CONVEX APPROACHES TO ISOLATE THE SHARED AND DISTINCT GENETIC STRUCTURES OF SUBPHENOTYPES IN HETEROGENEOUS COMPLEX TRAITS

Saikat Banerjee[1†] **Shane O'Connell**[2†] **Sarah M.C. Colbert**[2] **Niamh Mullins**[2] **David A. Knowles**[1,3]
[1]New York Genome Center, NY 10013, USA
{sbanerjee, daknowles}@nygenome.org
[2]Department of Psychiatry, Department of Genetics and Genomic Sciences, and Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, NY 10029, USA
{shane.oconnell, niamh.mullins,}@mssm.edu; sarah.colbert@icahn.mssm.edu
[3]Department of Computer Science, Columbia University, NY 10027, USA
dak2173@columbia.edu
[†]contributed equally

ABSTRACT

Groups of complex diseases, such as coronary heart diseases, neuropsychiatric disorders, and cancers, often display overlapping clinical symptoms and pharmacological treatments. The shared associations of genetic variants across diseases has the potential to explain their underlying biological processes, but this remains poorly understood. To address this, we model the matrix of summary statistics of trait-associated genetic variants as the sum of a low-rank component – representing shared biological processes – and a sparse component, representing unique processes and arbitrarily corrupted or contaminated components. We introduce `Clorinn`, an open-source Python software that uses convex optimization algorithms to recover these components by minimizing a weighted combination of the nuclear norm and of the L1 norm. Among others, `Clorinn` provides two significant benefits: (a) Convex optimization guarantees reproducibility of the components, and (b) The low-rank "uncorrupted" matrix allows robust singular value decomposition (SVD) and principal component analysis (PCA), which are otherwise highly sensitive to outliers and noise in the input matrix. In extensive simulations, we observe that `Clorinn` outperforms state-of-the-art approaches in capturing the shared latent factors across phenotypes. We apply `Clorinn` to estimate 200 latent factors from GWAS summary data of 2,110 phenotypes measured in European-ancestry Pan-UK BioBank individuals ($N = 420{,}531$) and 14 psychiatric disorders.

1 Introduction

Complex diseases often present as heterogeneous groups, encompassing multiple subtypes with distinct clinical manifestations, prognoses, and treatment responses. Such heterogeneous groups of diseases, such as coronary heart diseases, neuropsychiatric disorders and cancers, represent a broad continuum of severity and range of manifestations of several clinical symptoms. For instance, several clinical symptom-based subtypes have been identified using data-driven methods in major depressive disorder (MDD), one of the most common disorders worldwide. Within MDD, almost 1500 symptom combinations are possible [1]. Despite phenotypic diversity, these subtypes may share overlapping genetic and environmental risk factors. For instance, genetic and neuroimaging studies demonstrate that MDD is a genetically and biologically heterogeneous disorder involving multiple genes and brain circuits [2]. Sex and age contribute to the heterogeneity of MDD and are associated with differing symptomatology, risk factors, and responses to treatment [3–9]. Understanding the genetic predisposition of such heterogeneous disorders is crucial for elucidating the biological mechanisms driving both shared susceptibility and subtype-specific divergence.

Genome-wide association studies (GWAS) have estimated the effects of millions of single-nucleotide polymorphisms (SNPs) on thousands of diverse disorders in an effort to better understand these phenotypes. The degree to which phenotypes overlap genetically can inform on their underlying biology. For example, the magnitude of correlation between pairs of trait test statistics represented as Z scores corrected for linkage disequilibrium (LD) can indicate how similar traits are genetically [10]. Modeling the genetic covariance matrix can be used to model distinct underlying sources of genetic variance per trait. Examples include bivariate causal mixture modeling [11] and genomic structural equation modeling (gSEM) [12]. Both of these approaches model distinct genetic variance components which can be functionally informative depending on the traits considered. Disorder-specific variants may represent more unique biological processes compared to pleiotropic SNPs. However, approaches to date have several limitations. Bivariate causal mixture modeling does not return SNP-level parameters, instead estimating the degree of causal variant overlap between pairs of traits. While a trivariate version of the method has been developed [13], causal SNP properties across more than three traits cannot be modeled jointly. Methods such as gSEM require that a factor structure be specified, meaning that exploratory analyses or prior hypotheses are required. Finally, gSEM works via pairwise genetic covariance estimates, the estimation of which can be computationally intensive where traits are numerous.

For biobank scale analyses, several factor analysis methods based on matrix factorization have been proposed for extracting shared latent genetic components [14–17]. These approaches aim to factorize the matrix of Z scores of associations of SNPs with multiple traits into latent components / factors and their corresponding loadings; sometimes enforcing desired properties on the factors and their loadings. Tanigawa *et al*. [14] used truncated singular value decomposition (truncated SVD, tSVD), a reduced rank approximation of SVD, to identify the latent factors in a computationally tractable and interpretable fashion. Zhang *et al*. [15] provided a Bayesian framework for the matrix factorization, `FactorGo`, where prior distributions of effect sizes are specified and factors are estimated probabilistically. Another recent approach, `GLEANR`, was developed to estimate regularized latent genetic factors in the presence of sample sharing [16], by enforcing sparsity on the factors and their loadings. However, existing approaches suffer from two main shortcomings: (a) lack of sensitivity with respect to grossly corrupted or outlying observations, which are an inherent feature of genetic association matrices; and (b) nonconvex formulations, which yield different results at each run. This poses an issue from a reproducibility standpoint; results are sensitive to the addition of phenotypes, variants, or simply refitting from a different initialization.

Here, we circumvent these issues and introduce *Convex LOw Rank Inference via Nuclear Norm*, `Clorinn`, which derives a low-rank "outlier-free" matrix X from a matrix Z comprising N traits and P variant Z-scores using rank minimization, as opposed to matrix factorization. Using X, standard factor analysis methods can be applied to examine trait loadings, important SNPs, and global phenotype structure with greater latent factor resolution than methods applied directly to Z. Additionally, the low-rank matrix factorization approach of `Clorinn` is convex, meaning that our method is robust to large amounts of noise or the addition of extra phenotypes/variants. We demonstrate that `Clorinn` is computationally efficient and capable of identifying latent factors with greater accuracy in the presence of varying degrees of noise than other methods in extensive simulations. We apply `Clorinn` to $2,110$ real traits from the UK BioBank and identify several trait-group specific factors, allowing us to validate the results from `Clorinn` with the existing literature. Finally, we apply `Clorinn` to 14 psychiatric traits and characterize shared genetic latent factors to demonstrate our method's utility in a more focused setting.

2 Results

2.1 Method overview

We assume that the noisy matrix of observed Z scores for N traits and P SNPs, $\mathbf{Z} \in \mathbb{R}^{N \times P}$, can be decomposed as a sum of a low-rank matrix $\mathbf{X} \in \mathbb{R}^{N \times P}$ and some error $\mathbf{M} \in \mathbb{R}^{N \times P}$,

$$\mathbf{Z} = \mathbf{X} + \mathbf{M}\,. \tag{1}$$

$\mathbf{X}$ corresponds to the shared genetic components of the traits. The problem to recover $\mathbf{X}$ from noisy $\mathbf{Z}$ is called *low-rank matrix approximation* (LRMA). It seeks the best rank-k estimate of $\mathbf{X}$ by solving the problem,

$$\begin{array}{ll} \text{minimize} & \|\mathbf{Z} - \mathbf{X}\| \\ \text{subject to} & \text{rank}\,(\mathbf{X}) \leq K\,. \end{array}$$

There are two popular approaches for LRMA: *low-rank matrix factorization* (LRMF) and rank minimization.

LRMF seeks to recover latent structures in the data by factorizing the original data matrix $\mathbf{Z}$ into a linear combination of K shared trait-specific loadings $\mathbf{L} \in \mathbb{R}^{N \times K}$ with latent factors $\mathbf{F} \in \mathbb{R}^{P \times K}$,

$$\mathbf{Z} = \mathbf{L}\mathbf{F}^{\mathsf{T}} + \mathbf{M}\,. \tag{2}$$

It uses the estimates $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{F}}$ to reconstruct the low rank estimate $\widehat{\mathbf{X}} \simeq \widehat{\mathbf{L}}\widehat{\mathbf{F}}^{\mathsf{T}}$. Intuitively, the hidden factors $\mathbf{F}$ are some unknown underlying latent embeddings of the variants in K dimensions. These hidden factors can capture interesting biological insights or pathways that drive complex diseases. The above factorization implicitly assumes that factor-trait and variant-factor relationships are all linear. The number of factors (or dimensions) K needed for faithful representation of $\mathbf{Z}$ will depend on the complexity of the data. As in classical principal component analysis (PCA) [18–20], (2) can be efficiently solved via SVD if $\mathbf{L}$ and $\mathbf{F}$ are unregularized and the noise $\mathbf{M}$ is small and *i.i.d.* Gaussian. Under this regime SVD enjoys a number of optimality properties. Other common strategies, such as independent components analysis (ICA) and non-negative matrix factorization (NMF) enforce desired properties on $\mathbf{L}$ and $\mathbf{F}$. More recent approaches includes DeGAs [14] and FactorGo [15]. DeGAs applies SVD and truncates the columns of $\mathbf{L}$ and $\mathbf{F}$ to obtain low rank. FactorGo uses Bayesian priors on $\mathbf{L}$ and $\mathbf{F}$ to enforce sparsity. However, these methods are susceptible to the shortcomings mentioned above. There are two main issues with these existing approaches:

- **Outliers and noise.** One major shortcoming of classical PCA is its brittleness with respect to grossly corrupted or outlying observations [20]: A single grossly corrupted entry in $\mathbf{M}$ could render the estimated $\widehat{\mathbf{X}}$ arbitrarily far from the true $\mathbf{X}$. These methods assume that the noise is Gaussian, *i.e.*, $\mathbf{M} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ is a matrix of Gaussian residuals with variance $\mathbf{\Sigma}$. In many cases, particularly for heterogeneous diseases, the Z scores may have non-Gaussian outliers and discrepancies.
- **Non-convex optimization.** With the exception of PCA, matrix factorization models require non-convex optimization, which cannot be fully reproducible. Many practitioners fix the random seeds, which is only a bandaid solution as tiny perturbations in the data may yield very different results.

In this paper, we focus on rank minimization methods to obtain a low-rank, "outlier-free" matrix $\mathbf{X}$, before applying the matrix factorization via SVD. In recent computer vision literature, a class of robust PCA methods have been developed to handle outliers and heavy noise [21–25]. Following these methods, we assume that the errors $\mathbf{M}$ can be arbitrary in magnitude but sparsely supported, affecting only a fraction of the entries of $\mathbf{Z}$. The outliers and discrepancies captured by the sparse component $\mathbf{M}$ may include: (a) the private contribution of a few key variants to specific phenotypes, or (b) measurement or manual errors in the available summary statistics.

Therefore, instead of using Gaussian noise, we assume that $\mathbf{M}$ is sparse for the rank minimization problem. Since direct rank minimization is itself a nonconvex problem, it is often approximated by minimizing the *nuclear norm* of the estimated matrix $\mathbf{X}$, which is a convex relaxation of minimizing the matrix rank [26, 27]. This methodology is called nuclear norm minimization (NNM). The nuclear norm of a matrix $\mathbf{X}$, denoted by $\|\mathbf{X}\|_*$, is the sum of its singular values. NNM approximates $\mathbf{Z}$ by $\mathbf{X}$, while constraining $\|\mathbf{X}\|_*$. NNM is analogous to Lasso regression which approximates minimizing the number of non-zero regression coefficients.

Here, we use different convex optimization algorithms to perform LRMA:

1. Robust PCA (RPCA) using inexact alternating Lagrangian multipliers (IALM) algorithm,
2. Nuclear Norm Minimization (NNM) using Frank-Wolfe (FW) algorithm, and its variant,
3. Sparse Nuclear Norm Minimization (NNM-Sparse) using FW algorithm.

For Robust PCA, we used Algortihm 5 from Lin *et al*. [28]. We derived novel Frank-Wolfe algorithms for the NNM problem, as described in the Methods section.

Having estimated $\widehat{\mathbf{X}}$ via one of these methods, we use tSVD to obtain $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{F}}$ Fig. 1. Our convex algorithms guarantees reproducible estimates of $\widehat{\mathbf{X}}$. By applying tSVD on $\widehat{\mathbf{X}}$ instead of $\mathbf{Z}$, we avoid the shortcomings of classical PCA and obtain more robust estimates of the latent factors. In a statistical genetics context, this is especially pertinent; technical noise, measurement error, and population structure can have subtle impacts on association study test statistics. Characterization of latent components representing shared genetic effects via PCA is sensitive to variation arising from these non-biological sources. The estimation of $\widehat{\mathbf{X}}$ facilitates the description of accurate and reproducible latent genetic factors robust to the presence of outliers, allowing us to mitigate these issues.

2.2 Numerical experiments on simulated data

We characterize how well `Clorinn` is able to recover the underlying hidden factors under different conditions and compare its performance to other similar approaches using simulated data. We performed simulations under the model of Eqn 2. We start by generating K orthonormal vectors [29] of length P (number of variants), which form the columns of $\mathbf{F} \in \mathbb{R}^{P \times K}$. Each column of $\mathbf{F}$ (denoted by $\mathbf{F}_{:k}$) is an unit vector with mean 0, standard deviation $1/\sqrt{P}$ and is orthogonal to all other columns. They represent the underlying orthogonal hidden factors. For each hidden factor, we

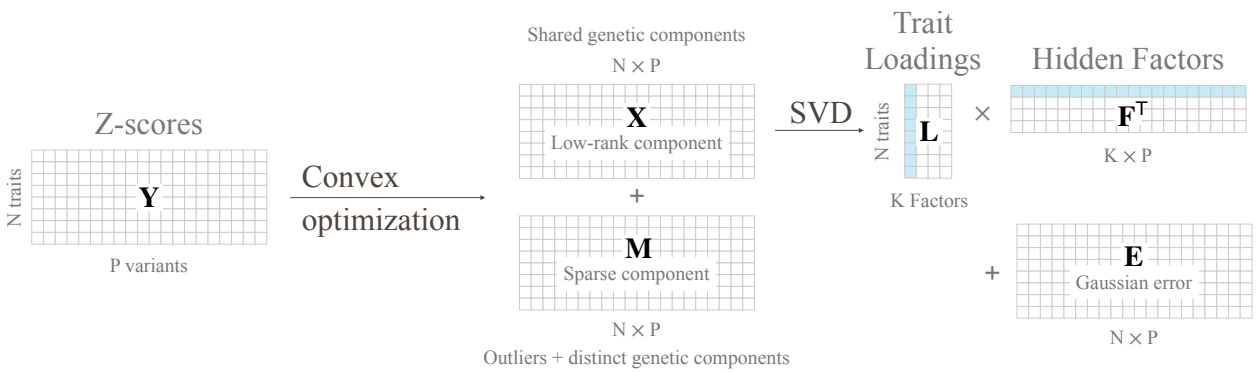

Figure 1: Schematic representation of `Clorinn`.

generate the latent embeddings for N studies (traits), which form the columns of $\mathbf{L} \in \mathbb{R}^{N\times K}$. To mimic the shared component of traits in real data, we sample each column of $\mathbf{L}$ from a multivariate normal distribution $\mathbf{L}_{:k} \sim \mathcal{N}(\mathbf{0}_{\mathrm{N}}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \left(\mathbf{g}\mathbf{g}^{\mathsf{T}} \odot \mathbf{A}\right)/K$, where $\odot$ denotes element-wise multiplication. Here, $\mathbf{g} \in \mathbb{R}^N$ is a column vector where element g_n is the square root of the heritability g_n^2 contributed by the low rank component $\mathbf{L}\mathbf{F}^{\mathsf{T}}$ to the n^{th} trait. The total heritability of the n^{th} trait is h_n^2; the remaining heritability $h_n^2 - g_n^2$ is contributed by the sparse component $\mathbf{M}$. $\mathbf{A}$ is a block diagonal matrix which encodes the strength of sharing between between different traits within and across disease categories. Each block can be considered as representing a broad disease category with multiple traits. Let Q be the number of disease categories (number of blocks) in $\mathbf{A}$ and A_q be the set of indices for the q^{th} block. Then, each element $\mathbf{A}_{ij}$ is given by,

$$\mathbf{A}_{ij} = \begin{cases} 1, & \text{if } i = j\,, \\ \alpha_q \in (0,1)\,, & \text{if } i \neq j \text{ and } i,j \in A_q\,, \\ \alpha_0 \in (0,1)\,, & \text{otherwise.} \end{cases} \tag{3}$$

Here, α_q is the amount of sharing between the traits of the q^{th} disease category, whereas α_0 is a measure of sharing between the different disease categories.

We then generate $\mathbf{M} \in \mathbb{R}^{N\times P}$ by sampling each column $\mathbf{M}_{:p}$ from a Laplace distribution,

$$\mathbf{M}_{:p} \sim \mathrm{Laplace}\left(\mathbf{0}_{\mathrm{N}}, \boldsymbol{\lambda}\right)\,, \quad \boldsymbol{\lambda} = \sqrt{\frac{\mathbf{h}^{\mathsf{T}}\mathbf{h} - \mathbf{g}^{\mathsf{T}}\mathbf{g}}{2P}} \tag{4}$$

to ensure that the total heritability for the traits is $\mathbf{h}^{\mathsf{T}}\mathbf{h}$. Each column of $\mathbf{M}$ is the portion of the Z score of a SNP which cannot be explained by shared factors. The true effect sizes are then calculated as,

$$\mathbf{Z}_0 = \mathbf{L}\mathbf{F}^{\mathsf{T}} + \mathbf{M}\,. \tag{5}$$

Finally, we add some measurement noise to the true Z scores to obtain the observed Z scores,

$$\mathbf{Z} = \mathbf{Z}_0 + \mathbf{E}\,, \quad \mathbf{E}_{n:} \sim \mathcal{N}\left(\mathbf{0}_{\mathrm{P}}, s_n^2 \mathbb{I}_P\right)\,. \tag{6}$$

We assume that the variants are independent (i.e., no LD) and that the genotype and phenotype have been standardized, so that the standard error s_n for the effect sizes depend only on the n^{th} study (trait). In particular, we can assume $s_n \approx 1/\sqrt{\nu_n}$ where ν_n is the sample size of study n. If the genotypes were not standardized, then the standard error for the effect sizes would also depend on the minor allele frequency of each variant.

An example simulation with $Q = 3$ disease categories is shown in Fig. 2a. The covariance of the loadings is block-diagonal (Fig. 2a right panel) and the loadings of the first two factors show 3 different heterogeneous disease groups (left panel). We compare `Clorinn` versions with tSVD and FactorGo as we vary the number of variants, number of hidden factors and the heritability, respectively (Fig. 2b-d). When the assumed number of factors (10) matches the ground truth, `Clorinn` and tSVD provides better estimates of the factors and their corresponding loadings, whereas FactorGo provides a better estimate of $\widehat{\mathbf{L}}\widehat{\mathbf{F}}^{\mathsf{T}}$. As we increase the number of ground truth factors (keeping the number of inferred factors at 10), all methods get worse at estimating $\mathbf{L}$ and $\mathbf{F}$, but the trend remains similar. The trend however reverses when the number of ground truth factors is less than the assumed K. That is, overspecification of number of factors K penalizes `Clorinn` more than other methods, as compared to underspecification of number of factors. In the top row, we show the adjusted mutual information score, which measures whether the estimated disease groups match the ground truth. The disease groups were estimated by clustering $\widehat{\mathbf{L}}$. Clearly, `Clorinn` performs much better in recovering the disease groups compared to other methods. This is important as a key goal is to disentangle the different heterogeneous disease groups.

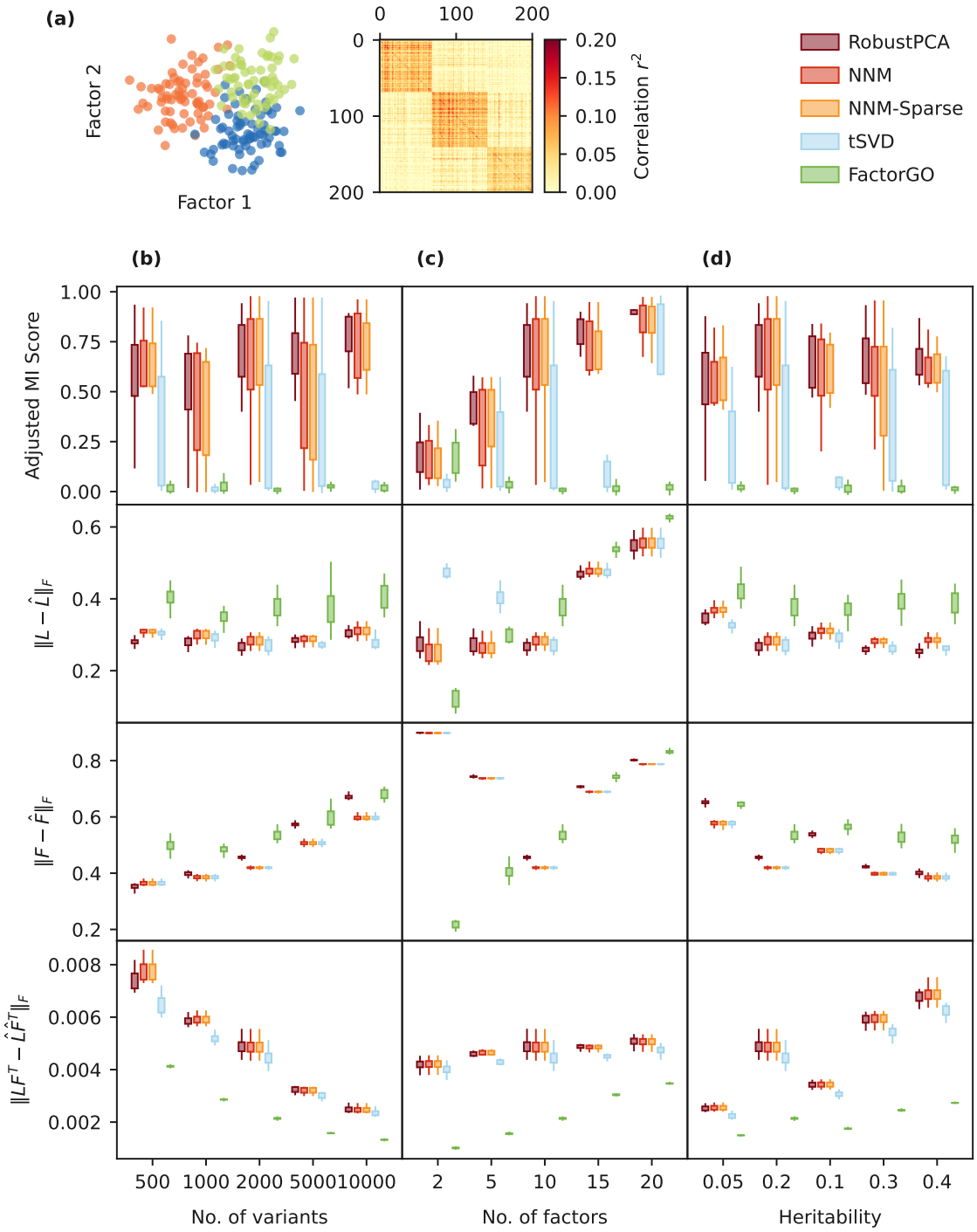

Figure 2: `Clorinn` accurately classifies subtypes of heterogeneous traits in realistic simulations.(a) Left panel shows the loadings of the first two factors in our ground truth, right panel heatmap shows the covariance of the corresponding loadings. (b) Comparison of the different methods as we vary the number of variants used in the simulations. The comparison metrics for each row are shown on the left. $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{L}}$ are the estimates of $\mathbf{F}$ and $\mathbf{L}$ by each of the methods. Adjusted mutual information score measures whether the estimated disease groups match the ground truth. Each boxplot comprise of 20 simulation replicates. (c) Same as column (b) but with varying number of hidden factors. Each method used $K = 10$ for estimating the hidden factors irrespective of the ground truth. (d) Same as column (b) but with varying heritability of the traits.

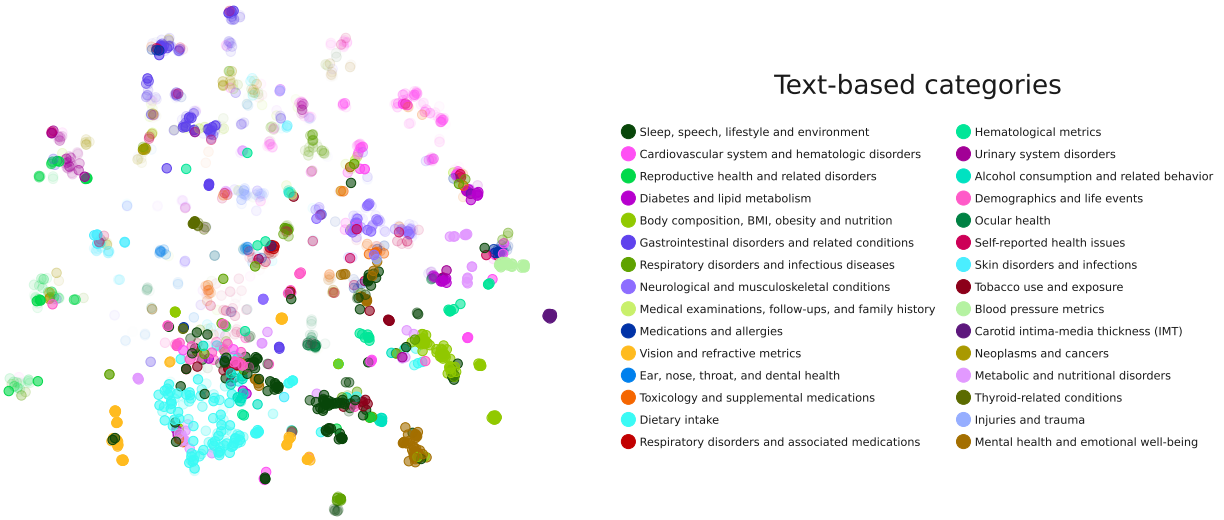

Figure 3: 2D tSNE embedding of the `Clorinn` low rank shared genetic matrix of 2,110 PanUKB traits group into clusters similar to their text descriptions. We classified the 2,110 traits into 30 categories from their the text descriptions, using a publicly available sentence-transformers model. Each point is a trait, colored by its disease category as mentioned in the legend. The opacity of each point is proportional to their estimated heritability.

2.3 Application to 2,110 UKB traits

Having demonstrated the performance in simulations, we used `Clorinn` to analyze 2,110 real traits from the Pan-UKB data. For prefiltering (see Methods) the Pan-UKB data, we followed the same procedure described by Zhang *et al*. [15], but we excluded the prescription traits. We constructed a matrix of GWAS Z scores at 51,368 HLA LD-pruned SNP variants across 2,110 traits. We applied `Clorinn` on this input matrix to obtain the low rank matrix $\widehat{\mathbf{X}}$. For NNM-Sparse, we used the NN constraint of $r \leq 155872$ and ℓ_1 constraint of $l \leq 1770$, which were chosen by cross-validation (see Supplementary Figure S1). We then used tSVD to obtain the top 200 factors and their corresponding loadings.

Our first goal was to analyze how the low rank matrix captures the shared genetic components. We observed that the nuclear norm of the low rank matrix, $||\widehat{\mathbf{X}}||_* = 106525$, is almost 25% of the input matrix, $\|\mathbf{Z}\|_* = 431234$. The matrix rank of $\widehat{\mathbf{X}}$ was 560. We classified the 2,110 traits into 30 broad disease categories from their text descriptions, using a publicly available sentence-transformers model (`ls-da3m0ns/bge_large_medical`). We observed that trait clusters obtained from $\widehat{\mathbf{X}}$ matches closely with their text-based unsupervised classification (Fig. 3). The clustering of traits using tSNE is based solely on their genetic associations, while the color-coded classification in the figure is based solely on their text description. Thus, the genetic association matrix contains similar information about the traits as their text descriptions and allows clustering into meaningful categories. However, we emphasize that the distance between the clusters in Fig. 3 has little meaning.

To illustrate how `Clorinn` captures the shared genetic components of related traits, we characterized the latent components/factors $\widehat{\mathbf{F}}$ and their corresponding loadings $\widehat{\mathbf{L}}$; and how they capture related sets of phenotypes, genes and variants in genetic associations. We used the top 200 factors for our analyses. We selected Type 2 diabetes as a focal trait to explore, given its significant genetic heritability. We quantified the relative importance of the factors using squared cosine scores [30]. The top 4 components of Type 2 diabetes (PC2, PC26, PC1 and PC17) explained over 55% of the genetic associations (23.6%, 14.4%, 10.7% and 6.5% respectively). The PCA plots in Fig. 4a and b show the loadings of all traits corresponding to these factors. The top 20 traits for each factor are annotated on the plots (top 40 traits for PC1). We found that height, weight and BMI related traits are the major contributors to PC1 and PC2. PC26 and PC17 are driven by diabetes-related traits and other major risk factors of Type 2 diabetes, including cholesterol levels, indicating that these factors capture the shared genetic components of these phenotypes.

Next, we identified the genetic variants contributing to PC26 using the variant contribution score. In Fig. 4, we show the names of theses variants and their nearest genes. Using the OpenTargets API [31], we obtained the association score of each gene with Type 2 diabetes. OpenTargets associations aim to aggregate all evidence referring to the target and disease and the association scores range between 0 and 1, with higher scores indicating stronger associations. We find

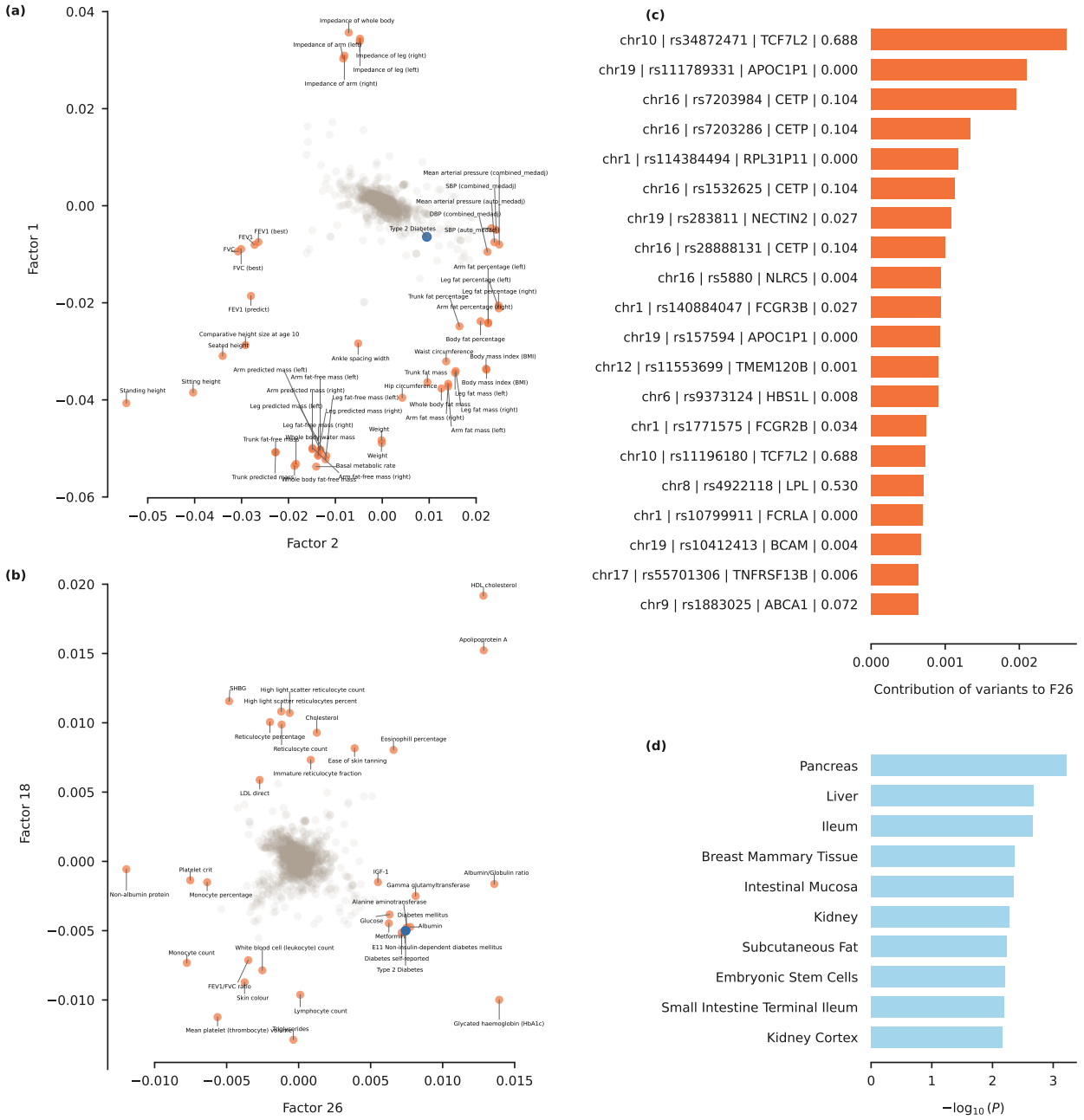

Figure 4: Characterization of hidden factors of Type 2 diabetes, as obtained from the PanUKB analysis. (a) Loadings of all traits corresponding to PC1 and PC2, traits with highest variation are annotated. (b) Same as (a) for PC17 and PC26. (c) Contribution score of variants to PC26. Labels on the y-axis include chromosome number, rsid of the variant, name of the nearest gene and association score of the gene with Type 2 diabetes as obtained from the Open Targets API. (d) Significantly enriched LDSC-SEG annotations for PC26.

that many genes that contribute to PC26 (*e.g.*, TCF7L2, LPL, CETP, NECTIN2, CGR3B) have known associations with Type 2 diabetes, thus establishing consistency with previous molecular findings. Finally, using stratified-LD score regression [10, 32] on 205 tissue- and celltype-specific annotation, we found that PC26 was significantly enriched in pancreas, liver and ileum tissues. This is concordant with the etiology of Type 2 diabetes.

2.4 Application to 14 psychiatric traits

`Clorinn` results in a decomposition of the input association matrix into the sum of two matrices, $\mathbf{X}$ representing 'uncorrupted' entries and $\mathbf{M}$ representing a noise component. This can be conceptualized in a genomics context as a 'shared' space and a 'disorder-specific' space. We sought to compare application of `Clorinn` to results of a recent analysis of 14 psychiatric traits using gSEM to examine the concordance in latent factor embeddings and solution properties [33]. We obtained GWAS summary statistics for each of the 14 analyzed traits using the best-powered available study (Table 1). SNPs were clumped using a subset of European individuals from the Haplotype Reference Consortium (HRC) as a reference with a p-value threshold of $5e^{-5}$. We excluded SNPs with missing association statistics for ≥ 4 traits, resulting in a final matrix $\mathbf{Z}$ of 14 phenotypes with $5,815$ SNP effects each.

Applying `Clorinn` with $\lambda = 0.016$ to $\mathbf{Z}$, we obtained an $\mathbf{X}$ matrix with nuclear norm of 867, marking a decrease from the nuclear norm of the input matrix ($\|Z\|_* = 1843$). The λ value was chosen to yield a $\mathbf{M}$ matrix with approximately 50% non-zero values. We applied PCA to the resultant $\mathbf{X}$ matrix and plotted the first two PCs (Figure 5A). We found that phenotypes clustered into the categories assigned from [33]. Further, we observed that the Silhouette score was higher for the first two PCs of $\mathbf{X}$ compared to that of $\mathbf{Z}$, indicating strong clustering via application of `Clorinn` (0.21 vs. 0.06, Supplementary materials).

The relative contribution of traits to PCs can be assessed using the $\cos^2$ score. We derived this quantity for the first 10 PCs and aggregated contributions by phenotype category to compare results to latent factor embeddings from previous gSEM results [33]. We find that every factor from the 5-factor model in Figure 2B of [33] is represented by at least one PC in our analysis. gSEM fits a model to a genetic covariance matrix of traits using genome-wide information from approximately 1.6 million HapMap3 variants per phenotype, whereas `Clorinn` approximates the same factor structure in this phenotype panel using $5,815$ SNP effects per phenotype. We visualized the contribution of individual phenotypes to PCs in Figure 5B.

Given the representation of our two decomposed matrices, we hypothesized that differences between the shared $\mathbf{X}$ and disorder-specific $\mathbf{M}$ matrices could be informative about SNP properties. We identified 1,539 SNPs where the average absolute value across phenotypes was greater in $\mathbf{M}$ than in $\mathbf{X}$, suggesting these SNPs were more disorder-specific. We expected these SNPs to be nominally significant (Z score ≥ 1.95) in fewer traits compared to SNPs with higher values in $\mathbf{X}$, which we considered more shared. This was confirmed: disorder-specific SNPs were significant in fewer traits on average than shared SNPs, and this difference was statistically significant (Figure 5C, $p = 4.9 \times 10^{-10}$). To more rigorously assess statistical significant, we repeated this operation using $1,000$ random re-samplings of $1,536$ SNPs from $\mathbf{Z}$ to examine if our observed P-value was due to chance rather than property differences between SNP sets. We found no test statistics more extreme than our observed results in these re-samplings. Similarly, we expected that shared SNPs were likely to be more pleiotropic than disorder-specific SNPs. To examine this, we calculated the HOrizontal Pleiotropy Score (HOPS) [34] in each set across our 14 phenotypes and tested for a difference in means. We found a statistically significant difference in pleiotropy magnitude between SNP sets, with disorder-specific SNPs having a lower mean pleiotropy score than shared SNPs (Figure 5D). Across $1,000$ random samplings of $1,536$ SNPs from $\mathbf{Z}$ to act as permuted disorder-specific SNPs, we found no test statistic as extreme as the observed p-value.

The degree to which PCs load on variables (traits) is given by multiplying eigenvectors by the square root of their eigenvalues. A method to conduct statistical hypothesis testing of loading values was previously introduced in [35]. Briefly, the relationship between raw variable values across samples and the PC value across samples is expected to be t-distributed under the null assumption of no correlation; this intuition stems from the fact that larger variable loadings imply stronger correlation of variables to respective PC values. The association of a variable with PC activity can be formally investigated by regressing PC values across samples on input variable values. Following this procedure, we regressed scaled PC1 values on each SNP individually. We found that 18 SNPs had genome-wide significant (GWS) test statistics ($p \leq 5 \times 10^{-8}$, Supplementary Note). We expected that these SNPs would be primarily composed of SNPs significant for phenotypes in the internalizing or compulsive disorder categories based on Figure 5B. 14 of the 18 SNPs were GWS for either PTSD or MDD, and 4 were not significant in any of the top 4 phenotypes loaded on by PC1. These SNPs were rs1604060, rs2196150, rs2865303, and rs865020. rs1604060 and rs2196150 were both included as clumped SNPs associated with SCZ [36], with rs2196150 also previously reported as a significant association of leukocyte quantity [37]. rs2865303 was a clumped SNP associated with TUD [38] and was previously reported as a significant association of externalizing behavior [39]. Finally, rs865020 was a clumped SNP from MDD [40] with no previously reported associations; it was also not GWS in MDD.

We used FUMA [41] to perform a suite of post-GWAS analyses on PC1. We found that genes significant via a gene-based association test carried out in MAGMA were nominally enriched for brain tissues, but these results were not significant after Bonferroni correction.

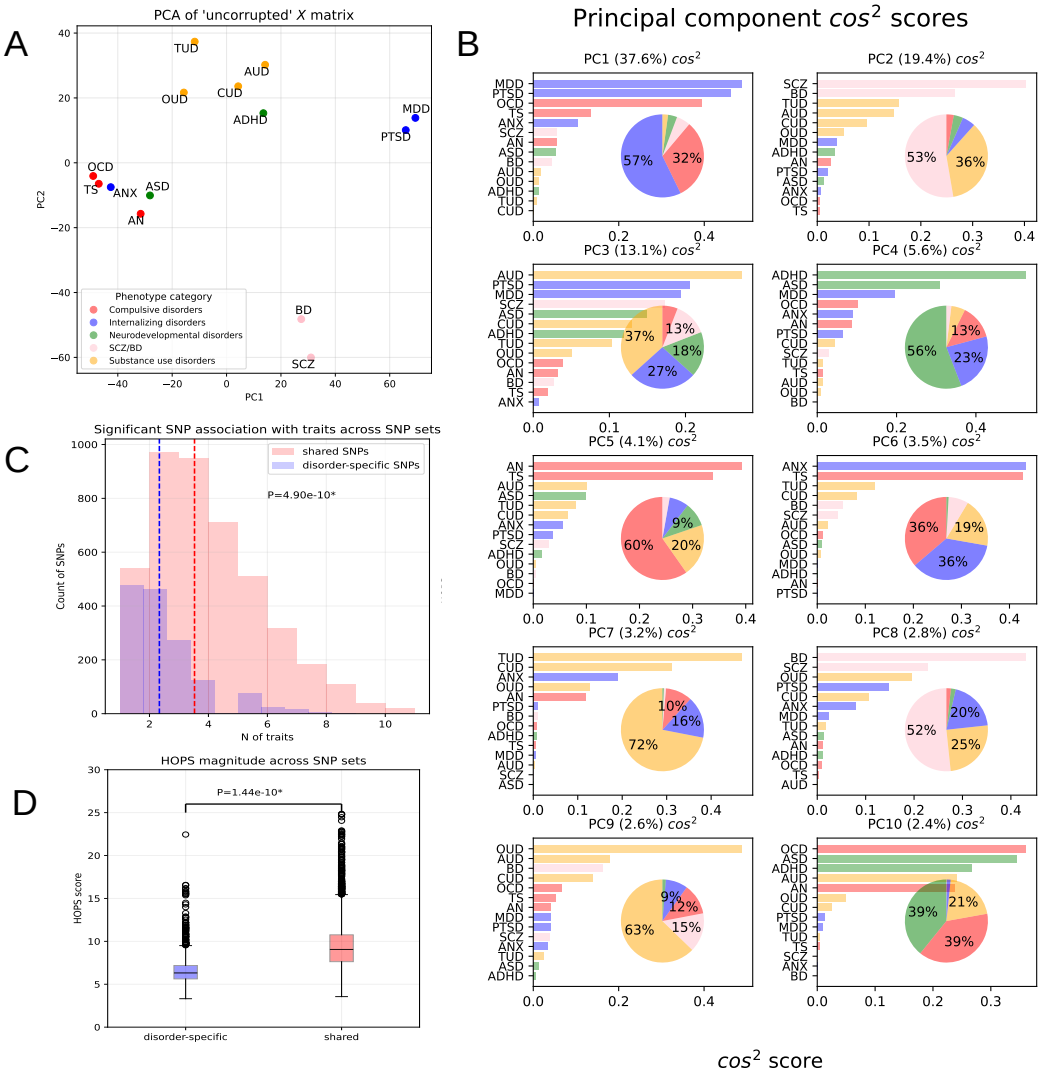

Figure 5: **A**: First two PCs from PCA of 'uncorrupted' $\mathbf{X}$ matrix of 14 psychiatric traits. Phenotype categories are derived from the 5 factor gSEM model from [33]. Phenotypes are labeled by their abbreviations: AUD (alcohol-use disorder), AN (anorexia nervosa), SCZ (schizophrenia), ANX (panic disorder), BD (bipolar disorder), TS (Tourette's syndrome), ASD (autism spectrum disorder), TUD (tobacco-use disorder), CUD (cannabis-use disorder), PTSD (post-traumatic stress disorder), OCD (obsessive-compulsive disorder), MDD (major-depressive disorder), ADHD (attention-deficit hyperactivity disorder), OUD (opioid-use disorder). **B**: $\cos^2$ scores per PC with associated aggregated contribution of phenotype categories in pie charts. Category contribution is measured as the normalized $\cos^2$ score of each phenotype to a PC aggregated by phenotype category, ordered from left to right (ascending). The variance explained by each PC is described in parentheses. The color code corresponds to the legend of A. **C**: Respective distributions of number of traits where SNPs from each set are nominally significant ($\bar{Z} > 1.95$). Shared SNPs are those with mean absolute $\mathbf{M}$ values less than their mean absolute $\mathbf{X}$ values; disorder-specific SNPs are those with mean absolute $\mathbf{M}$ values greater than their mean $\mathbf{X}$ values. Mean values of each distributions are denoted by dashed vertical lines, and the p-value was derived from an independent t-test of means between distributions. SNP counts are described on the y-axis and number of traits where SNPs are nominally significant are described on the x-axis. **D**: Boxplot of HOPS scores for each set of SNPs, representing their pleiotropy magnitude [34]. The p-value was derived from an independent t-test of means of pleiotropy magnitude across SNP sets.

Disorder name	Citation	N. SNPs	SNP h^2 (%)	Category
AUD (Alcohol use disorder)	[42]	536	12.7	Substance use disorders
AN (Anorexia nervosa)	[43]	233	11.0	Compulsive disorders
SCZ (Schizophrenia)	[36]	1001	24.0	SCZ/BD
ANX (Panic disorder)	[44]	97	28.0	Internalizing disorders
BD (Bipolar disorder)	[45]	664	22.0	SCZ/BD
TS (Tourette's syndrome)	[46]	107	21.0	Compulsive disorders
ASD (Autism spectrum disorder)	[47]	166	11.8	Neurodevelopmental disorders
TUD (Tobacco use disorder)	[38]	528	5.0	Substance use disorders
CUD (Cannabis use disorder)	[48]	295	6.7	Substance use disorders
PTSD (Post-traumatic stress disorder)	[49]	730	5.3	Internalizing disorders
OCD (Obsessive compulsive disorder)	[50]	80	4.1	Compulsive disorders
MDD (Major depressive disorder)	[40]	844	8.4	Internalizing disorders
ADHD (Attention-deficit hyperactivity disorder)	[51]	387	14.0	Neurodevelopmental disorders
OUD (Opioid use disorder)	[52]	223	12.8	Substance use disorders

Table 1: Table of phenotypes included in the cross-disorder phenotype analysis. N. SNPs refers to the number of clumped SNPs in the final pruned matrix contributed by that phenotype. SNP h^2 estimates are taken from the *LDSC* parameter estimate from each respective study in European populations on the liability scale [10]. Phenotype categories are taken from the cross-disorder group analysis [33].

Finally, we applied the NNM-sparse model to examine how reproducible the factor structure of our derived $\mathbf{X}$ matrix was across algorithms. We observed broadly similar clustering behavior and factor structure; further we observed $R^2 = 0.611$ between $\mathbf{X}$ derived by NNM-sparse compared to $\mathbf{X}$ derived by robust PCA (Supplementary Note).

3 Methods

3.1 Frank-Wolfe algorithms for rank minimization

We use the Frank-Wolfe (FW) algorithm [53–55] to solve the constrained optimization problem. Other methods [56] for solving NNM, including singular value thresholding [27, 57] and proximal methods [58], are computationally more expensive than the FW-type methods. Frank-Wolfe uses significantly less expensive linear subproblems per iteration compared to the quadratic problems in the latter, which can make certain problems intractable; for *e.g.*, the dual of structural SVMs [59], or an order of magnitude increase in iteration cost for the trace norm (leading eigenvector vs. SVD) [60]. For the FW algorithm, the loss function for the NNM problem is given by,

$$f(\mathbf{X}) = \frac{1}{2}\sum_i \sum_j (z_{ij} - x_{ij})^2 \ , \tag{7}$$

and its gradient is given by,

$$\nabla f(\mathbf{X}) = \mathbf{X} - \mathbf{Z} \tag{8}$$

We outline the steps for FW minimization in Algorithm 1. The main bottleneck in implementing the algorithm is solving the linear optimization subproblem in Line 2. The objective of the subproblem is linear even though the constraint set Ω may not be. Since Ω is compact, the solution to the subproblem always exists. After finding a solution $\mathbf{S}_t$ to the linear optimization subproblem, the rest of the algorithm concerns finding the appropriate step size to move in the direction dictated by $\mathbf{D}_t := \mathbf{X}_{t-1} - \mathbf{S}_t$. The new iterate is a convex combination of the previous iterate and $\mathbf{S}_t$.

The term g_t is called the *surrogate duality gap* which is also a *certificate* of the current approximation quality, *i.e.*, $g_t \geq f(\mathbf{X}) - f(\widehat{\mathbf{X}})$ where $\widehat{\mathbf{X}}$ is the desired solution of the optimization. Although $\widehat{\mathbf{X}}$ is not known, we can easily compute $\mathbf{X}$ and g_t, which can be used as a convergence criterion.

Linear optimization subproblem. Denote the singular value decomposition of $\nabla f(\mathbf{X}_{t-1})$ as $\mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T}$, where $\mathbf{U} \in \mathbb{R}^{N\times K}$ and $\mathbf{V} \in \mathbb{R}^{P\times K}$ are orthogonal matrices, $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbb{I}_K$. The columns of $\mathbf{U}$ and the columns of $\mathbf{V}$ are called left-singular vectors and right-singular vectors, respectively. $\mathbf{D} \in \mathbb{R}^{K\times K}$ is a diagonal matrix whose diagonal entries are the singular values. The rank of the matrix is $K = \min\{N, P\}$. Then the solution to the subproblem in Line 2 is the convex hull of the singular vectors with maximal singular value, appropriately scaled. In practice, we often take the solution $\mathbf{S}_t \approx -r\mathbf{u}_1\mathbf{v}_1^\mathsf{T}$, where $\mathbf{u}_1$ and $\mathbf{v}_1$ denote the first columns of $\mathbf{U}$ and $\mathbf{V}$ respectively [54, 55].

Algorithm 1: Frank-Wolfe algorithm for nuclear norm minimization

Input: $f, \nabla f, r$

Initialize: $\mathbf{X}_0$ such that $\|\mathbf{X}_0\|_* \leq r$

1: **for** $t = 1, 2, \ldots, T$ **do**
2: $\quad \mathbf{S}_t = \underset{\mathbf{S} \in \Omega}{\arg\min}\, \mathrm{Tr}\left(\nabla f\left(\mathbf{X}_{t-1}\right)^{\mathsf{T}} \mathbf{S}\right)$, where $\Omega = \left\{\mathbf{S} \in \mathbb{R}^{N \times P} \mid \|\mathbf{S}\|_* \leq r\right\}$
3: $\quad \mathbf{D}_t \coloneqq \mathbf{X}_{t-1} - \mathbf{S}_t$
4: $\quad g_t = \mathrm{Tr}\left(\mathbf{D}_t^{\mathsf{T}} \nabla f\left(\mathbf{X}_{t-1}\right)\right)$
5: $\quad$ **if** $g_t \leq \epsilon$ **then**
6: $\quad\quad$ **break**
7: $\quad$ **end if**
8: $\quad \gamma_t = \dfrac{2}{2+t}$ or $\gamma_t = \underset{\gamma}{\arg\min}\, f\left(\mathbf{X}_{t-1} - \gamma \mathbf{D}_t\right)$
9: $\quad \mathbf{X}_t = \mathbf{X}_{t-1} - \gamma_t \mathbf{D}_t$
10: **end for**

Step size. The step size for the subproblem in Line 8 can be derived analytically. Define

$$\mathbf{X}^* \coloneqq \mathbf{X}_{t-1} - \gamma \mathbf{D}_t \,. \tag{9}$$

From (7), we have,

$$f\left(\mathbf{X}^*\right) = \frac{1}{2} \sum_i \sum_j \left(z_{ij} - x_{ij}^*\right)^2 \tag{10}$$

The derivative of this function is,

$$\begin{aligned} \frac{\partial}{\partial \gamma} f\left(\mathbf{X}^*\right) &= \sum_{i,j} \left(z_{ij} - x_{ij}^*\right) d_{ij} \\ &= \sum_{i,j} \left(z_{ij} - x_{ij}\right) d_{ij} + \gamma \sum_{i,j} d_{ij}^2 \\ &= -\,\mathrm{Tr}\left(\mathbf{D}_t^{\mathsf{T}} \nabla f\left(\mathbf{X}_{t-1}\right)\right) + \gamma \|\mathbf{D}_t\|_F^2 \\ &= -g_t + \gamma \|\mathbf{D}_t\|_F^2 \,. \end{aligned} \tag{11}$$

Setting this equal to 0, we obtain the optimum step size as,

$$\gamma_t = \frac{g_t}{\|\mathbf{D}_t\|_F^2} \,. \tag{12}$$

NNM-Sparse. As discussed in Section 2.1, we want to extract sparse outliers from the input data, and propose our problem as,

$$\min_{\mathbf{X},\mathbf{M}} \frac{1}{2} \|\mathbf{Z} - \mathbf{X} - \mathbf{M}\|_F^2 \quad \text{subject to} \quad \|\mathbf{X}\|_* \leq r \,, \|\mathbf{M}\|_1 \leq l \,. \tag{13}$$

We identify $(\mathbf{X}, \mathbf{M})$ as the estimand and use the Frank-Wolfe algorithm to solve (13). The loss function and the gradients are given by,

$$\begin{aligned} f\left(\mathbf{X}, \mathbf{M}\right) &= \frac{1}{2} \sum_i \sum_j \left(z_{ij} - x_{ij} - m_{ij}\right)^2 \,, \\ \nabla_{\mathrm{X}} f &= \mathbf{X} + \mathbf{M} - \mathbf{Z} \,, \\ \nabla_{\mathrm{M}} f &= \mathbf{X} + \mathbf{M} - \mathbf{Z} \,, \end{aligned} \tag{14}$$

Algorithm 2: Frank-Wolfe algorithm for sparse nuclear norm minimization (13)

Input: $f(\mathbf{X}, \mathbf{M})$, gradient map $\nabla f(\mathbf{X}, \mathbf{M}) = (\nabla_{\mathrm{X}} f, \nabla_{\mathrm{M}} f)$, r, l

Initialize: $\mathbf{X}_0$, $\mathbf{M}_0$ such that $\|\mathbf{X}_0\|_* \leq r$, $\|\mathbf{M}_0\|_1 \leq l$

1: **for** $t = 1, 2, \ldots, T$ **do**
2: $\quad \mathbf{S}_{\mathrm{X}} = \underset{\mathbf{S}_{\mathrm{X}} \in \Omega_{\mathrm{X}}}{\arg\min} \operatorname{Tr}\left(\nabla_{\mathrm{X}} f(\mathbf{X}_{t-1})^{\mathsf{T}} \mathbf{S}_{\mathrm{X}}\right)$, where $\Omega_{\mathrm{X}} = \left\{\mathbf{S}_{\mathrm{X}} \in \mathbb{R}^{N \times P} \mid \|\mathbf{S}_{\mathrm{X}}\|_* \leq r\right\}$
3: $\quad \mathbf{S}_{\mathrm{M}} = \underset{\mathbf{S}_{\mathrm{M}} \in \Omega_{\mathrm{M}}}{\arg\min} \operatorname{Tr}\left(\nabla_{\mathrm{M}} f(\mathbf{M}_{t-1})^{\mathsf{T}} \mathbf{S}_{\mathrm{M}}\right)$, where $\Omega_{\mathrm{M}} = \left\{\mathbf{S}_{\mathrm{M}} \in \mathbb{R}^{N \times P} \mid \|\mathbf{S}_{\mathrm{M}}\|_1 \leq l\right\}$
4: $\quad \mathbf{D}_t \coloneqq (\mathbf{D}_{\mathrm{X}}, \mathbf{D}_{\mathrm{M}})$ where $\mathbf{D}_{\mathrm{X}} \coloneqq \mathbf{X}_{t-1} - \mathbf{S}_{\mathrm{X}}$ and $\mathbf{D}_{\mathrm{M}} \coloneqq \mathbf{M}_{t-1} - \mathbf{S}_{\mathrm{M}}$
5: $\quad g_t = \operatorname{Tr}\left(\mathbf{D}_t^{\mathsf{T}} \nabla f(\mathbf{X}_{t-1}, \mathbf{M}_{t-1})\right)$
6: $\quad$ **if** $g_t \leq \epsilon$ **then**
7: $\quad\quad$ **break**
8: $\quad$ **end if**
9: $\quad \gamma_t = \dfrac{2}{2+t}$ or $\gamma_t = \underset{\gamma}{\arg\min}\, f(\mathbf{X}_{t-1} - \gamma \mathbf{D}_{\mathrm{X}}, \mathbf{M}_{t-1} - \gamma \mathbf{D}_{\mathrm{M}})$
10: $\quad \mathbf{X}_t = \mathbf{X}_{t-1} - \gamma_t \mathbf{D}_{\mathrm{X}}$
11: $\quad \mathbf{M}_{t-\frac{1}{2}} = \mathbf{M}_{t-1} - \gamma_t \mathbf{D}_{\mathrm{M}}$
12: $\quad \mathbf{M}_t = \mathcal{P}_{\|\cdot\|_1 \leq r_{\mathrm{M}}}\left[\mathbf{M}_{t-\frac{1}{2}} - \nabla_{\mathrm{M}} f\left(\mathbf{X}_t, \mathbf{M}_{t-\frac{1}{2}}\right)\right]$
13: **end for**

so that the corresponding gradient map can be obtained $\nabla f(\mathbf{X}, \mathbf{M}) = (\nabla_{\mathrm{X}} f, \nabla_{\mathrm{M}} f)$. In Algorithm 2, we outline the optimization steps. Here, we need to solve two linear optimization subproblem, corresponding to the two constraints. The nuclear norm regularization for the subproblem in Line 2 is exactly the same as in Algorithm 1. The ℓ_1 norm regularization for the subproblem in Line 3 can be solved by noting that the dual of the ℓ_1 norm is the in ℓ_∞ norm, *i.e.*, $\|\mathbf{Z}\|_\infty \coloneqq \max_{(i,j)} |z_{ij}| = \max_{\|\mathbf{X}\| \leq 1} \operatorname{Tr}\left(\mathbf{Z}^{\mathsf{T}} \mathbf{X}\right)$. Therefore, one minimizer is $\mathbf{S}_{\mathrm{M}} = -r_{\mathrm{M}} \mathbf{e}_{i*} \mathbf{e}_{j*}^{\mathsf{T}}$ where $(i*, j*) \in \arg\max_{i,j} |x_{ij} + m_{ij} - z_{ij}|$, *i.e.*, $\mathbf{S}_{\mathrm{M}}$ has exactly one non-zero element. In the standard FW algorithm, the update for the sparse component updates only one element of the matrix (Line 11). To improve the speed of the algorithm, we include an additional step within each loop to minimize the sparse component, as suggested by Mu *et al.* [61]. Given a threshold $\beta > 0$, the projection onto the ℓ_1 ball minimizes the sparse component,

$$\mathcal{P}_{\|\cdot\|_1 \leq \beta}[\mathbf{Z}] = \underset{\|\mathbf{M}\|_1 \leq \beta}{\arg\min} \frac{1}{2} \|\mathbf{M} - \mathbf{Z}\|_F^2 . \tag{15}$$

As shown previously [61], the projected gradient step in Line 12 ensures that the objective is less than that produced by the standard FW step, *i.e.*, $f(\mathbf{X}_t, \mathbf{M}_t) \leq f\left(\mathbf{X}_t, \mathbf{M}_{t-\frac{1}{2}}\right)$.

3.2 Application to 14 psychiatric traits

We clumped SNPs per disorder using genotypes from 16,886 European individuals from the HRC as a reference with the following parameters in PLINK1.9: $P < 5e^{-5}$, window = 10,000kb, $r^2 < 0.1$ [62]. For OUD summary statistics, we obtained the full panel of tested SNPs in European cohorts from the authors [52]. We then annotated SNPs by position and allele using the 1000 Genomes European reference panel (SNP build 151) [63]. We filtered SNPs for missingness across phenotypes, whereby SNPs not present in more than 4 phenotypes were excluded. We verified that SNP values across traits were centered around zero by visualizing their Z-score distributions. We used the Robust PCA algorithm from `Clorinn` with a λ value of 0.016 to obtain a solution with approximately 50% non-zero values in M. We performed PCA on X with 10 PCs using the *scikit-learn* package [64].

Silhoutte scores can be used to index the accuracy of a clustering solution based on comparing the similarity between points within a cluster to those in other clusters. The score can be defined per cluster k via

$$S_k = \frac{(b_i - a_i)}{max\,(a_i, b_i)}, \tag{16}$$

where a is the mean intra-cluster distance between sample i and members of its assigned cluster, and b is the smallest distance between sample i and a member of its nearest non-assigned cluster. The output of this score is bounded between -1 and 1, with higher scores indicating less overlapping clusters. Distances between points were measured in the euclidean space. Results reported are the mean Silhouette score across samples per solution.

We obtained $\cos^2$ scores via

$$\cos_i^2 = \frac{\sigma_i^2}{\sum \sigma_i^2}, \tag{17}$$

where σ is the standard deviation of PC i.

We tested for a difference in means between numbers of traits where SNPs are significant across sets using an independent t-test of means. We calculated pleiotropy scores using a Python implementation of [34] in 'magnitude' mode, and tested for a difference in means between SNP sets using an independent t-test of means. We did not correct the score for LD as minimal differences were reported in the original paper when carrying out this operation.

We derived an association test statistic per SNP for their association with PC values using the procedure described in [35]. Briefly, we performed a singular value decomposition of the input Z matrix to obtain the right singular vectors V^T. The PC score is the product of the input data multiplied by the right singular vectors, yielding a $N \times N$ matrix (where N is the number of traits). The association score was obtained by regressing the scaled values of PC 1 from this matrix across phenotypes on the SNP association Z-scores across phenotypes. We repeated this operation for every SNP and used *FUMA* [41] to carry out functional enrichment of GWAS results.

4 Discussion

`Clorinn` showed comparable or better performance to similar methods for factor discovery in a range of simulation scenarios (Figure 2). This is likely due to the fact that `Clorinn` first applies LRMA before factor discovery to derive an outlier-free version of the input matrix. This added step ensures that subsequent analyses are not confounded by large amounts of noise, or the addition of extra phenotypes/variants. This property is likely to be useful for consortium-level analyses, whereby re-analysis with additional cohorts is required. Further, the Frank-Wolfe convex optimization procedure guarantees reproducibility of the $\mathbf{X}$ matrix regardless of initialization. This convexity feature is attractive from a solution stability and reproducibility standpoint. Secondly, we find that hidden factors derived via `Clorinn` can be used to accurately cluster large groups of diverse phenotypes in a biobank-scale application, as evidence from our tSNE embedding plot of 30 broad phenotype categories (Figure 3). Of particular note is the large groupings of dietary intake and mental health traits respectively (light blue and brown points in Figure 3). This suggests that granular categorical distinctions of phenotypes are possible even when considering large numbers of traits. In a phenotype-focused example, we found PC26 as the factor loading on Type 2 diabetes to the largest degree. An examination of the underlying biology via squared cosine scores establishes good concordance with previous literature on tissue enrichment and associated genes (Figure 4). For example, the top variant for PC26 was mapped to *TCF7L2*, which has been previously identified as a significant GWAS association of the condition [65]. We also observe that 4 of the top 8 variants for PC26 map to *CETP*, a plasma protein responsible for cholesterol transport. Recent research has described the potential of *CETP* inhibitors for the treatment of new-onset diabetes, demonstrating that our analysis can identify variants with therapeutic potential [66]. We can also leverage the fact that traits can be loaded on by multiple factors for biological insight, as evidenced through Figure 4B, whereby we examined variable traits along PC18 and PC26, which are the 2 top factors of Type 2 Diabetes. We find that several blood traits have large variance in this plot, including albumin/globulin ratio and white blood cell leukocyte count. Unsurprisingly, these traits have been previously reported as having significant genetic correlations with diabetes [67]. Additionally, we find that S-LDSC results implicate the pancreas as the main tissue of interest, which is relevant given that the molecular basis of Type 2 diabetes originates in this tissue [68] (Figure 4D). These experiments verify that biologically meaningful factors can be identified from biobank-scale data in a computationally tractable and reproducible manner. Furthermore, it demonstrates that biological discovery is possible at a broad scale for the identification of large category groupings, and at a more granular level via a focus on specific phenotype-factor relationships.

Additionally, we find that `Clorinn` recapitulates gSEM results in our psychiatric trait application. Firstly, phenotypes are shown to group based on their first two PCs according to categories defined in previous work after running PCA on

the resultant $\mathbf{X}$ matrix [33] (Figure 5A). This is interesting given that `Clorinn` makes use of information from just $5,815$ SNPs per trait. Furthermore, we found that the mean Silhouette score was improved via application of `Clorinn`, suggesting that the accuracy of cluster resolution is increased through derivation of an "outlier-free" SNP matrix. Our squared cosine embeddings describing loading of a factor on traits showed good agreement with results of the 5-factor model from [33] (Figure 5B). Specifically, we found that all factors from the 5-factor gSEM model are represented in the $\cos^2$ score embeddings in at least one PC. An advantage of this approach is the reduced amount of SNP information required per phenotype; additionally, a hypothesized phenotype factor structure is not required. This is useful from a computational and factor discovery standpoint. For instance, we find that internalizing disorders and compulsive disorders load primarily on PC1 in Figure 5B. We also observe that categories tend to contribute to PC activity primarily in pairs. For example, internalizing disorders in PC1 and PC6, and SCZ/BD and substance use disorders in PC2 and PC8. Further, certain phenotype categories have more specific contribution patterns to factors, such as in neurodevelopmental disorders, which contributes largely to PC4 and PC10, but with no discernible category pairing. This visualization technique has great potential for identifying correlated sets of traits belonging to broader groupings, especially where the number of traits is large. An advantage of `Clorinn` from this standpoint is its potential in both biobank-scale factor analysis and more controlled applications. Comparable methods are limited to the derivation of factors from larger numbers of phenotypes. We have demonstrated that `Clorinn` can be applied to thousands of of traits.

We also demonstrate the discovery potential of `Clorinn` in our psychiatric application via statistical hypothesis testing of factor loadings and the definition of pleiotropic and disorder-specific SNP sets. We found that 18 SNPs were GWS after correlating variant Z scores to factor activity across phenotypes, with 4 variants reported which were not GWS in any of the top 4 phenotypes loaded on by PC1. This framework has the potential to obtain proxy genetic associations with unobserved latent variables, which is a conceptually similar goal to approaches such as gSEM. Our approach is an attractive alternative owing to the computational ease of analysis and the sparsity of genetic information required. The definition of disorder-specific SNP sets could also be used to describe properties related to the residual variance per phenotype from gSEM and FactorGO, although an application of this was not explored in this work.

5 Limitations

Our simulation results did not include a direct comparison of our method to *GLEANR*. Our work also did not seek to define residual variance of phenotypes, but the information returned by `Clorinn` could be used for this application. Finally, while our LRMA approach has many useful properties, our resultant factors do not have a direct SNP-mapping (except through the outlier-free $\widehat{\mathbf{X}}$). This abstraction may makes it more difficult to interpret associated factors and their constituent elements. Similarly, an important point of future research will be developing a thorough interpretation of SNP $\widehat{\mathbf{X}}$ values relative to the input SNP-trait associations.

It has recently been shown that assortative mating inflates estimates of genetic correlation[69]. Factorization approaches like `Clorinn` will be influenced by this inflation, and it remains an open question how such analyses might be appropriately corrected.

6 Conclusion

Here, we have developed a computationally efficient and flexible modeling approach to describe robust latent genetic factors of an arbitrary number of GWAS results. Our method is reproducible and does not constitute a direct analysis of the traits-effects matrix for the definition of latent factors, which has several attractive properties. We anticipate that `Clorinn` can be used for analysis at the trait and variant level in a variety of applications, especially as the number of phenotypes scale.

7 Code availability

All code is publicly available at `https://github.com/banskt/colormann`.

8 Acknowledgments

Research reported in this manuscript was supported by the National Institute of Mental Health of the National Institutes of Health under award number R01MH130879. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] S. D. Østergaard, S. O. W. Jensen, and P. Bech. "The heterogeneity of the depressive syndrome: when numbers get serious". *Acta Psychiatrica Scandinavica* 124.6 (2011), pp. 495–496. DOI: https://doi.org/10.1111/j.1600-0447.2011.01744.x.

[2] L. Cui et al. "Major depressive disorder: hypothesis, mechanism, prevention and treatment". *Signal Transduction and Targeted Therapy* 9.1 (2024), p. 30. DOI: 10.1038/s41392-024-01738-y.

[3] T. A. LeGates, M. D. Kvarta, and S. M. Thompson. "Sex differences in antidepressant efficacy". *Neuropsychopharmacology* 44.1 (2019), pp. 140–154. DOI: 10.1038/s41386-018-0156-z.

[4] S. Wagner et al. "Effects of age on depressive symptomatology and response to antidepressant treatment in patients with major depressive disorder aged 18 to 65 years". *Comprehensive Psychiatry* 99 (2020), p. 152170. DOI: https://doi.org/10.1016/j.comppsych.2020.152170.

[5] S. M. Marcus et al. "Sex differences in depression symptoms in treatment-seeking adults: confirmatory analyses from the Sequenced Treatment Alternatives to Relieve Depression study". *Comprehensive Psychiatry* 49.3 (2008), pp. 238–246. DOI: https://doi.org/10.1016/j.comppsych.2007.06.012.

[6] E. Tedeschini et al. "Efficacy of Antidepressants for Late-Life Depression: A Meta-Analysis and Meta-Regression of Placebo-Controlled Randomized Trials". *The Journal of Clinical Psychiatry* 72.12 (2011), p. 5967. DOI: 10.4088/JCP.10r06531.

[7] M. Fellinger et al. "Seasonality in Major Depressive Disorder: Effect of Sex and Age". *Journal of Affective Disorders* 296 (2022), pp. 111–116. DOI: https://doi.org/10.1016/j.jad.2021.09.051.

[8] S. Li et al. "Sex difference in incidence of major depressive disorder: an analysis from the Global Burden of Disease Study 2019". *Annals of General Psychiatry* 22.1 (2023), p. 53. DOI: 10.1186/s12991-023-00486-7.

[9] A. R. Slan et al. "The role of sex and age in the differential efficacy of 10 Hz and intermittent theta-burst (iTBS) repetitive transcranial magnetic stimulation (rTMS) treatment of major depressive disorder (MDD)". *Journal of Affective Disorders* 366 (2024), pp. 106–112. DOI: https://doi.org/10.1016/j.jad.2024.08.129.

[10] B. K. Bulik-Sullivan et al. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". *Nature genetics* 47.3 (2015), pp. 291–295.

[11] O. Frei et al. "Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation". *Nature communications* 10.1 (2019), p. 2417.

[12] A. D. Grotzinger et al. "Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits". *Nature human behaviour* 3.5 (2019), pp. 513–525.

[13] A. A. Shadrin et al. "Dissecting the genetic overlap between three complex phenotypes with trivariate MiXeR". *medRxiv* (2024). DOI: 10.1101/2024.02.23.24303236.

[14] Y. Tanigawa et al. "Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight adipocyte biology". *Nature communications* 10.1 (2019), p. 4064.

[15] Z. Zhang et al. "A scalable approach to characterize pleiotropy across thousands of human diseases and complex traits using GWAS summary statistics". *The American Journal of Human Genetics* 110.11 (2023), pp. 1863–1874. DOI: https://doi.org/10.1016/j.ajhg.2023.09.015.

[16] A. R. Omdahl et al. "Sparse matrix factorization robust to sample sharing across GWAS reveals interpretable genetic components". *bioRxiv* (2024), pp. 2024–11.

[17] D. I. Chasman et al. "Pleiotropy-Based Decomposition of Genetic Risk Scores: Association and Interaction Analysis for Type 2 Diabetes and CAD". *The American Journal of Human Genetics* 106.5 (2020), pp. 646–658. DOI: 10.1016/j.ajhg.2020.03.011.

[18] C. Eckart and G. Young. "The approximation of one matrix by another of lower rank". *Psychometrika* 1.3 (1936), pp. 211–218. DOI: 10.1007/BF02288367.

[19] H. Hotelling. "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology* 24.6 (1933), pp. 417–441. DOI: 10.1037/h0071325.

[20] C. Goodall. "Principal Component Analysis". *Technometrics* 30.3 (1988), pp. 351–352. DOI: 10.1080/00401706.1988.10488412.

[21] J. Wright et al. "Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices via Convex Optimization". *Advances in Neural Information Processing Systems*. Ed. by Y. Bengio et al. Vol. 22. Curran Associates, Inc., 2009.

[22] E. J. Candès et al. "Robust principal component analysis?" *Journal of the ACM* 58.3 (2011). DOI: 10.1145/1970392.1970395.

[23] Q. Ke and T. Kanade. "Robust L1 norm factorization in the presence of outliers and missing data by alternative convex programming". *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. 2005, pp. 739–746. DOI: 10.1109/CVPR.2005.309.

[24] T. Bouwmans and E. H. Zahzah. "Robust PCA via Principal Component Pursuit: A review for a comparative evaluation in video surveillance". *Computer Vision and Image Understanding* 122 (2014), pp. 22–34. DOI: https://doi.org/10.1016/j.cviu.2013.11.009.

[25] H. Xu, C. Caramanis, and S. Sanghavi. "Robust PCA via Outlier Pursuit". *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*. NIPS'10. Vancouver, British Columbia, Canada: Curran Associates Inc., 2010, pp. 2496–2504.

[26] M. Fazel. "Matrix rank minimization with applications". PhD thesis. PhD thesis, Stanford University, 2002.

[27] J.-F. Cai, E. J. Candes, and Z. Shen. *A Singular Value Thresholding Algorithm for Matrix Completion*. 2008. arXiv: 0810.3286 [math.OC].

[28] Z. Lin, M. Chen, and Y. Ma. *The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices*. 2010. DOI: 10.48550/arXiv.1009.5055. arXiv: 1009.5055v3 [math.OC].

[29] F. Mezzadri. *How to generate random matrices from the classical compact groups*. 2007. arXiv: math-ph/0609050 [math-ph].

[30] H. Abdi and L. J. Williams. "Principal component analysis". *WIREs Computational Statistics* 2.4 (2010), pp. 433–459. DOI: https://doi.org/10.1002/wics.101.

[31] A. Buniello et al. "Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery". *Nucleic Acids Research* 53.D1 (2024), pp. D1467–D1475. DOI: 10.1093/nar/gkae1128.

[32] H. K. Finucane et al. "Partitioning heritability by functional annotation using genome-wide association summary statistics". *Nature Genetics* 47.11 (2015), pp. 1228–1235. DOI: 10.1038/ng.3404.

[33] A. D. Grotzinger et al. "The Landscape of Shared and Divergent Genetic Influences across 14 Psychiatric Disorders". *medRxiv* (2025), pp. 2025–01.

[34] D. M. Jordan, M. Verbanck, and R. Do. "HOPS: a quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases". *Genome biology* 20 (2019), pp. 1–18.

[35] H. Yamamoto et al. "Statistical hypothesis testing of factor loading in principal component analysis and its application to metabolite set enrichment analysis". *BMC bioinformatics* 15 (2014), pp. 1–9.

[36] V. Trubetskoy et al. "Mapping genomic loci implicates genes and synaptic biology in schizophrenia". *Nature* 604.7906 (2022), pp. 502–508.

[37] D. Vuckovic et al. "The polygenic and monogenic basis of blood traits and diseases". *Cell* 182.5 (2020), pp. 1214–1231.

[38] S. Toikumo et al. "Multi-ancestry meta-analysis of tobacco use disorder identifies 461 potential risk genes and reveals associations with multiple health outcomes". *Nature human behaviour* 8.6 (2024), pp. 1177–1193.

[39] R. Karlsson Linnér et al. "Multivariate analysis of 1.5 million people identifies genetic associations with traits related to self-regulation and addiction". *Nature neuroscience* 24.10 (2021), pp. 1367–1376.

[40] M. J. Adams. "Genome-wide study of half a million individuals with major depression identifies 697 independent associations, infers causal neuronal subtypes and biological targets for novel pharmacotherapies". *medRxiv* (2024).

[41] K. Watanabe et al. "Functional mapping and annotation of genetic associations with FUMA". *Nature communications* 8.1 (2017), p. 1826.

[42] H. Zhou et al. "Multi-ancestry study of the genetics of problematic alcohol use in over 1 million individuals". *Nature Medicine* 29.12 (2023), pp. 3184–3192.

[43] H. J. Watson et al. "Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa". *Nature genetics* 51.8 (2019), pp. 1207–1214.

[44] A. J. Forstner et al. "Genome-wide association study of panic disorder reveals genetic overlap with neuroticism and depression". *Molecular psychiatry* 26.8 (2021), pp. 4179–4190.

[45] K. S. O'Connell et al. "Genomics yields biological and phenotypic insights into bipolar disorder". *Nature* (2025), pp. 1–12.

[46] D. Yu et al. "Interrogating the genetic determinants of Tourette's syndrome and other tic disorders through genome-wide association studies". *American Journal of Psychiatry* 176.3 (2019), pp. 217–227.

[47] J. Grove et al. "Identification of common genetic risk variants for autism spectrum disorder". *Nature genetics* 51.3 (2019), pp. 431–444.

[48] D. F. Levey et al. "Multi-ancestry genome-wide association study of cannabis use disorder yields insight into disease biology and public health implications". *Nature Genetics* 55.12 (2023), pp. 2094–2103.

[49] C. M. Nievergelt et al. "Genome-wide association analyses identify 95 risk loci and provide insights into the neurobiology of post-traumatic stress disorder". *Nature genetics* 56.5 (2024), pp. 792–808.

[50] N. I. Strom et al. "Genome-Wide Association Study of Obsessive-Compulsive Symptoms including 33,943 individuals from the general population". *Molecular Psychiatry* 29.9 (2024), pp. 2714–2723.

[51] D. Demontis et al. "Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains". *Nature genetics* 55.2 (2023), pp. 198–208.

[52] J. D. Deak et al. "Genome-wide association study in individuals of European and African ancestry and multi-trait analysis of opioid use disorder identifies 19 independent genome-wide significant risk loci". *Molecular Psychiatry* 27.10 (2022), pp. 3970–3979.

[53] M. Frank and P. Wolfe. "An algorithm for quadratic programming". *Naval Research Logistics Quarterly* 3.1-2 (1956), pp. 95–110. DOI: 10.1002/nav.3800030109.

[54] L. Ding and M. Udell. "Frank-Wolfe Style Algorithms for Large Scale Optimization". *Large-Scale and Distributed Optimization*. Ed. by P. Giselsson and A. Rantzer. Lecture Notes in Mathematics. Cham: Springer International Publishing, 2018, pp. 215–245. DOI: 10.1007/978-3-319-97478-1_9.

[55] M. Jaggi. "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization". *Proceedings of the 30th International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 1938-7228. PMLR, 2013, pp. 427–435.

[56] D. Yu and D. Kong. "Nuclear Norm Regularization". *WIREs Computational Statistics* 17.1 (2025). e70013 EOCS-615.R1, e70013. DOI: https://doi.org/10.1002/wics.70013.

[57] J.-F. Cai, E. J. Candès, and Z. Shen. "A Singular Value Thresholding Algorithm for Matrix Completion". *SIAM Journal on Optimization* 20.4 (2010), pp. 1956–1982. DOI: 10.1137/080738970.

[58] K.-C. Toh and S. Yun. "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems". *Pacific Journal of optimization* 6.615-640 (2010), p. 15.

[59] S. Lacoste-Julien et al. "Block-Coordinate Frank-Wolfe Optimization for Structural SVMs". *Proceedings of the 30th International Conference on Machine Learning*. Ed. by S. Dasgupta and D. McAllester. Vol. 28. Proceedings of Machine Learning Research 1. Atlanta, Georgia, USA: PMLR, 2013, pp. 53–61.

[60] M. Jaggi and M. Sulovský. "A simple algorithm for nuclear norm regularized problems". *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010, pp. 471–478.

[61] C. Mu et al. "Scalable Robust Matrix Recovery: Frank–Wolfe Meets Proximal Methods". *SIAM Journal on Scientific Computing* 38.5 (2016), A3291–A3317. DOI: 10.1137/15M101628X.

[62] S. Purcell et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". *The American journal of human genetics* 81.3 (2007), pp. 559–575.

[63] G. P. Consortium et al. "A global reference for human genetic variation". *Nature* 526.7571 (2015), pp. 68–74.

[64] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[65] L. del Bosque-Plata et al. "The role of TCF7L2 in type 2 diabetes". *Diabetes* 70.6 (2021), pp. 1220–1228.

[66] K. Dangas, A.-M. Navar, and J. J. Kastelein. "The effect of CETP inhibitors on new-onset diabetes: a systematic review and meta-analysis". *European Heart Journal-Cardiovascular Pharmacotherapy* 8.6 (2022), pp. 622–632.

[67] D. P. Howrigan et al. *Nealelab UK Biobank GWAS: v2*. 2023. DOI: 10.5281/ZENODO.8011557.

[68] M. Macauley et al. "Altered volume, morphology and composition of the pancreas in type 2 diabetes". *PloS one* 10.5 (2015), e0126825.

[69] R. Border et al. "Cross-trait assortative mating is widespread and inflates genetic correlation estimates". en. *Science* 378.6621 (2022), pp. 754–761.

Supplementary note

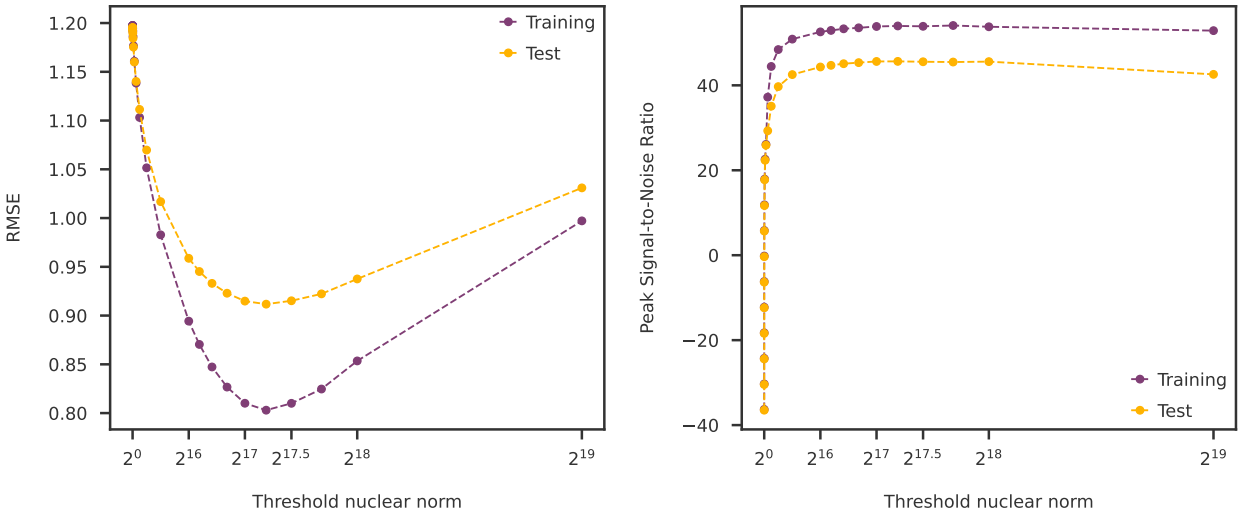

Figure S1: Cross-validation to obtain rank threshold for NNM-Sparse FW algorithm on PanUKB data. Both error metrics, root mean square error (RMSE, left panel) and peak signal-to-noise ratio (right panel) suggests a threshold of $r = 2^{17.25} = 155872$

.

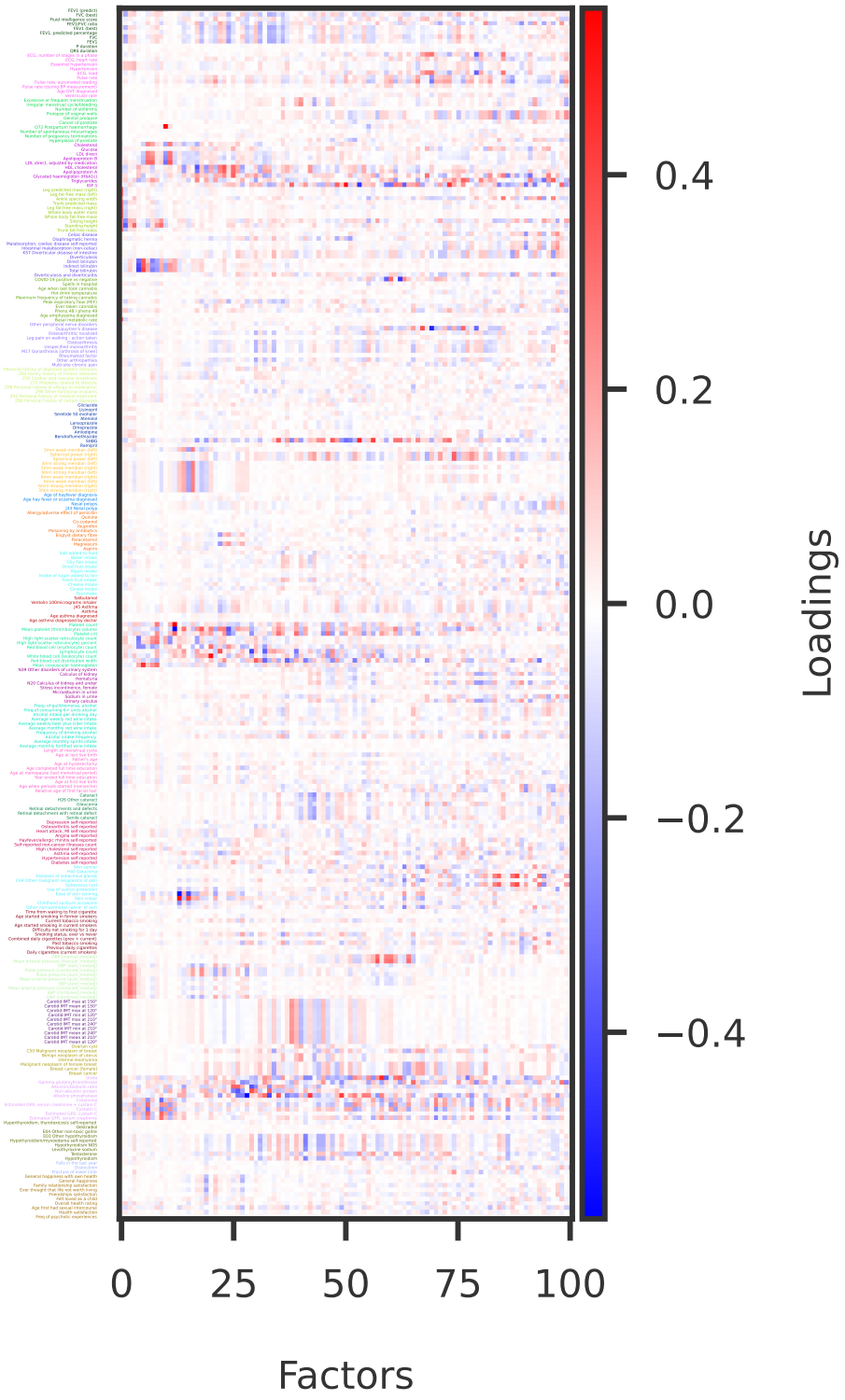

Figure S2: Contribution of some selected traits (shown on y-axis) on the top 100 factors (shown on x-axis) obtained from the application of `Clorinn` on the PanUKB dataset. The heatmap shows the magnitude of the loadings (color scale shown on the right). The traits were selected based on their estimated heritability ($h^2 \geq 0.001$), and a maximum of 10 traits were retained for each text-based disease category, see Figure 3 in main text.

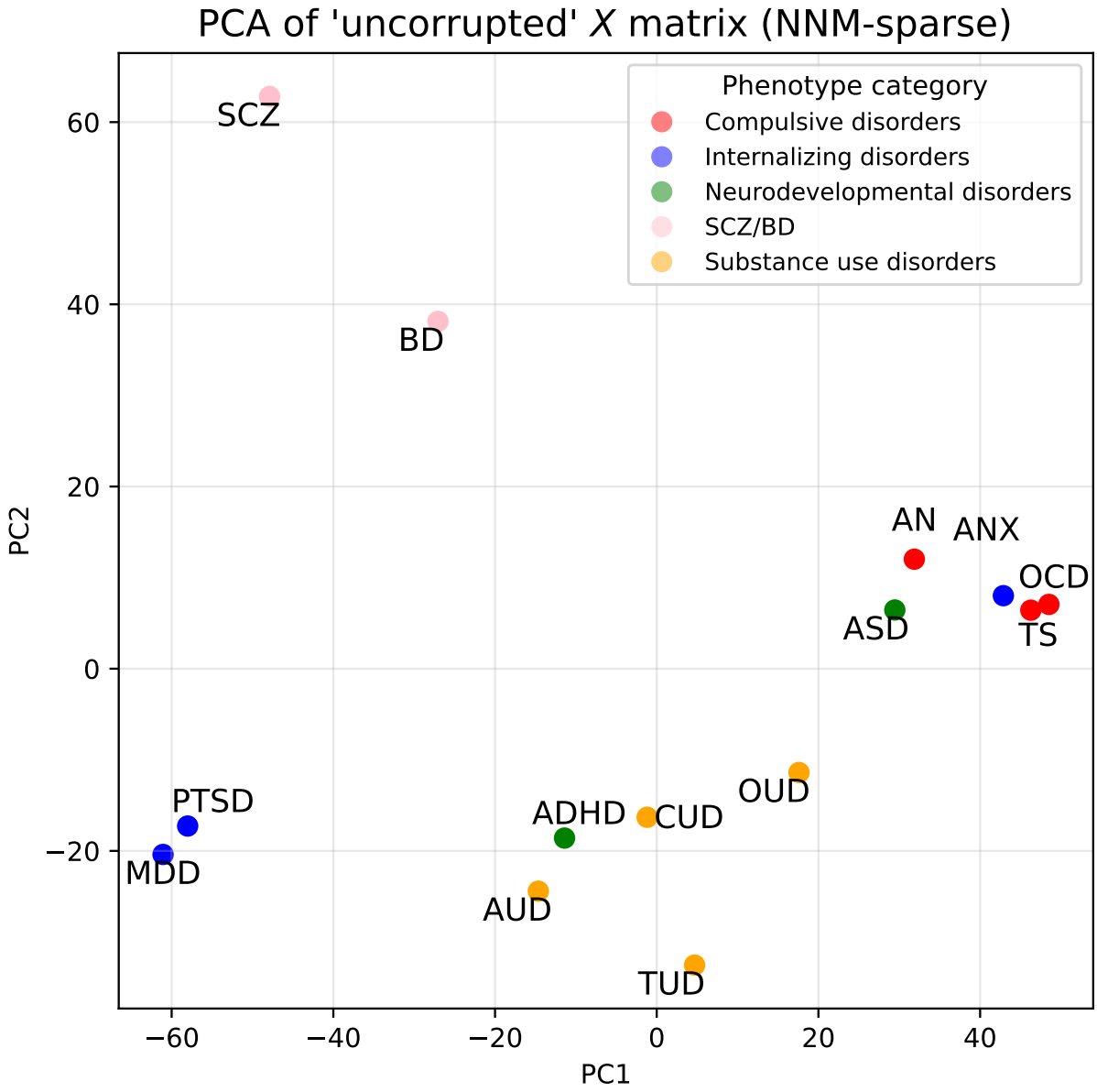

Figure S3: PC1 and PC2 of NNM-sparse-derived X for the application of `Clorinn` to 14 psychiatric traits. The nuclear norm constraint was chosed with a 5 fold cross validation procedure. Points are colored by their phenotype categories as described previously.

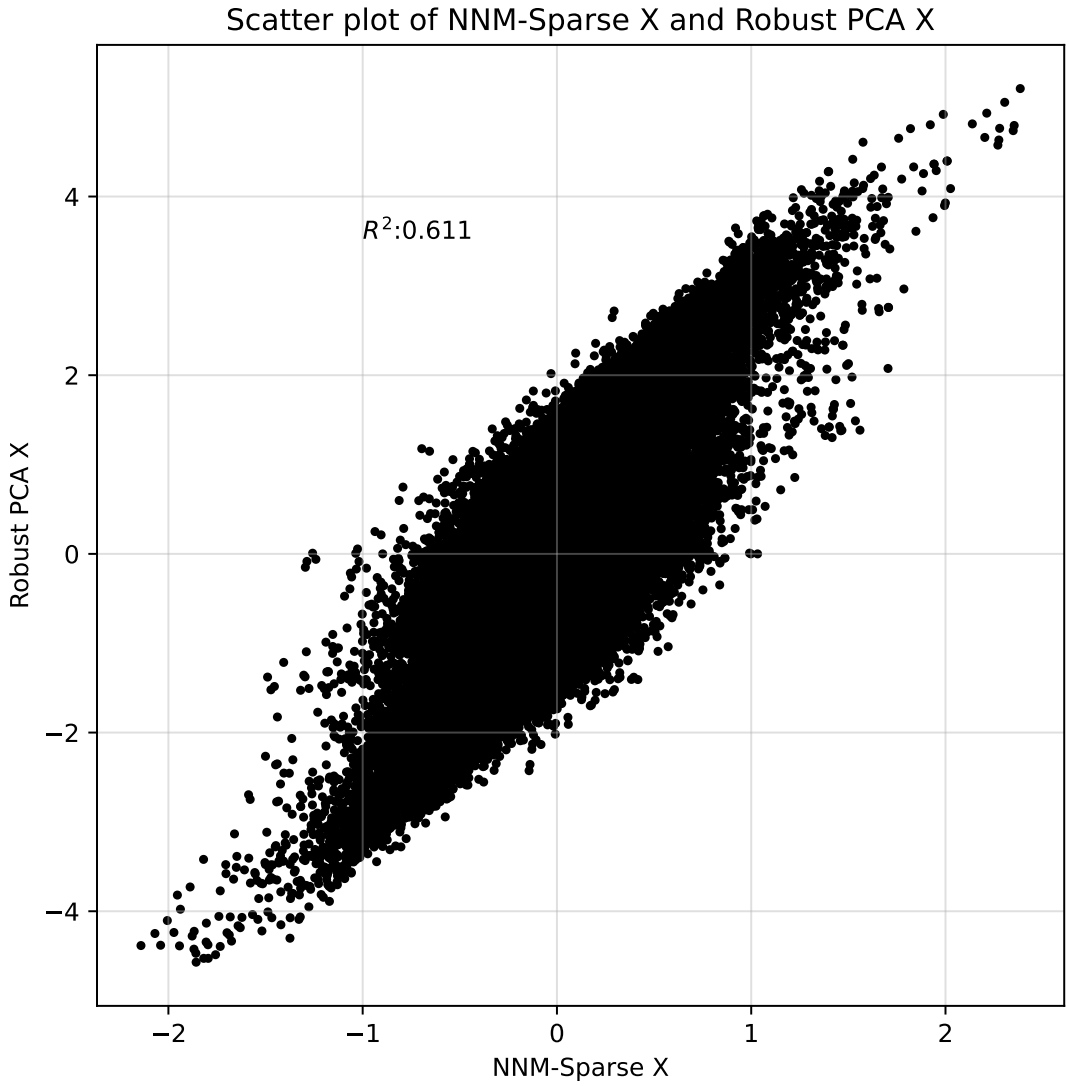

Figure S4: Scatter plot of X derived from robust PCA vs. X derived using NNM-sparse.

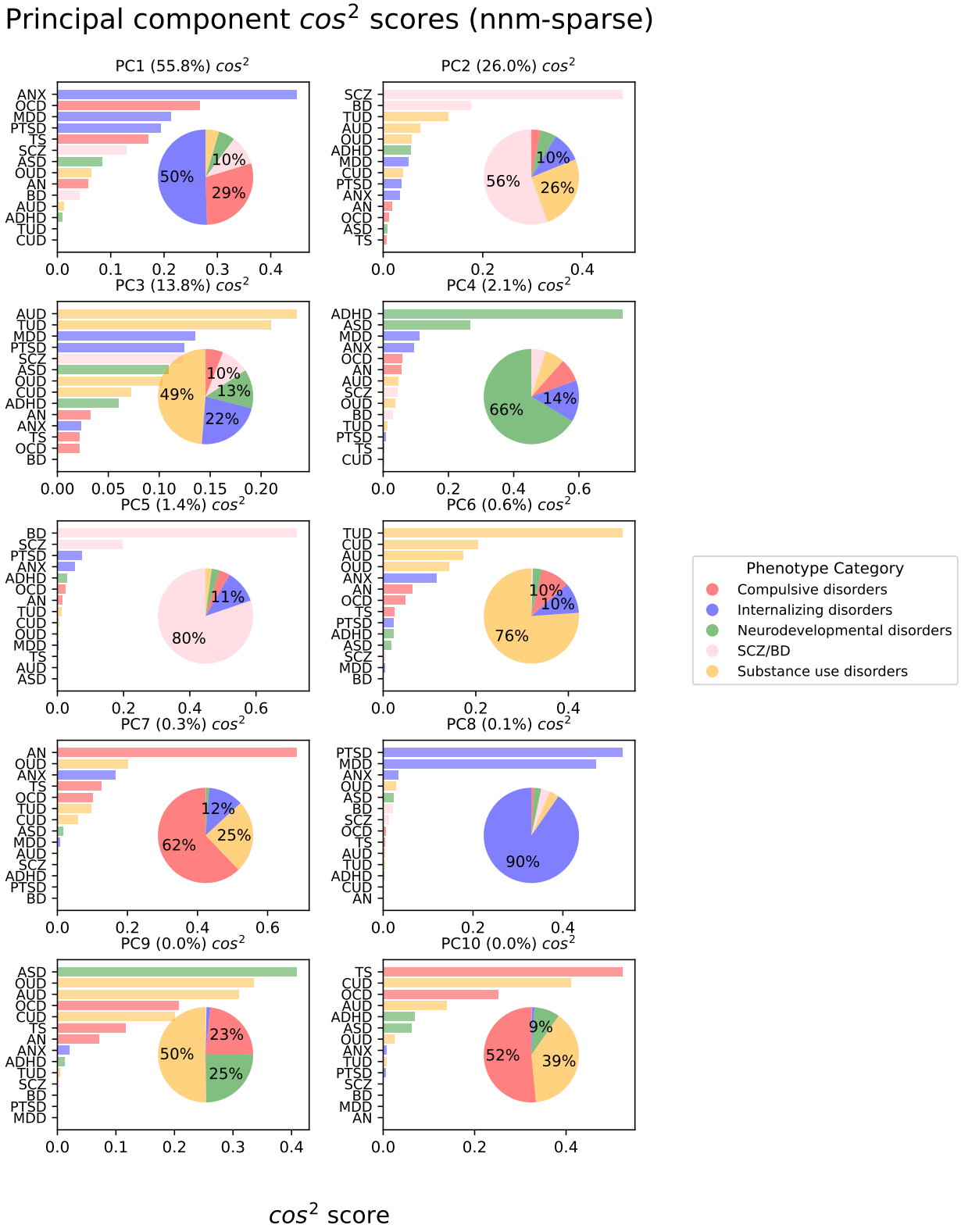

Figure S5: Squared cosine scores for factors derived from NNM-sparse X. Legend and details are inherited from Figure 5.

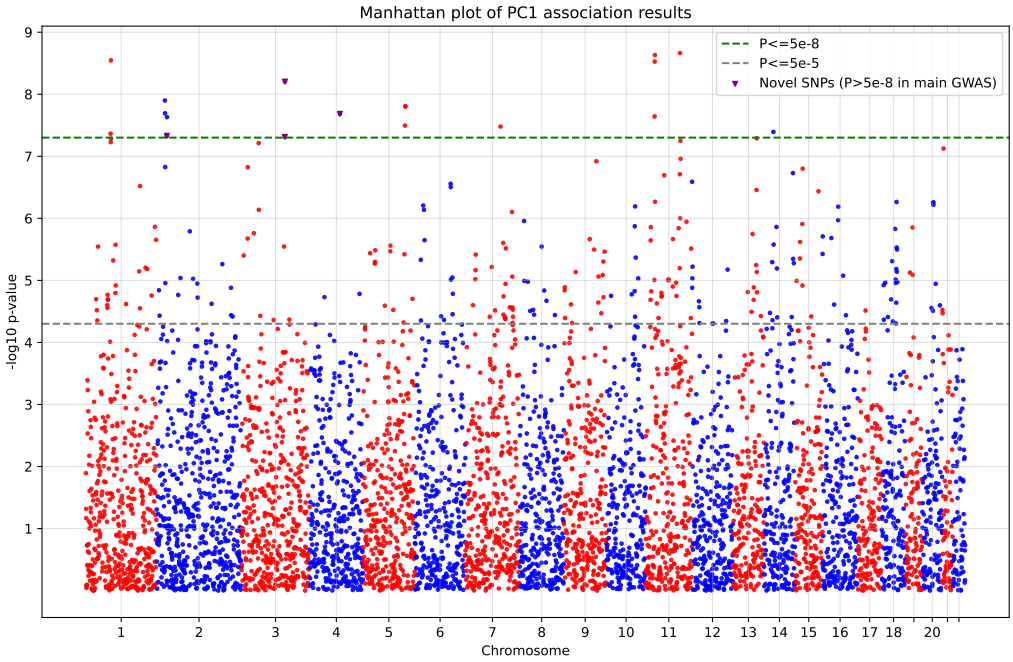

Figure S6: Manhattan plot of PC1 association statistics for $5,815$ SNPs from the application of `Clorinn` with the robust PCA algorithm. Statistics were derived by regressing PC score per phenotype against raw variant values across phenotypes from the input. Nominal and genome-wide significance lines are drawn with dashes; SNPs found to be GWS for PC1 activity but were not GWS for any of the top 4 loaded phenotypes (MDD, PTSD, OCD, and TS) for PC1 are colored in purple and marked with a triangle. These SNPs are discussed in the main text.