Vision Transformer Autoencoders for Unsupervised Representation Learning: Capturing Local and Non-Local Features in Brain Imaging to Reveal Genetic Associations

Samia R. Islam, MS, Ziqian Xie, PhD, Wei He, Degui Zhi, PhD
The University of Texas Health Science Center at Houston
D. Bradley McWilliams School of Biomedical Informatics

Abstract
The discovery of genetic loci associated with brain architecture can provide deeper insights into neuroscience and improved personalized medicine outcomes. Previously, we designed the Unsupervised Deep learning-derived Imaging Phenotypes (UDIPs) approach to extract endophenotypes from brain imaging using a convolutional (CNN) autoencoder, and conducted brain imaging GWAS on UK Biobank (UKBB). In this work, we leverage a vision transformer (ViT) model due to a different inductive bias and its ability to potentially capture unique patterns through its pairwise attention mechanism. Our approach based on 128 endophenotypes derived from average pooling discovered 10 loci previously unreported by CNN-based UDIP model, 3 of which were not found in the GWAS catalog to have had any associations with brain structure. Our interpretation results demonstrate the ViT's capability in capturing non-local patterns such as left-right hemisphere symmetry within brain MRI data, by leveraging its attention mechanism and positional embeddings. Our results highlight the advantages of transformer-based architectures in feature extraction and representation for genetic discovery.

Keywords: vision transformer, deep learning, brain MRI, medical imaging, GWAS

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, such as brain region volumes, cortical surface areas, and cortical thickness, estimated using traditional software tools like FSL and FreeSurfer[1–5], or statistical and mathematical representations of anatomical modeling and population genetics[6–11]. 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. Deep learning has been widely applied to medical imaging tasks such as disease prediction, detection and diagnosis, segmentation of abnormalities, and medical image synthesis[12]. Deep neural network models are capable of recognizing meaningful features and patterns from medical images that are difficult for clinicians to observe, leading to more accurate and efficient medical analyses[13].

Previous works on brain imaging GWAS have primarily focused on image-derived phenotypes (IDPs) via image processing pipelines[14,15]. IDPs are generally used to capture high level features of brain anatomy, such as structural, functional, or connectivity patterns. These features are typically extracted through complex image processing steps to quantify specific aspects of brain structure or function. However, IDPs can miss subtle or complex patterns due to

their reliance on predefined features and model assumptions, even more so in case of limited data availability. Additionally, variability in image quality and challenges in interpretation can impact the accuracy and applicability of these phenotypes across different populations or imaging modalities. IDPs often rely on segmentation as a critical step and in the case of brain imaging, this is difficult to achieve because of the brain's complex anatomy, boundary ambiguity, and artifacts or noise induced from motion during the scanning process.

Recent advances in machine learning techniques offer more flexible methods that can learn important features from raw imaging data[16–20]. There are attempts using supervised deep learning to derive IDPs through labeled MRI data for brain imaging GWAS[21]. The primary challenge with this approach is that labeled data for supervised deep learning carry the same biases from the human labeler. Also, the labeling is a labor intensive task, especially for large datasets.

Using unsupervised deep learning to discover phenotypes for GWAS of brain imaging 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) and then performs GWAS on these UDIPs[22]. We propose a similar approach where an unsupervised learning vision transformer (ViT)[23] model is used for brain imaging GWAS instead of convolutional neural network (CNN). This addresses the limitations of CNNs which are known to capture local patterns but struggle to model global relationships across an image[24]. ViTs are a particular type of neural network and have performed efficiently in feature extraction from medical imaging data, such as MRI[25], because of its ability to understand relationships among features from local regions, as well as global context[23,26]. Self-attention mechanism in transformer architecture allows the ViT to capture information across the entire image, even in the early layers[23].

We designed a pipeline that first, utilized a ViT autoencoder (ViT-AE) to extract UDIPs from brain MRI scans and second, performed GWAS to discover the genetic signals that were associated with these UDIPs. Henceforth, these UDIPs will be referred to as ViT-UDIPs. The ViT-AE consisted of an encoder to learn contextual embeddings for each patch, an average pooling layer that induced a bottleneck by collapsing all the patch representations into one, and a decoder to generate outputs from the compact representations. The model was trained to minimize the mean squared error (MSE) on patches within the brain mask. A perturbation-based decoder interpretation (PerDI) step mapped the ViT-UDIPs to specific brain regions[22]. These extracted ViT-UDIPs were then passed through a linear layer to perform univariate GWAS to find statistically significant single nucleotide polymorphisms (SNPs) that were clumped into loci using Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA GWAS)[27]. We make the following contributions: our pipeline discovered new loci associated with brain structure using unlabeled data and brain imaging GWAS.

Methods

Our overall pipeline is shown in Figure 1. The four major steps are data selection, ViT-AE model development, GWAS, and finally, SNP-ViT-UDIP to brain mapping via the PerDI approach.

Data Selection

The MRI scans were obtained from UKBB. The preprocessing done by UKBB involved brain extraction and bias field correction. Linear registrations were performed to align the imaging

data into MNI152 space. Disjoint datasets were used for training, validation and inferencing 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 inferencing consisted of 37,376 scans.

The data-loading stage included pre-processing steps (see Supplementary Figure 1) on the linearly registered T1 brain MRI scans of dimensions 182 x 218 x 182. In order to make them suitable to be divided into equal-sized patches, they were padded to be of size 182 x 224 x 182. Masking enabled isolation of the brain region, effectively removing the background and retaining only the foreground containing relevant anatomical features. The resulting brain mask was normalized using z-transform to standardize the intensity values, ensuring consistency across scans and facilitating better model performance.

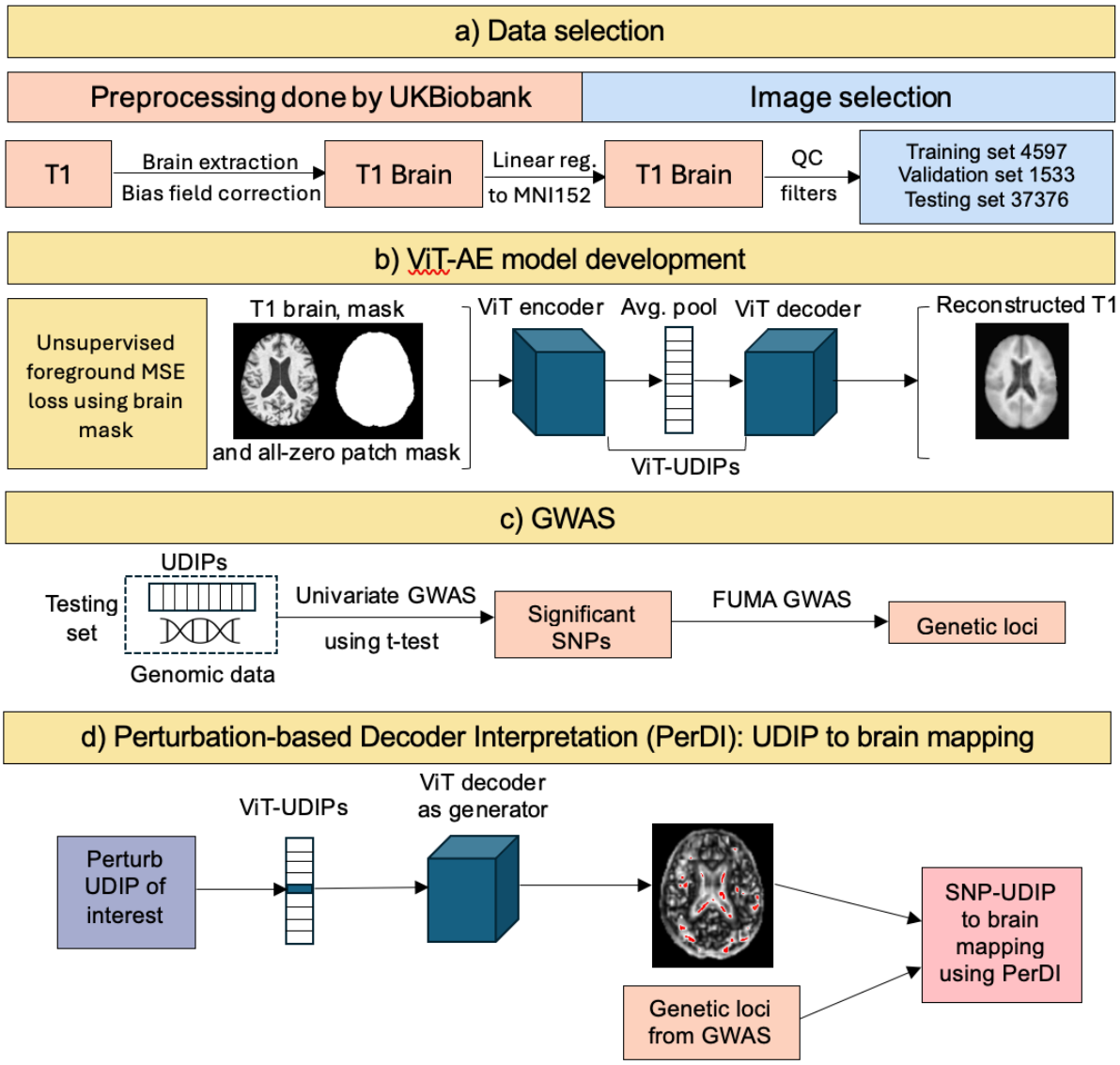

Figure 1 | Overall pipeline of the study. a) T1 brain MRI preprocessed by UKBB were used for training, validation and testing. b) ViT-AE trained by background masked mean square error

(MSE) loss of non-zero patches. c) Genetic loci discovered by univariate GWAS and FUMA GWAS. d) SNP-ViT-UDIP to brain mapping performed using Perturbation-based Decoder Interpretation (PerDI).

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 were treated as a token embedding that fed into the decoder layers along with empty token embeddings solely containing positional encoding. These token embeddings passed through multiple transformer layers and were decoded by a linear projection to reconstruct the input patches. Figure 2 shows the block diagram of the overall ViT architecture. The T1 brain MRI images were segmented into non-overlapping patches of size 14 x 16 x 14 for a total of 13 x 14 x 13 patches. We optimized the model by eliminating those patches which were zero across all three axes of the images[28]. Notably, this reduced the encoder workload by approximately 60%. For the purpose of distinguishing these masked patches from brain mask in our pipeline, we referred to them as *batch mask*. The hyperparameters used in the ViT-AE model are given in Table 1.

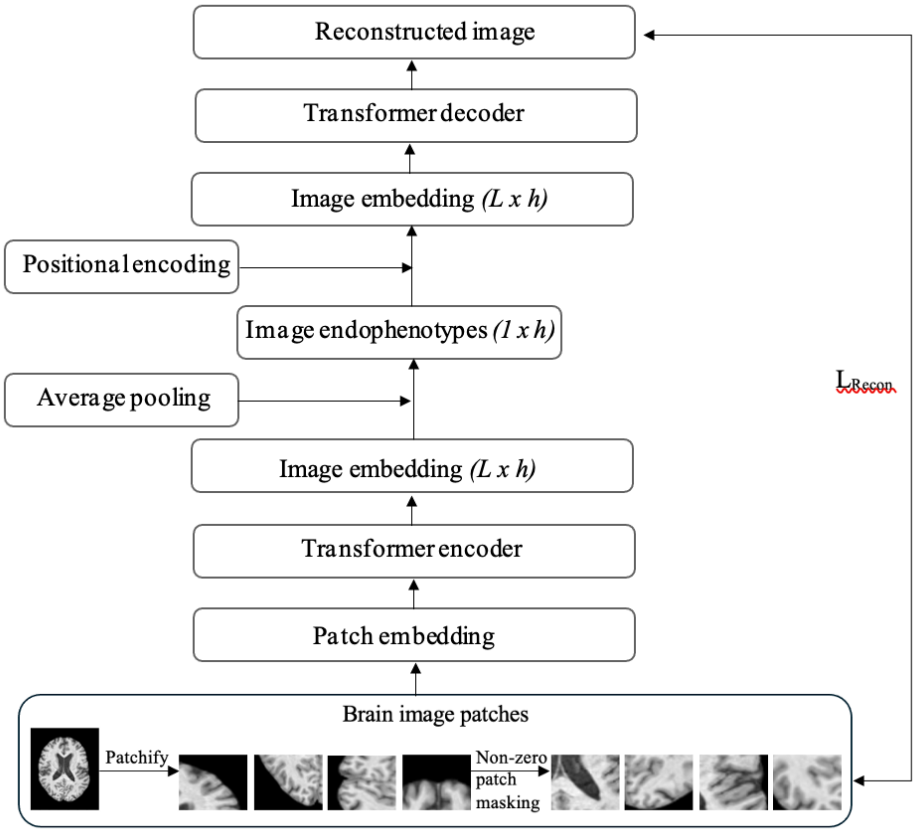

Figure 2 | Block diagram of ViT-AE model, showing the generation of the endophenotypes from the average pooling layer which are then used by the ViT decoder for prediction of output patches. The encoder creates image embeddings of size *L x h* which are averaged across *h* dimensions to create the ViT-UDIPs. Empty patch embeddings with positional encoding are used to re-create *L x h* image embeddings for the decoder. Reconstruction loss is denoted by L_{Recon}.

Parameter	Setting
Patch size	14 x 16 x 14
Total number of patches	12 x 14 x 13
Batch size	2
Encoder embedding dimension	128
Number of attention heads	8
Number of layers	12 encoder, 12 decoder
Optimizer	AdamW
Learning rate	LR scheduler, lr=0.001, gamma=0.5, step size=3, minimum clipped at 1e-5
Decoder embedding dimension	64

Table 1 | ViT-AE hyperparameters

The main modules within the ViT-AE model were encoder, average pooler, and decoder. The ViT encoder produced learned representations of input data from flattened patches. These latent outputs were then average pooled to reduce their dimensionality and capture a global representation of the data. The ViT decoder performed patch-wise processing of extracted ViT-UDIPs for prediction of output patches. MSE loss was computed between predicted patches and original patches, according to batch mask. Figure 3 shows the ViT-AE modules and high-level pipeline of image and latent flow from encoder to decoder, as well as the final stage of loss computation. MSE loss was subsequently backpropagated to optimize model parameters during training.

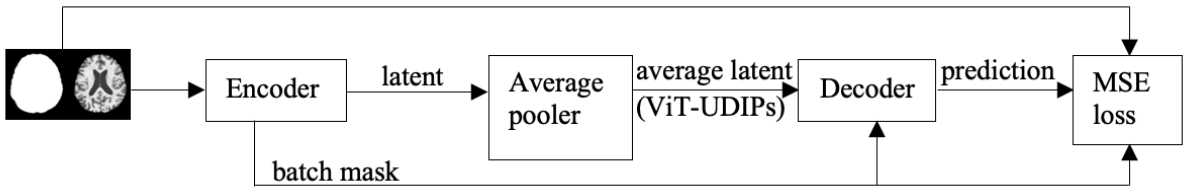

Figure 3 | Image and latent flow through ViT-AE modules

The reconstructed layer comprised of one-dimensional predicted patches from the decoder being reorganized into 3D images of size 182 x 224 x 182, as shown in Figure 4.

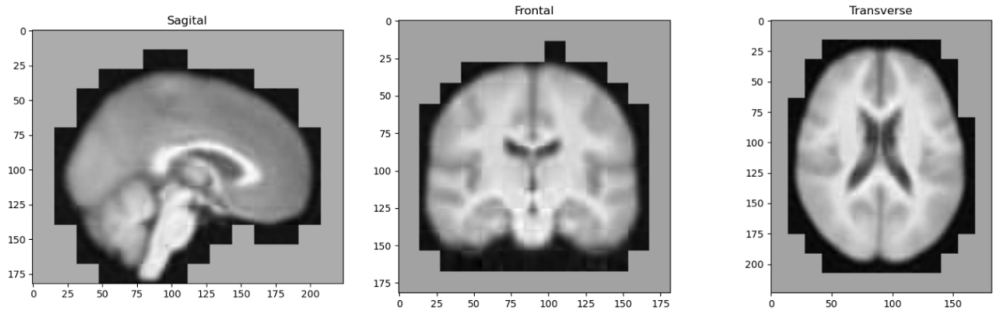

Figure 4 | Reconstructed brain images with batch mask. Note that the gray areas in the images correspond to the patches which were masked due to containing all-zero values across all three image axes.

The ViT-AE mode was trained for 400 epochs on training set and validated on a separate validation set. Inferencing was performed on a disjoint test set using the trained model and the ViT-UDIPs extracted were saved and used for GWAS in the downstream pipeline.

GWAS

Related individuals and multiple visits were further filtered out, where only scans from last visits were retained. A total of 22,250 subjects were used for performing GWAS on the extracted ViT-UDIPs. Linear regression was performed between the SNPs and the ViT-UDIPs, with age (UKBB field 21003), sex (UKBB field 31), 10 genetic principal components (UKBB field 22009), ethnicity (UKBB field 21000), T1 inverted contrast to noise ratio (UKBB field 25735), assessment center (UKBB field 54) as covariates, using a GPU-based GWAS tool developed in-house, where the p-value 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 GWAS SNP2GENE tool[27] and the generated GenomeRiskLoci.txt file was then used to obtain the SNPs clumped into loci.

The top five ViT-UDIPs with the highest number of significant SNPs were used for the next step, where noise was added to perturb these ViT-UDIPs for decoder interpretation.

PerDI

To visualize the brain regions associated with the extracted ViT-UDIPs, we utilized a perturbation-based approach[22]. For 500 subjects from the test set, we added 1 standard deviation of noise to each of the five dimensions out of the 128 ViT-UDIPs, with the highest number of significant SNPs. The decoder was then used to reconstruct the images based on the perturbed ViT-UDIPs, followed by a voxel based paired t-test as shown in Figure 5.

A mask based on a brain template linearly registered to the T1 MNI152 space was applied to the t-test results, followed by a Gaussian filter (sigma=1) for smoothing. The five resulting images, one for each of the perturbed ViT-UDIPs, were saved as NIFTI files for visualization using FSLeyes. These NIFTI files were also used to find the regions of intersection with the Harvard-Oxford Cortical and Subcortical Structural Atlases (RRID:SCR_001476)[29–33] and MNI Structural Atlas[29,34,35].

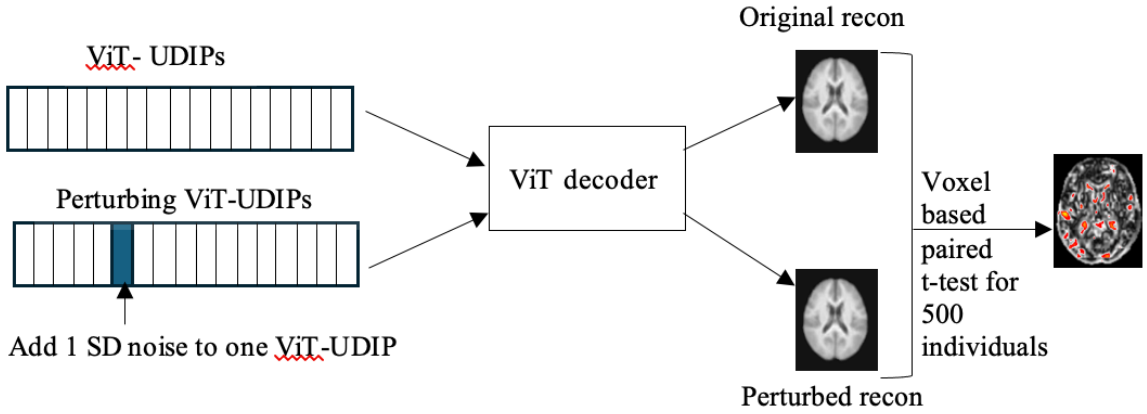

Figure 5 | Perturbation-based Decoder Interpretation (PerDI)

Attention Head Analysis

We analyzed patches attended by the attention heads in our ViT-AE model in order to visually determine the positions of these patches and the relative distances. The primary motive behind this was to identify pairs of patches attending to each other, thereby uncovering relationships between different regions of the brain. By examining the attention maps, we aimed to understand how the model prioritizes certain brain structures and identifies spatial correlations across the brain.

For each attention head, we obtained attention scores for all patches within the batch mask and considered only the patches with the highest attention scores. From these, we found the voxel coordinates of patches and reconstructed them to the voxel coordinates of the MRI scans, taking into account the corresponding batch masks. These patches were then superimposed on brain atlases to visualize the regions of attention.

Results

Model Performance

The average validation loss of the ViT-AE model remained at approximately 0.253 after 180 epochs, as shown in Figure 6. Average loss on test set was 0.2598. The goal of our ViT-AE model was to extract features which would yield the highest number of loci through GWAS. Training was stopped when we did not see any improvement in the number of discovered loci.

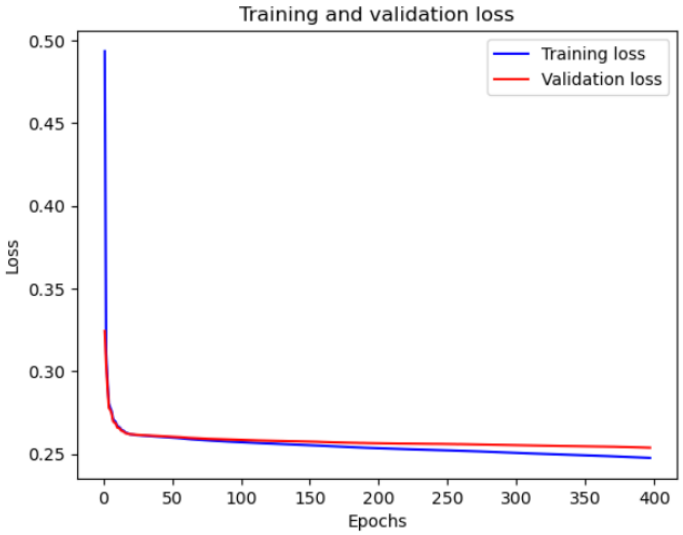

Figure 6 | Training and validation loss of ViT-AE

A linear reconstruction layer was added to the model architecture to enable visualization of reconstructed images and compare them with the original images. Figure 7 displays brain images of two individuals, highlighting visible differences in anatomical structure, along with their respective reconstructed images by the ViT-AE model, where these differences appear to be preserved.

As a comparison, the CNN-based UDIP model[22] achieved an average validation loss of approximately 0.28 after 15 epochs that increased to 0.3 at 25 epochs (see Supplementary Figure 2). Visual observation of reconstruction was similar to that of the ViT-AE model (see Supplementary Figure 3).

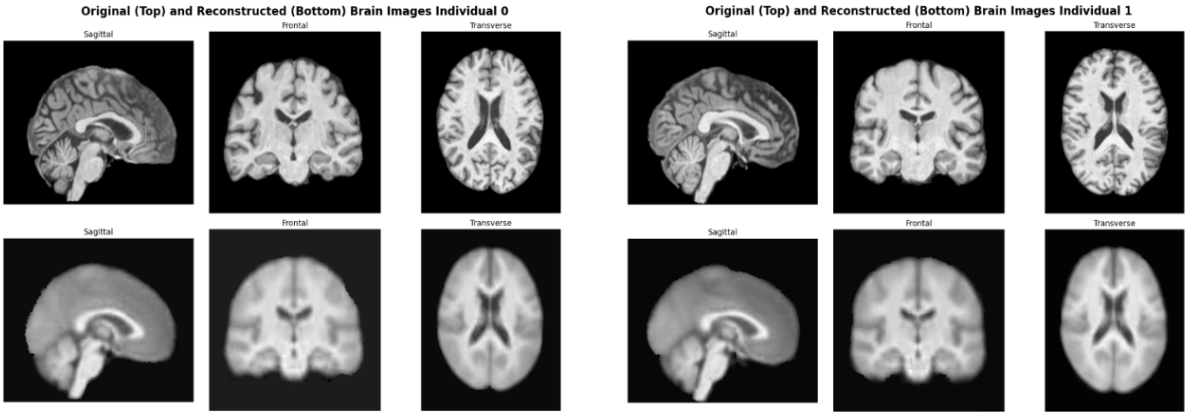

Figure 7 | Original and reconstructed brain images for two individuals

Our ViT-AE model was developed for representation learning from brain MRI scans and extraction of ViT-UDIPs. Our workflow pipeline did not have a particular focus on model performance improvement by lowering reconstruction losses via techniques like hyperparameter tuning. However, initial implementation of the ViT-AE model consisted of encoder embedding dimension of 384 and this was later changed to 128 (see Supplementary Note 1 and Supplementary Figures 8A, 8B). The final pipeline included a learning rate scheduler and AdamW optimizer (see Supplementary Note 2).

There has been widespread discussion on how ViTs perform with limited data. This is especially a concern in healthcare because of challenges in data availability[37]. Our dataset included the original MRI scans only, which were further filtered as described in the previous sections. As a means of experimentation, we trained our ViT-AE model on combined training and testing datasets and validated it on a disjoint validation set. The model performance considerably improved with an average validation loss of 0.18 after 300 epochs, compared to 0.253 on training set only (see Supplementary Figure 9A). The number of loci discovered was also significantly higher, with 96 associated loci found at 221 epochs (see Supplementary Figure 9B). Although the trained model identified additional associated loci, these findings were excluded from the final pipeline because the GWAS analysis was conducted on the testing dataset, which was also part of the training process. This overlap posed a risk of data leakage, potentially exaggerating the model's performance and compromising the validity of the results. However, we present these findings to showcase the ViT-AE model's potential to perform better with a larger training dataset, demonstrating its capability to uncover meaningful associations given sufficient data.

Furthermore, the pooling layer used in our ViT-AE architecture averaged the latent embeddings across all patches. This can potentially lead to diminished local information. Some modifications that we would like to explore further includes using adaptive average pooling, where some spatial information could be retained. Notably we tried to use attentive pooling / learned queries to pool the embeddings into multiple representations. However, the model experienced dimensional collapse, resulting in significantly worse performance compared to using average pooling to extract the embedding. This would enable better focus on the more important regions of the images.

Performing SmoothGrad-based saliency mapping highlighted the regions of the brain most influential in the ViT-AE model's predictions, allowing for clearer visualization of key structural features. We observed both local and non-local brain features being highlighted by the model.

In our results, non-local patterns refer to relationships in brain MRI scans that extend beyond neighboring regions. While the ViT-AE captured local patterns by attending to nearby structures, it also learnt non-local patterns by integrating information from distant brain regions. This allowed the ViT-AE to detect structural connections and variations that were not confined to a single localized area with closely connected structural features. This contributed to a more comprehensive understanding of brain organization. As shown in Figure 8A, the ViT-AE identified mirrored patterns across the left and right hemispheres, emphasizing its ability to capture structural symmetries across the brain. Figure 8B shows the distant regions of the brain were captured, revealing non-local dependencies and spatial relationships across the entire brain image.

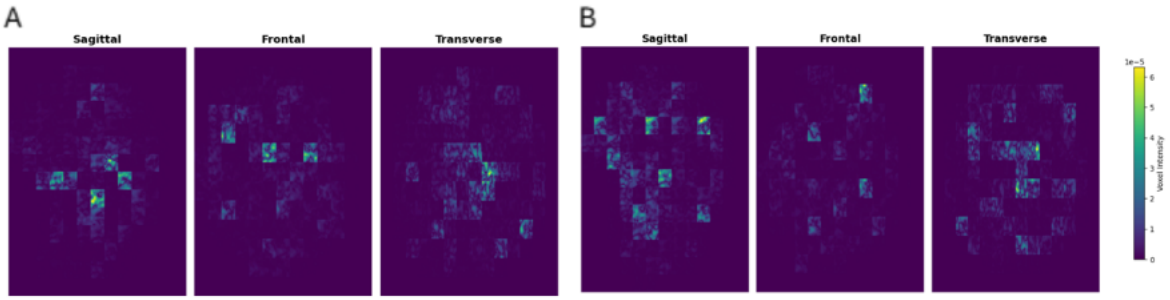

Figure 8 | Saliency maps of subjects

Statistical Analysis of Reconstruction Loss

We performed basic statistical analysis of inferencing results to determine distribution of reconstruction loss and outliers in the dataset (see Supplementary Note 3). Although the primary goal of our work was to design a ViT-AE-based pipeline for brain feature extraction and finding associated genetic signals, this analysis provided insights into the model's performance and helped detect anomalies and unusual deviations that could indicate potential challenges in reconstruction.

Overall reconstruction loss was found to be distributed normally, suggesting the model performed consistently across most scans in the dataset (see Supplementary Figure 2A). Boxplot and scatter plots indicated outliers that might have contributed to higher average reconstruction loss for some scans (see Supplementary Figure 2B, 2C). For the sake of efficiency, we considered only extreme outliers, which we defined to be outside of lower quartile (Q1) and upper quartile (Q3) by three times the inter-quartile range (IQR) (see Supplemental Figure 2D) and visualized those with the highest reconstruction losses to determine if there were significant anomalies in the brain structures or artefacts introduced by motion during scanning (see Supplementary Figure 3).

Motion correction and advanced preprocessing techniques like denoising filters were not applied to any of the datasets.Motion correction and advanced preprocessing techniques like denoising filters were not applied to any of the datasets.

Analysis of ViT-UDIPs obtained from inferencing demonstrated normalized distribution for each ViT-UDIP (see Supplementary Figure 4). Lines in violin plot for the ViT-UDIPs represented median, Q1 and Q3 values and each violin width represented kernel density estimate (KDE) of the data. The overall shapes indicated overall normality of the ViT-UDIPs.

GWAS

The primary motivation behind our work was to identify the maximum number of significant SNPs clustered into loci that were associated with ViT-UDIPs. Inference was run using

the test dataset on trained models saved at various intervals. We conducted linear regression to determine the p-value for each SNP-ViT-UDIP pair, then used FUMA to clump the significant SNPs (those with the minimum p-value passing the Bonferroni-corrected threshold) into loci. Our goal was to identify which model(s) yielded the highest number of loci associated with brain structure. During training, we recorded the number of loci discovered every 10 epochs. The model was trained until 100 consecutive epochs showed no further increase in the number of loci identified. The highest number of loci was discovered when inferencing was performed on model trained for 200, 220 and 300 epochs (see Supplementary Figure 5).

We only analyzed FUMA GWAS results for epochs 200, 220 and 300, as they represent the epochs with the most loci discovered. Referring to the GenomeLociRisk.txt file from FUMA GWAS results, we matched loci positions with corresponding dimension on which those minimum p-values were computed in the GWAS step. Notably, we discovered 10 new loci associated with brain structure that were not previously reported by CNN-based UDIP approach[22]. Using the dbSNP tool[36], we investigated the lead SNPs in these loci. Figure 9 shows the previously unreported loci and the corresponding dimension locations of UDIPs on a Manhattan plot. Chromosomes 7 and 11 each had two significantly associated loci. An example of how to interpret the plot is as follows: the gene SLC6A20 was identified as significantly associated with the ViT-UDIP located on dimension 94, indicating a relationship between the genetic expression of SLC6A20 and the feature represented in dimension 94 of the learned embeddings.

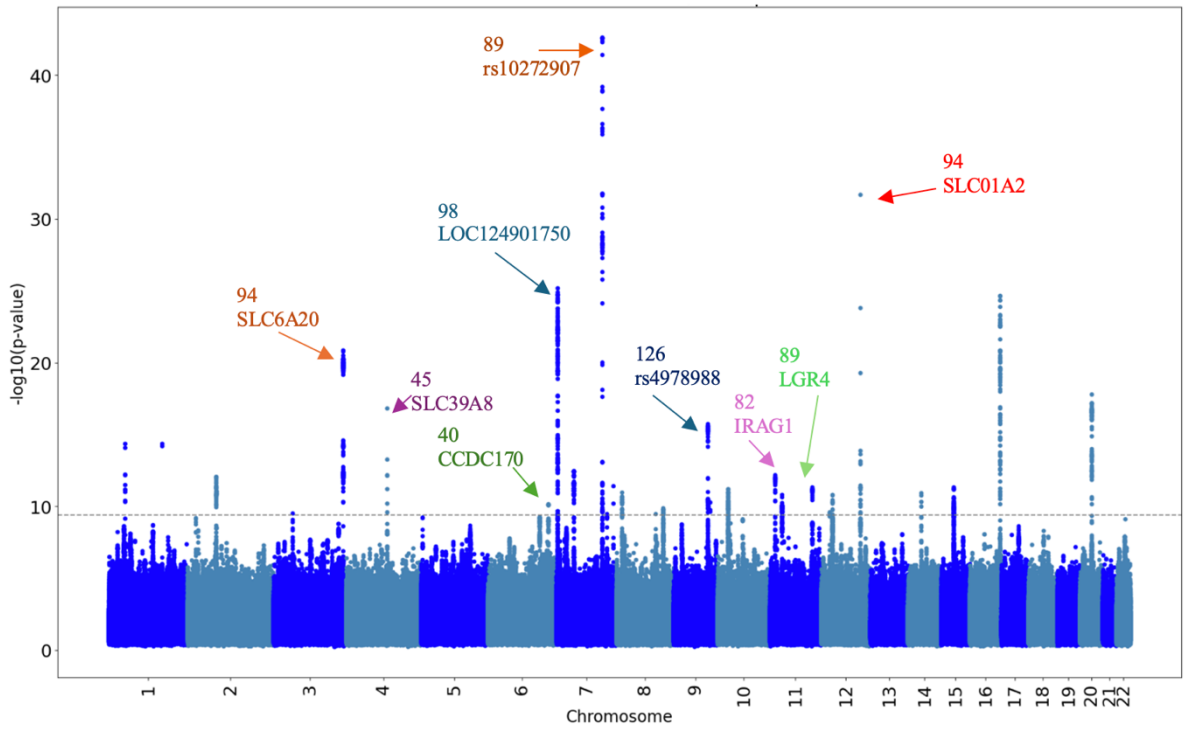

Figure 9 | UDIP dimension location and new loci discovered from ViT-based UDIP approach

GWAS revealed a total of 29 loci (see Supplementary Table 1 for all loci and associated lead SNPs and genes[38]), out of which 10 loci were not previously reported in the CNN-based UDIP approach[22]. These were cross-checked with the GWAS catalog and our findings are shown in **Error! Reference source not found.**.

Loci ID	Lead SNP	Gene	ViT dim.	Previous association with brain structure from GWAS catalog
4	rs17279437	SLC6A20	94	Vertex-wise cortical metrics[39], Brain morphology[40], Cortical phenotypes[41]
6	rs13107325	SLC39A8	45	Subcortical nuclei volumes, Subcortical nuclei intensity, T1-based brain region volumes, Thalamic nuclei volumes[42] Brain region volumes[43], Cortical surface area[5], Subcortical nuclei volume[44], Brain imaging measurements[14], Voxel-wise brain imaging measurements[45]
7	rs4869744	CCDC170	40	N/A
9	rs10272907	N/A	89	N/A
11	rs6973339	LOC124901750	98	Cortical sulcal depth[39]
14	rs2514524	GDF6	71	Cortical thickness[39]
17	rs4978988	N/A	126	N/A
20	rs10770131	IRAG1	82	Brain region volumes[43], Cortical thickness[41], Cortical sulcal depth[39]
21	rs4074516	LGR4	89	Brain region volumes[42]
23	rs74651308	SLC01A2	94	Cortical thickness[39]

Table 2 | Loci cross-checked with GWAS catalog

Several loci had been previously associated with brain structure. For brevity, we included only those loci which were discovered by ViT-AE, but not by CNN-based UDIP approach. Furthermore, we limited our findings to traits derived from T1-weighted MRI only. Investigating the above loci, we found that loci 4, 6, 11, 14, 20, 21 and 23 were already present in the GWAS catalog.

Loci 7, 9 and 17 had not been previously documented in the catalog as being associated with brain architecture. Locus 7, characterized by the lead SNP rs4869744 and linked to the gene CCDC170 on chromosome 6, was identified through ViT-AE dimension 40. Prior research had primarily associated this locus with bone density. Similarly, locus 9, marked by the lead SNP rs10272907 on chromosome 7, had also shown connections predominantly related to bone density.

Locus 17 with lead SNP rs4978988 on chromosome 9 had previously been found related to increased signal intensity in white matter regions, typically observed in T2-weighted MRI or FLAIR sequences[46]. However, brain phenotypes derived from T1-weighted MRI had not been associated with this locus.

Visualizing endophenotype activation patterns by PerDI

We performed PerDI on dimensions that a) included the highest number of SNPs significantly associated with the ViT-UDIPs (dimensions 18, 34, 40, 124, 126), and b) revealed new loci previously not discovered by CNN-based UDIP approach (dimensions 45, 82, 89, 94, 98).

Using FSLeyes, we visualized NIFTI files for each of these dimensions. We identified activated regions by overlaying the images with the Harvard-Oxford Cortical and Subcortical, and MNI Structural Atlases.

Applying PerDI to ViT-UDIPs revealed learned representations from local and non-local features of the brain MRI scans. Furthermore, symmetry was also somewhat observed in these visualizations. Figure 10 demonstrates the activation regions for dimension 98. At the subcortical level, the thalamus region was seen highlighted on both right and left sides of the brain, along with cerebral white matter and cerebral cortex. At the structural level, the cerebellum, along with the frontal and parietal lobes, was observed, capturing a comprehensive range of structural patterns across the hindbrain and forebrain. Similar patterns were identified across the other dimensions, highlighting the model's ability to generalize its representations effectively.

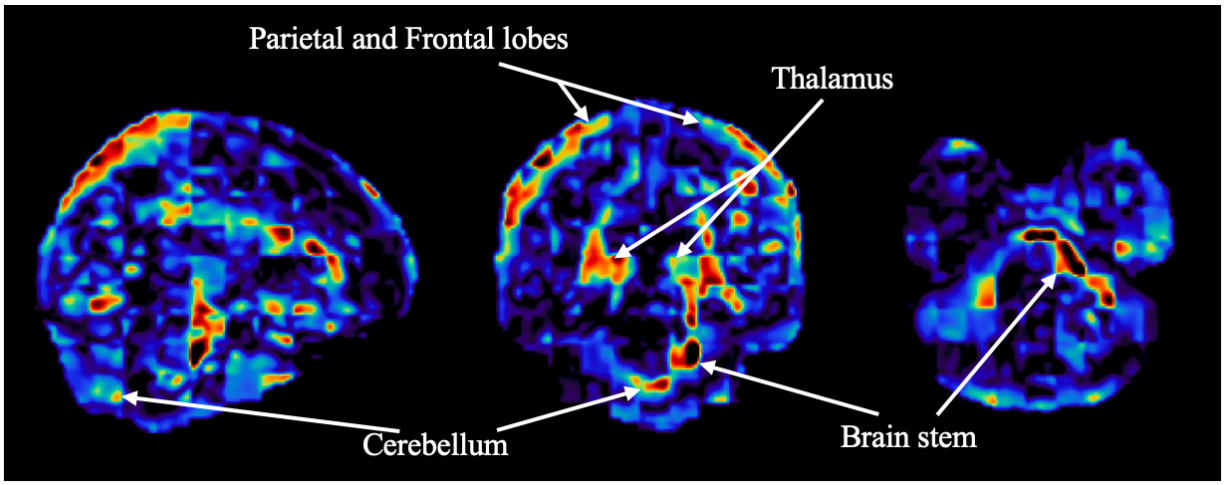

Figure 10 | PerDI activation on dimension 98

PerDI activation regions were investigated for loci 7, 9 and 17, which were previously unreported to have associations with brain structure. These associations were identified by dimensions 40, 89 and 126, respectively. Figure 11 provides the activations regions from each of these dimensions.

Dimension 40 (Figure 11A), once activated using PerDI, highlights the pallidum, hippocampus, and accumbens in a symmetrical pattern. The frontal lobe region at the structural level, as well as the caudate in the central subcortical region of the brain is highlighted by dimension 89 (Figure 11B). In a similar way, dimension 126 (Figure 11C) activates the ventricles and thalamus in the central subcortical region. We also see activation in the cerebellum at the

structural layer. In all three dimensions, we also see activation in regions corresponding to the subcortical regions of cerebral cortex and cerebral white matter.

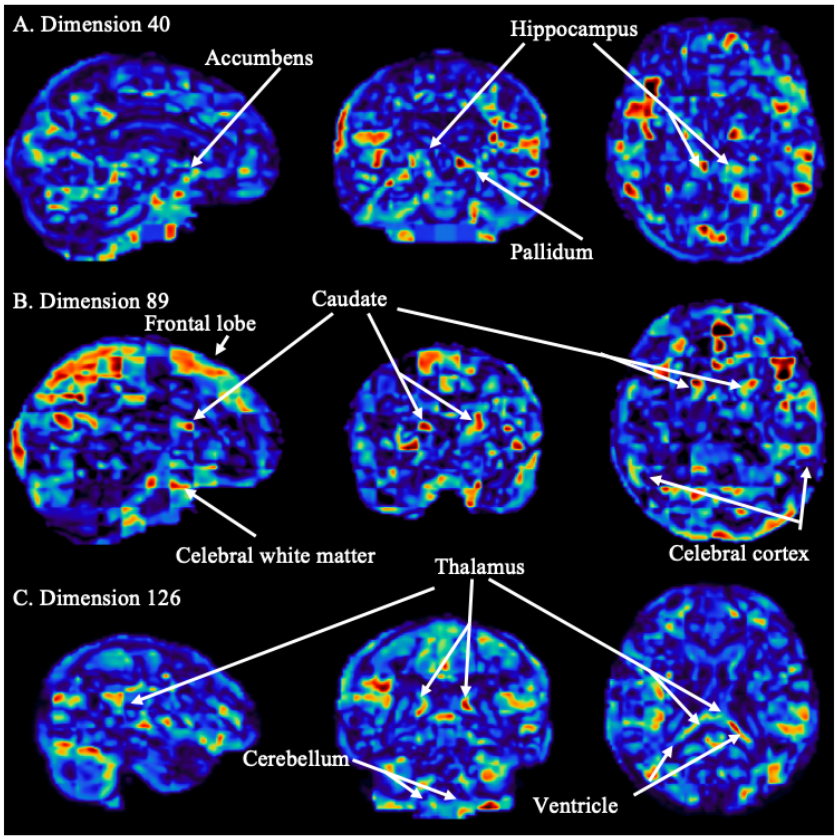

Figure 8 | Activation regions for dimensions 40, 89, 126

The PerDI ViT-UDIP SNP-to-brain mapping step demonstrated the ViT-AE model's ability to capture activations across both local and non-local features of the brain, revealing structural patterns at varying levels of detail. At the local level, specific regions such as the pallidum, hippocampus, accumbens, thalamus, and the cerebral cortex were highlighted, showing activation within central subcortical and cortical areas. Additionally, activations were observed in key structures such as the ventricles and cerebral white matter, which play crucial roles in brain communication and the protection of neural tissue. These activations reflected the model's capacity to identify symmetrical patterns across bilateral structures, consistent with the brain's inherent anatomical symmetry, and to recognize fine-grained structural details.

On a broader scale, the model's activations extended across the entire brain image, integrating information from both cortical and subcortical regions, including the ventricles, cerebral white matter, and cerebral cortex. This comprehensive activation reflected the ViT-AE's ability to map complex interconnections between different brain areas, providing a holistic view of the brain's structure. At the structural level, these activations underscored the model's capability to capture a wide array of brain features, from the deep structures of the thalamus and subcortical regions to the superficial cortical regions and white matter pathways, allowing for an enriched understanding of the brain's anatomy and its functional relationships.

Attention Head Analysis

We visualized the attention patches on brain atlases using FSLeyes. Our analysis provided insights into the model's interpretability, highlighting the regions of the brain that are most influential in the feature extraction process and offering a deeper understanding of how the ViT-AE model captures complex spatial dependencies in brain MRI scans.

Our observations revealed long-range dependencies and symmetrical attention across the brain regions. The model was able to attend to distant patches across different brain regions simultaneously, demonstrating its capacity to capture non-local relationships within the brain (see Supplementary Figure 10A). Additionally, we observed symmetrical attention, with patches in one hemisphere of the brain attending to corresponding patches in the opposite hemisphere (see Supplementary Figure 10B). This symmetry suggests that the ViT-AE model is effectively capturing bilateral relationships, reflecting the natural symmetry of brain anatomy. These findings highlight the model's ability to integrate both local and non-local features, providing a comprehensive understanding of brain architecture and uncovering complex interconnections between brain regions.

Furthermore, 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 (Figure 12), 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.

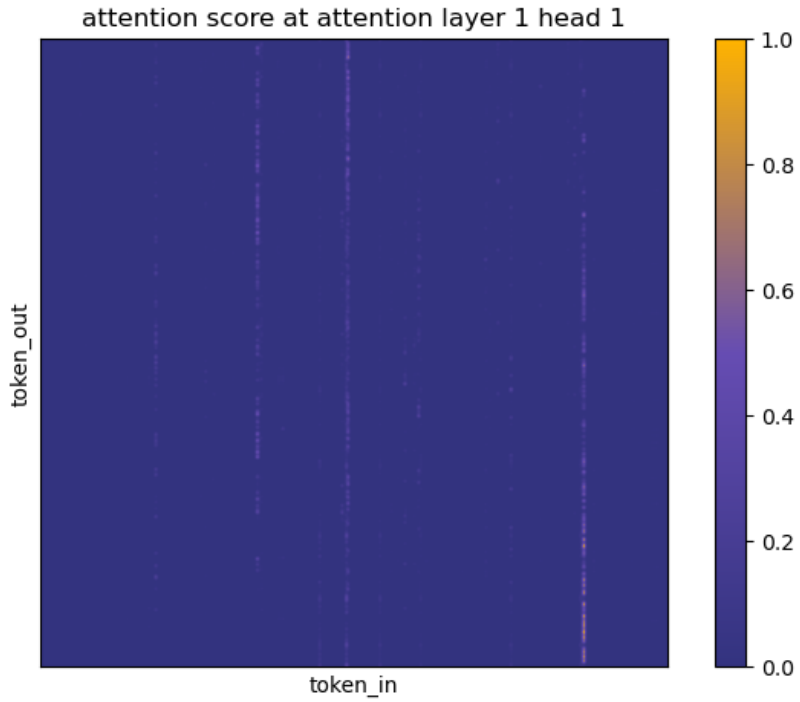

Figure 9| Softmax attention score at attention layer 1 head 1 between 957 tokens, image is blurred using Gaussian filter with sigma=2 for better visual quality.

We found that this behavior is driven by the specific embeddings learned by the model. The input embedding matrix projects the majority of patches to a similar direction, while a small

subset remains weakly correlated with each other and the majority (Figure 13). Henceforth, we will refer to them as audience tokens and spotlight tokens for brevity.

At each attention head, the (unnormalized) attention score between two tokens X_1 and X_2 can be written as $X_1^T W_K^T W_Q X_2$, this can be interpreted as first rotating them differently (based on the rotation matrices formed by the left and right singular vectors of $W_K^T W_Q$), followed by projecting X_1, X_2 to a low-dimensional space and finally computing a weighted inner product. Specifically, singular value decomposition of $W_K^T W_Q = U\Sigma V^T$. U, V are rotation matrices and Σ is a diagonal scaling matrix. The interaction between X_1 and X_2 can be written as $(PU^T X_1)^T \Sigma_r (PV^T X_2)$, where P is a wide identity matrix that selects the first d_h coordinates of the input, Σ_r is the reduced singular value matrix which only consists of nonzero entries of Σ of shape $d_h \times d_h$, the dimension of the head. 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.

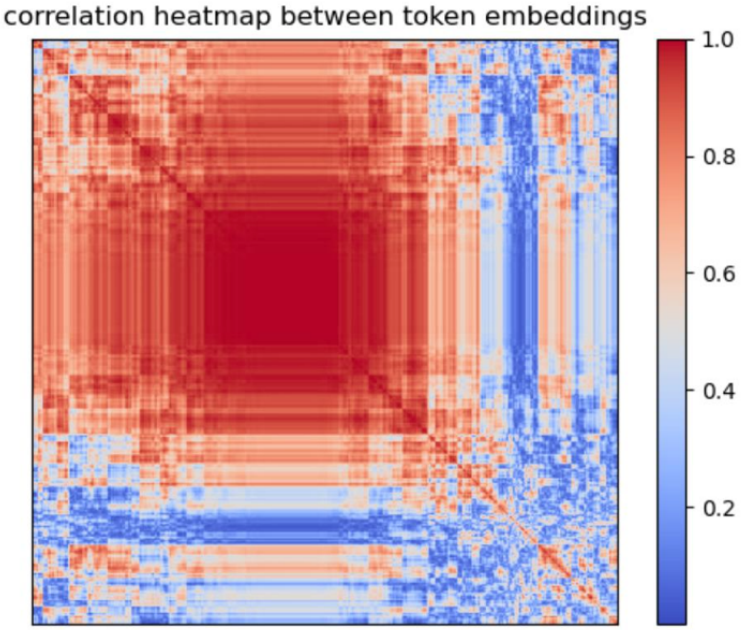

Figure 10 | Correlation heatmap between 957 input embeddings, we use hierarchical clustering with average linkage to reorder the dimensions of the heatmap.

We then visualize the distribution of spotlight tokens across different brain slices. The highlighted regions in Figure 14 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.

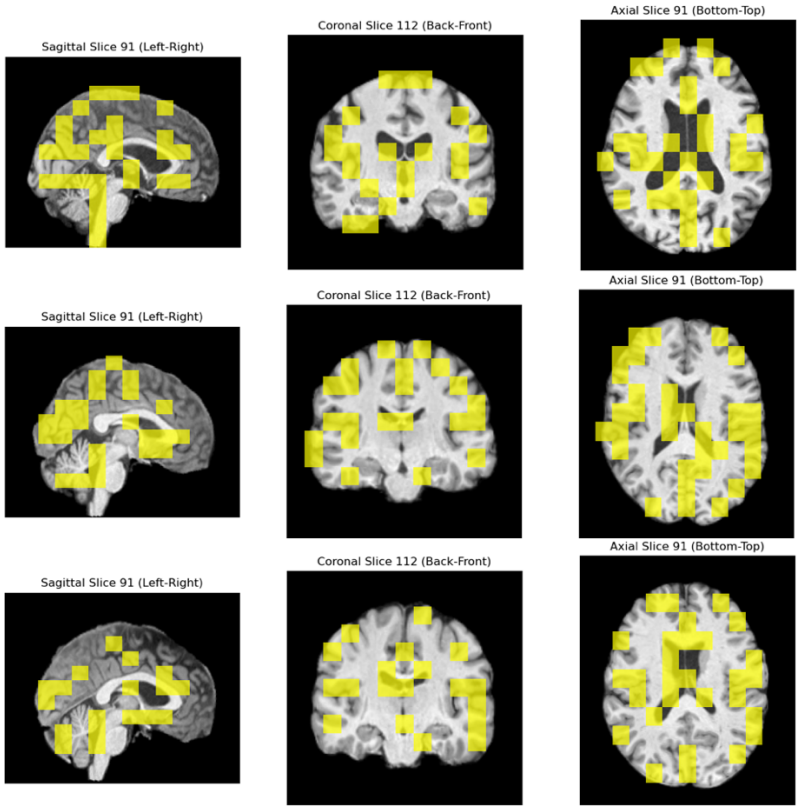

Figure 14 | Visualization of spotlight tokens across brain slices. Highlighted regions (yellow)

Conclusion

We designed an approach that integrated unsupervised deep representation learning, specifically using a ViT-AE model, for brain imaging GWAS. By leveraging the ViT's ability to capture both local and non-local ~~global~~ structural patterns, our pipeline overcame the limitations of traditional methods relying on predefined features and assumptions. A few CNN-based models addressing these challenges in turn have limitations, either requiring large amounts of labeled data, or unable to uncover a wider range of features. Our ViT-AE pipeline allowed us to identify 10 loci previously unreported by CNN-based UDIP pipeline, out of which loci 7, 9, and 17 had not been found in the existing GWAS catalog to have any associations with brain structures. Our design highlighted the ViT's superior capability in detecting complex relationships in brain imaging data. Notably, the ViT-AE model revealed activation patterns in key brain regions such as the pallidum, hippocampus, accumbens, and thalamus, which are integral to motor coordination, cognition, and sensory processing. The model's ability to identify symmetrical features across bilateral brain regions further emphasizes its potential for capturing intricate anatomical details.

Our pipeline also incorporated a perturbation-based decoder interpretation approach to map these loci and their associated imaging phenotypes to specific brain regions. The ViT-AE model's non-local attention mechanism enabled the model to capture complex spatial relationships across the entire brain, including important structures like the ventricles, cerebral cortex, and cerebral white matter. By linking genetic markers to their anatomical locations, our results provide new

insights into the genetic underpinnings of brain structure. Additionally, SmoothGrad-based saliency mapping highlighted the regions of the brain most influential in the ViT-AE model's predictions, allowing for clearer visualization of key structural features. We observed both local and non-local brain features were highlighted by the model. Lastly, we compared original and reconstructed images for two individuals and visualized the absolute differences (see Supplementary Figure 11). Our model was able to distinguish between them, showing its capacity to capture subtle structural differences in brain morphology. This ability highlights the model's effectiveness in learning unique, individualized patterns that contribute to distinguishing between different brain structures. The success of this transformer-based approach underscores its potential for improving the sensitivity and accuracy of brain imaging GWAS, offering a deeper understanding of the brain's complex genetic architecture.

We analyzed the dataset and presented how model performance improved with a larger dataset. In future work, we suggest incorporating data augmentation and synthetic data generation techniques to further enhance model performance. Data augmentation methods, such as random rotations, flips, and intensity variations, could be employed to create more diverse variations of the original brain MRI scans, helping the model learn more robust and generalized features[25]. Additionally, utilizing synthetic data generation techniques like Generative Adversarial Networks (GANs) could allow for the creation of realistic, high-quality brain images, addressing challenges related to limited data availability. These strategies could significantly improve the model's ability to capture complex patterns and relationships within brain structure, ultimately enhancing the accuracy and robustness of our pipeline in future applications.

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, it provides a powerful tool for identifying previously unknown genetic loci associated with brain architecture. As we explore future improvements, such as incorporating data augmentation and synthetic data generation, this pipeline has the potential to further enhance the accuracy and depth of our understanding of the genetic underpinnings of brain structure, paving the way for more personalized approaches in neuroscience and medicine.

Ethics oversight

Our analysis was approved by UTHealth 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.

Acknowledgements

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

References

1. García-Marín, L. M. *et al.* Investigating the genetic relationship of intracranial and subcortical brain volumes with depression and other psychiatric disorders. *Imaging Neuroscience* **2**, 1–16 (2024).
2. Hibar, D. P. *et al.* Novel genetic loci associated with hippocampal volume. *Nat Commun* **8**, 13624 (2017).
3. The Alzheimer's Disease Neuroimaging Initiative *et al.* Common genetic variants influence human subcortical brain structures. *Nature* **520**, 224–229 (2015).
4. Elvsåshagen, T. *et al.* The genetic architecture of human brainstem structures and their involvement in common brain disorders. *Nat Commun* **11**, 4016 (2020).
5. Grasby, K. L. *et al.* The genetic architecture of the human cerebral cortex. *Science* **367**, eaay6690 (2020).
6. Hulshoff Pol, H. E. *et al.* Genetic Contributions to Human Brain Morphology and Intelligence. *J. Neurosci.* **26**, 10235–10242 (2006).
7. Thompson, P. M. *et al.* Genetic influences on brain structure. *Nat Neurosci* **4**, 1253–1258 (2001).
8. Thompson, P., Cannon, T. D. & Toga, A. W. Mapping genetic influences on human brain structure. *Annals of Medicine* **34**, 523–536 (2002).
9. Glahn, D. C., Paus, T. & Thompson, P. M. Imaging genomics: Mapping the influence of genetics on brain structure and function. *Human Brain Mapping* **28**, 461–463 (2007).
10. Toga, A. W. & Thompson, P. M. GENETICS OF BRAIN STRUCTURE AND INTELLIGENCE. *Annu. Rev. Neurosci.* **28**, 1–23 (2005).
11. Brun, C. C. *et al.* Mapping the regional influence of genetics on brain structure variability — A Tensor-Based Morphometry study. *NeuroImage* **48**, 37–49 (2009).

12. Kim, M. *et al.* Deep Learning in Medical Imaging. *Neurospine* **16**, 657–668 (2019).

13. Mall, P. K. *et al.* A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. *Healthcare Analytics* **4**, 100216 (2023).

14. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562**, 210–216 (2018).

15. Liu, Y. *et al.* Associations between brain imaging–derived phenotypes and cognitive functions. *Cerebral Cortex* **34**, bhae297 (2024).

16. Hussain, L. *et al.* Prostate cancer detection using machine learning techniques by employing combination of features extracting strategies. *CBM* **21**, 393–413 (2018).

17. Sharma, A. *et al.* Bone Cancer Detection Using Feature Extraction Based Machine Learning Model. *Computational and Mathematical Methods in Medicine* **2021**, 1–13 (2021).

18. Rasheed, J. Analyzing the Effect of Filtering and Feature-Extraction Techniques in a Machine Learning Model for Identification of Infectious Disease Using Radiography Imaging. *Symmetry* **14**, 1398 (2022).

19. Chabert, S. *et al.* Applying machine learning and image feature extraction techniques to the problem of cerebral aneurysm rupture. *RIO* **3**, e11731 (2017).

20. Giger, M. L. Machine Learning in Medical Imaging. *Journal of the American College of Radiology* **15**, 512–520 (2018).

21. Yu, S. *et al.* A novel classification framework for genome-wide association study of whole brain MRI images using deep learning. *PLoS Comput Biol* **20**, e1012527 (2024).

22. Patel, K. *et al.* Unsupervised deep representation learning enables phenotype discovery for genetic association studies of brain imaging. *Commun Biol* **7**, 414 (2024).

23. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Preprint at https://doi.org/10.48550/ARXIV.2010.11929 (2020).

24. Zhang, T., Zeng, Y. & Xu, B. HCNN: A Neural Network Model for Combining Local and Global Features Towards Human-Like Classification. *Int. J. Patt. Recogn. Artif. Intell.* **30**, 1655004 (2016).

25. Dhinagar, N. J., Thomopoulos, S. I., Laltoo, E. & Thompson, P. M. Efficiently Training Vision Transformers on Structural MRI Scans for Alzheimer's Disease Detection. Preprint at https://doi.org/10.48550/arXiv.2303.08216 (2023).

26. Al-hammuri, K., Gebali, F., Kanan, A. & Chelvan, I. T. Vision transformer architecture and applications in digital health: a tutorial and survey. *Vis. Comput. Ind. Biomed. Art* **6**, 14 (2023).

27. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).

28. He, K. *et al.* Masked Autoencoders Are Scalable Vision Learners. Preprint at https://doi.org/10.48550/arXiv.2111.06377 (2021).

29. FSL Atlases. https://ftp.nmr.mgh.harvard.edu/pub/dist/freesurfer/tutorial_packages/centos6/fsl_507/doc/wiki/Atlases.html (2014).

30. Makris, N. *et al.* Decreased volume of left and total anterior insular lobule in schizophrenia. *Schizophrenia Research* **83**, 155–171 (2006).

31. Frazier, J. A. *et al.* Structural Brain Magnetic Resonance Imaging of Limbic and Thalamic Volumes in Pediatric Bipolar Disorder. *AJP* **162**, 1256–1265 (2005).

32. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31**, 968–980 (2006).

33. Goldstein, J. M. *et al.* Hypothalamic Abnormalities in Schizophrenia: Sex Effects and Genetic Vulnerability. *Biological Psychiatry* **61**, 935–945 (2007).

34. Collins, D. L., Holmes, C. J., Peters, T. M. & Evans, A. C. Automatic 3-D model-based neuroanatomical segmentation. *Human Brain Mapping* **3**, 190–208 (1995).

35. Mazziotta, J. *et al.* A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Phil. Trans. R. Soc. Lond. B* **356**, 1293–1322 (2001).

36. Phan, L. *et al.* The evolution of dbSNP: 25 years of impact in genomic research. *Nucleic Acids Research* gkae977 (2024) doi:10.1093/nar/gkae977.

37. Mao, J., Zhou, H., Yin, X. & Xu, Y. Chang. B. Nie. R. Masked autoencoders are effective solution to transformer data-hungry. Preprint at https://doi.org/10.48550/ARXIV.2212.05677 (2022).

38. GWAS Catalog. https://www.ebi.ac.uk/gwas/home.

39. Van Der Meer, D. *et al.* The genetic architecture of human cortical folding. *Sci. Adv.* **7**, eabj9446 (2021).

40. Van Der Meer, D. *et al.* Boosting Schizophrenia Genetics by Utilizing Genetic Overlap With Brain Morphology. *Biological Psychiatry* **92**, 291–298 (2022).

41. Shadrin, A. A. *et al.* Vertex-wise multivariate genome-wide association study identifies 780 unique genetic loci associated with cortical morphology. *NeuroImage* **244**, 118603 (2021).

42. Smith, S. M. *et al.* An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nat Neurosci* **24**, 737–745 (2021).

43. Zhao, B. *et al.* Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nat Genet* **51**, 1637–1644 (2019).

44. Satizabal, C. L. *et al.* Genetic architecture of subcortical brain structures in 38,851 individuals. *Nat Genet* **51**, 1624–1636 (2019).

45. Luo, Q. *et al.* Association of a Schizophrenia-Risk Nonsynonymous Variant With Putamen Volume in Adolescents: A Voxelwise and Genome-Wide Association Study. *JAMA Psychiatry* **76**, 435 (2019).

46. Sargurupremraj, M. *et al.* Cerebral small vessel disease genomics and its implications across the lifespan. *Nat Commun* **11**, 6285 (2020).