular value decomposition (SVD) to a matrix of summary statistics produces a diagonal matrix containing the singular values that is left- and right-multiplied by orthogonal matrices. Truncating these two orthogonal matrices based on a user-specified number of latent factors gives a set of orthonormal vectors that—viewing this arrangement as a three-layer network (Fig. 1)—may be considered as the normalized weights from a layer of single nucleotide polymorphisms (SNPs) to a new layer of latent factors and from the latter to a layer of traits. Applications of DeGAs have already included, for example, newer measures of polygenic risk scores that are informed by these factors [7].

A shortcoming of DeGAs, however, stems from the fact that orthogonal matrices produced by SVD are not unique: for example, one can multiply by any orthogonal matrix and its inverse (the transpose in this case) to get another possible decomposition. In other words, this is essentially the problem of choosing a basis for the latent factors that gives a set of maximally sparse, independent, and interpretable factors.

Independent Component Analysis (ICA) is a statistical method for decomposing a mixed signal into a set of statistically independent ones (*e.g.*, separating individual conversations from a recording of multiple conversations in the famous cocktail party problem) [8, 9, 10, 11]. It has been successfully applied in many applications in genomics as a denoising or dimensionality reduction alternative to principle components analysis (PCA) [12], most recently to analyze the GWAS summary statistics of brain image-derived phenotypes [13]. Our method, Genetic Unmixing by Independent Decomposition (GUIDE), uses TSVD and a two-sided application of ICA—simultaneously for both the SNPs-to-latents and latents-to-traits weights—to determine a set of sparse and independent latent factors given GWAS summary statistics. Unlike previous approaches, GUIDE also includes an additional method, also based on ICA, for estimating the number of latent factors.

The latent factors determined by GUIDE can illuminate relations between different diseases, help interpret results from the GWAS of any trait, and perhaps motivate a reclassification of diseases that more accurately reflects their interrelations via, for example, shared symptoms or common disease mechanisms. Such interaction networks have been proposed as a new paradigm for the classification of psychiatric disorders [14, 15].

We apply GUIDE to several curated subsets of summary statistics from the UK Biobank, including

those used in [5], allowing for a direct comparison with DeGAs. An interactive browser presenting the GUIDE model for this dataset can be found at guide-analysis.hail.is, allowing for the analysis of traits of choice from the UK Biobank using GUIDE latent factors.

Using GUIDE to analyze Alzheimer's disease, for example, we show how GUIDE, driven solely by data without any user-specified modeling choices or parameters, prioritizes *APOE* and other nearby genes as contributing to a distinct mechanism leading to Alzheimer's disease, and, moreover, how it derives a known connection between Alzheimer's disease and high cholesterol that is still an active area of research [16, 17]. Applying GUIDE to another complex trait, forced expiratory volume in one second (FEV1), showcases its ability to disentangle molecular from psychosocial or epidemiological effects on a given trait, identifying latent factors that influence FEV1 by affecting smoking behavior, cardiovascular health, immune or developmental processes, standing height, and thoracic cavity volume.

We provide a brief overview of GUIDE in Section 1.1 of Results (with full details in Methods), followed by a review of the salient properties of GUIDE latent factors in Section 1.2. Section 1.3 focuses on the GUIDE analysis of Alzheimer's disease, while Section 1.4 is a deep dive into FEV1 and related traits.

1 Results

1.1 Overview and motivation for the GUIDE method

As described in full detail in Methods, GUIDE simultaneously applies Independent Component Analysis (ICA) to both the SNP and trait weights matrices corresponding to a given set of summary statistics after obtaining an initial decomposition of the summary statistics matrix using singular value decomposition (SVD). This is motivated by the following insights.

Entropy is a measure of randomness or uniformity of a probability distribution, and its value is lowered by the presence of information or constraints, which make some events more probable than others. Therefore, on a finite support (*e.g.*, $[0, 10]$) the distribution with the highest entropy is the uniform distribution, while on an infinite support (*e.g.*, the real number line, $\mathbb{R}$), among distributions with a finite variance, the Gaussian distribution maximizes entropy [18]. Solving a sixty-year-old problem posed by Claude Shannon, it was shown [19] that normalized sums of independent random variables have a higher entropy with every random variable added, eventually converging to the maximum entropy value as the sum becomes normally distributed in the limit. In other words, these sums are not only Gaussian in the limit, per the central limit theorem, but are also 'more Gaussian', in the sense of having a higher entropy, at every single step.

The GUIDE model assumes the existence of independent latent modules mediating the effects of genetic variants on traits (see Fig. 1). Since the effects of these latent factors are summed over, the factors themselves are hidden in the data, and may seem normally distributed due to the above considerations. The solution, therefore, would be not only to decompose the summary statistics using SVD, but also to find a basis for the network weights that *minimizes* the entropy—one that is maximally non-Gaussian—in order to recover the original latent factors. Luckily, ICA does exactly that. ICA is built on the idea of maximizing non-Gaussianity via the minimization of entropy (or, equally, maximization of negentropy). Although entropic measures are the optimal cost functions from a mathematical perspective [19, 8, 9], in practice, ICA implementations use computationally favorable proxies of entropy, such as kurtosis or some notions of sparsity, which do not require the estimation of a probability distribution function at every step.

The implementation of ICA we use, FastICA, is based on a particular set of optimal cost functions derived using the maximum entropy principle that were found to outperform previous proxies based on sparsity or kurtosis [10, 11].

Nevertheless, these older proxies are still useful in assessing the performance of our solutions (SI, Sec. 4.2) and in gaining intuition for the kinds of solutions that GUIDE finds. More specifically, as we will see in the next section, since we apply ICA to both the $\mathcal{X} \to \mathcal{L}$ and $\mathcal{T} \to \mathcal{L}$ weights, the weight matrices to and from the GUIDE latent factors are maximally sparse and kurtotic, that is, with few, sharp peaks of signal that capture most of the interlayer relationships of a given latent factor with the relevant SNPs and traits.

An outline of the GUIDE method can be found in Algorithm 1 in Methods.

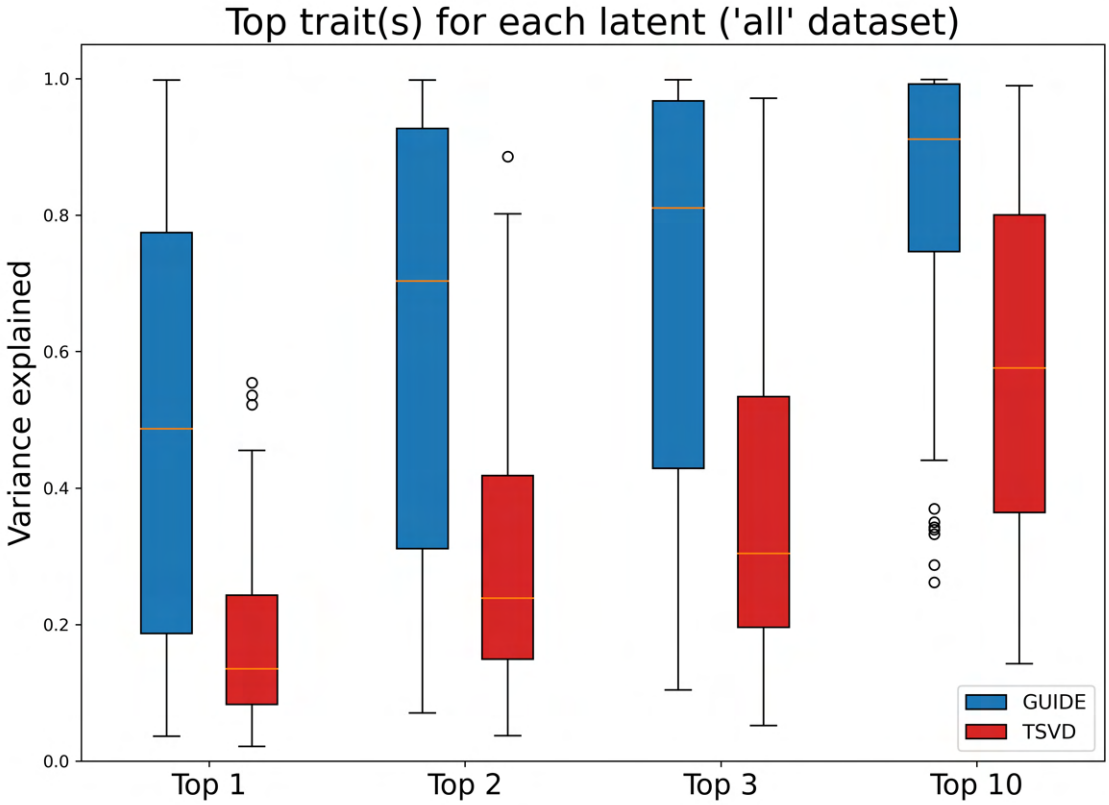


Figure 2: Contributions of the top 1, 2, 3, and 10 traits to the normalized variance explained by each latent factor for both GUIDE and TSVD model weights. See Fig. S22 for the variance explained curves for all 100 latent factors for GUIDE and for TSVD.

1.2 Properties of GUIDE factors

For direct comparison with DeGAs, we used the 'all' dataset consisting of 235,907 variants and 2,138 phenotypes from [5], which originated from the UK Biobank. We will only refer to DeGAs as such when applying it to the 'all' dataset on which it was tested by the authors in [5] and for the particular analyses they had used it for; otherwise, especially in subsequent sections using new curated subsets of the UK Biobank or novel analyses, we will write 'TSVD'.

As shown in Figs. 2 and 3, GUIDE latent factors explain the variance in the traits using far fewer factors than those of TSVD, confirming GUIDE's ability to pick a basis for the latent factors that sparsifies the weights from the latent factors to both the SNPs and the traits (see also Figs. S16 and S17). Put differently, the top GUIDE latent factors always explain more of the variance than their TSVD counterparts. This effect is also qualitatively seen in Figs. S20 and S21, which explicitly compare the SNPs-to-latents and the latents-to-traits weight matrices for GUIDE and DeGAs, showing that, in both cases, the GUIDE matrices are sparser with much more concentrated signal as compared with the DeGAs matrices, for which the weight is more uniformly spread across the latent factors.

Table 1 compares the fraction of the genetic variance explained (see Methods for the definition) by the top three latent factors of DeGAs and of GUIDE for six traits: BMI, myocardial infarction (MI), cholelithiasis (gall stones), Alzheimer's disease, high blood pressure, and diabetes. Fig. 3 displays these results for other traits from the same dataset. Both show that the GUIDE latent factors explain a larger fraction of the genetic contribution to the given traits using the same underlying set of SNPs.

To more specifically characterize the biology of certain traits, we applied GUIDE to eight subsets of the UK Biobank consisting of SNPs significant for a given trait with $p < 5 \times 10^{-8}$ and all other traits correlated to the given trait with $r^2 > 0.5$ (see Table S5 in SI). Figs. S18 and S19 show the GUIDE and TSVD $\mathcal{L} \to \mathcal{T}$ heat maps for models with $L = 100$ and $L = 10$, respectively, for the high blood pressure subset in Table 5. While the $L = 100$ GUIDE model significantly out-performed the corresponding TSVD model in terms of both sparsity and concentration of signal, the difference was not significant for the $L = 10$ models. Motivated by these observations, and by the fact that the qualitative discrimination

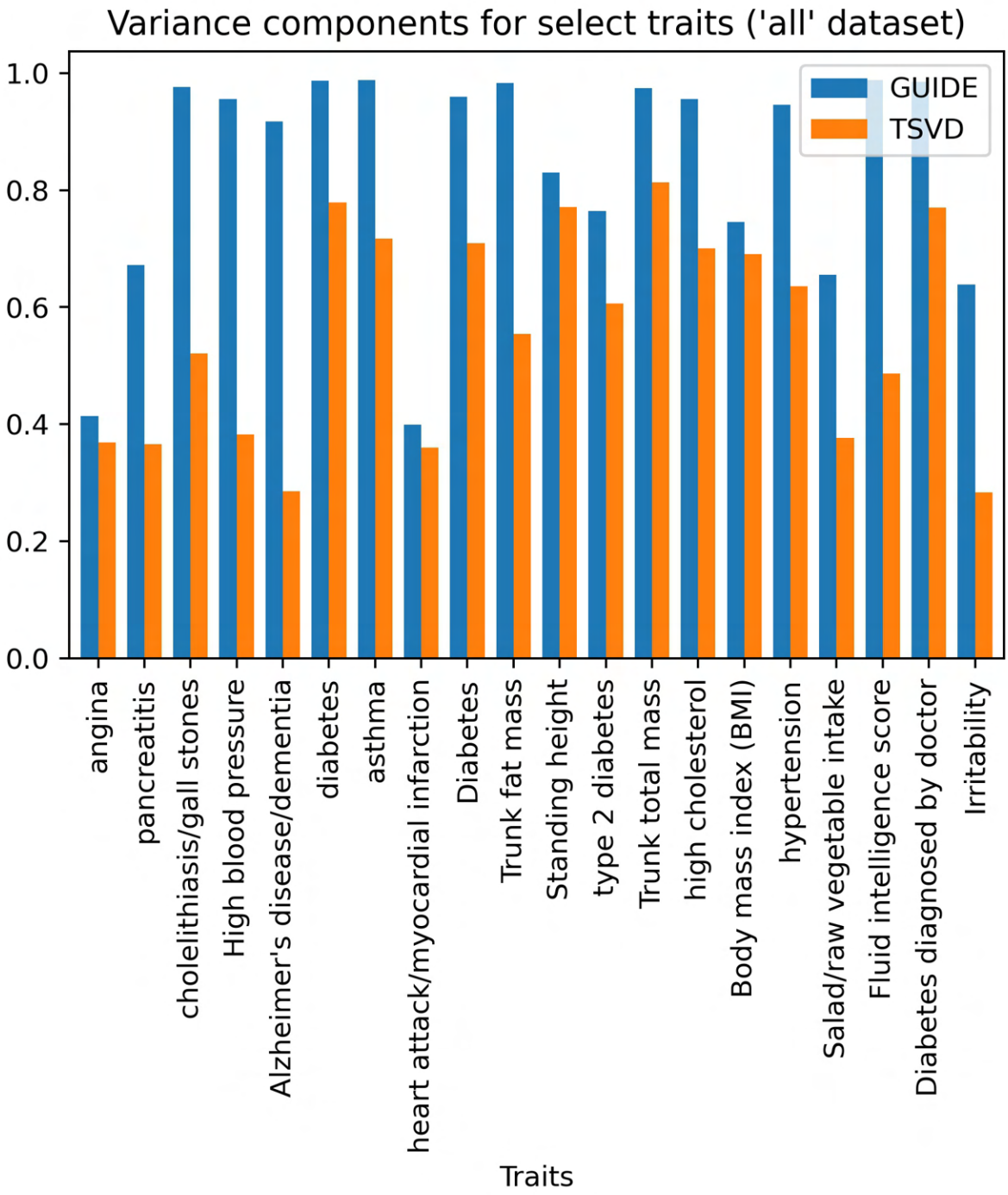


Figure 3: Genetic variance components for select trait using the 'all' dataset from [5]. We report the sum of the scores for the top three latent factors for each trait.

Table 1: Percent of traits explained by latents

Trait ('all' dataset)	DeGAs	GUIDE
BMI	69.0%	93.5%
MI	36.0%	41.0%
Gall stones	52.0%	97.3%*
Alzheimer's	28.5%	88.6%*
High blood pressure	38.2%	90.7%*
Diabetes	77.9%	97.8%*

Percent of the genetic variance explained from the 'all' subset of UK Biobank curated by [5]. All values were the sums of the three highest latent factors for each trait, except for values with superscripted $(*)$, which correspond to the single top GUIDE latent. The DeGAs values for the first three traits were reported in [5] and verified by us, while the last three were solely computed by us.

of models afforded by these types of heat maps is not as precise and will not be feasible for vastly larger values of M or T, we turned to more quantitative measures of GUIDE's performance as a function of the number of latent factors, L.

While, as mentioned above, using entropy is computationally expensive in the execution of the ICA portion of our method, it can still be used ex post facto to compare the performance of GUIDE relative to TSVD. We note that a lower entropy corresponds to a more informative signal. We also look at the entropy gap, defined as the difference between the TSVD model's entropy and the GUIDE model's entropy, with the understanding that the higher the entropy gap, the more information is gained by GUIDE relative to the TSVD model.

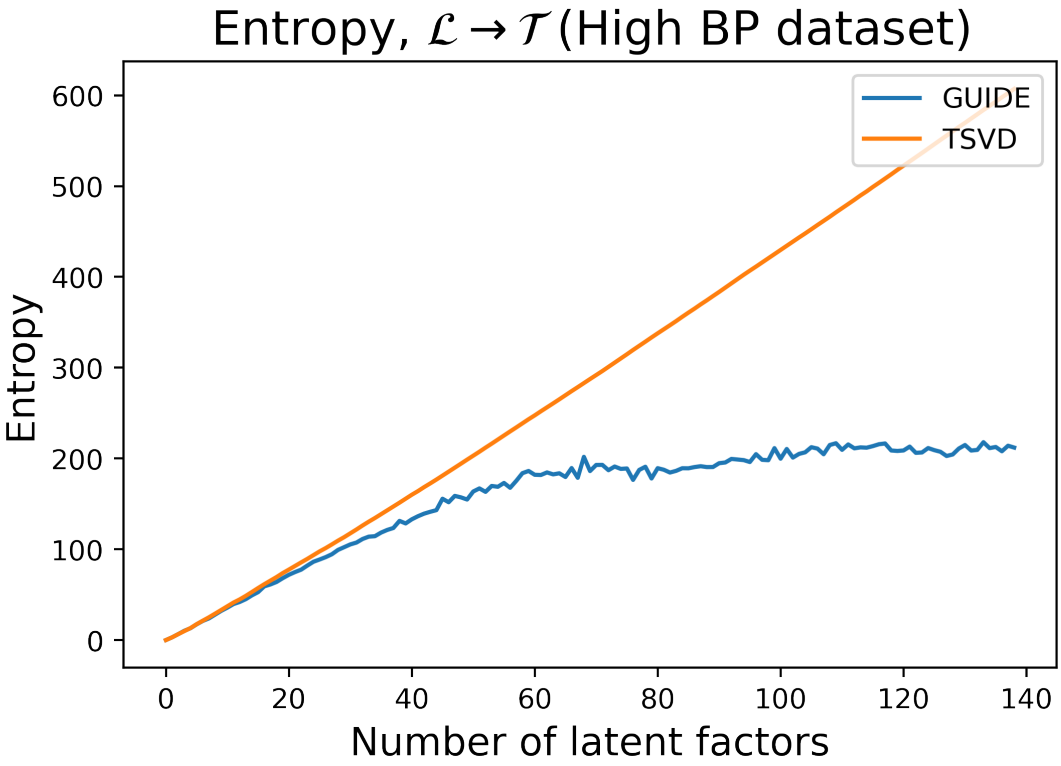
The first plot in Fig. 4 compares the entropies for the TSVD and GUIDE models as a function of the number of latent factors in each model, while the second shows the entropy gap, defined as the difference between the TSVD model's entropy and the GUIDE model's entropy, also as a function of the number of latent factors. Fig. 4 makes it clear that the difference in the entropies is negligible for $L = 10$ and quite significant for $L = 100$, which explains our previous qualitative observation comparing GUIDE's performance between Fig. S18 and Fig. S19. Moreover, referring to the $\mathcal{L} \to \mathcal{T}$ entropy plot in Fig. 4, we see that the entropy plateaus at around $L \approx 50$, suggesting that GUIDE models with at least that number of latent traits should capture most of the information contained in the traits of that dataset given the same dataset's SNPs. The $\mathcal{X} \to \mathcal{L}$ entropy gap in Fig. 4 shows that the largest entropy gap is for $L \in [40, 70]$, further suggesting that a choice of L in this range would allow for more information gain (for the $\mathcal{X} \to \mathcal{L}$ weights) and a more discernible advantage for GUIDE. (Referring again to the entropy plot in Fig. 4, it is clear that the corresponding entropy gap would be monotone increasing, but that this would be driven purely by the increasing entropy of the TSVD model, which would nonetheless mean that for larger value of L, the GUIDE weights would be consistently sparse and concentrated.) Figure S25 shows the $\mathcal{L} \to \mathcal{T}$ entropy plot for the UKB dataset.

We also prepared a subset of the UK Biobank PanUKB dataset (pan.ukbb.broadinstitute.org) by filtering to only include genotyped SNPs, traits with a z-score greater than 4, and excluding missing regression coefficients, consisting of 635,964 SNPs and 677 traits, which we will refer to as the 'UKB $Z > 4$ dataset'. We built an $L = 100$ GUIDE model for this dataset.

Using the contribution scores between the $\mathcal{L}$ and $\mathcal{T}$ layers, and a threshold $1/T$, which is the value corresponding to a completely uniform distribution of the weights (Methods), we defined the sets of traits for each latent factor with squared weights greater than $1/T$ as an indication of which traits are significantly represented by each latent factor.

For the UKB $Z > 4$ dataset, we found that there are 13.94 traits per GUIDE latent on average, compared to 67.38 traits per TSVD latent, even though GUIDE latent factors load onto 68% of all traits with greater than uniform weight, compared to 54% for the TSVD latent factors. Thus, not only are the GUIDE latent factors more sparse (with almost 5 times less traits per latent), but they nevertheless account for more of the traits than TSVD.

For the $\mathcal{X}$ and $\mathcal{L}$ layers, we used a $1/M$ threshold for the same reasons. Fig. 5 shows the number of latent factors each SNP load onto with squared W_{XL} weights greater than $1/M$, confirming GUIDE's ability to deconvolve shared SNP effects by sparsifying interlayer weights, as will be discussed in more detail in the next section.



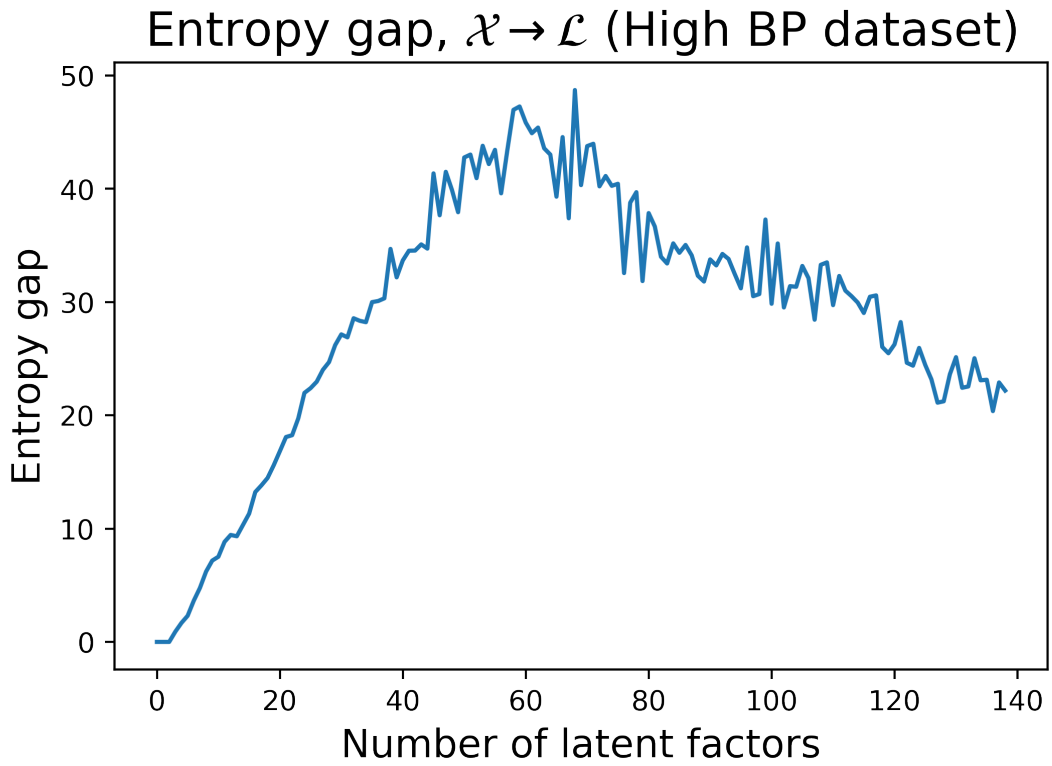


Figure 4: Plots of the entropy for the $\mathcal{L} \to \mathcal{T}$ weights and entropy gap for the $\mathcal{X} \to \mathcal{L}$ weights (defined as the difference between the TSVD model entropy and the GUIDE model entropy) as a function of the number of latent factors in the each model for the High Blood Pressure dataset, defined in Table S5, consisting of 1183 SNPs significant for high blood pressure ($p < 5 \times 10^{-8}$) and 344 traits that are correlated to it with $r^2 > 0.5$. The entropy plot for the $\mathcal{X} \to \mathcal{L}$ weights and the entropy gap plot for the $\mathcal{L} \to \mathcal{T}$ can be found in the Supplementary Information.

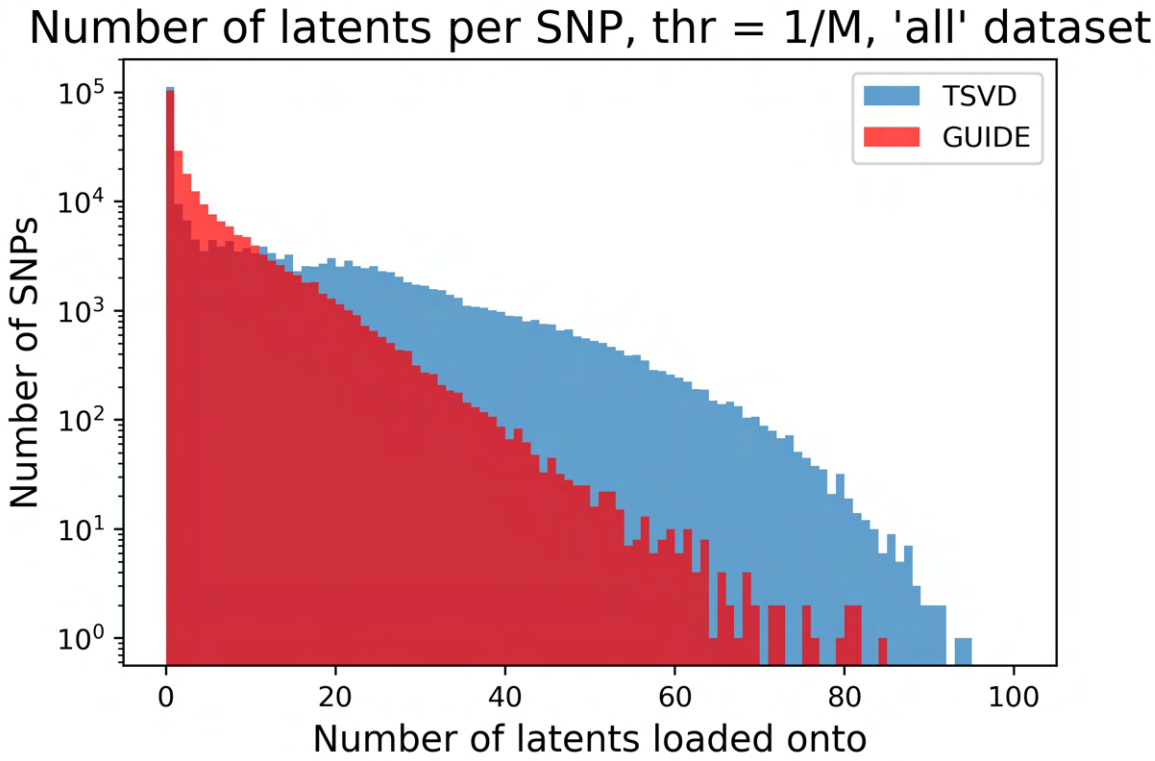


Figure 5: Number of latent factors each SNP loads on to with a squared weight greater than $1/M$. SNPs load onto fewer GUIDE factors compared to TSVD factors.

1.3 Isolating molecular pathophysiological pathways of Alzheimer's disease

The analyses in this section and the next use the GUIDE model built using the 'all' dataset described in Sec. 1.2.

In order to characterize and contextualize each of the top latent factors for Alzheimer's disease (AD), we also considered the traits, "high cholesterol" and "peripheral vascular disease," for which GUIDE latent factors revealed common pathophysiological mechanisms.

For each of these traits we identify the top latent factors based on their *genetic variance components*, defined as a normalized sum of $\mathcal{X} \to \mathcal{L}$ or $\mathcal{L} \to \mathcal{T}$ weights (Methods). We use Genomic Regions Enrichment of Annotations Tool (GREAT) [20] to interpret these latent factors and explore the top genes that are enriched for them.

Table 2 lists the top two or three latent factors for Alzheimer's disease, high cholesterol, and peripheral vascular disease for both GUIDE and TSVD together with their genetic variance components. Fig. 7 presents the top traits and loci contribution scores for the GUIDE and TSVD latent factors for AD and high cholesterol. To improve interpretability, we used ANNOVAR [21] to assign the top 1-3 genes for each locus implicated by our method in Figure 7 and in Figure 9 in the next section, though there is some uncertainty in the assignment of some of the genes (*e.g.*, for loci with several nearby genes). Fig. 6 reports the p-values and binomial fold enrichments given by GREAT for the shared GO processes of GUIDE latent 72 and TSVD latent 71.

For Alzheimer's disease, GUIDE latent 72, which has a variance contribution score of 87.5% (see subsection 3.3 in Methods), is highly enriched for the *APOE*, *NECTIN2*, and *TOMM40* genes, with 49 of the top 54 loci in the 19:45020859-45728059 region near those genes having a contribution score of 50.9% and with many other SNPs in this same region also implicated by GUIDE latent 72, further contributing to the total contribution score due to genes in this locus [22]. Moreover, GUIDE prioritizes genes, such as *RBFOX1* (chr. 16) and *CSMD1* (chr. 8), which have only recently been implicated in AD [23, 24], as well as loci such as 20q13.33, which have previously been found to be associated with AD [25].

On the other hand, GUIDE latent 10 loads on other cholesterol genes, such as *APOA*, *APOB*,

APOC, *LDLR*, *LPA*, *LPL*, *LPLR*, and *PCSK9* [26], and can thus be interpreted as mediating the effects of genes involved in cholesterol metabolism more generally (even though it also loads onto *APOE*, but less significantly than latent 72). GUIDE latent 72 can then be understood as mediating disease processes involving *APOE*, but not as part of the cholesterol metabolic pathway per se. Some of the genes implicated by latent 10, such as *SORT1* (chr. 1) [27], have only recently been found to be significant for AD.

In contrast, TSVD latent 71, the top TSVD factor for Alzheimer's disease with a variance contribution score of 10.2%, mostly picks up a host of related genes such as *APOC*, *APOB*, *LDLR*, among others, without significant enrichment for *APOE*, and with the top 50 loci—nonspecifically spanning chromosomes 1-4, 7-15, 17, 19, and 20—having a contribution score of only 12.7%. Unsurprisingly, TSVD latent 71 significantly implicates GO processes that are not exclusively related to Alzheimer's pathophysiology, such as "positive regulation by host of viral process" ($p = 4.8 \times 10^{-8}$), "drug transport" ($p = 1.16 \times 10^{-7}$), and "spleen development" ($p = 2.8 \times 10^{-6}$).

For high cholesterol, GUIDE latent 10 contributes a variance contribution score of 84.2%, and its high contribution is consistent with its earlier interpretation as a factor significant for a larger group of cholesterol metabolism genes. GUIDE latent 72, highly enriched for the *APOE* gene is next, this time contributing a variance contribution score of 10.2%.

Unlike GUIDE, TSVD does not pick up on the connection with the cholesterol latent factors first identified in connection with Alzheimer's disease, instead prioritizing two other, unrelated factors.

The case of peripheral vascular disease shows GUIDE's usefulness even when its latent factors have low genetic variance component scores that are comparable with TSVD's factors. The genetic variance components for the top three latent factors for GUIDE and TSVD are both around 27%, with each top factor (latent 10 for GUIDE and latent 21 for TSVD, Table 2) matching the top latent factor for high cholesterol. Nevertheless, while the second GUIDE latent ($L8$) prioritizes vascular and cardiac muscular processes, such as "myofibril assembly" ($p = 3.97 \times 10^{-16}$) and "sarcomere organization" ($p = 4.13 \times 10^{-19}$), and with a significant number of SNPs implicating genes such as *TTN*, *GJA1*, *SCN5A*, *WDR1*, *MYH7*, and *CSK*, among others [28, 29, 30, 31, 32, 33], the second TSVD latent (latent 0) does not seem to identify processes directly related to vascular disease pathophysiology, prioritizing GO biological processes such as "regulation of glucose metabolic process" ($p = 1.03 \times 10^{-23}$) and "neural plate pattern specification" ($p = 2.87 \times 10^{-18}$). And while TSVD latent 53, the third contributing factor to peripheral vascular disease, does include GO processes related to cardiovascular morphology and function, it only predominantly prioritizes *PITX2*, and none of the other important genes implicated by GUIDE latent 8. For GUIDE, the third highest factor, GUIDE latent 64, significantly prioritizes genes such as *AHR*, *PML*, *GDNF*, and *IRX1-3*, which were not tagged by TSVD latent factors, but were recently implicated in important mechanisms contributing to vascular disease, such as vascular inflammatory pathways, atherosclerosis, endothelial dysfunction, and oxidative stress [34, 35, 36, 37].

Table 2: Top latent factors for Alzheimer's disease, high cholesterol, and peripheral vascular disease

Trait	GUIDE latent (var. comp.)	TSVD latent (var. comp.)
Alzheimer's disease	L72 (0.875)	L71 (0.102)
	L10 (0.022)	L64 (0.095)
High cholesterol	L10 (0.842)	L21 (0.615)
	L72 (0.102)	L19 (0.045)
Peripheral vascular disease	L10 (0.114)	L21 (0.107)
	L8 (0.101)	L0 (0.099)
	L64 (0.054)	L53 (0.070)

Running a GUIDE model on an LD-pruned set of FinnGen data [38] ($r^2 < 0.2$, with 188,140 SNPs and 2408 traits) revealed additional insights not found in the UKB models. For example, the FinnGen model had one top latent for AD that predominantly implicated 14q24.2, the locus where *PSEN1* is located, another implicated the 9p21 locus associated with vascular dementia and AD [39]. We were unable, however, to replicate latent factors from UKB GUIDE model in the FinnGen GUIDE model, likely due to GUIDE not accounting for effects such as those due to LD, genetic covariance, stratification, and ancestry.

Figure 8 shows the percent shared SNPs for the top 500 SNPs associated with each latent factor

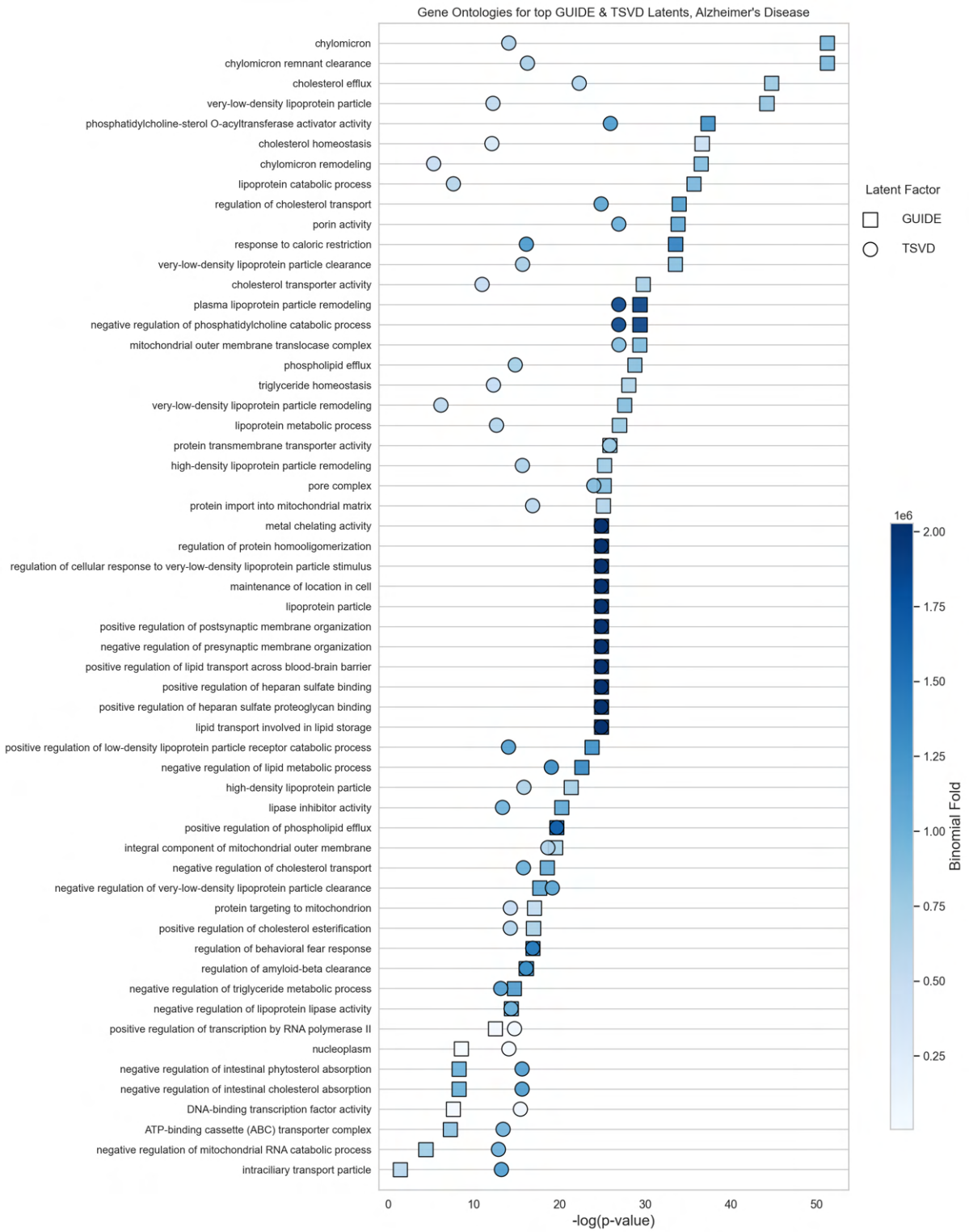


Figure 6: Gene Ontology (GO) Biological Processes for the top latent factors for Alzheimer's disease identified by each method, GUIDE latent 72 and TSVD latent 71. This factor is highly enriched for loci near the *APOE* genes, and exclusively prioritizes processes involved in cholesterol metabolism and processes implicated in Alzheimer's disease, in particular. Although this figure shows only shared processes, we note that TSVD latent 71 also prioritizes many processes that are not related to Alzheimer's pathophysiology (see main text).

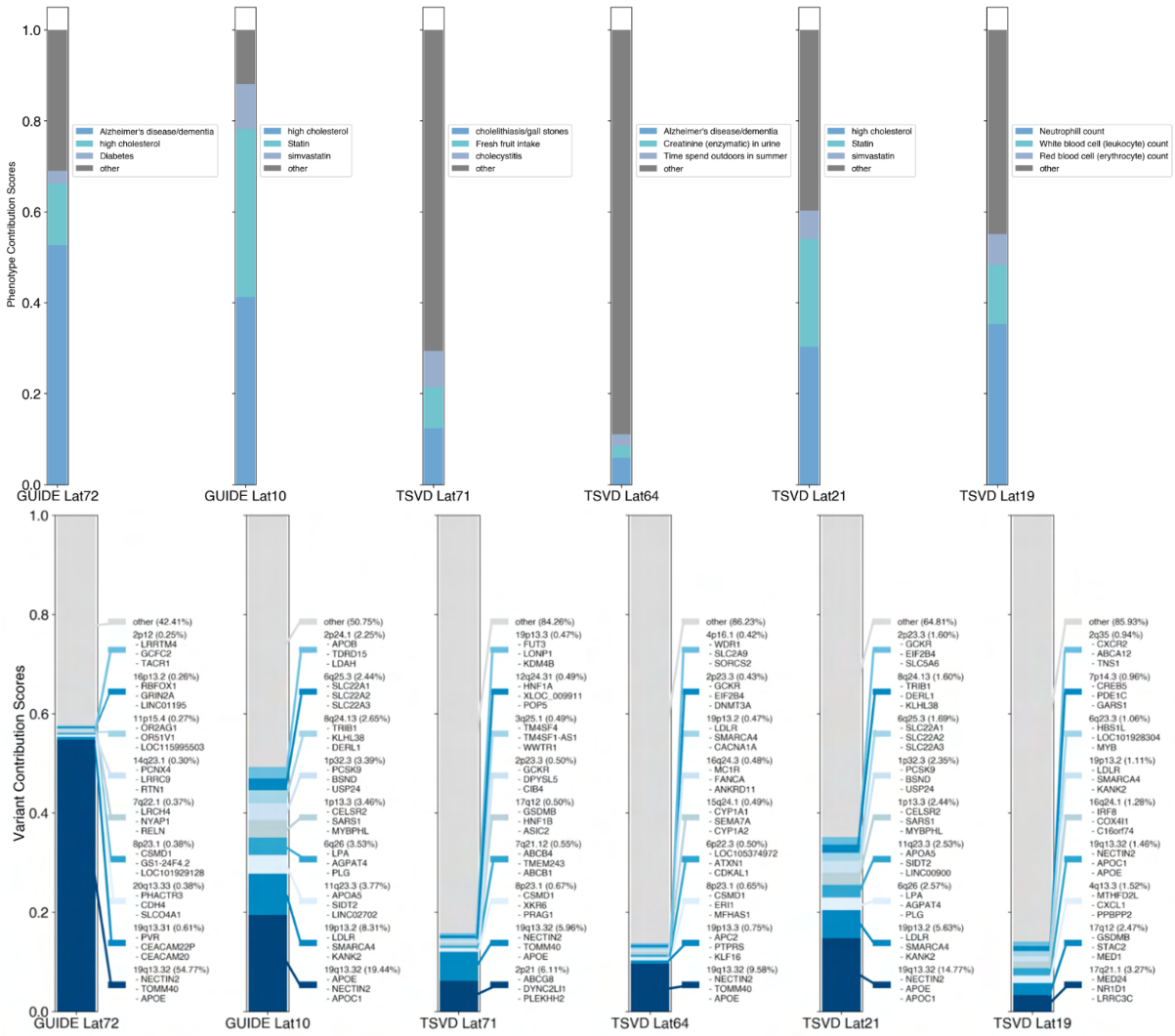


Figure 7: Top three contributions from traits and top nine contributions from loci onto the relevant GUIDE and TSVD latent factors as defined in Table 2. We determined the genes for the top SNPs using ANNOVAR [21]. The corresponding contribution scores for each locus are in parentheses. The contribution score due to each gene within a given locus is based on the ANNOVAR annotations.

in Table 2, measured using the Jaccard index (Methods). Among the GUIDE latent factors, there is only an overlap of 8.0% between latents 72 and 10, due to the *APOE*-associated SNPs in latent 10, as mentioned earlier. Among the TSVD latents, there is more overlap between factors 71, 64, 21, and 19, suggesting that a more significant part of the genetic variance component—as compared with the sparser GUIDE factors—is contributed by TSVD factors due to double counting of certain contributing SNPs, which inflates the score.

GUIDE latent 10 and TSVD latent 21, both of which correspond to nonspecific genes involved in cholesterol metabolism, share 73.4% of their top 500 SNPs. This kind of overlap is to be expected, especially for highly effective TSVD factors, since TSVD can, for some factors, pick out a basis that happens to closely align with that derived by GUIDE. TSVD latent 21 is quite effective given that, unlike the factor comparison with Alzheimer's disease where none of the top TSVD latents were shared, we do see that latent 21 is again implicated in peripheral vascular disease, like its GUIDE counterpart, latent 10. Figure 7 summarizes some of the above discussion, showing the top contributing genes and traits for each of the relevant latent factors identified in Table 2.

All in all, Fig. 8 summarizes some of the trends we have seen in this section: GUIDE latent factors are more independent (in the sense of lower mutual information; Methods) and sparse (except where certain processes naturally overlap), allowing them to home in on a wider range of important genes while differentiating between distinct biological processes and disease mechanisms.

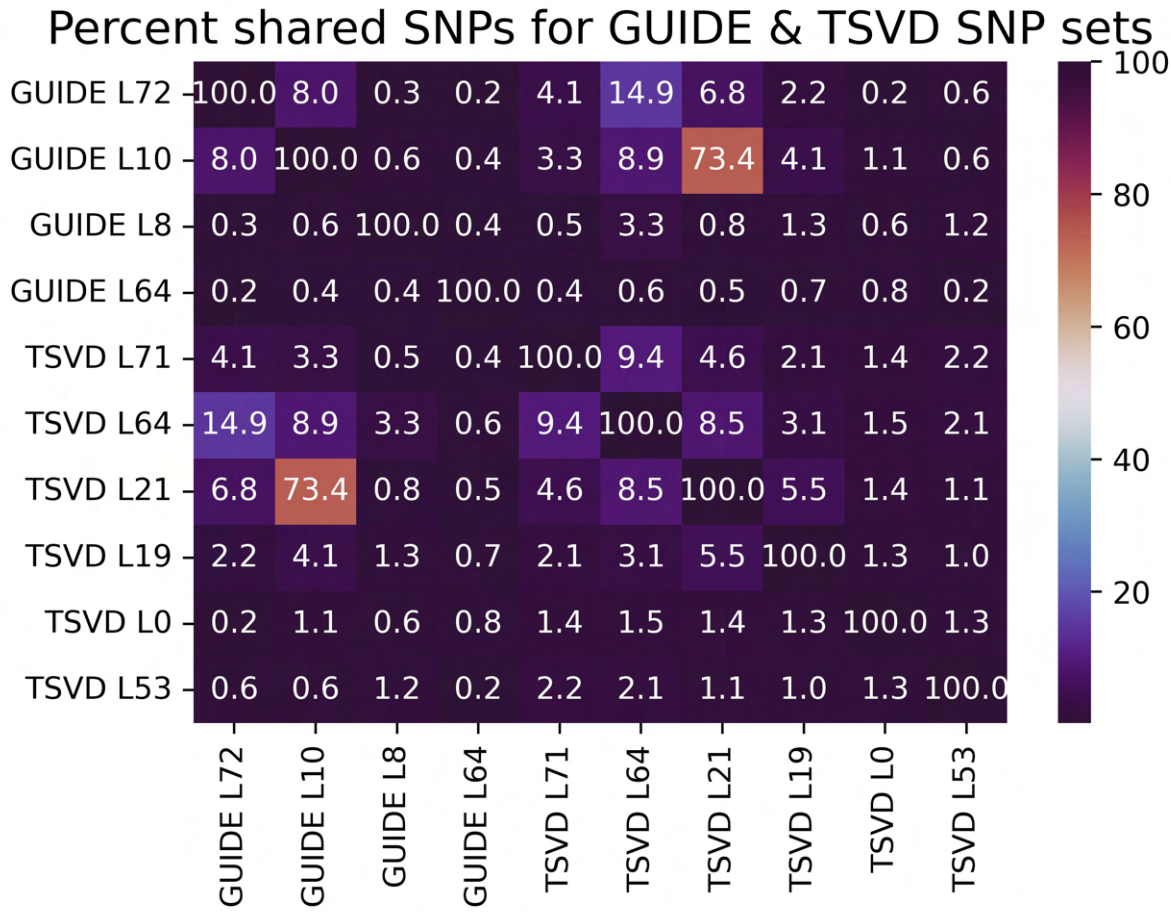


Figure 8: Percent common SNPs (as measured by the Jaccard index, Methods) for the top 500 SNPs corresponding to the GUIDE and TSVD latents in Table 2. While the general cholesterol factors (GUIDE latent 10 and TSVD latent 21) share a many SNPs, all other factors do not.

1.4 Deconvolution of FEV1 reveals environmental and molecular factors

Whereas the last section showed GUIDE's ability to deconvolve the molecular mechanisms contributing to a given trait, in this section we show, using the example of forced expiratory volume in one second (FEV1), GUIDE's ability to disentangle and stratify the behavioral or environmental mechanisms from the molecular ones contributing to a particular end trait.

Fig. 9 shows the stacked plots for the top latent factors for FEV1 in descending order by the magnitude of their genetic variance components for FEV1. The top two panels show the breakdown of the contributions to each latent factor by loci and traits, respectively, while the bottom panel shows the Manhattan plot for 'FEV1, best measure', where the loci tagged by the top three latent factors are colored in.

Latent factor 7 contributes to FEV1 through immunological and epigenetic regulatory mechanisms, tagging genes such as *KANSL1, WNT3* and *GYPA* [40, 41]. Latent factors 80 and 27 in Figure 9 (see also Fig. S26, Supplementary Information) are the top contributing factors for smoking behavior, implicating loci with genes such as *CHRNA3*, which encodes the alpha 3 subunit of the nicotinic acetylcholine receptor, which is known to increase the risk for initiation of smoking behavior [42]. Interestingly, after building a GUIDE model using the same traits but removing traits related to smoking behavior, the corresponding latent factors—determined by matching contributions other traits and variants—in the new model do not implicate those same genes, as seen in Figure S27.

The phenotypic contributions to latent 65 are dominated by the left and right hand grip strength traits, while the variants tagged by latent 65 are associated with genes such as *TTN, UQCC1, TSC2, CARNS1*, and *IGF1R*, suggesting that this factor represents cardiovascular health and fitness as mechanisms affecting FEV1. Latent 5 contributes to FEV1 by way of developmental processes that lead to increased lung volume by virtue of having a larger thoracic cavity, implicating genes such as *UQCC1, PTCH1*, and *CDK2AP1* [43]. Unsurprisingly, latent 5 is the top contributing latent factor for sitting height and the third for standing height.

As in the last section, we also compared these results with the same GUIDE model on the LD-pruned set of FinnGen data [38] ($r^2 < 0.2$, with 188,140 SNPs and 2408 traits). Given that the FinnGen trait set is completely different from that of the UKB—and, in particular, FinnGen does not have any FEV1 traits—we instead looked at all the traits related to respiratory disorder, such as asthma, chronic obstructive pulmonary disease (COPD), interstitial lung disease (ILD), or respiratory distress of the newborn. The top latent factors implicated for these traits loci such as 1p31.1 associated with lung cancer in never-smokers [44], 2q11.2 (which seems to be implicated in immune processes/cancer in the lungs) [45], and 2p24, which seems to be associated with FEV1, asthma, and COPD [46, 47, 48].

The lower panel in Fig. 9 shows a Manhattan plot for the trait "FEV1, best measure", where we highlighted the variants tagged by the top three GUIDE latent factors contributing to this trait.

Let g_{sig} be the number of variants implicated by GUIDE that are above the significance threshold, g_{total} be the total number of variants implicated by GUIDE, n_{sig} be all variants above the significance threshold and n_{total} be the total number of variants. We can then define the enrichment of the top three GUIDE latent factors by,

$$\text{En} = \frac{g_{\text{sig}}/g_{\text{total}}}{n_{\text{sig}}/n_{\text{total}}}.$$

For the Manhattan plot in Fig. 9, En = 24.5, indicating, by a naive interpretation of this metric, that the top three GUIDE latent factors tag 24.5 times more significant variants as would be expected were variants picked at random. However, given that LD is nonuniformly spread across variants, with a preference for significant variants, this metric cannot be taken at face value. Conversely, stringent LD pruning—which would presumably fix this problem—may make it more difficult for GUIDE to find significant variants given that most of them would be removed. The most sensible approach thus seems to appreciate—qualitatively—that En > 1 indicates that the latent factors have picked up more significant variants than would be expected were they to be chosen randomly, but not to overinterpret the actual value of En (though perhaps allowing for comparisons).

2 Discussion

We developed GUIDE as a method for disentangling polygenicity into more coherent subsets of genetic effects that might relate to specific pathways or exposures relevant to outcome traits under study. GUIDE uses summary statistics data from GWAS and PheWAS to build a sparse and interpretable three-layer network consisting of SNPs, traits, and an intermediate layer of latent factors. GUIDE not only uses a

Figure 9: Top latent factors for FEV1 traits, showing the top variants' (top panel) and phenotypes' (middle panel) contributions to each latent factor. The latent factors are ordered in descending order by their contribution to the trait, "FEV1, Best measure". Bottom panel: Manhattan plot highlighting the variants tagged by the top three latent factors contributing to FEV1 traits. The significance threshold, given by the dotted red line, has been set at $p = 5 \times 10^{-8}$.

simultaneous two-sided application of ICA to find the most optimal basis (informed by both the genetic and phenotypic data) for these latent factors, but also estimates the number of required latent factors using an approach that, unlike current model selection methods, works better when applied to larger, noisier data sets.

The independence of the latent factors and the sparsity of the weights connecting them with the SNP layer $\mathcal{X}$, on one side, and the trait layer $\mathcal{T}$, on the other, are not only mathematically and computationally convenient properties. More importantly, as we have shown, these properties make the GUIDE latent factors more biologically interpretable and allow them to extract more information from the underlying genetic architecture. For example, as seen in the application to Alzheimer's disease in Section 1.3, GUIDE latent factors were able to separate out some of the main genetic contributions to the disease due to the *APOE* genes, thus discriminating between the pathophysiological pathways due to these genes versus other ones in the cholesterol metabolic pathway, which were captured by another latent factor.

It is important to note, however, that the success of the ICA portion of the GUIDE algorithm rests on the assumption that there exist discrete modules mediating between the SNPs and traits (see Fig. 1). This assumption fails, for example, for genetic variants with infinitesimal effects on given traits, which induce a continuous variance-covariance structure that cannot be written as a sum of discrete non-Gaussian modules. Here, TSVD would be the better approach. In fact, using the approach outline in Methods (3.8) GUIDE can be used to build models that capitalize on this phenomenon by extracting the discrete modules (the latent factors) as before, but using the remaining Gaussian components (which ICA cannot distinguish) to analyze the continuous contributions to a trait's variance.

Moreover, the results of GUIDE are influenced by the traits included in the data, as we showed in Sec.1.4 in the analysis of FEV1 traits using GUIDE models with and without smoking traits. Naturally, since GUIDE latent factors are informed by the given genetic variants and traits, larger datasets produce better results, suggesting that broader and better powered systematic GWAS would help build better informed GUIDE models. Though we focused on diagnoses rather than course, outcome or treatment response traits in this paper, we hope that using larger sets of traits that include the latter will also allow GUIDE to glean new insights that will improve our understanding of the dynamics of human illnesses.

Of course, the expectation that using larger datasets would better inform GUIDE latent factors is mitigated by the presence of linkage disequilibrium (LD) and genetic covariance.

Given that GUIDE does not explicitly account for LD or genetic covariance, the latent factors it derives are biased by the presence of both, although to a somewhat lesser degree than TSVD. For example, in Fig. S24, we plotted the genetic variance components for the 2,692,064 million SNPs by 4,357 traits UKB dataset GUIDE model with $L = 200$ (as estimated by our model selection algorithm, Methods). We noticed that for this set, the genetic variance components for the top three TSVD latent factors had a score close to 1 for nearly all traits, while the top three GUIDE latent factors had much lower scores (even lower than for the curated subsets of the UK Biobank used in Results). This is likely due to large LD and genetic covariance inflating the weights of the TSVD latent factors (which, as mentioned above in Sec. 1.3, usually have a much larger number of above-average-loading traits per latent) to a larger extent than the sparser GUIDE weights, an effect we have observed in Fig. 8. These problems can be partially avoided by LD and genetic covariance pruning, and running a GWAS of many traits form the same cohorts, as opposed to using GWAS from different studies.

As mentioned in Sec. 1.4, overly aggressive LD pruning may hinder GUIDE's ability to find significant variants since most of them would be removed. The definitive solution, which we hope to pursue in future work, therefore seems to require explicitly incorporating the effects of both LD and genetic covariance, allowing for models that accurately account for both intra- and inter-layer effects. This would allow using the information contained in these intra-layer relationships to inform the model instead of biasing its results. Such models also hold promise to implicate causal variants, serving as a new avenue for fine-mapping. By directly accounting for LD and genetic covariance, while also perhaps incorporating newer information-maximizing algorithms [49], we hope that future iterations of our method will be even better equipped to learn biologically-relevant latent modules, while also accounting for population differences and ancestry.

In summary, GUIDE does provide a principled way to decompose genetic associations for complex traits into more coherent subunits. The interpretation of the factors that emerge from a GUIDE analysis is facilitated by the strength of the relationship to the outcome traits as well as the biological annotation of the genetic variants driving the signal. These resulting factors may facilitate the interpretation of genetic analyses and suggest groups of SNP effects to pursue together for functional follow-up studies.

3 Methods

3.1 Preliminaries

In our network model, we start with a set of SNPs, $\mathcal{X}$, and a set of traits, $\mathcal{T}$. Then the matrix of estimated regression coefficients ("betas"), $\hat{B}$, gives the weights of the edges connecting every SNP-trait pair. Our goal is to determine the weights to and from a layer of intermediate latent factors, $\mathcal{L}$, that best represent and summarize the relationships between $\mathcal{X}$ and $\mathcal{T}$. In other words, our basic problem is to find a decomposition,

$$\hat{B} = B_{XL} B_{LT}, \tag{1}$$

where B_{XL} is the $M \times L$ matrix of $\mathcal{X}$ to $\mathcal{L}$ weights and B_{LT} is the $L \times T$ matrix of $\mathcal{L}$ to $\mathcal{T}$ weights, with $L < T < M$. Intuitively, the 'true' latent layer $\mathcal{L}$ is an intermediate, lower dimensional space that maximally extracts the genetic information encoded in $\mathcal{X}$ and is thus a better, more biologically informed representation of the resulting traits in $\mathcal{T}$.

Let u_ℓ be the ℓth column vector of B_{XL}, and v'_ℓ be the ℓth row vector of B_{LT}:

$$B_{XL} = \begin{bmatrix} | & & | & & | \\ u_1 & \cdots & u_\ell & \cdots & u_L \\ | & & | & & | \end{bmatrix}, B_{LT} = \begin{bmatrix} — & v'_1 & — \\ & \vdots & \\ — & v'_\ell & — \\ & \vdots & \\ — & v'_L & — \end{bmatrix}.$$

Then we can write (1) as the sum

$$\hat{B} = \sum_{\ell=1}^{L} u_\ell v'_\ell, \tag{2}$$

where, in the context of our network model, each product of vectors gives the matrix of weights to and from a single latent factor.

Therefore, the number of latents traits L is, in theory, equal to the rank of $\hat{B}$:

$$\text{rank}(\hat{B}) = L \leq T, \tag{3}$$

where the inequality here is due to the property that the rank of a matrix is at most equal to the least of its dimensions (in our case $\hat{B}$ is a $M \times T$ matrix, with $T < M$). In reality, however, most datasets are sufficiently noisy so that $\text{rank}(\hat{B}) = K \approx T$. Therefore, more sophisticated methods are needed to estimate L (see Sec. 3.8).

Our task is essentially to decompose $\hat{B}$ as in Eq. (1). Of course, if B_{XL}, B_{LT} satisfy (1) then so would

$$B_{XL} A \leftarrow B_{XL} \tag{4}$$

$$A^{-1} B_{LT} \leftarrow B_{LT}, \tag{5}$$

since

$$\hat{B} = B_{XL} A A^{-1} B_{LT} = B_{XL} B_{LT}, \tag{6}$$

where $A \in \text{GL}(L, \mathbb{R})$ is a transformation matrix—a member of the family of all $L \times L$ invertible matrices with entries in $\mathbb{R}$—that chooses the proper scaling and orientation of the weights. (Of course, for $L = 1$, A and A^{-1} are simply multiplicative constants.) By whitening $\hat{B}$ we instead look for $A \in \text{O}(L, \mathbb{R})$, that is for an $L \times L$ orthogonal matrix.

Our solution to this decomposition problem, GUIDE, will be outlined in the next section. We note here that we can easily extend GUIDE to allow one to include predefined latent factors in the latent layers by defining a *discrepancy matrix*, $\hat{B}_* = \hat{B} - \tilde{B}_{XL} \tilde{B}_{LT}$, where the second term includes the weights to and from the predefined set of latent factors. Starting from $\hat{B}_*$ we then derive the rest of the latent factors to fill out the network using our method as before.

3.2 An optimal basis for the latent factors

We start by using truncated singular value decomposition (TSVD) to obtain an initial decomposition of $\hat{B}$:

$$\hat{B} = USV', \tag{7}$$

where S is an $L \times L$ diagonal matrix consisting of the nonzero eigenvalues of $\hat{B}\hat{B}'$, U is an $M \times L$ orthogonal matrix consisting of the eigenvectors of $\hat{B}\hat{B}'$ corresponding to the (squared) eigenvalues in S, and V is an $T \times L$ orthogonal matrix consisting of the same for $\hat{B}'\hat{B}$. In the next section we address how to estimate the appropriate value for L using our method.

Let C_n be the $n \times n$ centering matrix, an orthogonal projection given by

$$C_n = I_n - \frac{1}{n} 1_n 1_n',$$

where I_n is the $n \times n$ identity matrix and 1_n is the vector of all ones. Even though the nonzero entries in $\hat{B}$ may be drawn from a distribution with a mean of zero, the sampled values will not be exactly mean-centered. Thus, we first mean center both the rows and columns of $\hat{B}$ by applying left- and right-multiplying by the appropriate centering matrix:

$$C_M \hat{B} C_T \tag{8}$$

Define the $L \times (M+T)$ matrix by concatenating U' and V':

$$\mathcal{G} = (U' | V'). \tag{9}$$

As a consequence of the centering, U and V are mean-centered as well. Then, using the fact that $A^{-1} = A'$, we may write (4) and (5) as

$$\mathcal{G}_u = A\mathcal{G}, \tag{10}$$

where $\mathcal{G}_u$ is the $L \times (M+T)$ unmixed matrix each of whose rows consists of the $M+T$ weights to and from a latent trait. Written this way, it is clear that we are looking for the unmixing matrix A that would pick a basis for the latent factors that is simultaneously informed by both the SNPs in $\mathcal{X}$ and the traits in $\mathcal{T}$. The weight matrices W_{XL}, W_{LT} to and from the latent factors in $\mathcal{L}$ are then given by (modulo scaling of the weights),

$$W_{XL} = UA \tag{11}$$

$$W_{LT} = A'V', \tag{12}$$

where we solve for A by applying Independent Component Analysis (ICA)to (10) using the FastICA algorithm [11, 8]. Note that $\mathcal{G}_u = (W_{XL}' \mid W_{LT})$.

With this in mind, we rewrite (6), now in the new basis, as

$$\hat{B} = W_{XL} S_* W_{LT} \tag{13}$$

$$= \underbrace{UA}_{\substack{\text{normalized}\\ \text{weights in}\\ \text{new basis}}} \overbrace{A'SA}^{\substack{\text{scaling relative}\\ \text{to new basis}}} \underbrace{A'V'}_{\substack{\text{normalized}\\ \text{weights in}\\ \text{new basis}}}, \tag{14}$$

where

$$S_* = A'SA.$$

Of course, S_* is not diagonal, which reflects the fact that the components of the new basis are generally not orthogonal in the original (prewhitened) feature space (if they were, then A would be a permutation matrix, and S_* would be diagonal with elements equal S but reordered). For example, let $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$, the permutation matrix that swaps the second and third components, and write

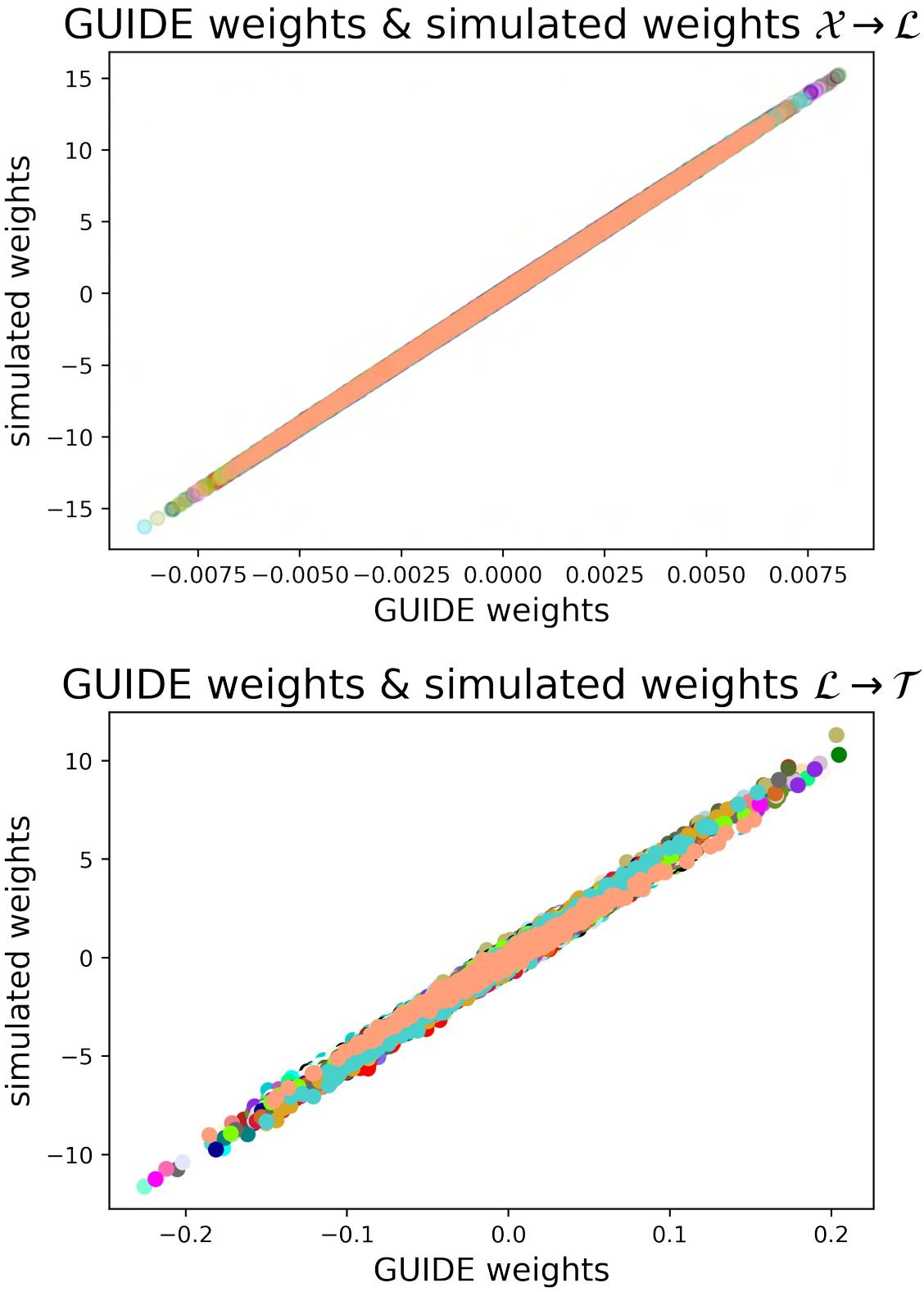


Figure 10: Scatter plot comparing the real (simulated) weights with the estimated ones using GUIDE. In this simulation, we used $M = 10^6$, $T = 1000$ and $L = 50$, and corrected for the permutation and sign differences between the real and estimated latent factors. Each color corresponds to a different latent factor.

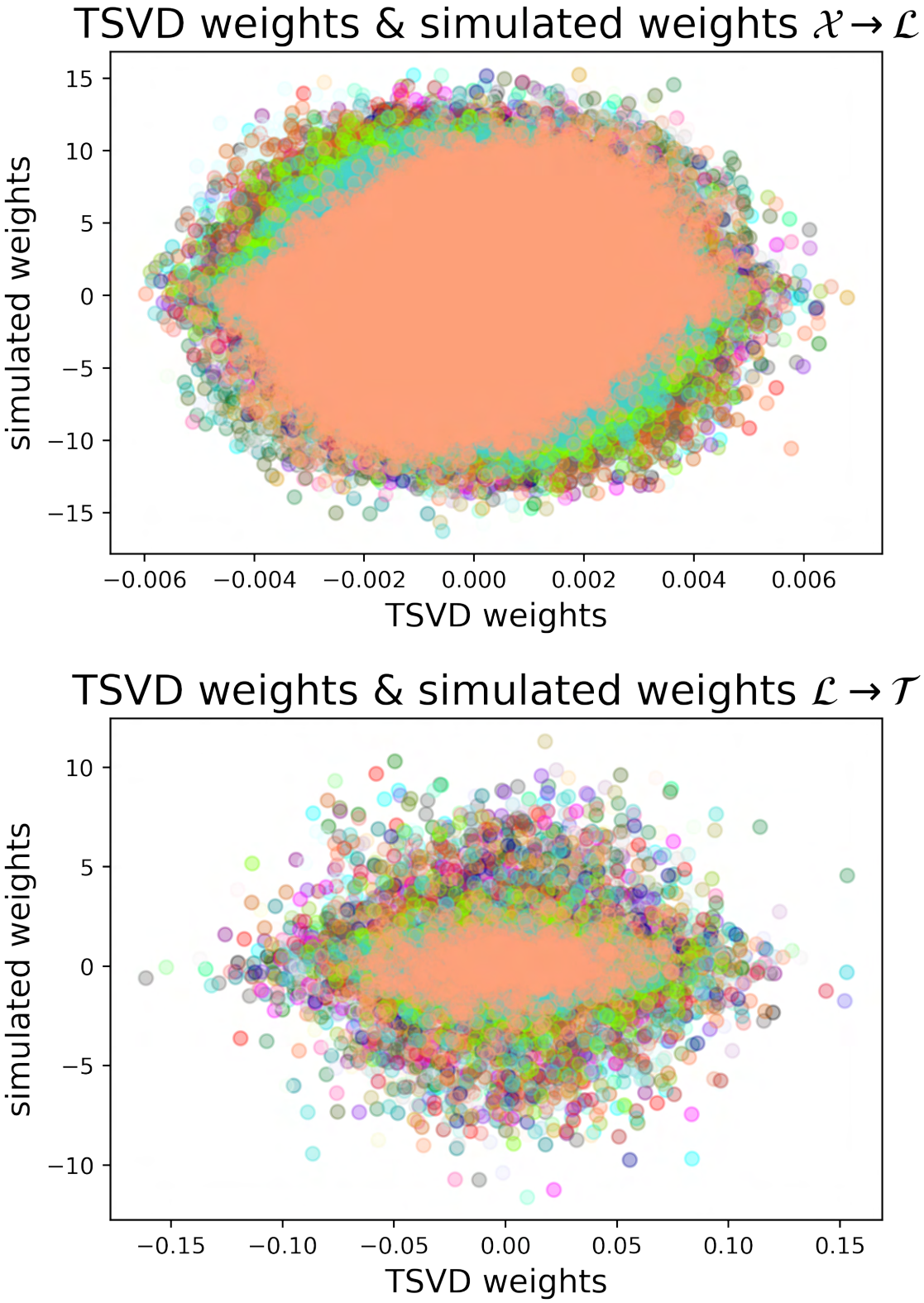


Figure 11: Scatter plot comparing the real (simulated) weights with the estimated ones using TSVD only (as in DeGAs).The simulation parameters and procedure are otherwise the same as for Fig. 10.

$S = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}$. Then $S_* = ASA' = \begin{bmatrix} a & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & b \end{bmatrix}$. As this simple example illustrates, S_* is nothing more than a scaling relative to the new basis. While a permutation keeps the mutual orthogonality of the components intact, a rotation generally does not, which may provide some intuition for the existence of nonzero off-diagonal entries in S_* when A is a rotation matrix.

Also note that $S'_* = (ASA')' = AS'A' = ASA' = S_*$. In the whitened feature space, the ICA components are orthogonal, and in that space A is a permutation matrix.

Alternatively, we can view the decomposition as

$$\hat{B} = \underbrace{UA}_{BW'_{LT}}\overbrace{A'SA}^{W'_{XL}B}A'V',$$

which is further addressed in Section 3.4

The method detailed in this section is summarized below in Algorithm 1. Fig. 10 shows how weights estimated using GUIDE can recover the real (simulated) weights modulo scaling for both $\mathcal{X} \to \mathcal{L}$ and $\mathcal{L} \to \mathcal{T}$ weights, while Fig. 11 shows how using TSVD alone fails to do so.

Algorithm 1

1: $B \leftarrow C_M B C_T$ ▷ mean-centering rows and columns of B
2: $B = USV'$ ▷ TSVD of B
3: $\mathcal{G} = (U'|V')$
4: $\mathcal{G}_u = A\mathcal{G}$ ▷ finding unmixing matrix A using ICA
5: $W_{XL} = UA$, $W_{LT} = A'V'$.

3.3 Contribution scores

We define the contribution scores for the SNPs and traits, respectively, as

$$W_{LT}^{\circ 2} = (w_{\ell t}^2), \;\; (V')^{\circ 2} = (v_{\ell t}^2), \;\; \ell = 1, \cdots, L, t = 1, \cdots, T,$$

where $W_{LT} = (w_{\ell t})$ and $V' = (v_{\ell t})$ are the $L \times T$, $\mathcal{L} \to \mathcal{T}$ weight matrices for GUIDE and TSVD, respectively, and the superscripted $\circ$ denotes the Hadamard (entrywise) exponent. By construction, both W_{LT} and V' are semiorthogonal, in the sense of satisfying $W_{LT}W'_{LT} = I_L$ and $V'V = I_L$, where I_L is the identity matrix of size L, but $W'_{LT}W_{LT} \neq I_T$ and $VV' \neq I_T$. Therefore, if the total weight were uniformly distributed across W_{LT} and V', then every entry of $W_{LT}^{\circ 2}$ and $(V')^{\circ 2}$ would be given by $1/T$. Thus, one way to characterize a latent is to record all traits for a given latent such that $w_{\ell t}, v_{\ell t} > 1/T$. This is the default criterion we use for the GUIDE network visualization module in our software package. Taking the union of all traits with greater than uniform weights can also indicate the fraction of traits that are significantly represented by the latent factors, as we have done in Section 4.4.

3.4 Genetic variance components

Given the truncated singular value decomposition (TSVD) $\hat{B} = USV'$, the unmixing matrix defined in Algorithm 1, A, we define the TSVD factor scores,

$$\begin{aligned} F_{XL} &= US \\ F_{LT} &= VS, \end{aligned} \tag{15}$$

and the GUIDE factor scores,

$$\begin{aligned} G_{XL} &= USA \\ G_{LT} &= VSA. \end{aligned} \tag{16}$$

The corresponding definition for the GUIDE factor scores can be justified in several ways. First, the unmixing (rotation) matrix can only be applied after the scaling due to S is applied, since otherwise (if we had defined it as UAS, for example) A would have likely permuted at least some of the factors, so the singular values would have been applied to the wrong factors. Alternatively, we replace each of the matrices in 15 with their GUIDE counterparts: $U \mapsto W_{XL} = UA$, $V \mapsto W_{LT} = VA$, and $S \mapsto A'SA$. Thus, $G_{XL} = W_{XL}T = UAA'SA = USA$ and $G_{LT} = W_{LT}T = VAA'SA = VSA$.

Writing $F_{XL} = (f_{x\ell})$, $F_{LT} = (f_{\ell t})$, $G_{XL} = (g_{x\ell})$, $G_{LT} = (g_{\ell t})$, the genetic variance components for each factor are then defined as,

$$\mathrm{Comp}(f_{x\ell}) = \frac{f_{x\ell}^2}{\sum_{\ell=1}^{L} f_{x\ell}^2}, \quad \mathrm{Comp}(f_{\ell t}) = \frac{f_{\ell t}^2}{\sum_{\ell=1}^{L} f_{\ell t}^2}, \tag{17}$$

$$\mathrm{Comp}(g_{x\ell}) = \frac{g_{x\ell}^2}{\sum_{\ell=1}^{L} g_{x\ell}^2}, \quad \mathrm{Comp}(g_{\ell t}) = \frac{g_{\ell t}^2}{\sum_{\ell=1}^{L} g_{\ell t}^2}. \tag{18}$$

These scores are referred to as 'cosine squared scores' in [5].

3.5 Genetic analysis of traits using GUIDE factors

Given a GUIDE model and a trait of interest, the top contributing latent factors are first determined based on their genetic variance components. Each latent factor is then characterized and interpreted by rank ordering the SNPs and traits that load onto it with the highest contribution scores, which can be visualized using stacked plots, as in Fig. 7, for example.

For further analysis, one may next 'forward propagate' from the given latent factor to the top contributing trait and then 'back propagate' to that trait's top loading latent factor to compare the effects of these two latent factors. Alternatively, starting again from the first latent factor, one may back propagate to the top loading SNP (or group of SNPs) and then forward propagate to its top loading latent factor for a similar analysis. Fig. 12 summarizes the first round of this procedure. One may alternatively start with a given SNP or set of SNPs instead of a trait and proceed as above.

To isolate mechanisms that do not involve some other trait or process—*e.g.*, as was done in Sec. 1.4 with FEV1 and smoking behavior—one may rank order, based on the genetic variance components, the latent factors for both traits and subtract the latter from the former. In Sec. 1.4 this was used to identify latent factors contributing to FEV1 that do not involve influencing risk for smoking behavior or addictive behavior.

3.6 Stability of GUIDE latent factors

To check for consistency of the latent factors we divided in half the set of SNPs in the original UKB dataset, random assigning each SNP to one of two partitions, and ran GUIDE on each partition. We compared each partition's GUIDE model with GUIDE model on the full dataset, which we considered the ground truth. We rank ordered the genetic variance components for every trait to account for the permutation of the latent factor labels between every run of the model (the sign is not a problem in this case since genetic variance components are computed by squaring the weights, and so are positive by definition). To quantify the error in the genetic variance components (and by extension the network weights) we used the weighted absolute percent error (WAPE), given by,

$$WAPE(X_{ij}, \hat{X}_{ij}) = \frac{\sum_{i,j} |X_{ij} - \hat{X}_{ij}|}{\sum_{i,j} |X_{ij}|},$$

where X_{ij} are the ground truth values and $\hat{X}_{ij}$ are those for the partitioned data.

For the randomly partitioned data the mean WAPE was around 8.5%, but seemed to have been driven up by a few outliers. As an example, for Alzheimer's disease (AD), the WAPE was 2.0%, while for diabetes it was 1.3%.

We also partitioned the data based on even versus odd chromosome number, finding that the error, while on average higher for this nonrandom partition, was still fairly low for polygenic traits, but was higher for traits with top latent factors loading strongly on particular loci. For example, for AD, the

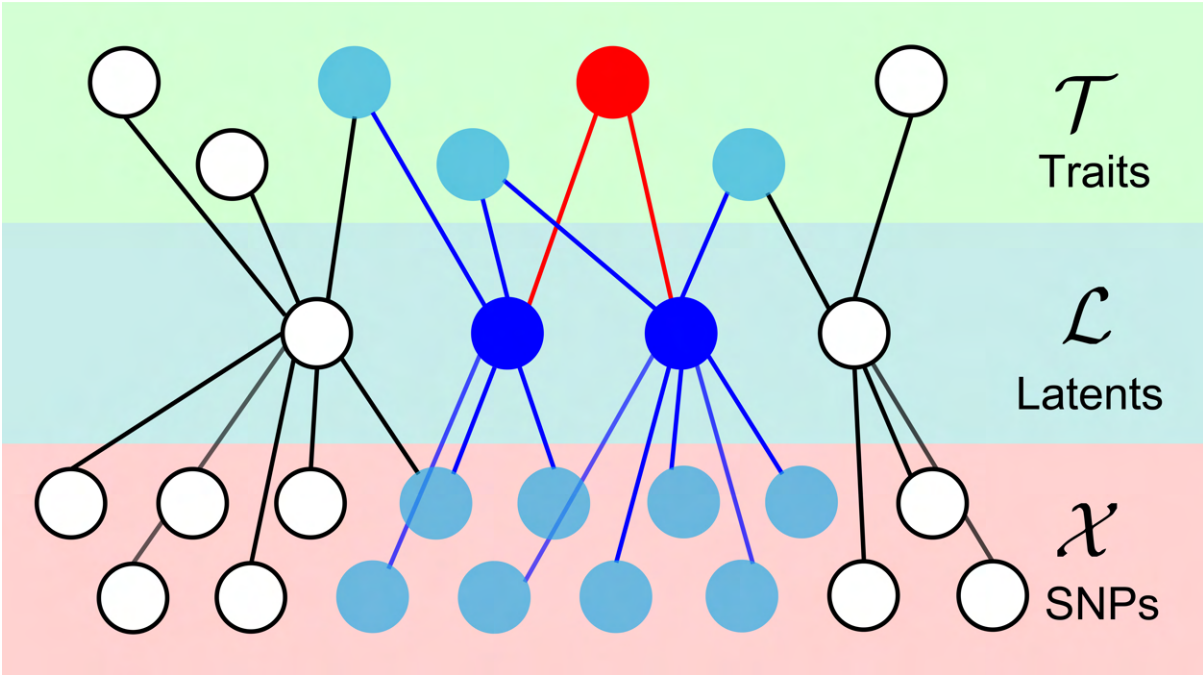


Figure 12: Given a trait of interest (red node), we determine the latent factors explaining the most variance (red edges). These latent factors (blue nodes) are in turn decomposed based on the highest contributing traits and SNPs (blue edges and cyan nodes).

WAPE for the odd chromosome partition (which includes the *APOE* gene on chromosome 19) was 6.5%, while for the even chromosome partition it was 24.0%. By contrast, for diabetes these were 11.7% and 10.6%, respectively.

3.7 Measuring SNP set overlap

In order to compare the percent overlap of two SNP sets, S_1 and S_2, we used the Jaccard index, defined as,

$$\text{jac}(S_1, S_2) = \frac{S_1 \cap S_2}{S_1 \cup S_2}. \tag{19}$$

This was used in constructing Fig. 8.

3.8 Model selection with GUIDE

The core idea is that for a mixture of signals that includes non-Gaussian components as well as Gaussian ones, ICA will find a rotation to reconstruct the non-Gaussian components, but will include some random rotation for the Gaussian ones [8, 9]. Given this, if we have a dataset described by L latent factors corrupted by some Gaussian noise, and in our initial run of GUIDE we pick some $K > L$, the resulting unmixing matrix will include some components needed to rotate the L-dimensional signal (and noise), and some random rotation for the remaining $K - L$ components (which only have noise). Now if we run ICA a second time, we get a different unmixing matrix. If we look at the cross-correlation of the two unmixing matrices, it will be a signed permutation for the parts that actually reconstruct the signal, but closer to 0 for the random rotation for the $K - L$ "extra" components we added. This works because FastICA, the implementation of ICA we're using, uses a random matrix to initialize the algorithm for finding the unmixing matrix [8, 9]. This gives a way to estimate the proper value of L to use in our model. To increase accuracy we run ICA a third time and compare the resulting unmixing matrices in the same way.

Acknowledgements

The authors would like to thank Luke O'Connor, Raymond Walters, Mark J. Daly, and Nikolas Baya for helpful discussions and for their comments on the draft, and Yosuke Tanigawa and Manual A. Rivas for freely sharing their data. This work is supported by the Novo Nordisk Foundation (NNF21SA0072102), and US National Institutes of Health grants U01MH115727, R01MH101244, and R37MH107649. This research has been conducted using the UK Biobank Resource under Application Number 31063.

4 Supplementary Information

4.1 Analysis of the errors in the GUIDE decomposition

Given a sum of statistically independent rank one matrices, can one recover the terms? How, and under what conditions? This decomposition is of the form

$$Z = \sum_{k=1}^{K} t_k x_k y_k' = XTY', \tag{20}$$

where Z is an $M \times N$ data matrix (with $\text{rank}(Z) = K$), X is an $M \times K$ matrix, T is a $K \times K$ diagonal matrix, Y is an $N \times K$ matrix, and x_k and y_k are the M- and N-dimensional column vectors of X and Y, respectively. The latter three matrices are the estimands in our problem. Note that since x_k and y_k are independent high dimensional vectors, they are approximately orthogonal, but not exactly orthogonal due to sampling variation.

As an initial decomposition, we find the compact singular value decomposition (SVD), given by

$$Z = USV', \tag{21}$$

where S is a $K \times K$ diagonal matrix consisting of the nonzero eigenvalues of ZZ', U is an $M \times K$ orthogonal matrix consisting of the eigenvectors of ZZ' corresponding to the eigenvalues in S, and V is an $N \times K$ orthogonal matrix consisting of the same for $Z'Z$.

Unlike the columns of X and Y, those of U and V are guaranteed to be exactly orthogonal by the SVD. If X and Y were orthogonal, then $X = U$ and $Y = V$, and the SVD would give the correct decomposition into statistically independent components.

Nevertheless, since $XTY' = USV'$, we can write

$$U = X \underbrace{TY'VS^{-1}}_{\mathcal{M}_X} \tag{22}$$

$$V = Y \underbrace{TX'US^{-1}}_{\mathcal{M}_Y} \tag{23}$$

where $\mathcal{M}_X$ and $\mathcal{M}_Y$ are the true mixing matrices. These two equations define two ICA problems. After running ICA on U and V, respectively, we obtain

$$U = \hat{X}\mathcal{M}_{\hat{X}} \tag{24}$$

$$V = \hat{Y}\mathcal{M}_{\hat{Y}} \tag{25}$$

where the hatted matrices denote the estimated quantities obtained for the corresponding estimands. Unlike the estimands, the estimators $\hat{X}$, $\hat{Y}$, $\mathcal{M}_{\hat{X}}$ and $\mathcal{M}_{\hat{Y}}$ have orthonormal columns by construction.

The purpose of our analysis in this section is to find the estimators for X, T, Y in (20) so that $\hat{X}\hat{T}\hat{Y}' \approx XTY'$. Two of these, $\hat{X}$ and $\hat{Y}$, are given by the ICA in (24) and (25). How do we estimate $\hat{T}$? By the above we have $USV' = XTY' \approx \hat{X}\hat{T}\hat{Y}'$. By (24) and (25) we obtain

$$\begin{aligned} USV' &= \hat{X}\mathcal{M}_{\hat{X}} S \mathcal{M}_{\hat{Y}}' \hat{Y}' \approx \hat{X}\hat{T}\hat{Y}' \\ \hat{T} &\approx \mathcal{M}_{\hat{X}} S \mathcal{M}_{\hat{Y}}'. \end{aligned} \tag{26}$$

A small error term, one for each of X and Y, and reflecting the effects of the sampling variation, allows us to relate each estimand to its estimator:

$$\begin{aligned} X &= \hat{X}(I_K + \mathcal{E}_X) \\ Y &= \hat{Y}(I_K + \mathcal{E}_Y) \end{aligned} \tag{27}$$

Using these relations and the fact that the estimated quantities $\hat{X}\hat{T}\hat{Y}' \approx XTY'$, we obtain a relation for $\hat{T}$ in terms of the errors and the true T:

$$\begin{aligned} XTY' &= \hat{X}(I_K + \mathcal{E}_X)T(I_K + \mathcal{E}_Y')\hat{Y}' \\ &= \hat{X}(T + \mathcal{E}_X T + T\mathcal{E}_Y' + \cancel{\mathcal{E}_X T \mathcal{E}_Y'})\hat{Y}' \approx \hat{X}\hat{T}\hat{Y}' \\ \Rightarrow \hat{T} &\approx T + \mathcal{E}_X T + T\mathcal{E}_Y'. \end{aligned} \tag{28}$$

Formula (26) provides a way to estimate $\hat{T}$, and (28) shows that its off-diagonal elements are only due to the sample variation since T is diagonal. Note that while the off-diagonal elements of $\hat{T}$ are purely due to the error terms ($\mathcal{E}_X$ and $\mathcal{E}_Y$), its diagonal entries, while containing the true values of T, are also affected by the same.

We can also find the error term relating $\mathcal{M}_{\hat{X}}$ to $\mathcal{M}_X$ in terms of $\mathcal{E}_X$. Write

$$\begin{aligned} \mathcal{M}_X &= \mathcal{M}_{\hat{X}}(I_K + \mathcal{E}_{\mathcal{M}_X}) \\ \mathcal{M}_Y &= \mathcal{M}_{\hat{Y}}(I_K + \mathcal{E}_{\mathcal{M}_Y}). \end{aligned} \tag{29}$$

We will only explicitly write the calculations for X, as those for Y are analogous. From (22), (27), and (29) we have

$$\begin{aligned} U = X\mathcal{M}_X &= \hat{X}(I_K + \mathcal{E}_X)\mathcal{M}_{\hat{X}}(I_K + \mathcal{E}_{\mathcal{M}_X}) \\ &= \hat{X}\mathcal{M}_{\hat{X}} + \hat{X}\mathcal{E}_X\mathcal{M}_{\hat{X}} + \hat{X}\mathcal{M}_{\hat{X}}\mathcal{E}_{\mathcal{M}_X} + \cancel{\hat{X}\mathcal{E}_X\mathcal{M}_{\hat{X}}\mathcal{E}_{\mathcal{M}_X}} \\ &\approx U + \hat{X}\mathcal{E}_X\mathcal{M}_{\hat{X}} + \hat{X}\mathcal{M}_{\hat{X}}\mathcal{E}_{\mathcal{M}_X} \end{aligned}$$

where we used (24) in the last step, and canceled the term that is second order in the error in the penultimate one. This gives an expression for the error term for the estimated mixing matrix in terms of the error for the estimated X:

$$\begin{aligned} \hat{X}\mathcal{M}_{\hat{X}}\mathcal{E}_{\mathcal{M}_X} &\approx -\hat{X}\mathcal{E}_X\mathcal{M}_{\hat{X}} \\ \mathcal{E}_{\mathcal{M}_X} &\approx -\mathcal{M}'_{\hat{X}}\mathcal{E}_X\mathcal{M}_{\hat{X}} \end{aligned} \tag{30}$$

Plugging (30) into (29) we obtain:

$$\begin{aligned} \mathcal{M}_X &\approx \mathcal{M}_{\hat{X}} - \mathcal{M}_{\hat{X}}\mathcal{M}'_{\hat{X}}\mathcal{E}_X\mathcal{M}_{\hat{X}} \\ \mathcal{M}_X &\approx (I_K - \mathcal{E}_X)\mathcal{M}_{\hat{X}} \end{aligned} \tag{31}$$

Using formulas (27) and (31) we can check that the approximations we made are justified by verifying that the two expressions for U in (22) and (24) match up to a negligible error term:

$$\begin{aligned} X\mathcal{M}_X &\approx \hat{X}(I_K + \mathcal{E}_X)(I_K - \mathcal{E}_X)\mathcal{M}_{\hat{X}} \\ &= \hat{X}(I_K - \mathcal{E}_X^2)\mathcal{M}_{\hat{X}}, \end{aligned}$$

so using the above approximations does indeed reproduce the relation $X\mathcal{M}_X = \hat{X}\mathcal{M}_{\hat{X}}$ up to $O(\mathcal{E}_X^2)$.

4.2 Kurtosis as a measure of performance

Table 3 and Fig. 13 compare the kurtosis for the latent factors produced by DeGAs versus that of GUIDE for the 'all' dataset. Kurtosis is a computationally favorable proxy measure for non-Gaussianity (or negentropy), which is the cost function maximized (or minimized) by ICA algorithms [8, 9]. As discussed in more depth in Section 4.4, this is one way of assessing the performance of the ICA step of our algorithm.

Table 3

Kurtosis ('all' dataset)	DeGAs	GUIDE
Mean	185.87	813.51
Median	116.63	685.93
Range	826.14	2108.86
Std Dev	184.04	635.65
Max	838.18	2126.20
Min	12.03	17.33

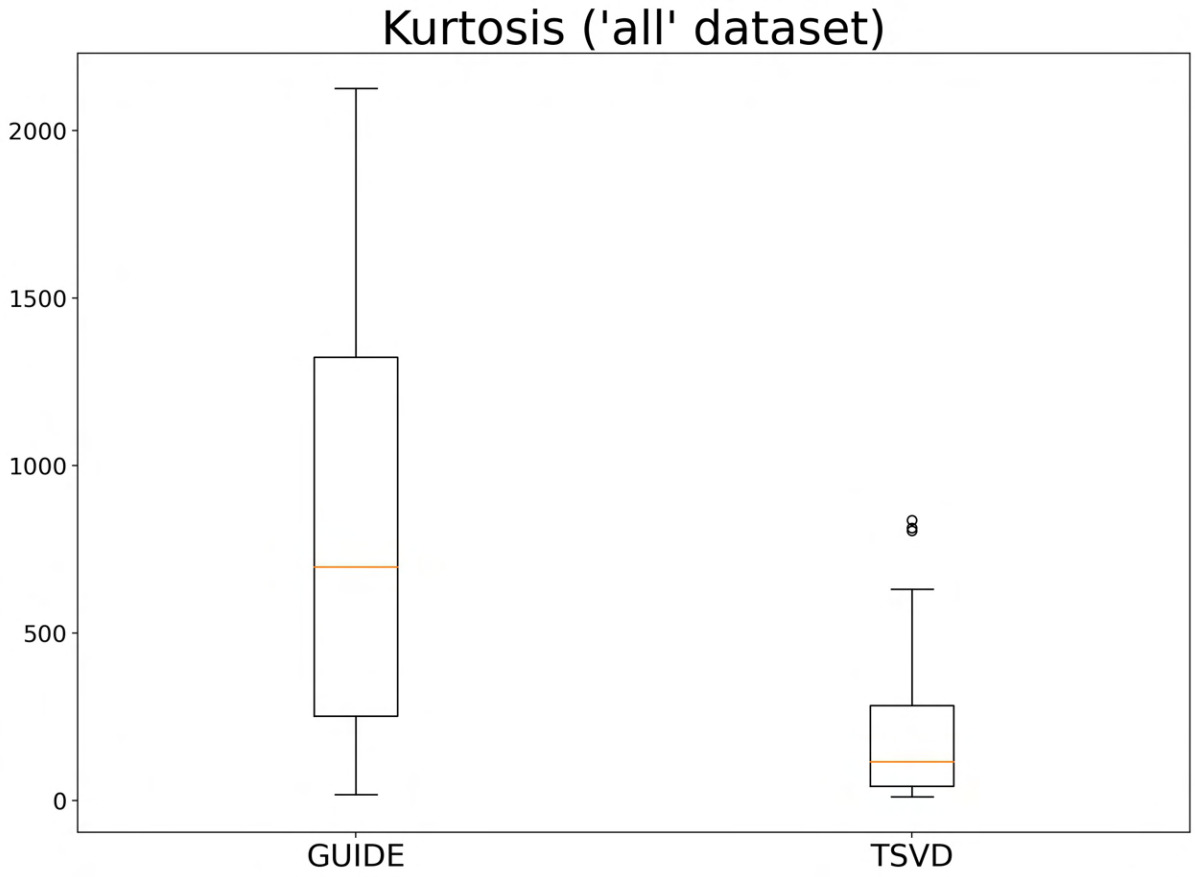


Figure 13

4.3 Independence of latent factors

As mentioned in the main text, ICA maximizes nongaussianity by maximizing negentropy, defined as the difference between the entropy of a Gaussian distribution (which is the maximum entropy among all distributions with equal variance) and the entropy of the given distribution. ICA can equally be interpreted as minimizing mutual information, given by

$$I(X,Y) = \sum_{x,y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right),$$

where $p(x,y)$ is the joint distribution for the random variables X, Y and $p(x), p(y)$ are the corresponding marginal distributions [9]. Intuitively, mutual information measures how far $p(x,y)$ is from being independent. Thus, in the context of GUIDE, ICA finds latents factors that mutually minimize mutual information, and are in this sense 'maximally independent'. Notably, this notion of independence is stronger than merely decorrelation of factors, since the former includes higher order (as opposed to only second-order) and nonlinear effects [50].

4.4 Heritability and trait enrichment of GUIDE factors

We prepared a set of minimally pruned UK Biobank data with 2,692,064 million SNPs and 4,357 phenotypes (see Methods for details), which we will refer to as the "UKB dataset". We then built a GUIDE model for this dataset with $L = 200$ (see Section 3.8 in Methods for a discussion on how this number was estimated using GUIDE's model selection algorithm).

Table 4 and Fig. 14 compare the heritabilities of the latent factors against those of the traits for the UKB dataset. The latent factors carry approximately the same total heritability as the traits, and so on average have a much higher heritability per trait (Methods).

Table 4: Statistics for the (normalized) heritabilities of the final traits versus the latent factors.

	Final trait h^2	Latent factor h^2
Mean	0.0002295157	0.0057803467
Median	9.5080×10^{-8}	3.2220×10^{-4}
Mode	9.5080×10^{-8}	2.7972×10^{-7}
Standard deviation	0.007022	0.01813
Maximum	0.26688	0.13398
Minimum	9.3536×10^{-8}	2.7972×10^{-7}

Fig. 15 shows the genetic genetic variance component values for the UKB $Z > 4$ dataset, for which close to 90% of GUIDE weights are larger than the corresponding TSVD weights (for the other 10% they are comparable, likely reflecting that this dataset does not include the necessary SNPs to inform the latent factors for those specific traits). Moreover, 29.8% of GUIDE latent weights have genetic genetic variance component scores above 0.9, as compared to 3.5% of TSVD latents factors.

4.5 Supplementary figures and tables

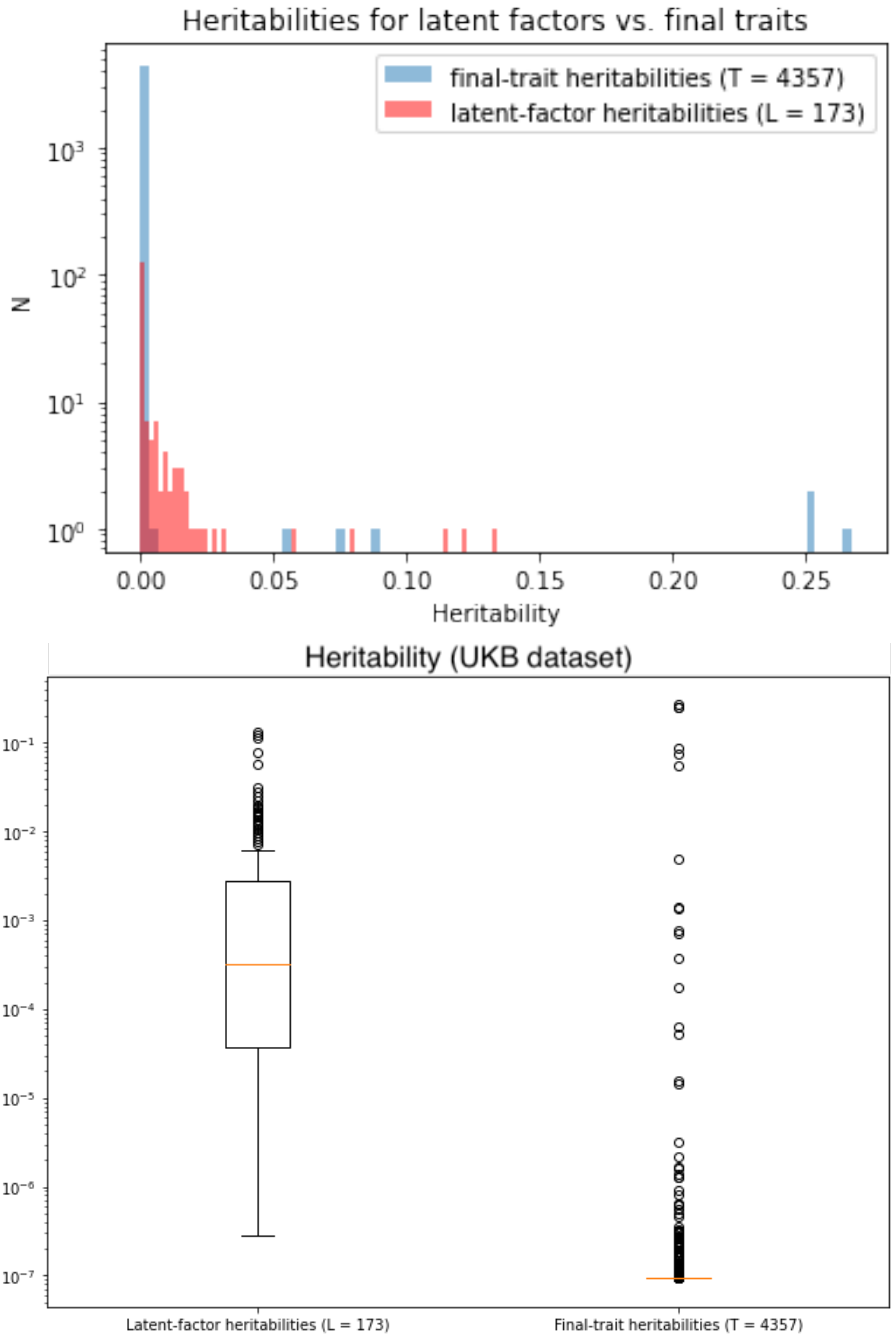


Figure 14: A comparison of the heritabilities for the final traits versus those for the latent factors, showing that latent factors capture more heritability on average, and moreover, despite being far fewer in number, capture the same total heritability for all traits.

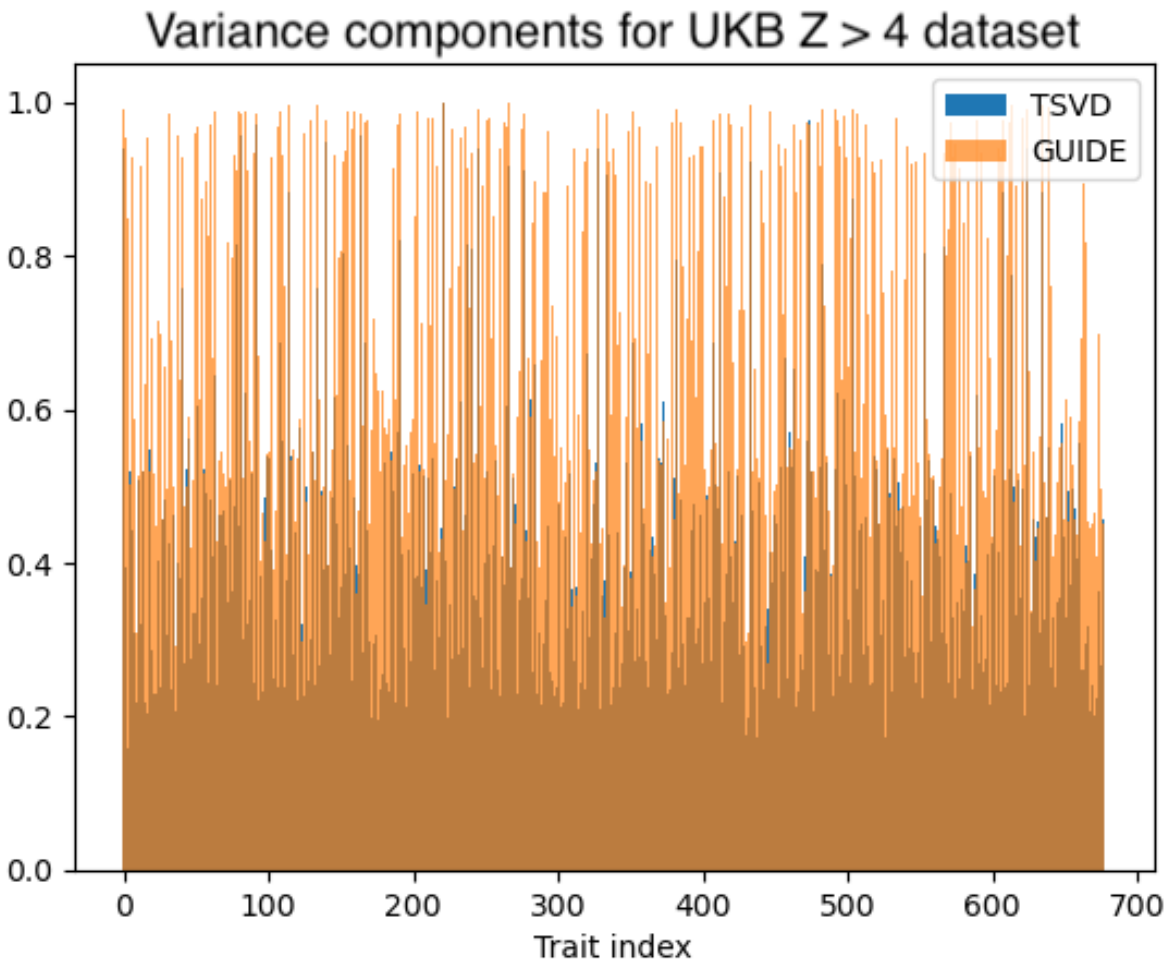


Figure 15: Here we plot, as in Fig. 3, the genetic variance components for traits in the 'all' dataset from [5] for the top three latent factors, but this time for all 2,138 traits. In this plot, the GUIDE latent factors were larger than their TSVD counterparts in 89.5% of cases.

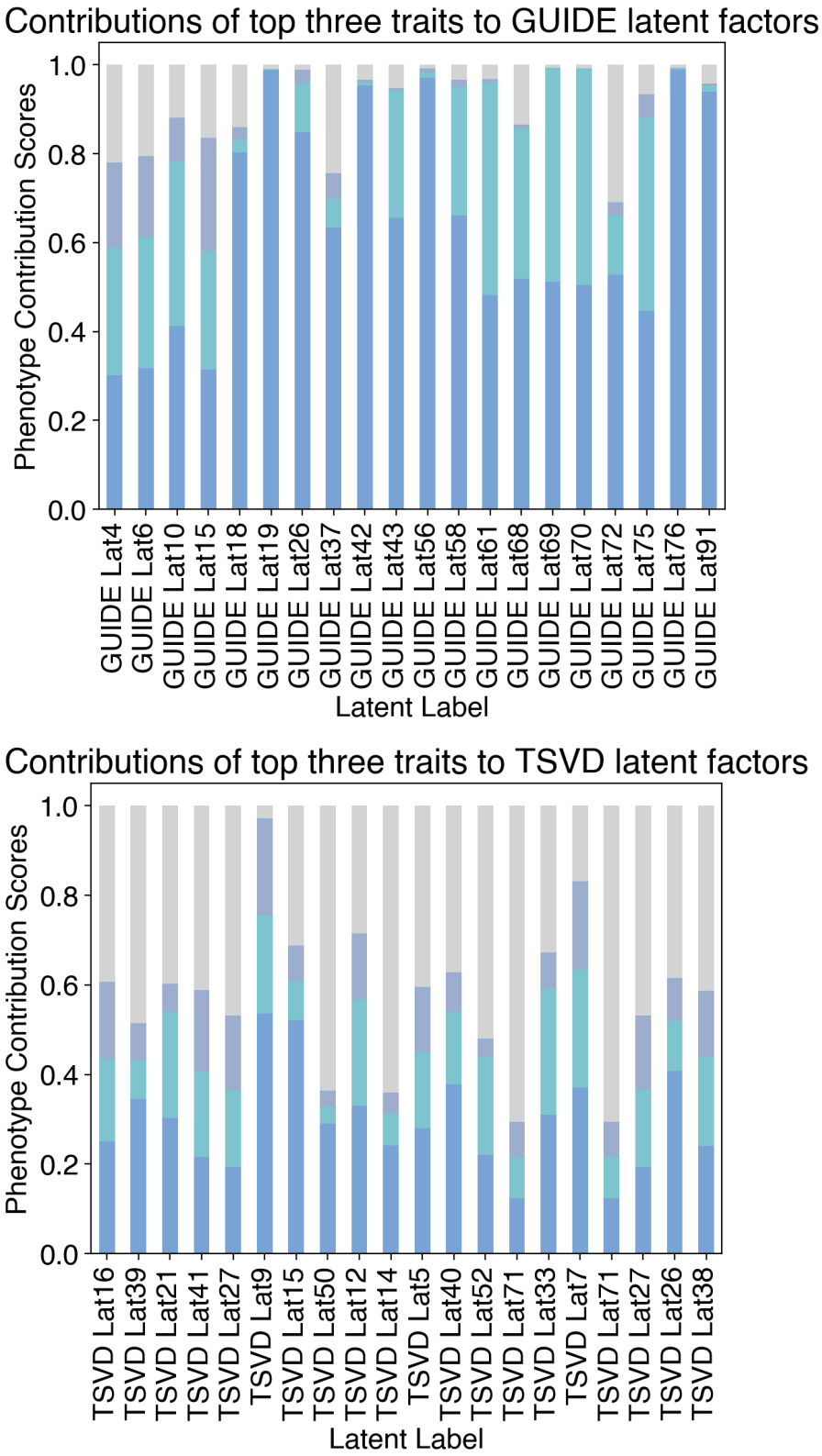


Figure 16: Stacked bar plots comparing, for each latent factor, the contributions from the top loading traits. See Fig. S22 for the variance explained curves for all 100 latent factors for GUIDE and for TSVD.

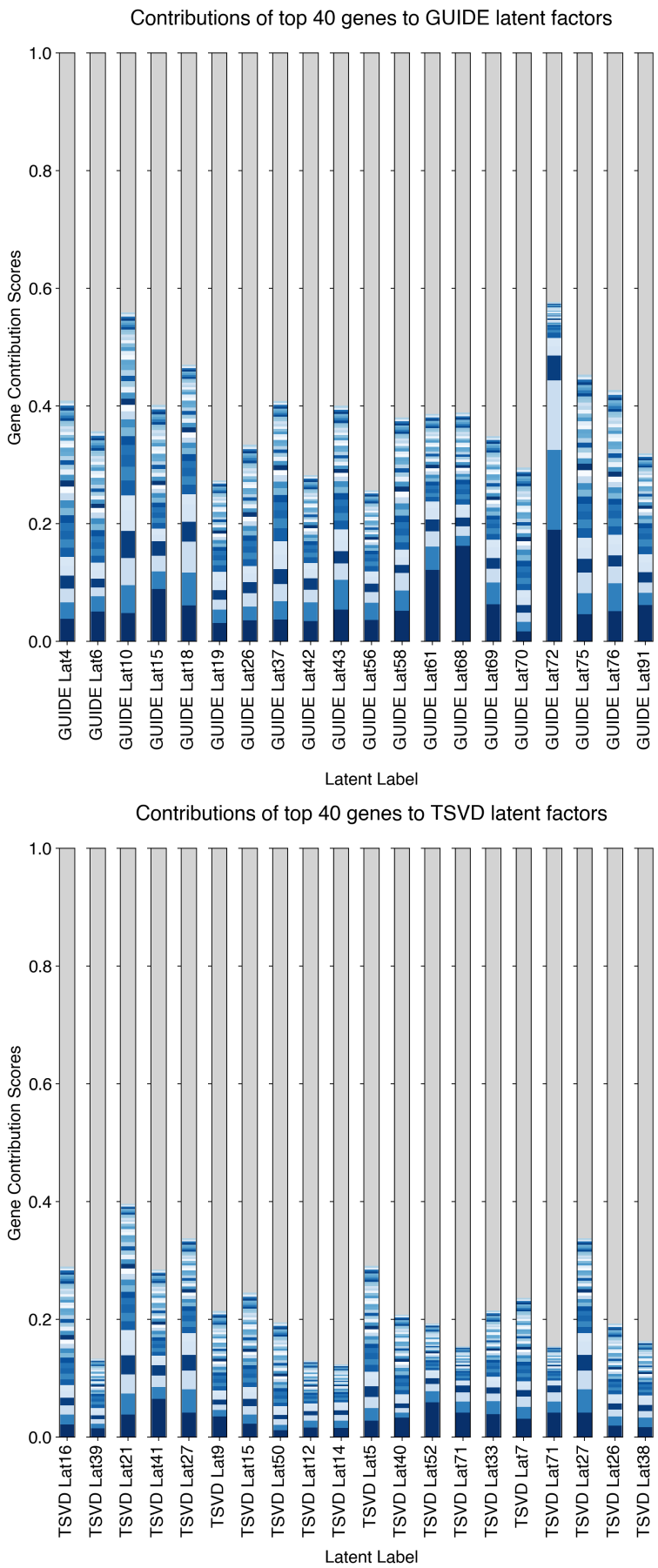


Figure 17: Stacked bar plots comparing, for each latent factor, the contributions from the top loading SNPs.

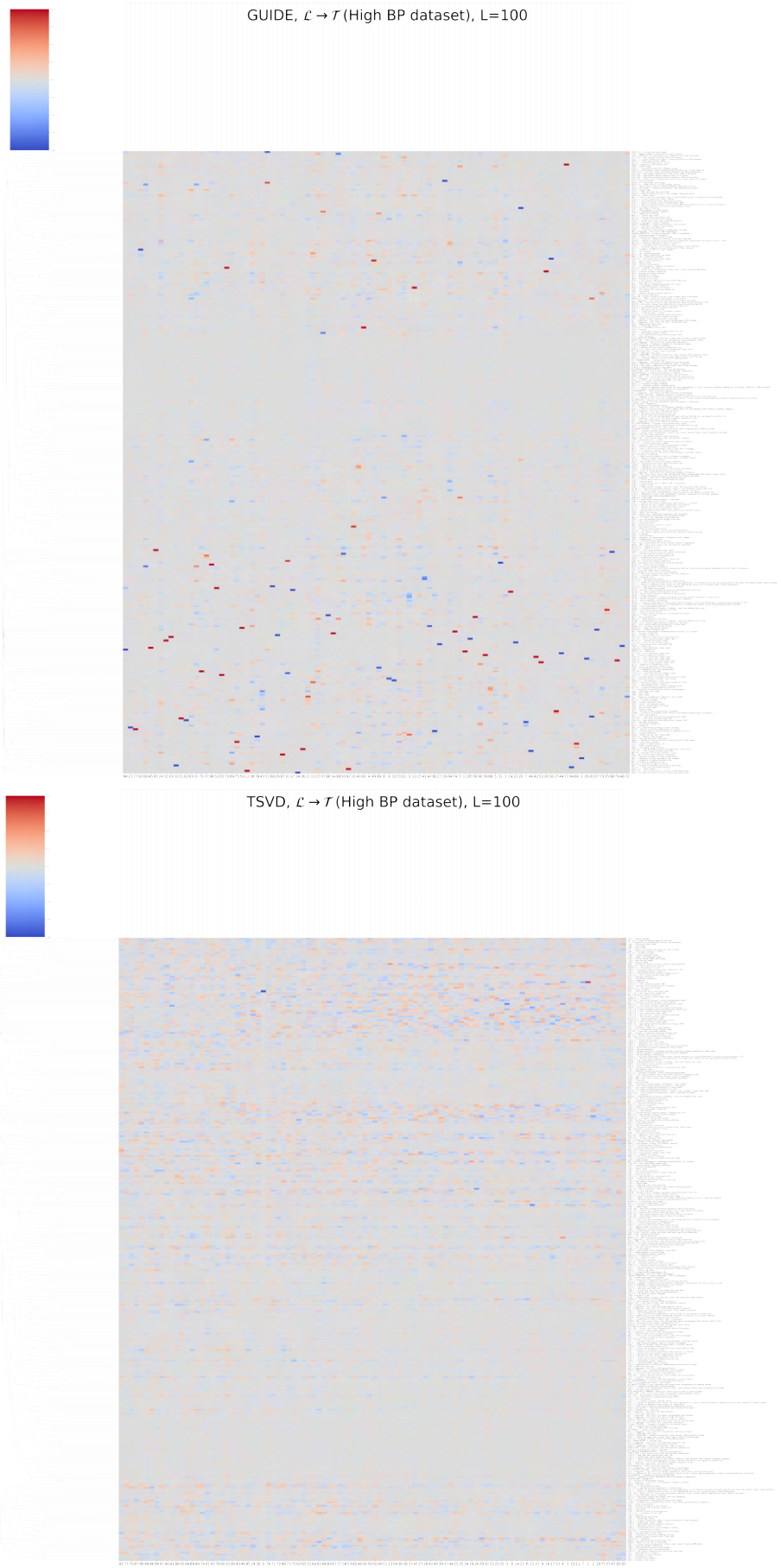


Figure 18: GUIDE $\mathcal{L} \to \mathcal{T}$ weights for loci significant for the "High Blood Pressure" dataset for for $L = 100$ compared with the corresponding TSVD $\mathcal{L} \to \mathcal{T}$ weights.

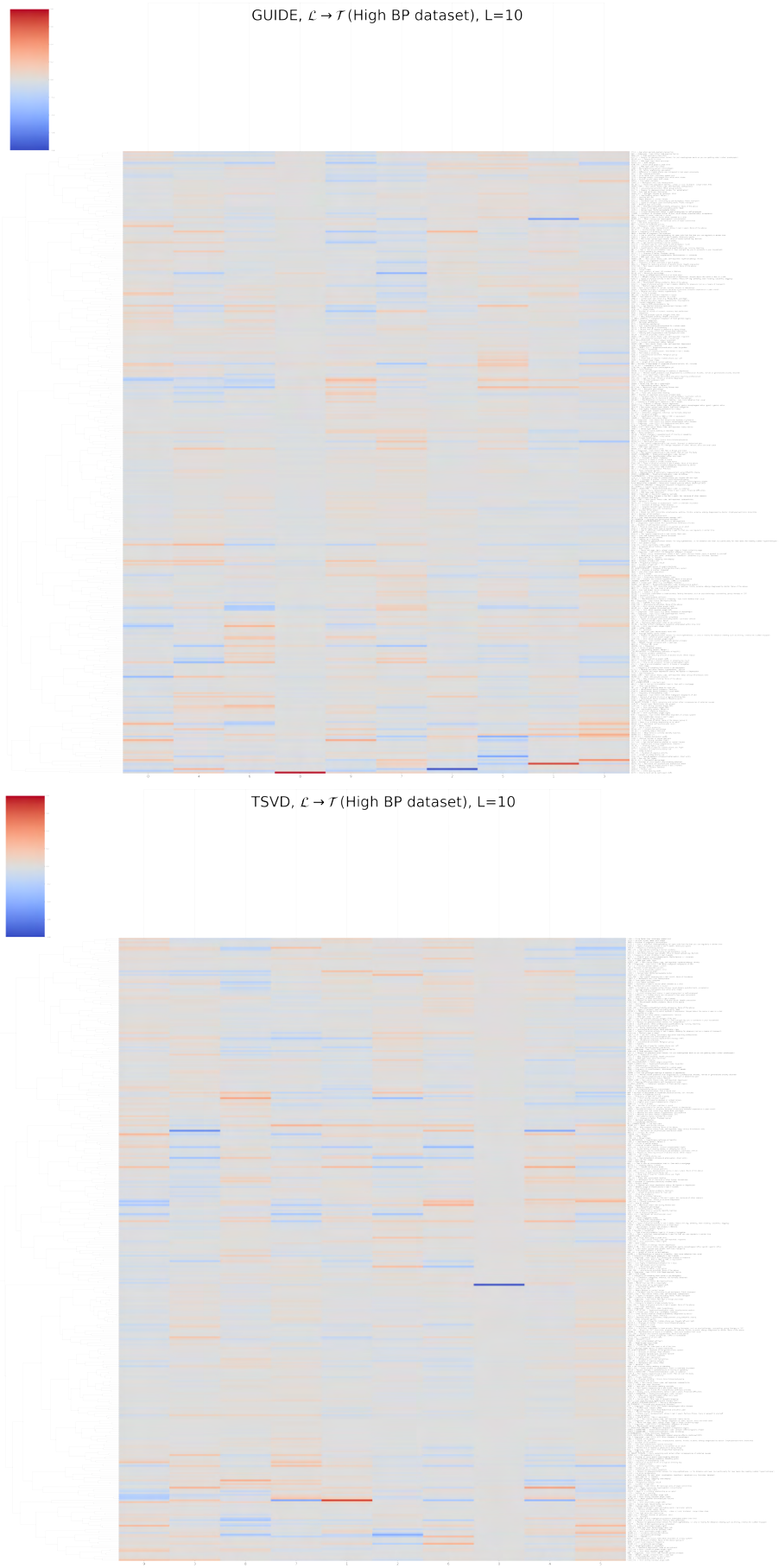


Figure 19: GUIDE $L \to T$ weights for loci significant for the "High Blood Pressure" dataset for for $L = 10$ compared with the corresponding TSVD $\mathcal{L} \to \mathcal{T}$ weights.

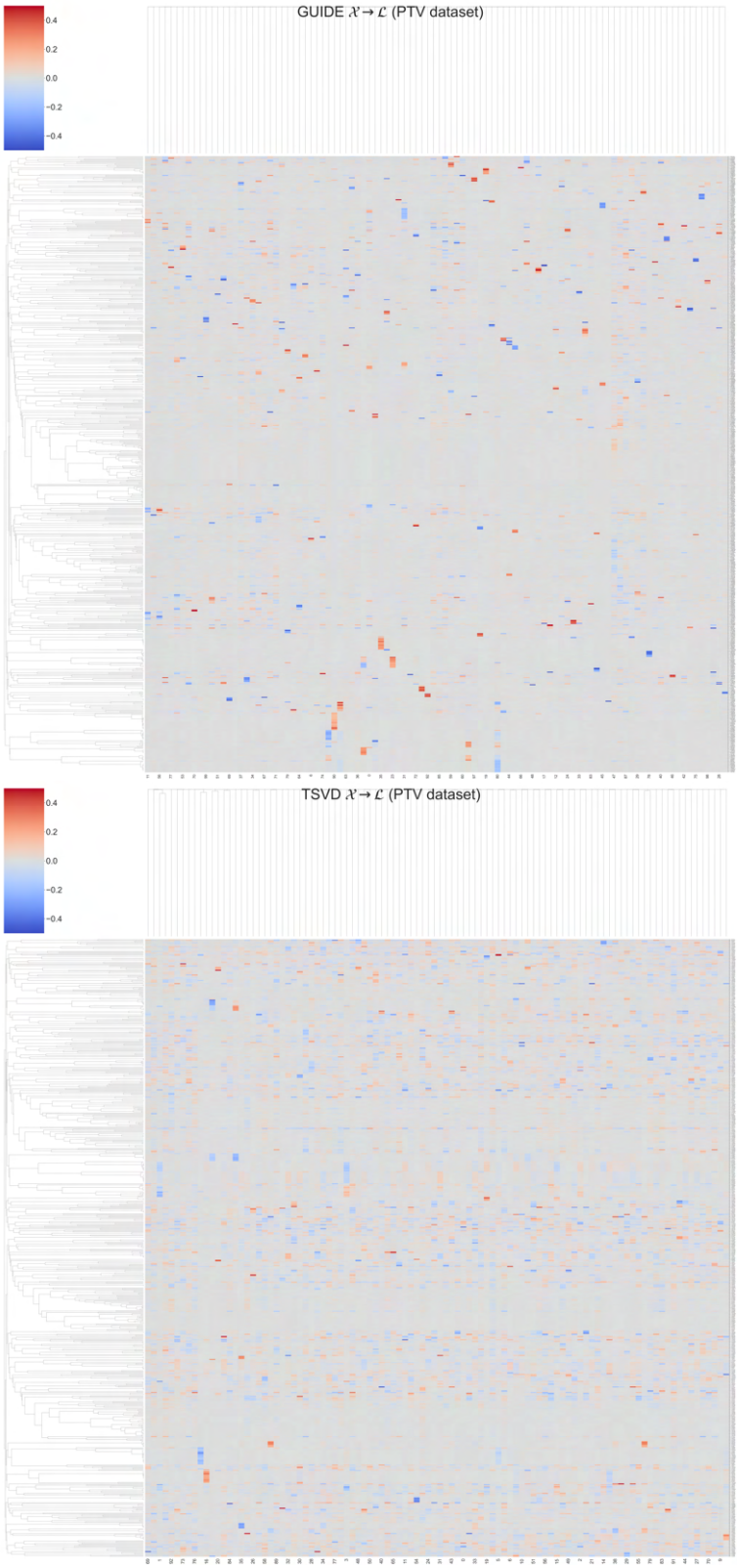

GUIDE $\mathcal{X} \to \mathcal{L}$ (PTV dataset)
0.4
0.2
0.0
−0.2
−0.4
TSVD $\mathcal{X} \to \mathcal{L}$ (PTV dataset)
0.4
0.2
0.0
−0.2
−0.4

Figure 20

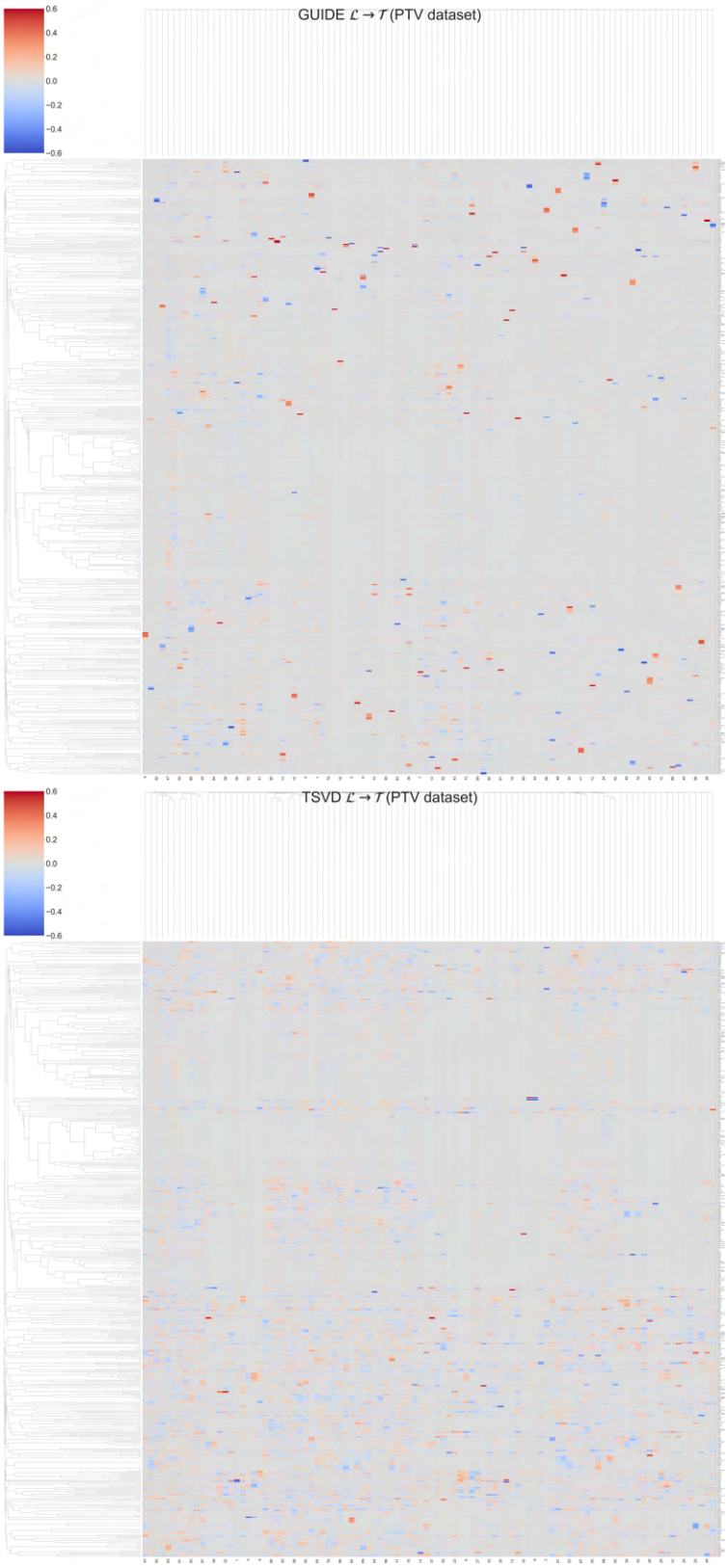

GUIDE $\mathcal{L} \to \mathcal{T}$ (PTV dataset)
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
TSVD $\mathcal{L} \to \mathcal{T}$ (PTV dataset)
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Figure 21

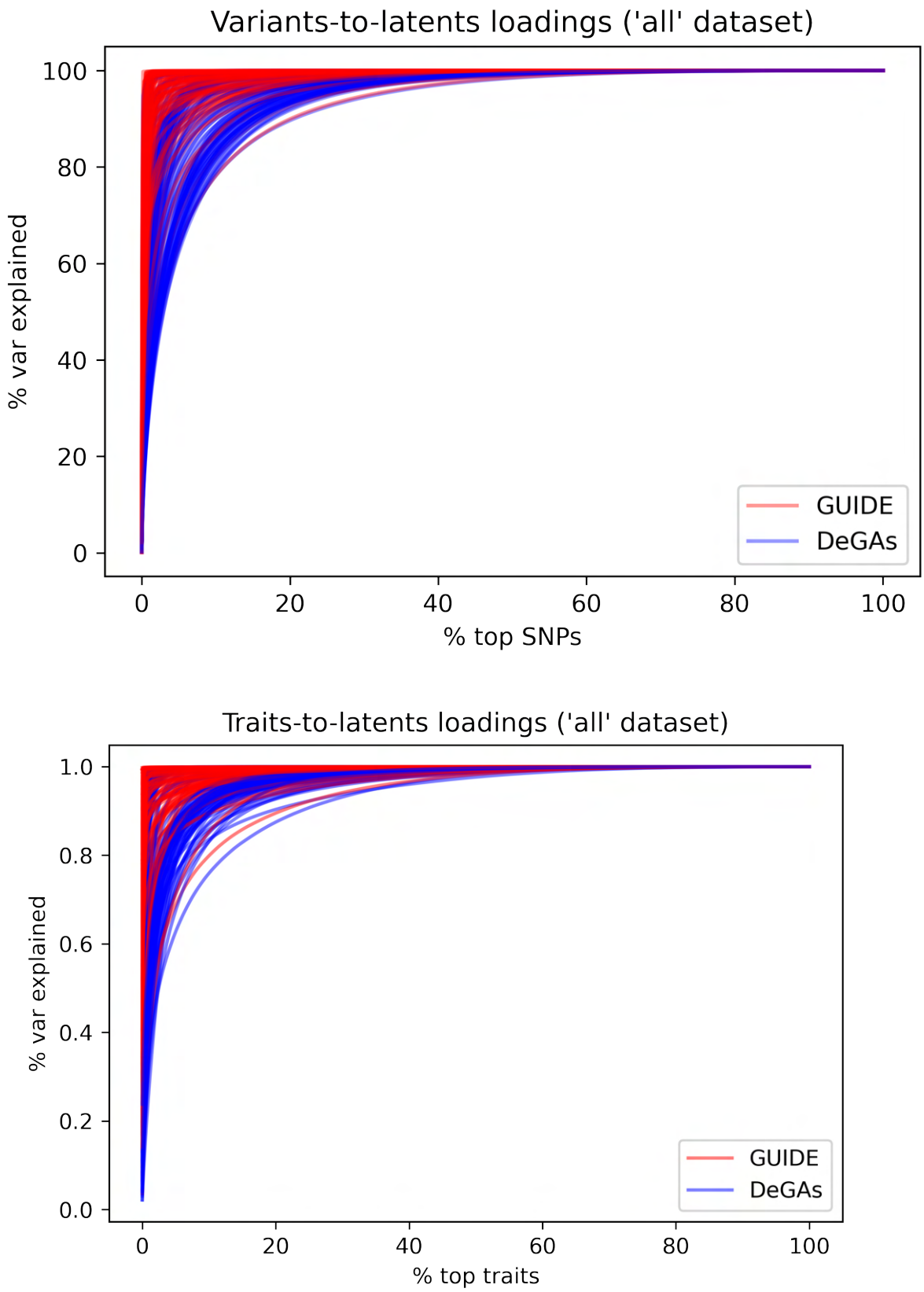


Figure 22: Variance explained plots comparing GUIDE with DeGAs as a function of the rank-ordered (based on $\mathcal{X} \to \mathcal{L}$ or $\mathcal{T} \to \mathcal{L}$ weights) SNPs or traits for every latent factor. Each line corresponds to a different latent factor. For GUIDE, most of the variance is explained by fewer SNPs and fewer traits for most of the latent factors. Results shown for the 'all' dataset [5] consisting of 235,907 variants and 2,138 traits, and both models were built using 100 latent factors.

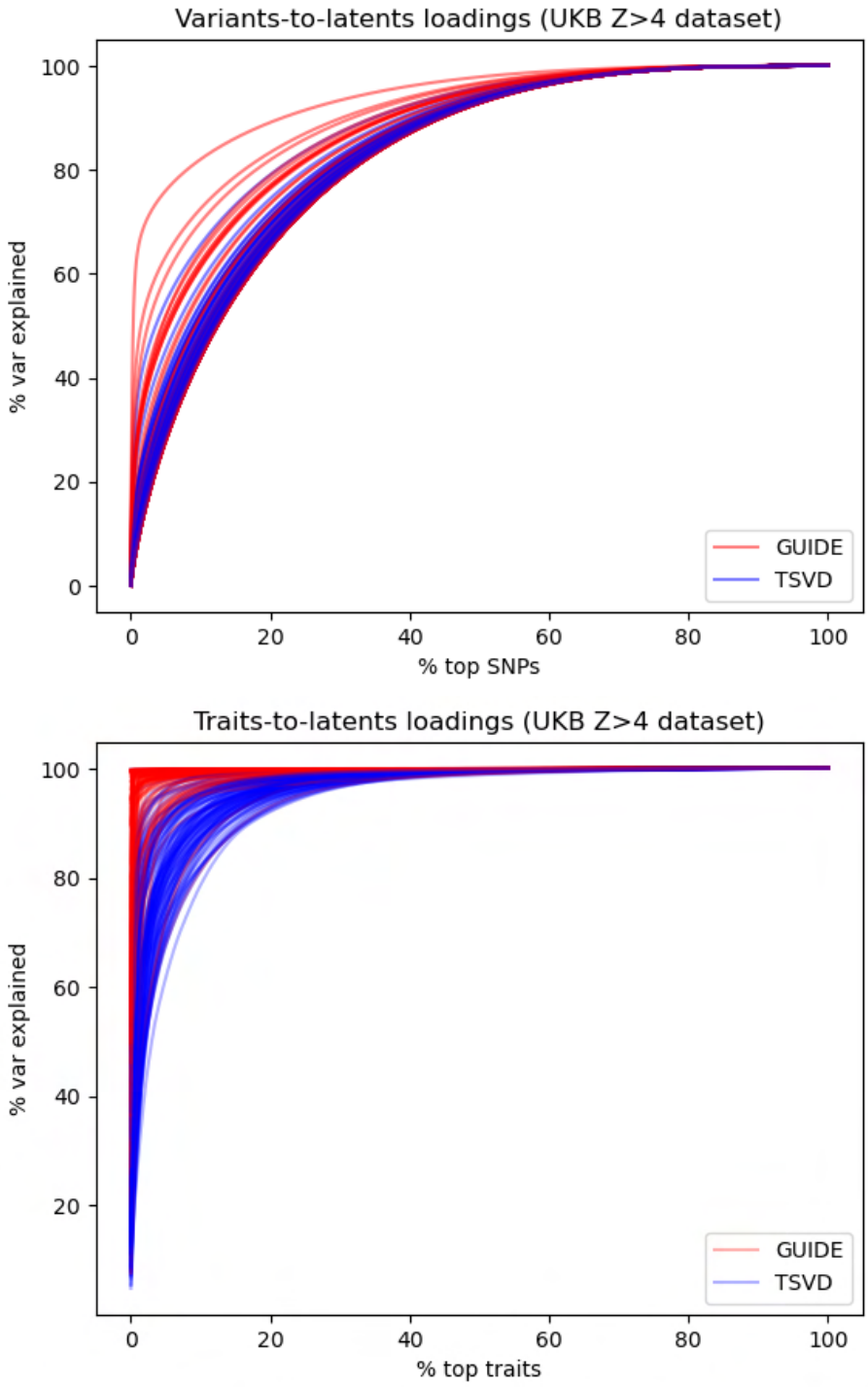

Variants-to-latents loadings (UKB Z>4 dataset)
% var explained
% top SNPs
GUIDE
TSVD
Traits-to-latents loadings (UKB Z>4 dataset)
% var explained
% top traits
GUIDE
TSVD

Figure 23

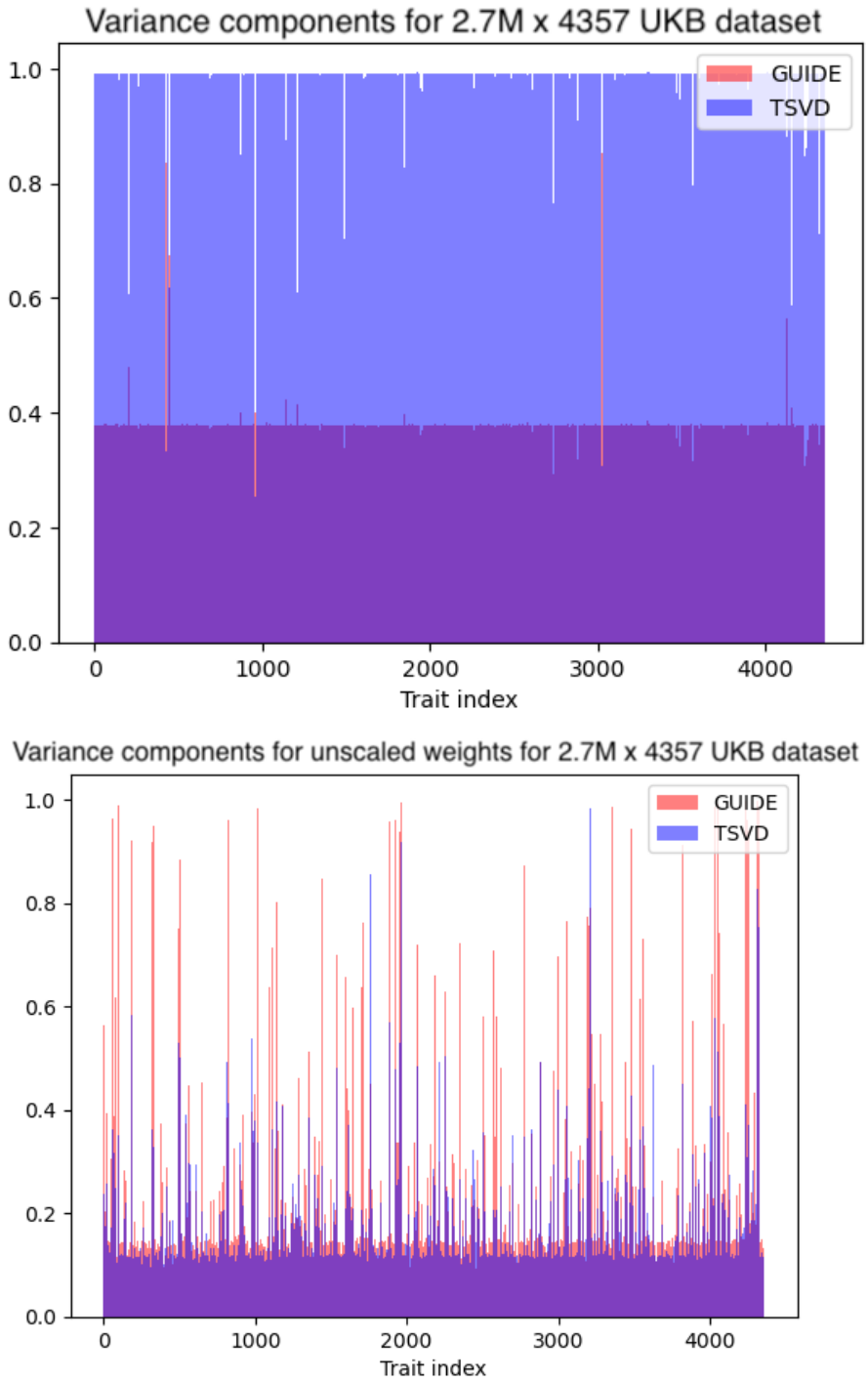


Figure 24: For the unpruned UK Biobank dataset, the TSVD latent factors have genetic variance components that are higher than their GUIDE counterparts in 98.9% of cases, unlike the results for $Z > 4$ dataset in Fig. 15. As discussed in the main text, this is likely due to LD and genetic covariance, which future work will hopefully address. Unsurprisingly, when we calculate the same plot but using the variance contribution score that is unweighted by the singular values (that is, V instead of VS and VA instead of VSA in Equations 17 and 18), the opposite is the case, with the GUIDE factors outweighting the TSVD ones 85.1% of the time. This is expected since the ICA basis was found for the unscaled data.

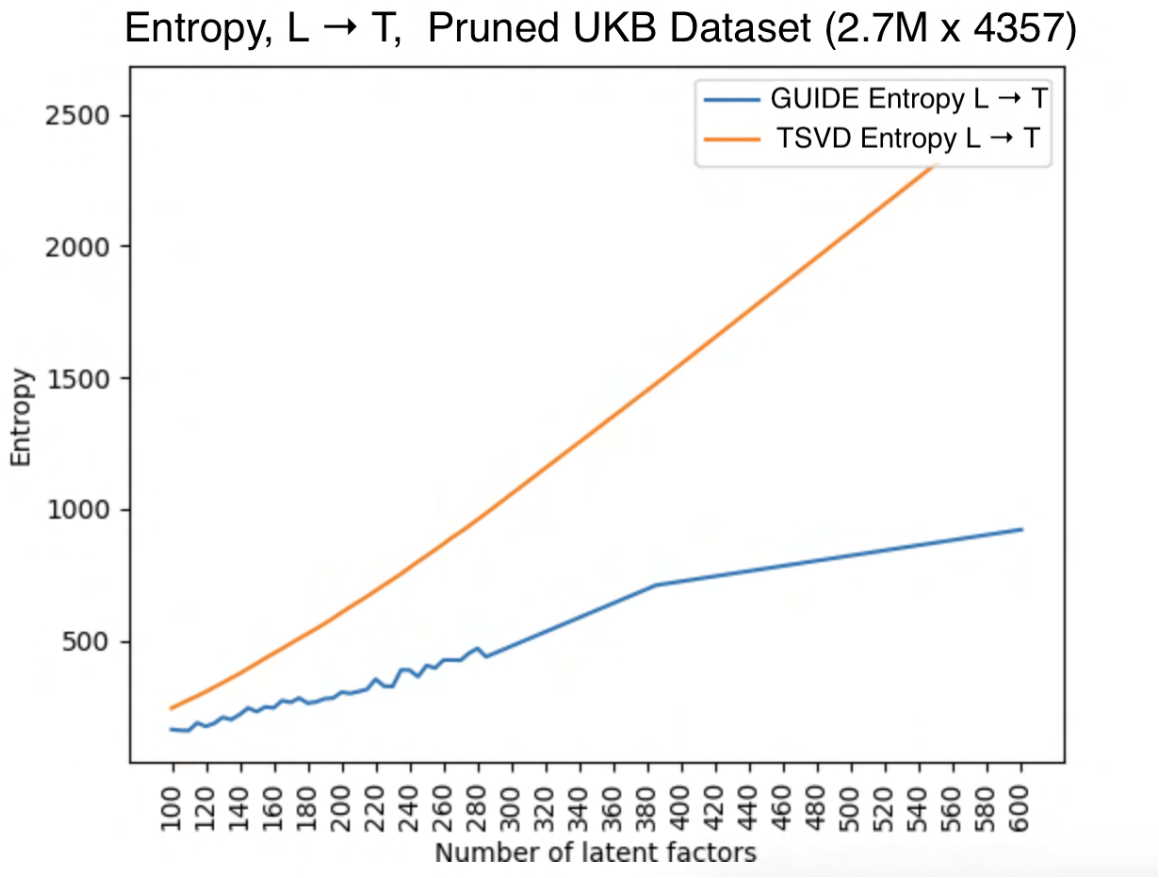


Figure 25: The entropy of the $\mathcal{L} \to \mathcal{T}$ weights as a function of the number of latent factors in the each model for the full UKB dataset (2,692,064 million SNPs and 4,357 phenotypes).

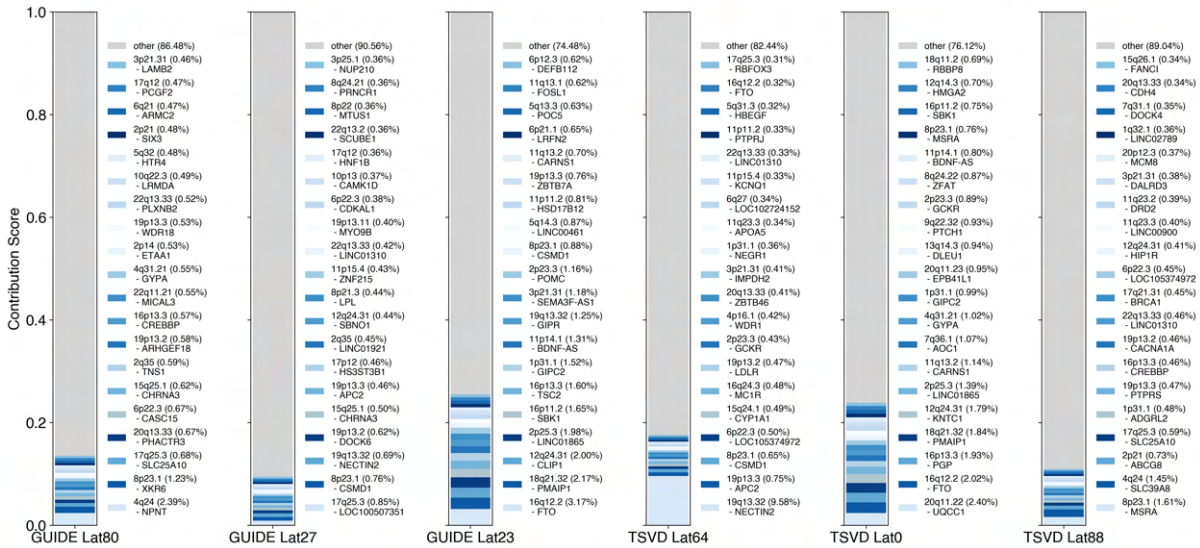


Figure 26: Top loci implicated by the leading GUIDE and TSVD latent factors for the traits 'Number of cigarettes smoked daily' and 'Number of cigarettes previously smoked daily'.

Table 5: A group of 8 traits with the corresponding number of variants that are significant for each trait ($p < 5 \times 10^{-8}$) and the number of traits that are correlated to each trait with $r^2 > 0.5$

Variants Significant For	No. Variants	No. Traits
BMI	2123	335
Ever Smoked	137	212
Standing Height	12163	414
Diabetes (diagnosed by doctor)	1007	194
Alcohol Intake Frequency	520	121
High Blood Pressure	1183	344
Miserableness	403	66
Seen Psychiatrist for Nerves, Anxiety, Tension or Depression	376	60

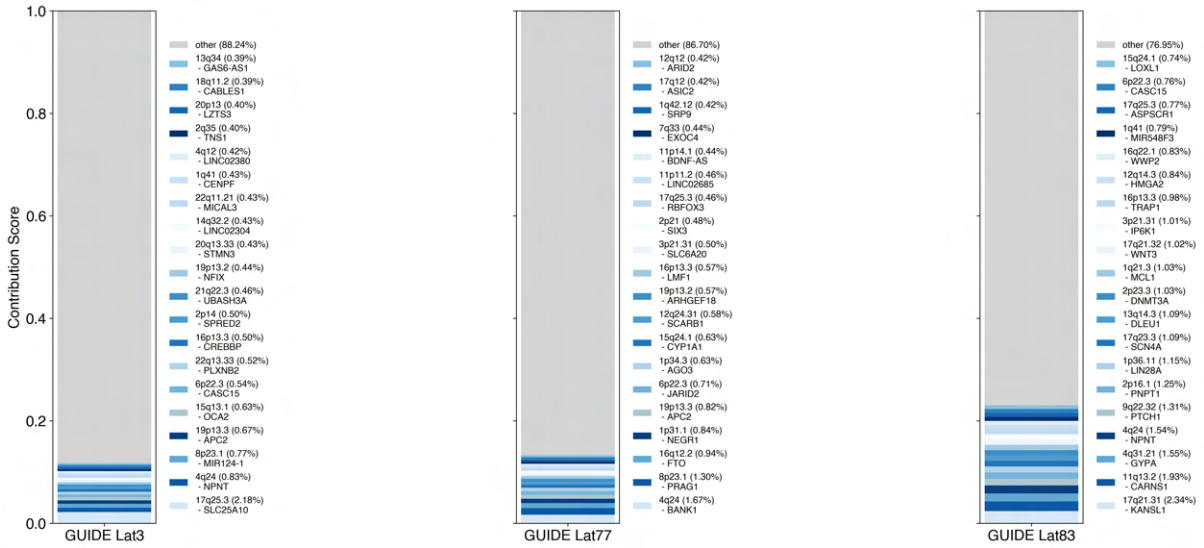


Figure 27: GUIDE top latent factors for the 'all' dataset without the smoking-related traits. GUIDE latent factors for the with- and without-smoking models were matched using the top loading traits for each factor, which was consistent between the top latent factors for each model.

References

[1] O'Connor, L. J. The distribution of common-variant effect sizes. *Nature Genetics* **53**, 1243–1249 (2021). URL `https://doi.org/10.1038/s41588-021-00901-3`.

[2] Grotzinger, A. D. *et al.* Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature Human Behaviour* **3**, 513–525 (2019).

[3] Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology* **17**, 132 (2016).

[4] Ballard, J. L. & O'Connor, L. J. Shared components of heritability across genetically correlated traits. *The American Journal of Human Genetics* **109**, 989–1006 (2022).

[5] Tanigawa, Y. *et al.* Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight adipocyte biology. *Nature Communications* **10** (2019).

[6] Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nature Genetics* **53**, 1415–1424 (2021). URL `https://doi.org/10.1038/s41588-021-00931-x`.

[7] Aguirre, M. *et al.* Polygenic risk modeling with latent trait-related genetic components. *European Journal of Human Genetics* **29**, 1071–1081 (2021). URL `https://doi.org/10.1038/s41431-021-00813-0`.

[8] Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. *Neural Networks* **13**, 411–430 (2000).

[9] Hyvärinen, A., Karhunen, J. & Oja, E. *Independent Component Analysis* (Wiley-Interscience, 2001).

[10] Hyvärinen, A. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing Systems*, vol. 10, 273–279 (MIT Press, 1998).

[11] Hyvärinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks* **10**, 626–634 (1999).

[12] Yao, F., Coquery, J. & LêCao, K.-A. Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets. *BMC Bioinformatics* **13**, 24 (2012).

[13] Oblong, L. M. *et al.* Principal and independent genomic components of brain structure and function. *Genes, Brain and Behavior* **23**, e12876 (2024).

[14] Borsboom, D. & Cramer, A. O. Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology* **9**, 91–121 (2013). PMID: 23537483.

[15] Boschloo, L. *et al.* The network structure of symptoms of the diagnostic and statistical manual of mental disorders. *PLoS One* **10**, e0137621 (2015).

[16] Varma, V. R. *et al.* Bile acid synthesis, modulation, and dementia: A metabolomic, transcriptomic, and pharmacoepidemiologic study. *PLoS Med* **18**, e1003615 (2021).

[17] Ferguson, E. L. *et al.* Low- and high-density lipoprotein cholesterol and dementia risk over 17 years of follow-up among members of a large health care plan. *Neurology* (2023).

[18] Cover, T. & Thomas, J. *Elements of Information Theory* (John Wiley, 1991).

[19] Artstein, S., Ball, K. M., Barthe, F. & Naor, A. Solution of Shannon's problem on the monotonicity of entropy. *J. Amer. Math. Soc.* **17**, 975–982 (2004).

[20] McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology* **28**, 495–501 (2010). URL `https://doi.org/10.1038/nbt.1630`.

[21] Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164–e164 (2010).

[22] Shao, Y. *et al.* DNA methylation of *TOMM40-APOE-APOC2* in Alzheimer's disease. *Journal of Human Genetics* **63**, 459–471 (2018).

[23] Raghavan, N. S. *et al.* Association Between Common Variants in *RBFOX1*, an RNA-Binding Protein, and Brain Amyloidosis in Early and Preclinical Alzheimer Disease. *JAMA Neurology* **77**, 1288–1298 (2020).

[24] Stepanov, V., Marusin, A., Vagaitseva, K., Bocharova, A. & Makeeva, O. Genetic Variants in *CSMD1* Gene Are Associated with Cognitive Performance in Normal Elderly Population. *Genetics Research International* **2017**, 6293826 (2017).

[25] Lee, J. H., Cheng, R., Graff-Radford, N., Foroud, T. & Mayeux, R. Analyses of the National Institute on Aging Late-Onset Alzheimer's Disease Family Study: implication of additional loci. *Arch. Neurol.* **65**, 1518–1526 (2008).

[26] Picard, C. *et al.* Proprotein convertase subtilisin/kexin type 9 (*PCSK9*) in Alzheimer's disease: A genetic and proteomic multi-cohort study. *PLOS ONE* **14**, e0220254– (2019). URL `https://doi.org/10.1371/journal.pone.0220254`.

[27] Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nature Genetics* **54**, 412–436 (2022).

[28] Loescher, C. M., Hobbach, A. J. & Linke, W. A. Titin (TTN): from molecule to modifications, mechanics, and medical significance. *Cardiovascular Research* **118**, 2903–2918 (2021). URL `https://doi.org/10.1093/cvr/cvab328`. `https://academic.oup.com/cardiovascres/article-pdf/118/14/2903/46948484/cvab328.pdf`.

[29] Huang, G.-Y. *et al.* Evaluating the role of connexin43 in congenital heart disease: Screening for mutations in patients with outflow tract anomalies and the analysis of knock-in mouse models. *Journal of Cardiovascular Disease Research* **2**, 206–212 (2011). URL `https://www.sciencedirect.com/science/article/pii/S0975358311240038`.

[30] Li, W. *et al.* SCN5A variants: Association with cardiac disorders. *Frontiers in Physiology* **9** (2018). URL `https://www.frontiersin.org/articles/10.3389/fphys.2018.01372`.

[31] Montenont, E. *et al.* Platelet WDR1 suppresses platelet activity and is associated with cardiovascular disease. *Blood* **128**, 2033–2042 (2016).

[32] de Frutos, F. *et al.* Natural History of *MYH7*-Related Dilated Cardiomyopathy. *Journal of the American College of Cardiology* **80**, 1447–1461 (2022).

[33] Lee, H.-J. *et al.* Gene Silencing and Haploinsufficiency of Csk Increase Blood Pressure. *PLOS ONE* **11**, 1–15 (2016). URL `https://doi.org/10.1371/journal.pone.0146841`.

[34] Seong, E. *et al.* Activation of Aryl Hydrocarbon Receptor by ITE Improves Cardiac Function in Mice After Myocardial Infarction. *Journal of the American Heart Association* **10**, e020502 (2021).

[35] Karle, W. *et al.* Promyelocytic leukemia protein promotes the phenotypic switch of smooth muscle cells in atherosclerotic plaques of human coronary arteries. *Clinical Science* **135**, 887–905 (2021).

[36] Queme, L. F., Weyler, A. A., Cohen, E. R., Hudgins, R. C. & Jankowski, M. P. A dual role for peripheral GDNF signaling in nociception and cardiovascular reflexes in the mouse. *Proceedings of the National Academy of Sciences* **117**, 698–707 (2020).

[37] Kim, K.-H., Rosen, A., Bruneau, B. G., chung Hui, C. & Backx, P. H. *Iroquois* Homeodomain Transcription Factors in Heart Development and Function. *Circulation Research* **110**, 1513–1524 (2012).

[38] Kurki, M. I. *et al.* Finngen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).

[39] Emanuele, E. *et al.* Chromosome 9p21.3 genotype is associated with vascular dementia and alzheimer's disease. *Neurobiol Aging* **32**, 1231–1235 (2011).

[40] Wain, L. V. *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir Med* **3**, 769–781 (2015).

[41] Wyss, A. B. *et al.* Multiethnic meta-analysis identifies ancestry-specific and cross-ancestry loci for pulmonary function. *Nat Commun* **9**, 2976 (2018).

[42] Ware, J. J., van den Bree, M. & Munafò, M. R. From Men to Mice: CHRNA5/CHRNA3, Smoking Behavior and Disease. *Nicotine & Tobacco Research* **14**, 1291–1299 (2012).

[43] Alsayegh, K. N., Sheridan, S. D., Iyer, S. & Rao, R. R. Knockdown of CDK2AP1 in human embryonic stem cells reduces the threshold of differentiation. *PLoS ONE* **13**, e0196817 (2018).

[44] Bossé, Y. *et al.* Transcriptome-wide association study reveals candidate causal genes for lung cancer. *Int J Cancer* **146**, 1862–1878 (2020).

[45] Li, C., Meng, J. & Zhang, T. Ncaph is a prognostic biomarker and associated with immune infiltrates in lung adenocarcinoma. *Sci Rep* **12**, 9578 (2022).

[46] Artigas, M. S. *et al.* Sixteen new lung function signals identified through 1000 genomes project reference panel imputation. *Nature Communications* **6**, 8658 (2015).

[47] Smolonska, J. *et al.* Common genes underlying asthma and copd? genome-wide analysis on the dutch hypothesis. *European Respiratory Journal* **44**, 860–872 (2014).

[48] Sakornsakolpat, P. *et al.* Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat Genet* **51**, 494–505 (2019).

[49] DeMeo, B. & Berger, B. SCA: recovering single-cell heterogeneity through information-based dimensionality reduction. *Genome Biology* **24**, 195 (2023). URL `https://doi.org/10.1186/s13059-023-02998-7`.

[50] Cardoso, J.-F. Dependence, Correlation and Gaussianity in Independent Component Analysis. *Journal of Machine Learning Research* **4**, 1177–1203 (2003).