Vision Transformer Autoencoders for Unsupervised Representation Learning: Revealing Novel Genetic Associations through Learned Sparse Attention Patterns

Samia R. Islam **Tian Xia** **Wei He** **Ziqian Xie**[*] **Degui Zhi**[*]
D. Bradley McWilliams School of Biomedical Informatics,
The University of Texas Health Science Center at Houston
{ziqian.xie, degui.zhi}@uth.tmc.edu

Abstract

Linking genetic variation to human brain structure is a key step toward understanding the biological basis of cognition and disease. Progress in this area, however, has been limited by two major challenges: genome-wide association studies (GWAS) require very large cohorts, and imaging features are often predefined, restricting the discovery of novel associations. Data-driven representation learning offers a way to overcome these barriers by extracting informative features directly from imaging data without strong prior assumptions. Here, we present a pipeline that applies a Vision Transformer (ViT)-based autoencoder to derive 128-dimensional representations from T1-weighted brain MRI scans of 6,130 UK Biobank participants. These representations were used in GWAS to identify single nucleotide polymorphisms, which were further aggregated into genetic loci. To evaluate model performance, we introduce two metrics: recovery of previously reported brain-structure-associated loci and the total number of loci discovered. The ViT-based approach outperformed alternative methods on both measures. Feature interpretation also revealed that the model captured non-local anatomical patterns, such as hemisphere symmetry. Together, these results demonstrate the value of transformer-based architectures in discovering novel and robust imaging phenotypes for genetic discovery.

1 Introduction

Insights into understanding brain structural anatomy and related genetic signals have the potential to pave the way for breakthroughs in neuroscience and precision medicine. Most genome-wide association studies (GWAS) in the past have focused on direct phenotype measurements [García-Marín et al., 2024, Hibar et al., 2017, 2015] or statistical and mathematical representations of anatomical modeling and population genetics [Brun et al., 2009, Pol et al., 2006]. While these approaches have led to interesting discoveries, they rely on prior knowledge and constraints and may overlook more nuanced or complex patterns in the data. Recent advances in machine learning techniques offer more flexible and data-driven methods that can learn important features from raw imaging data. T1-weighted magnetic resonance imaging (MRI) brain scans capture detailed structural architecture and may be used as a rich source of data for advanced machine learning models.

Previous works on brain imaging GWAS have primarily focused on image-derived phenotypes (IDPs) via image processing pipelines [Elliott et al., 2018, Smith et al., 2021]. However, IDPs can miss

[*] Co-corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

subtle or complex patterns due to their reliance on predefined features and model assumptions, even more so in case of limited data availability.

Recent advances in machine learning techniques offer more flexible methods that can learn important features from raw imaging data [Hussain et al., 2018, Giger, 2018]. Yu et al. [2024] demonstrated using supervised deep learning to derive IDPs through labeled MRI data for brain imaging GWAS. The primary challenge with this approach is that labor-intensive labeled data for supervised deep learning carry the same biases from the human labeler.

Patel et al. [2024] used unsupervised deep learning to discover phenotypes for GWAS of brain imaging because this approach can bypass the challenges of relying on predefined features, segmentation, and manual annotation by enabling models to learn complex and relevant features directly from the raw image data, where the model creates an n-dimensional representation known as Unsupervised Deep learning derived Imaging Phenotypes (UDIPs).

Here, we propose an approach where an unsupervised learning vision transformer (ViT) model [Dosovitskiy et al., 2021] is used for brain imaging GWAS. This utilizes a different inductive bias from CNN architectures, which typically build representations through stacked layers of template matching, whereas the ViT uses patch-based self-attention to directly model potentially long-range interactions within the image. ViTs have demonstrated strong performance in various medical imaging applications, including feature extraction from brain MRI [Dhinagar et al., 2023], their architectural design facilitating the direct modeling of global context and long-range dependencies, even in the early layers [Dhinagar et al., 2023, Al-Hammuri et al., 2023].

This paper presents a novel Vision Transformer autoencoder to derive a new class of phenotypes, ViT-UDIPs, directly from brain MRI data. A genome-wide association study performed on these ViT-UDIPs reveals novel genetic loci that were not captured by previous analyses. Additionally, an interpretability framework is introduced to probe the model's learned pattern, offering insights into the ViT-derived phenotypes.

2 Methods

Our overall pipeline is shown in 1. The three major steps are data selection, ViT-Autoencoder (ViT-AE) model development, and GWAS analysis.

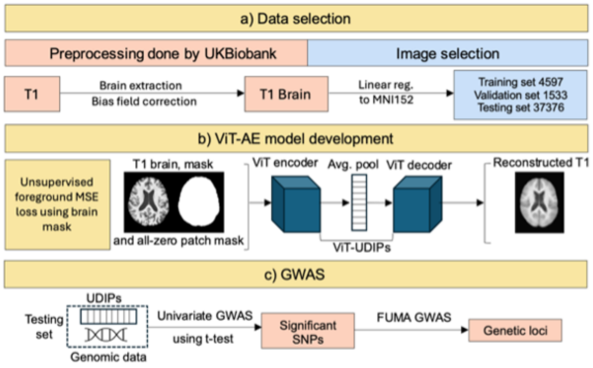

Figure 1: Overall pipeline of the study. a) T1 brain MRI preprocessed by UKBB were used for training, validation and testing. b) ViT-AE trained by background masked mean square error (MSE) loss of non-zero patches. c) Genetic loci discovered by univariate GWAS and FUMA GWAS.

2.1 Data Selection

The MRI scans were obtained from UKBB. Disjoint datasets were used for training, validation and GWAS to ensure robust model evaluation and prevent data leakage. 6,130 MRI scans were chosen for the model and split into 75/25 training/validation sets. 4,597 scans were used for training, and 1,533 scans were used for validation. The dataset used for GWAS consisted of 37,376 scans.

The linearly registered T1 brain MRI scans were of dimensions 182 x 218 x 182. To allow division into equal-sized patches, these scans were padded to 182 x 224 x 182. The background voxels were already set to 0 by UK Biobank preprocessing. For the foreground voxels, we applied a z-transform to standardize the intensity values, ensuring consistency across scans and facilitating better model performance.

2.2 ViT-AE Model

The ViT-AE architecture consisted of a ViT encoder, an average pooling layer and a ViT decoder. The features from the average pooling layer, which served as a concise representation of the input, became a distinct token embedding that fed into the decoder layers along with empty token embeddings solely containing positional encoding. These combined token embeddings passed through multiple transformer layers and were decoded by a linear projection to reconstruct the input patches. MSE loss was computed between predicted patches and original patches within the batch mask. The reconstructed layer assembled and reshaped the predicted patches from the decoder into 3D images of size 182 x 224 x 182.

2.3 GWAS

Related individuals were filtered out since our GWAS was conducted using a linear model, additionally, multiple visits were filtered where only scans from the first visits were retained. A total of 22,867 subjects were used for performing GWAS on the extracted ViT-UDIPs. GWAS was performed between the SNPs and the 128 ViT-UDIPs, with age (field ID 21003), age$\hat{2}$, sex (field ID 31), sex x age, sex x age$\hat{2}$, 10 genetic PC (field ID 22009), head size (field ID 25000), inverted contrast-to-noise ratio (field ID 25735), head position in scanner (field ID 25756-25758), scanner table position (field ID 25759), location of the assessment center (field ID 54) and date of attending assessment center (field ID 53) as covariates FastGWA[Jiang et al., 2019], implemented in GCTA (Genome-wide Complex Trait Analysis, Version 1.94.1), was used to perform linear mixed model association analyses based on a sparse kinship matrix provided by the UK Biobank. The Bonferroni-corrected significance threshold was set to 5e-8/128. We created a summary statistics file with only the most significant p-value (minP) for each SNP. This file was uploaded and run on FUMA SNP2GENE [Watanabe et al., 2017] and the generated GenomeRiskLoci.txt file was then used to obtain the clumped loci.

3 Results

3.1 Model Performance

The average validation loss of the ViT-AE model remained at approximately 0.22 after 200 epochs. Training was stopped at epoch 300 when we did not see any decrease in validation loss and improvement in the number of discovered loci. We further performed training optimizations for the ViT-AE model. For training, the optimal configuration was found to be the AdamW optimizer combined with a cosine learning rate scheduler and an enlarged batch size of 16. This setup yielded the highest number of discovered loci (63) and the greatest overlap with loci from the BIG40 panel. This configuration is adopted as the current setting, and all subsequent results are based on this setup.

Table 1: Model optimization

Base:ViT	Val loss	SSIM	Best number of loci discovered across epochs and replications	Number of loci hit on the IDP loci panel (BIG40)
StepLR scheduler	0.25	0.79	31	19
Cosine scheduler	0.21	0.81	56	28
Increase batch size	0.22	0.80	**63**	**29**

3.2 GWAS

We analyzed FUMA GWAS results for epochs 300, as they represented the epochs with the most number of loci discovered. Notably, we discovered 24 new loci associated with brain structure that were not previously reported by CNN-based UDIP approach by Patel et al. [2024]. We investigated the lead SNPs in these loci. 2 shows the previously unreported loci and the nearest gene to the leadSNP on a Manhattan plot. Four of the newly identified loci are located on chromosome 2, and three on chromosomes 4, 6, and 14, respectively. An example of how to interpret the plot is as follows: the SLC39A8 gene on chromosome 4 was found to be significantly associated with the ViT-UDIP feature at dimension 95, suggesting a potential link between the genetic activity of SLC39A8 and the representation captured in this embedding dimension.

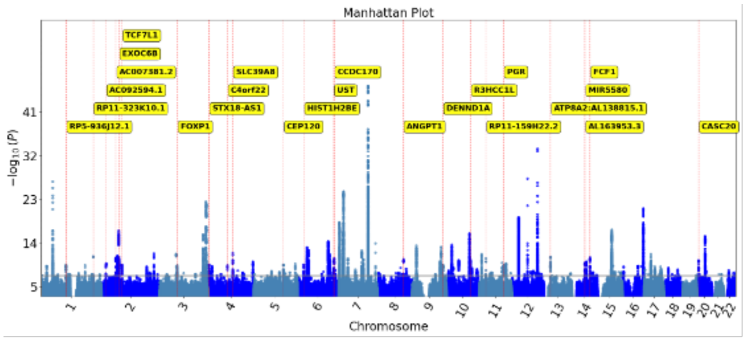

Figure 2: New loci discovered from ViT-based UDIP approach.

14 loci out of 24 had been previously associated with brain structure. 10 loci out of 24 had not been previously documented in the GWAS catalog as being associated with brain architecture.

3.3 Attention Head Analysis

We investigated the attention scores of the trained model and found some interesting patterns. From the very first attention layer, the model started to focus on a subset of patches across different heads. The attention pattern of each head was extremely sparse (3), while different heads at different layers focused on different patches. This subset of patches consisted of less than one-third of the total patches, with all other patches attending to them.

We found that this behavior arises from the specific embeddings learned by the model. The input embedding matrix maps most patches to a similar direction—referred to as **audience tokens**—while a small subset remains only weakly correlated with both the majority and each other, which we denote as **spotlight tokens**. (3).

At each attention head, the (unnormalized) attention score between two tokens $\mathbf{X}_1$ and $\mathbf{X}_2$ can be written as

$$\mathbf{X}_1^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{X}_2.$$

This can be interpreted as first rotating them differently (based on the rotation matrices formed by the left and right singular vectors of $\mathbf{W}_K^\top \mathbf{W}_Q$), followed by projecting the rotated $\mathbf{X}_1$ and $\mathbf{X}_2$ onto the dimensions with nonzero singular values and finally computing a weighted inner product.

Specifically, consider the singular value decomposition

$$\mathbf{W}_K^\top \mathbf{W}_Q = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top,$$

where $\mathbf{U}$ and $\mathbf{V}$ are rotation matrices, and $\boldsymbol{\Sigma}$ is a diagonal scaling matrix.

Then, the interaction between $\mathbf{X}_1$ and $\mathbf{X}_2$ can be written as

$$(\mathbf{P}\mathbf{U}^\top \mathbf{X}_1)^\top \boldsymbol{\Sigma}_r (\mathbf{P}\mathbf{V}^\top \mathbf{X}_2),$$

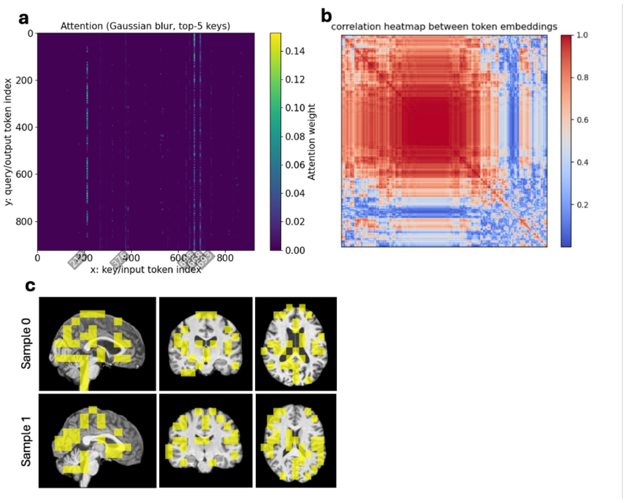

Figure 3: Investigation of unique patterns from ViT. a) Softmax attention score at attention layer 1 head 1 between 957 tokens, image is blurred using a Gaussian filter with sigma=2 for better visual quality. b) Correlation heatmap between 957 tokens input embeddings, we use hierarchical clustering with average linkage to reorder the dimensions of the heatmap. c) Visualization of spotlight tokens across brain slices. Highlighted regions (yellow).

where $\mathbf{P}$ is a wide identity matrix that selects the first d_h coordinates of the input (the dimension of the attention head), and $\mathbf{\Sigma}_r$ is the reduced singular value matrix consisting only of the nonzero entries of $\mathbf{\Sigma}$, with shape $d_h \times d_h$.

As a result, at each attention head, audience tokens tend to focus on one or a few nearby "spotlight" tokens that happen to align with them through rotations, leading to sparse attention patterns and accumulation of information from these spotlight patches.

We then visualize the distribution of spotlight tokens across different brain slices. The highlighted regions in 3 indicate key locations where the model focuses its attention. We observe that these spotlight patches tend to cluster around deep sulci, align along the boundaries of the ventricles, and appear in structures such as the cerebellum and brainstem. Additionally, some spotlight patches appear in corresponding regions across hemispheres, suggesting a tendency to capture bilaterally relevant anatomical features. The distribution of attention is sparse and concentrated around major structural landmarks, indicating that the model prioritizes these regions for information aggregation.

4 Conclusion

We developed an approach that integrated unsupervised deep representation learning, specifically using a ViT-AE model, for brain imaging GWAS. Our ViT-AE pipeline successfully identified 24 genetic loci associated with brain structure that were missed by a previous CNN-based approach, with 10 of these having no prior associations with brain morphology in the GWAS Catalog. Furthermore, our analysis of the model's attention mechanism revealed the data-driven emergence of "spotlight tokens," which receive a disproportionate amount of attention and correspond to anatomically informative regions with high structural variance, such as the boundaries between different types of tissue (e.g. gray and white matter), the edges of ventricles or the complex folding of deep sulci. The model appears to have learned to focus its attention on these information-rich patches.In conclusion, this robust pipeline demonstrates significant promise in uncovering genetic signals from brain MRI data by leveraging the ViT's ability to capture intricate structural patterns and spatial relationships.

Acknowledgments and Disclosure of Funding

Our analysis was approved by UTHealth Houston committee for the protection of human subjects under No. HSC-SBMI-20-1323. UKBB has secured informed consent from the participants in the use of their data for approved research projects. UKBB data was accessed via approved project 24247.

This work was supported by grants from the National Institute on Aging U01AG070112 and R01AG081398

References

Khaled Al-Hammuri, Fayez Gebali, Ahmad Kanan, and Indu T. Chelvan. Vision transformer architecture and applications in digital health: a tutorial and survey. *Visual Computing for Industry, Biomedicine, and Art*, 6, 2023. doi: 10.1186/s42492-023-00140-9.

C. C. Brun et al. Mapping the regional influence of genetics on brain structure variability - a tensor-based morphometry study. *Neuroimage*, 48, 2009.

N. J. Dhinagar, S. I. Thomopoulos, E. Laltoo, and P. M. Thompson. Efficiently training vision transformers on structural mri scans for alzheimer's disease detection. In *Proceedings of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 2023. doi: 10.1109/EMBC40787.2023.10341190.

Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Lloyd T. Elliott, Kevin Sharp, Fidel Alfaro-Almagro, Saad J. Shi, Kristoffer L. Miller, Gwenaëlle Douaud, Jonathan Marchini, and Stephen M. Smith. Genome-wide association studies of brain imaging phenotypes in uk biobank. *Nature*, 562(7726):210–216, 2018. doi: 10.1038/s41586-018-0571-7. URL `https://doi.org/10.1038/s41586-018-0571-7`.

L. M. García-Marín et al. Investigating the genetic relationship of intracranial and subcortical brain volumes with depression and other psychiatric disorders. *Imaging Neuroscience*, 2:1–16, 2024.

Maryellen L. Giger. Machine learning in medical imaging. *Journal of the American College of Radiology*, 15, 2018.

D. P. Hibar et al. Common genetic variants influence human subcortical brain structures. *Nature*, 520:224–229, 2015.

D. P. Hibar et al. Novel genetic loci associated with hippocampal volume. *Nature Communications*, 8:13624, 2017.

L. Hussain et al. Prostate cancer detection using machine learning techniques by employing combination of features extracting strategies. *Cancer Biomarkers*, 21, 2018.

Longda Jiang, Zhili Zheng, Tianqi Qi, Kathryn E. Kemper, Naomi R. Wray, Peter M. Visscher, and Jian Yang. A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, 51(12):1749–1755, 2019. doi: 10.1038/s41588-019-0530-8. URL `https://doi.org/10.1038/s41588-019-0530-8`.

K. Patel et al. Unsupervised deep representation learning enables phenotype discovery for genetic association studies of brain imaging. *Communications Biology*, 7:414, 2024.

H. E. H. Pol et al. Genetic contributions to human brain morphology and intelligence. *Journal of Neuroscience*, 26, 2006.

S. M. Smith et al. An expanded set of genome-wide association studies of brain imaging phenotypes in uk biobank. *Nature Neuroscience*, 24:737–745, 2021.

Kyoko Watanabe, Emre Taskesen, Arjen Van Bochoven, and Danielle Posthuma. Functional mapping and annotation of genetic associations with fuma. *Nature Communications*, 8, 2017.

S. Yu, J. Wu, Y. Shao, D. Qiu, and Z. S. Qin. A novel classification framework for genome-wide association study of whole brain mri images using deep learning. *PLoS Computational Biology*, 20:e1012527, 2024.