

Supplementary Notes for

New phenotype discovery method by unsupervised deep representation learning empowers genetic association studies of brain imaging

Khush Patel^{1*}, Ziqian Xie^{1*}, Hao Yuan⁵, Sheikh Muhammad Saiful Islam¹, Wanheng Zhang², Assaf Gottlieb¹, Han Chen^{2,3}, Luca Giancardo¹, Alexander Knaack⁶, Evan Fletcher⁶, Myriam Fornage^{2,4}, Shuiwang Ji⁵, Degui Zhi^{1**}

¹School of Biomedical Informatics, University of Texas Health Science Center, Houston, Texas 77030, USA;

²School of Public Health, University of Texas Health Science Center, Houston, Texas 77030, USA;

³Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas 77030, USA

⁴McGovern Medical School, University of Texas Health Science Center, Houston, Texas 77030, USA;

⁵Department of Computer Science and Engineering, Texas A&M University, College Station, Texas 77843, USA

⁶Department of Neurology and Imaging of Dementia and Aging (IDeA) Laboratory, University of California at Davis, Davis, California 95618, USA.

* Joint first-authors

**Corresponding author: Degui Zhi (Degui.Zhi@uth.tmc.edu)

Supplementary Note 1: Gene-based GWAS catalog analysis using FUMA

Rather than focusing on individual GWAS loci, we used FUMA pipelines to annotate genes and conduct functional enrichment of genes for ENDO GWAS summary statistics. For T2 ENDOs, first, by the positional gene prioritization pipeline (gene assignment pipeline) from FUMA, a total of 353 genes were found (Supplementary Table 10). Using the Gene2Func pipeline of FUMA which performed Gene set enrichment analysis (GSEA), we identified 402 gene sets (adjusted p-value <0.05) using 20260 background genes (protein coding genes) (Supplementary Table 11). To better understand which phenotypes are implicated by our ENDOs, we further analyzed the gene set enriched with various phenotypes of the GWAS Catalog used in FUMA. Out of all the phenotypes, the *Autism spectrum disorder or schizophrenia-related* gene set was the most significantly enriched gene set with an adjusted p-value of 6.74×10^{-38} . Additionally, gene sets from 19 other brain-related phenotypes from GWAS catalogs were also significantly enriched, including for example, General factor of neuroticism (adjusted p-value = 7.28×10^{-21}), Handedness (adjusted p-value = 1.52×10^{-10}), Parkinson's disease (adjusted p-value = 2.54×10^{-7}), and Alzheimer's disease in APOE e4- carriers (adjusted p-value = 3.93×10^{-4}).

The same procedure was followed for T1 ENDOs, and 432 genes were found (Supplementary Table 12). Three hundred forty-four gene sets were enriched (Supplementary Table 13). Overall, more brain-related phenotypes were associated with T1 ENDOs than T2 ENDOs (25 vs. 20). Similar to T2 ENDOs, *Autism spectrum disorder or schizophrenia-related* gene sets were the most significant gene sets. But more brain structure-related phenotypes like Hippocampal subfield CA4 volume, Total hippocampal volume, Hippocampal subfield CA3 volume, and Hippocampal subfield CA1 volume are enriched for T1 ENDOs. See **Gene annotation** for more information.

To aid further interpretation of the ENDOs with 25 T1 and 20 T2 brain-related GWAS Catalog phenotypes, we built a bipartite graph of ENDOs and phenotypes whose edges are weighted by the number of shared genes associated with the ENDO and the phenotype by FUMA. Although the bipartite graph is all interconnected (Supplementary Figures 26-27), we identified four robust bi-clusters of ENDOs and brain-associated phenotypes in T1 and three robust clusters in T2 using a mixture of clustering method to determine the most robust number of phenotype clusters across different resolutions (Quantum Clustering¹, see **Bi-Clustering ENDOs and phenotypes**) and bi-clustering of phenotypes and ENDOs (Spectral CoClustering², see **Bi-Clustering ENDOs and phenotypes**). Lateral ventricular volume in normal aging formed an independent cluster with multiple ENDOs associated only with it in both T1 and T2 (Supplementary Figures 26-29). Not surprisingly, *Schizophrenia*, *Autism spectrum disorder or schizophrenia* and *Social communication problems* cluster together in both T1 and T2. Interestingly, they cluster with *General cognitive activity*, *Intelligence* and *adventurousness* in T1 but *Intelligence* and *adventurousness* are in a different cluster in T2 (Supplementary Figures 26-29). Previous studies indicate that alleles for autism overlap broadly with alleles for high intelligence³. T1 has a unique cluster that does not appear in T2 and is driven by two ENDOs (65 and 115) and gene sets involved in several hippocampal locations (Dentate gyrus granule cell layer volume and molecular layer volume, Hippocampal subfield tail, CA1, CA3 and CA4 subfields and total Hippocampal volumes).

Gene annotation

FUMA⁴ was used to annotate our GWAS results functionally from February 26, 2022 till April 15, 2022. FUMA has integrated the functional annotation pipeline, including SNP annotation, gene mapping, and gene set enrichment analysis. We used the summary statistics of GWAS as input and identified the prioritized genes through functionally annotated variants. We used the default setting in the FUMA web interface to find positionally prioritized genes within a 10 kb window of functionally associated variants. To assess the functionally enriched phenotypes associated with our ENDOs, we performed gene set enrichment analysis in FUMA with all unique genes identified in our ENDOs.

Bi-clustering ENDOs and phenotypes

ENDOs were associated with phenotypes based on the number of shared genes that are mutually associated with them, where genes associated with ENDOs were retrieved from FUMA^{1,4} and genes associated with a phenotype were retrieved from GWAS Catalog. In order to bi-cluster the ENDOs and phenotypes, we used a two-step process. The first step was used to assess the number of robust clusters. We applied Quantum Clustering¹ with pre-processing involving top principal components from singular value decomposition and whitening of the data (normalizing to unit vector) across different resolutions, as described in Maignan et al⁵. Quantum clustering receives a resolution parameter (sigma - width of the potential function), ranging from zero to 1. We ran the algorithm with sigma ranging between 0.2 and 0.75, resulting in clusters between 10 with the highest resolution and one with the lowest. The number of clusters that remained the same across the widest set of sigma values was considered the most robust (four for T1, with sigma between 0.29 and 0.56 and three for T2 with sigma between 0.32 and 0.68). We then ran the Spectral Co-Clustering algorithm² with the number of robust clusters identified in step one to obtain the bi-clustering of the ENDOs and phenotypes. We used Cytoscape version 3.9.1 to plot the T1 and T2 networks formed by the association between ENDOs and phenotypes and color them based on the aforementioned bi-clusters.

References

1. Horn, D. & Gottlieb, A. Algorithm for data clustering in pattern recognition problems based on quantum mechanics. *Phys. Rev. Lett.* **88**, 018702 (2002).
2. Dhillon, I. S. Co-clustering documents and words using bipartite spectral graph partitioning. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01* Preprint at <https://doi.org/10.1145/502512.502550> (2001).
3. Crespi, B. J. Autism As a Disorder of High Intelligence. *Front. Neurosci.* **0**, (2016).
4. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
5. Maignan, A. & Scott, T. A comprehensive analysis of quantum clustering : Finding all the potential minima. *Int. J. Data Min. Knowl. Manag. Process* **11**, 33–54 (2021).

Supplementary Note 2: Learning rate for deep learning training

We used Adam optimizer with initial learning rate (lr) of 0.0005248074602497723 for T1 and 0.0003019951720402019 for T2 was used with a batch size of 62. Learning rate finder (LR finder) function of the Pytorch Lightning framework was used to obtain the initial optimal lr. LR finder increases lr after each mini-batch and plots a lr vs loss plot. The plot is then used to select optimum initial lr. Learning rate was reduced by half when validation set MSE loss plateaued for four epochs with a lower cap of lr/1000.