Medical Imaging and Applications

Master Thesis, July 2023

Neurodegeneration identification in Parkinson's Disease with Deep Learning models using 3T quantitative MRI maps

Alejandro Cortina Uribe, David Meder

Danish Research Centre for Magnetic Resonance, Copenhagen, Denmark

Abstract

Parkinson's disease (PD) is a neurodegenerative syndrome with diverse motor and non-motor symptoms. While clinical assessment is the primary diagnostic method, magnetic resonance imaging (MRI) has gained importance in aiding PD diagnosis and treatment planning. While researchers have identified spatial patterns of neurodegeneration related to iron and neuromelanin (NM) that correlate with specific symptoms at 7T field strength, the applicability of these insights at 3T remains uncertain. Quantitative MRI (qMRI) maps are commonly used to model parameters that are robust across imaging sites and acquisition times. In PD, R2* and quantitative susceptibility mapping (QSM) images, highly sensitive to iron, are frequently employed. From a cohort study in our centre, we acquired 3T scans from which we can obtain different qMRI maps. Since the 3T protocol was not developed for PD imaging, performing frequentist statistics may not be suitable, and a DL-based analysis could provide better insights leveraging more powerful feature extraction and representation techniques.

Our study aims to investigate the ability of 3T qMRI maps to identify neurodegenerative changes in PD patients by training a well-performing DL pilot model using limited data and employing different learning techniques. We pursued two strategies: a) transfer learning-based binary classification using a 3D convolutional neural network (CNN) and application of explainable artificial intelligence (XAI) algorithms to interpret model predictions, and b) normative modeling, where we derived anomalies from reconstruction error maps and conducted binary classification based on the percentage of anomaly within specific regions of interest (ROIs). Although the first strategy did not yield a high-performing model, XAI proved invaluable in detecting issues such as overfitting and shortcut learning. In the second strategy, we performed group average statistics on reconstruction error maps and identified relevant subcortical nuclei in the MTsat, PD*, and R2* maps. By leveraging these ROIs, we quantified the error distribution among healthy controls and discovered anomalies that facilitated classification between PD patients and controls. The most discriminatory ROIs were the left globus pallidus interna in the MTsat map (AUROC: 0.84, G-mean: 0.82) and the left subthalamic nucleus (AUROC: 0.84, G-mean: 0.85). Our results highlight the challenges of binary classification with a small dataset and a 3D model architecture, even when employing diverse transfer learning strategies. However, the use of XAI to assess model predictions and identify signs of shortcut learning is crucial. Additionally, other learning techniques, such as unsupervised normative modeling, exhibit promising results, but necessitate careful selection of generative models, enlargement of the controls dataset to better capture its distribution, and rigorous validation of results.

Keywords: Parkinson's Disease, quantitative MRI, Deep Learning, normative modeling, classification, explainable AI

1. Introduction

1.1. Parkinson's Disease and Imaging

Parkinson's disease is a neurodegenerative syndrome that affects multiple motor and non-motor neural circuits. It involves two primary pathological processes: the loss of dopamine neurons and the accumulation of Lewy bodies. However, the order of occurrence of these processes is still unclear (Rizek et al., 2016). The loss of dopaminergic function leads to a decline in motor func-

tion and the emergence of clinical symptoms. Since there is no definitive test for confirming the diagnosis of PD, clinical diagnosis relies on assessing symptoms and patient history (DeMaagd and Philip, 2015). Neuroimaging studies, such as transcranial Doppler ultrasonography, PET, SPECT, and MRI, are performed to aid in the differential diagnosis and exclude other parkinsonian disorders (Rizek et al., 2016).

Structural changes resulting from neurodegeneration can be reflected in alterations in the local iron and neuromelanin (NM) content within the dopaminergic substantia nigra pars compacta (SNc) and the noradrenergic locus coeruleus (Madelung et al., 2022; Zucca et al., 2017) . Specifically, NM accumulates with age in the SNc within dopamine and noradrenaline neurons, but it depletes in PD patients due to the loss of these NM-containing neurons. On the other hand, iron also accumulates with age, but its deposition is excessive in PD (Biondetti et al., 2020; Zucca et al., 2017). These changes are strongly associated with motor impairment, such as the volume decrease of SNc in iron-sensitive quantitative susceptibility mapping (QSM) correlating with the severity of bradykinesia and rigidity, especially in patients with longer disease duration (Poston et al., 2020). Additionally, they are related to non-motor impairment, such as orthostatic changes in systolic blood pressure and apathy in locus coeruleus spatial neurodegeneration assessed by NM-sensitive MRI (Madelung et al., 2022).

Nevertheless, the relationship between these structural changes and the complex pathophysiology of PD is still not fully understood (Zucca et al., 2017). Magnetic Resonance Imaging (MRI) has become a valuable tool for researchers and clinicians to localize these changes, utilizing techniques such as NM-MRI (Trujillo et al., 2017) and iron-sensitive MRI (Biondetti et al., 2021). In recent years, high-resolution images obtained with ultra-high field scanners (7 teslas) have provided new insights into the topographic patterns of disease-related structural changes within these small nuclei ((Madelung et al., 2022). Furthermore, task-related functional MRI (fMRI) has revealed alterations in brain activation patterns related to the complex interactions of dopaminergic neurodegeneration in target nuclei (Meder et al., 2019).

However, the current MRI modalities targeting NM and iron have not yet provided robust diagnostic biomarkers for PD, mainly because they lack specificity to the melanin-iron complex or its metabolic processes during disease progression and onset. Additionally, research-only ultra-high field scanners are not widely available compared to the more commonly used 3 tesla MRI scanners, and it remains unclear whether MRI images acquired at this field strength can reveal similar or different patterns of PD-related changes.

Therefore, there is growing interest in emerging techniques such as quantitative MRI (qMRI) mapping, which aim to image tissue microstructure by modeling specific parameters (e.g., relaxation rates R1 or R2*), providing absolute measures and facilitating inter-site comparability across different time points (Tabelow et al., 2019; Weiskopf et al., 2013; Wenger et al., 2021). The most widely used quantitative maps in recent PD research are based on iron quantification within tissues, including T2* relaxometry (R2*) and quantitative susceptibility mapping (QSM) that utilize local susceptibility and phase information from gradient-echo or SWI sequences (Arribarat and Péran, 2020; Bae et al., 2021). In terms of NM imaging, these sequences exploit the property of melanin to reduce T1 relaxation time, while magnetization transfer imaging (MTw) is used to improve the contrast to NM, resulting in high-intensity signals in NM-rich areas (Bae et al., 2021; Madelung et al., 2022). Although quantitative maps derived from these sequences have not been extensively utilized, it is expected that R1 and magnetization transfer saturation maps contain information sensitive to NM.

1.2. Data analysis

To gain a better understanding of the aforementioned structural changes or functional patterns and draw interpretable conclusions, the field of neuroscience research has focused on conducting frequentist statistics on smaller cohorts. These cohorts are often limited by factors such as the availability of image modalities, subject and patient recruitment, and the complexity of disease progression.

More recently, deep learning (DL) has emerged as an alternative approach by addressing the problem of representation learning. DL aims to disentangle high-dimensional data into a lower-dimensional representation, enabling the identification of meaningful patterns and anomalies. In other words, DL attempts to learn abstract patterns that are relevant to the data.

Among the various learning problems that DL can assist with, classification tasks have been widely implemented. By training models to automatically extract features and perform "patient versus healthy control" classification for different brain diseases, we can develop end-to-end computer-aided diagnosis (CAD) systems that demonstrate exceptional predictive power compared to traditional machine learning models (see Section 2 State-of-the-art).

However, as we increase the complexity and flexibility of DL models, their interpretability and explainability diminish, contributing to the general skepticism among clinical researchers towards the "black box" nature of DL models. To address this concern, numerous explainability algorithms have been developed to gain insights into the learned features and decision-making processes of the models (Chaddad et al., 2023). Furthermore, the application of DL models in the medical domain is limited by data scarcity, which hampers their performance in generalization across different domains.

To mitigate this limitation, various training methodologies, such as transfer learning, unsupervised learning, and self-supervised learning, have been widely employed (Chen et al., 2019; Kim et al., 2022; Taleb et al., 2020).

Another valuable application of DL is the creation of normative models. In this framework, we move away from the assumption that clinical groups are easily distinguishable and homogeneous, aiming to better understand differences in relation to a reference model (Rutherford et al., 2022). Normative models have been utilized in various clinical scenarios, ranging from growth charting in pediatrics to mental disorders (Marquand et al., 2019). In the context of brain imaging, normative modeling has been employed to identify regions of the brain affected by disease or specific pathological patterns (see Section 2 State-of-the-art).

1.3. Project proposal

In this thesis project, we aim to investigate the relevance of qMRI maps acquired at 3 teslas in identifying structural changes in PD patients using a data-driven approach. We explore the possibility of training a high-performing DL pilot model with various learning techniques on limited data and examine the explanations for their performance. Our main general hypothesis is as follows:

- The qMRI maps (R1, R2*, PD*, and MTsat) obtained at 3 teslas are sensitive to neurodegeneration markers in PD, such as iron accumulation and NM loss, as well as potentially other structural changes. We will evaluate the classification performance of the proposed DL models and utilize explainability methods to identify relevant regions of interest.

We propose two exploratory strategies:

a) Unimodal binary classification with transfer learning: From a best performing model amongst different experiments we will initially obtain a predictive performance metric. Subsequently, by employing explainability methods, we will generate attribution heatmaps to localize the most important brain regions for the model's predictions. This approach may help us identify known nuclei affected by neurodegeneration, such as the SNc, as well as other regions of interest.

b) Normative modeling with unsupervised learning: In contrast to the first strategy, from PD patients we will first generate a reconstruction error map to identify disease anomalies and their spatial distribution. Then, by determining optimal thresholds that differentiate PD patients from controls, we will derive a final classification performance metric.

These two strategies will enable us to assess the potential of qMRI maps at 3 teslas in detecting structural changes related to PD. It is important to note that, despite the obtained qMRI maps were not particularly developed to be sensitive to PD neurodegenerative markers, we are optimistic that R2* maps are indeed sensitive to iron and MTsat and R1 maps might be sensitive to NM. This motivated our data-driven exploratory project oriented to investigate the sensitivity of these novel maps to identify structural changes related to PD, through DL methods that are able to capture and extract complex features and information from the images.

1.4. Abbreviations

PD, Parkinson's Disease. HC, healthy control. NM, neuromelanin. SNc, substantia nigra pars compacta. qMRI, quantitative magnetic resonance imaging. XAI, explainable artificial intelligence. ROI, region of interest.

2. State of the art

In our literature review, we did not find specific approaches that predicted PD or assessed PD neurodegeneration using qMRI maps and DL models. Currently, the research on DL-based PD classification has predominantly utilized other MRI sequences, brain imaging modalities such as SPECT and ECG, clinical and genetic data, or combinations of them.

When dealing with PD, the options for utilizing DL models are limited due to requirements of the dataset size. Thus, researchers often resort to using large public datasets like the multi-modal longitudinal Parkinson's Progression Markers Initiative (PPMI) (Marek et al., 2018) to train the models and validate them on smaller in-house datasets. In Chaki and Woźniak (2023), a systematic review highlighted that DL has been extensively used for neurodegenerative disorders in recent years. However, for PD, they found a majority of papers focusing on classification using non-brain imaging datasets such as speech and handwriting, with only a few studies using brain imaging data.

More recently, an increasing number of papers have been published using the PPMI study and other datasets to perform classification and explainability analyses. For instance, Camacho et al. (2023) gathered 13 different datasets comprising T1-weighted MRI scans (over 2000 participants). They employed a convolutional neural network (CNN) to classify PD and healthy control (HC) subjects using Jacobian maps derived from deformation fields of MNI spatial normalization, along with basic clinical parameters. They achieved an AUROC of 0.86 in their independent test set and generated saliency maps using the SmoothGrad (Smilkov et al., 2017) algorithm, which identified frontotemporal regions, the orbital-frontal cortex, and multiple deep gray matter structures as the most important.

In Shinde et al. (2019), an in-house dataset of 80 subjects of NM-sensitive MRI was used to classify PD

patients versus HC subjects and PD versus parkinsonian syndromes (APS) patients. They employed a 2D ResNet50 model trained with axial slices of the brainstem region. Their classification results were compared with two other ML-based models using radiomics and contrast-ratio features, and they obtained an AUROC of 0.906 on their test set, outperforming the ML approaches (AUROC of 0.54). To explain the DL model's decisions, they created class activation maps (CAM) from the weights and feature maps of the last convolutional layers and assessed contra-lateral activations in the SNc. They found a significantly larger mean activation in the left SNc compared to the right in PD patients.

In Huang et al. (2023), to address the limited interpretability of DL models, the authors defined disease classification (prodromal PD versus HC) as a graph representation task. They obtained relevant clinical interpretations by highlighting key nodes. They used diffusion tensor imaging (DTI) and structural MRI data from 194 subjects in the PPMI dataset to track fiber tracts and construct structural brain networks (SBNs). By employing a graph neural network, they achieved promising classification performance compared to other DL-based and ML-based models. Furthermore, through parametric decomposition and leveraging embedded GNN characteristics, they identified salient structural regions of interest (ROIs) that occurred most frequently among subjects, highlighting diverse cortical structures such as the precentral gyrus-L and the superior frontal gyrus-orbital.

The normative modeling framework has gained interest in recent years for its application in medical imaging tasks such as segmentation and classification. Additionally, it has been explored as a means to detect anomalies and identify lesions in brain MRI (Tschuchnig and Gadermayr, 2021). For instance, in the study by Baur et al. (2019), a novel deep autoencoding model with adversarial training was proposed for the detection and delineation of multiple sclerosis (MS) lesions based on reconstruction error maps. The authors trained a variational autoencoder (VAE) using 2D slices of FLAIR images from an in-house dataset of 83 healthy subjects, achieving the highest dice score coefficient (DSC) compared to other model architectures.

Similarly, Pinaya et al. (2021a) employed autoencoders to identify deviations from normal brains in Alzheimer's disease (AD) patients and identify associated critical regions. They trained a conditional autoencoder on a large cohort of healthy controls using subregional volume features extracted from over 11,000 structural MRI images from the UK Biobank. The performance of the model was validated on five additional datasets, where the mean squared error (MSE) between the reconstructed and inputted data served as a metric for brain deviation. This approach demonstrated high discriminative performance in distinguishing between healthy controls and AD patients. In a subsequent study, Pinaya et al. (2021b) developed a novel model based on VAEs and transformers to automatically detect various types of lesions and their delineations. By training their normative model on 15,000 FLAIR images from the UK Biobank, they achieved superior performance in lesion detection, specifically for white matter hyperintensities and tumors, outperforming similar autoencoder-based models in terms of DSC.

Lastly, in Muñoz-Ramírez et al. (2022), they identified subtle anomalies in *de novo* Parkinsonian patients by training spatial autoencoders with healthy controls DTI scans from the PPMI dataset. By utilizing 2-channel hemisphere axial slices derived from mean diffusivity (MD) and fractional anisotropy (FA) parameter maps, the authors generated joint reconstruction error maps for both the healthy control test set and the Parkinson's disease (PD) set. By evaluating the error maps per ROIs, they performed classification between controls and patients, achieving the highest geometric mean (G-mean) value for the macro regions of white matter and temporal lobe, as well as subcortical structures including the globus pallidus interna and thalamus.

With all the previously mentioned approaches, we want to highlight the diverse MRI sequences and DL models that have been used, as well as the efforts to explain the model's decisions and find disease-relevant ROIs. The latter aspect is particularly crucial when employing DL-based classification models, as the localization of spatial neurodegenerative patterns is essential in the current clinical diagnostic strategy. Therefore, explainability algorithms that provide attribution heatmaps at the pixel-level are necessary. In normative approaches, this region localization is inherently obtained through the reconstruction error map. Moreover, it is evident from these studies that either large cohort datasets or the extraction of 2D slices from MRI images are commonly employed to account for the size of the used dataset. Given the limitations of a small dataset in the present project, we adopt various learning techniques to evaluate the potential of qMRI maps in identifying PD neurodegeneration compared to more widely used MRI sequences while preserving the 3D nature of MRI scans and exploit inter-slice information to extract valuable information.

3. Material and methods

The general structure of this project comprised the following. Initially, we used an existing internal dataset part of the 7TPD project of the Danish Research Centre for Magnetic Resonance (Madelung et al., 2022), composed of 7T and 3T structural MRI of PD patients and HC subjects. From the 3T data we obtain a series of qMRI maps and, according to the requirements of the following steps, we pre-processed them (e.g. intensity rescaling, skull stripping, etc.). Then, we developed the proposed strategies of work: a) a series of ex-

periments systematically designed to compare different pre-training techniques and perform binary classification with DL models with the general goal of obtaining the best-performing pilot model, and later implement XAI algorithms to obtain attribution heatmaps; and b) a series of experiments designed to perform normative modeling of 3D neuroimaging data of a healthy population, thus creating an anomaly detector for PD patients. Each branch posed different challenges and limitations that will be addressed accordingly.

3.1. Dataset

We had access to a dataset of MRI scans acquired at 3 teslas on a Siemens Magnetom Prisma 3T scanner, comprising the multi-parameter mapping (MPM) protocol proposed by Weiskopf et al. (2013). The MPM protocol includes three multi-echo 3D fast low-angle shot (FLASH) scans: proton density (PDw), T1w, and magnetization transfer (MTw), a map of the B0 field (double gradient-echo FLASH acquisition) and a series of 3D EPI acquisitions of spin-echo (SE) and stimulated echo (STE) to map the RF transmit field B1. The dataset includes 72 subjects, out of which 49 have been diagnosed with PD and 23 are healthy controls (HC). In the PD group, there are 21 (42.85%) females and 28 males, with a mean age of 65± 10.75 years, and in the HC group, there are 8 (34.78%) females and 15 males, with a mean age of 67 ± 9.07 years.

We used the hMRI toolbox (Tabelow et al., 2019) that is based on SPM12 to obtain 1 mm high-resolution qMRI maps (Fig. 1):

- Longitudinal relaxation rate (R1 = 1/T1)
- Effective proton density (PD*)
- Magnetization transfer saturation (MTsat)
- Effective transverse relaxation rate (R2* = 1/T2*)

We used the multi-echo (TE = 2.34, 4.68, 7.02, 9.36, ..., 14.04 ms) FLASH scans: six MTw, eight PDw echoes, and eight T1w, to model their signal by the Ernst Equation (Ernst and Anderson, 2004), thus obtaining R1, PD*, and MTsat maps. The R2* map was derived through log-linear weighted least squares (WLS). To correct the quantitative data for transmit bias, the B1 transmit bias field was determined using consecutive pairs of SE/STE images corresponding to different flip angle nominal values, as well as the B0 field magnitude and phase images. Also, we corrected the RF sensitivity bias through the Unified Segmentation method, since no RF sensitivity map from the body and/or head coil was available. For further explanation of the methodology, please refer to Tabelow et al. (2019).

After obtaining the qMRI maps from all subjects, visual assessment was performed and two subjects (i.e. PD group, both males) were discarded due to data corruption problems.

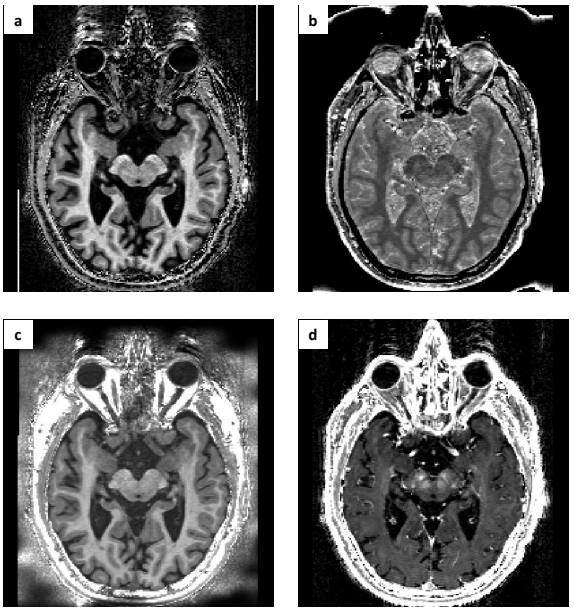

Figure 1: Quantitative MRI maps: a) MTsat, b) PD*, c) R1, d) R2*. Here displaying an axial slice at the SNc level after intensity scaling.

We preprocessed the obtained maps by first scaling the intensities to the recommended range per map: PD* = [50, 120] p.u., MTsat = [0, 2] p.u., R2* = [0, 70] s-1, R1 = [0, 1.4] s-1. After that, as required per each experiment level, we masked the volumes to obtain a region of interest accordingly. To obtain the skull-stripped volumes we utilized SynthStrip (Hoopes et al., 2022) and to obtain the brain parcellation we used SynthSeg (Billot et al., 2023), both tools available on FreeSurfer. From the brain parcellation, we had 33 labels from which we used the brainstem, left and right ventral diencephalon, left and right caudate, left and right thalamus, and left and right putamen, to create a binary mask of the brainstem and other nuclei of interest, which we will refer to from now on as the brainstem region.

For our comprehensive analysis and evaluation, we incorporated two labeled atlases: the previously mentioned SynthSeg atlas, which encompasses macroregions and selected subcortical parcellation regions, and the MNI PD25 atlas (Xiao et al., 2014), which specifically serves to MRI analysis and enables localization of pertinent PD regions (refer to the Appendix A.4 for a complete list of labels). The MNI PD25 atlas provides bilateral subcortical structure delineations, including the red nucleus (RN), substantia nigra compacta (SNc), subthalamic nucleus (STN), putamen, caudate, thalamus, and external and internal globus pallidus (GPi, GPe). To align each subject's R1 qMRI map with the PD25 T1 MPRAGE average atlas, we employed ANTs (Avants et al., 2011) for rigid, affine, and deformable spatial normalization. Cross-correlation served as the registration metric, and we used a multiresolution framework to enhance the accuracy of the

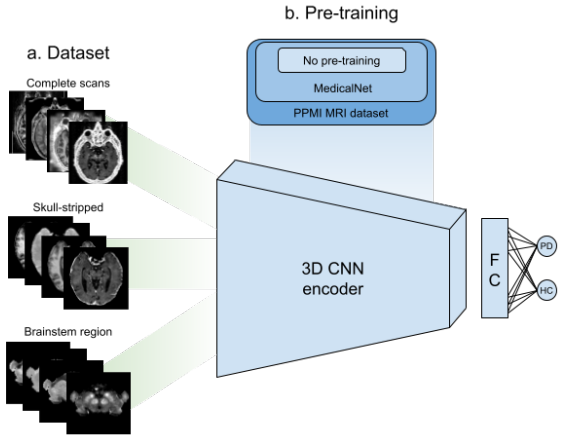

Figure 2: Binary classification strategy. a) Three different datasets per qMRI map type: complete scan, skull-stripped, and brainstem masked. b) Three different leves of model pre-training: no pre-training, using MedicalNet pre-trained model, and further pre-training using PPMI's MRI dataset of T1w images.

process.

3.2. Binary classification

We performed single-modality binary classification using transfer learning and a convolutional neural network (Fig. 2). We used the 3D Resnet (He et al., 2015) model architecture since it contains residual connections to tackle the vanishing gradient problem, and we chose the smallest version of that family to avoid over-parametrization.

Encoding full 3D scans and performing supervised binary classification requires a sufficient number of data samples to avoid overfitting the model or driving it to shortcut learning. Shortcut learning occurs when a model focuses on unintended easy-to-learn unrelated features, leading to a lack of generalization and unintuitive failures (Geirhos et al., 2020). To investigate this phenomenon, we conducted independent experiments where the model was trained with three distinct levels of region-of-interest (ROIs) (Fig. 2a). This strategy aims to limit the models to overfit to irrelevant spatial information or noise at each level.

Additionally, in the medical domain, transfer learning has emerged as a valuable technique to tackle limited data availability. This approach involves pre-training a model on a large-scale dataset, allowing it to extract general features, and subsequently fine-tuning it on a smaller dataset for the specific task at . We explored two levels of pre-training using medical datasets. Firstly, we leveraged the pre-trained models from MedicalNet (Chen et al., 2019), a framework trained on eight diverse medical image datasets (3DSeg-8), encompassing various imaging modalities such as MRI and computed tomography (CT). The authors have demonstrated notable performance improvements in segmentation and classification tasks using these models (Chen et al., 2019).

Subsequently, we extended the pre-training by incorporating MRI images from the PPMI dataset (Fig. 2b). The PPMI dataset (Marek et al., 2018) encompasses multimodal imaging data, including CT, fMRI, SPECT, PET, DTI, and MRI, collected at different time visits from two main cohorts: Parkinson's disease (PD) patients and healthy controls. For our purposes, we utilized the 3T 3D T1-weighted scans from the initial visit, resulting in a final dataset of 481 subjects (372 patients and 109 healthy controls). We utilized the MedicalNet pre-trained network and fine-tuned it using 60% of our PPMI dataset.

In this way, we have the same model architecture and three available sets of pre-trained weights (i.e. model parameters). We carried out transfer learning by replacing the pre-trained classification head with an adaptive max pool 3D layer and a single fully connected layer, with Xavier uniform parameter initialization. Because of this, for that group of parameters, we used an initial learning rate ten times larger than the group of parameters from the encoder.

Furthermore, we employed data augmentation techniques on the training set, a widely adopted approach to artificially expand the training set by applying various random transformations. The primary objective is for the model to encounter diverse variations of a single subject and learns robust features from them, for example, image orientation, rotations, or even changes in contrast. It is important to note that while traditional augmentation aims to create variations that align with the reality or the imaging technique's nature/domain, recent approaches have explored the opposite, generating synthetic data to enhance the model's robustness to various variations (Billot et al., 2023). Since our project focuses on assessing the predictive capabilities of qMRI maps, we employed simple affine transformations that would not modify the intensity content of the images as we were interested in preserving small or subtle contrast information.

For each experiment, we trained the model for a maximum of 150 epochs and implemented early stopping along with a reduce-on-plateau learning rate scheduler. To ensure optimal performance, we conducted a conservative hyperparameter tuning process, which involved evaluating different optimizers, loss functions, and initial learning rates. Considering the extensive number of experiments and the limitations of time and computational resources, we opted for a single train-validation stratified split of 80% and 20%, respectively.

3.2.1. Self-supervised learning

One of the most widely used strategies to face the limitations of data availability in the medical domain is to perform self-supervised learning (SSL). In SSL, opposite to traditional transfer learning strategies, the prelearnt features are derived from the same data through

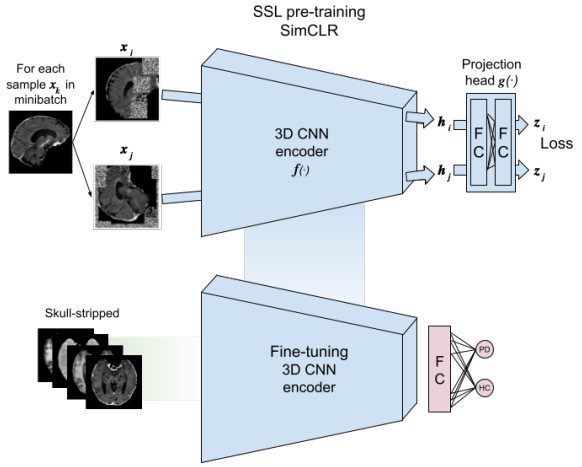

Figure 3: Self-supervised pre-training. Top: By using our own skull-stripped dataset, we train an encoder $f(\cdot)$ and a projection head $g(\cdot)$ using the SimCLR framework. Bottom: After pre-training, only the encoder is used for the downstream classification task using skull-stripped volumes.

proxy-task training. Subsequently, fine-tuning is performed in a supervised downstream task, reducing the need for a larger sample size and their annotations (Taleb et al., 2020).

The main goal of using this approach is to have the model learn an embedding space that is based on semantic similarity. In general, a spatial context proxy task is defined, such as predicting the relative position between image patches (Doersch et al., 2015), solving jigsaw puzzles (Noroozi and Favaro, 2017), or based on contrastive predictive coding (Hénaff et al., 2020; van den Oord et al., 2019). We chose to use the Simple Framework for Contrastive Learning of Representations (SimCLR) (Chen et al., 2020), since it has achieved state-of-the-art results in various computer vision tasks.

In the SimCLR framework we needed to follow two-steps. First, we created two different views from each image in the training dataset by using a heavy data augmentation composed of random flipping, affine transformations, and by masking regions of the image with noise (Fig. 3). With this, we were aiming for the model to encode information regarding the intensity distribution of different parts of the brain. Second, we trained a projection head in a contrastive manner by maximizing an agreement between differently augmented views of the same image while minimizing an agreement between views from different images. For this, we used the NT-Xent loss (Eq. 2) , which is a normalized temperature-scaled cross entropy loss that uses cosine similarity (Eq. 1).

$$\mathrm{sim}(z_i, z_j) = \frac{z_i^\top \cdot z_j}{\|z_i\| \cdot \|z_j\|} \tag{1}$$

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(z_i, z_k)/\tau)} \tag{2}$$

Where z_i, z_j are the embeddings output of the classification head coming from two augmented views of the same image, and τ is the temperature parameter.. We trained the model for 400 epochs using Adam optimizer and a learning rate of 0.001. After that, we used the SSL pre-trained network and fine-tune it for the unimodal binary classification task using only the skull-stripped volumes.

3.2.2. Explainability of Artificial Intelligence (XAI)

The primary objective of this project is to enhance the transparency of the model's predictions. To achieve this, we explored various XAI algorithms to gain insights into the model's behavior through attribution heatmaps and to gain a better understanding of disease-related spatial neurodegeneration. Typically, XAI methods are employed once a robust model with good performance and validated generalization is obtained, allowing for the assessment of any shortcut learning by visualizing relevant features.

To obtain feature importance attribution, we implemented two primary attribution algorithms: occlusion sensitivity (OS) and integrated gradients (IG), which evaluate the contribution of each input feature (voxel) to the model's output through image perturbation or manipulation.

OS is a method that involves masking or occluding parts of an input image to determine the contribution of each pixel to the output of a neural network. This method can help identifying the regions of an image that are most salient for a given classification task (Fig. 4a). With OS, we obtain an attribution heatmap at a pixel-level, meaning that we would know how much a region of specified size attributes to the model's final confidence score (Zeiler and Fergus, 2013). Because of this, the final resolution of the map depends solely on the sliding window size and stride, and furthermore, changing these parameters will influence directly the interpretation of the map regarding the relevance of the spatial information.

IG, on the other hand, computes the importance of each input feature for the neural network's output by integrating the gradient of the output with respect to the input along a straight path from a baseline input to the actual input (Fig. 4b). By integrating the gradient over this path, the method can capture the contribution of each feature to the final prediction (Sundararajan et al., 2017). In practice, the integral is efficiently approximated through summation, with the parameter m representing the number of steps between the baseline and the model's input.

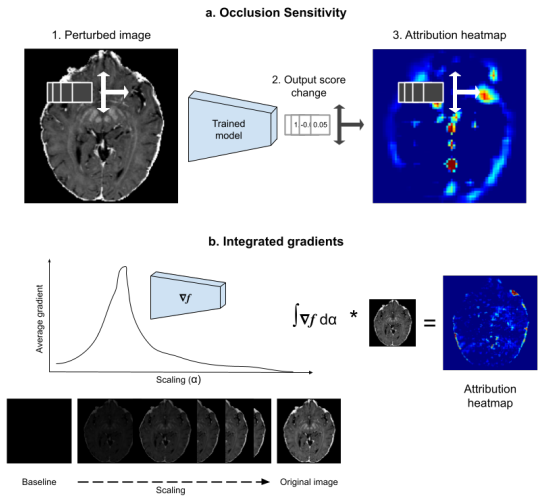

Figure 4: XAI. a. Occlusion sensitivity, by occluding parts of the image with a sliding window we measure how important that part is for the model, b. Integrated gradients, starting from an informationless baseline, the model gradients are computed and later integraded w.r.t. the scaling factor α.

For both algorithms, we utilized PyTorch's Captum implementation (Kokhlikyan et al., 2020). We determined a sliding window size of 8 voxels and a stride of 5 voxels as a suitable trade-off between granularity and computational cost for OS. Very small and overlapping patches significantly increase computation time when occluding a 3D volume. Regarding the IG algorithm, we used a zeros image as the baseline and approximated the integral using 200 steps.

For the best performing model, we obtained XAI maps for accurately predicted healthy control (HC) and Parkinson's disease (PD) subjects with the highest confidence scores, selecting a total of 8 subjects (4 HC and 4 PD) from the validation set. To identify the most significant between-group differences, we performed group average statistics per region of interest (ROI) using independent-samples t-tests and one-way ANOVA (F statistic) tests on the mean values derived from the normalized XAI maps.

3.3. Normative modelling

In our second line of investigation, we pursued an unsupervised learning approach using normative modeling to create a model of a healthy brain. Our goal was to identify variations from the norm in diseased brains. The basic concept involved constructing an identity model, where an original image served as input, and the model aimed to produce a reconstruction that closely resembled the original, thereby minimizing the reconstruction error (RE) between them. After the model is trained, when a pathological scan is provided as input, we expect to obtain a RE map indicating areas where the scan deviated from normality. This RE map functioned as an explanation heatmap for the model's predictions. Subsequently, by determining an optimal error threshold, we could evaluate the discriminative capabilities of different ROIs in distinguishing between diseased and control samples, enabling the computation of a performance metric.

As seen in section 2 State of the art, one of the most widely used architectures to perform normative modeling with brain imaging is the autoencoder (AE). In this simple structure, an image $x \in \mathbb{R}^{H\times W\times D}$ is fed through an encoder f_θ to obtain a latent space representation vector z, then a symmetrical decoder g_θ will then map z back to the reconstructed output $\hat{x} \in \mathbb{R}^{H\times W\times D}$. As concluded by Muñoz-Ramírez et al. (2022) and Baur et al. (2019), the dimensions of the latent space representation play a key role in the reconstruction error. Their experiments show that using a dense latent space $z \in \mathbb{R}^n$ performs significantly worse than having a 3D spatial latent space $z \in \mathbb{R}^{h\times w\times d}$, thus naming the autoencoder as spatial AE (sAE).

Although the AE can yield to high quality reconstructions, this type of model is not generative, meaning that as the model is allowed to create the latent space freely to output the best reconstruction, if we ever choose to create new synthetic images from a random latent embedding, we would obtain unrealistic noisy images. That is why variational autoencoders (VAE) were designed to mitigate this behaviour, as they map the original image to a latent space constraining it to follow a multivariate normal distribution, i.e. by encoding it into a mean μ and standard deviation σ latent variables. In this way, by sampling values from each variable we can obtain the latent space representation z.

To investigate how the latent space type and dimensions affects the reconstructions, we also implemented the vector-quantized VAE (VQ-VAE), a special type of VAE proposed by Oord et al. (2017). In it, the output of the encoder is mapped to the nearest point of a discrete latent space, so the latent embedding space is a codebook e of size K (i.e. vocabulary size) of vectors (i.e. words) with dimension D, $e \in \mathbb{R}^{K\times D}$. When training this framework (see details in Oord et al. (2017)), the codebook is learnt jointly with other model parameters. In order to obtain a final latent discrete representation, it would only be needed to replace each latent code by its index k from the codebook.

We employed fully convolutional 3D models to investigate the influence of depth and latent space size on the quality of reconstruction. In order to preserve the spatial information of the input data in the latent space, we opted for shallow encoders-decoders (Fig. 5a). Following the architecture of the VQ-VAE model proposed by Tudosiu et al. (Tudosiu et al., 2022), we implemented it using MONAI's Generative Models package (Cardoso et al., 2022). The VQ-VAE architecture incorporates residual units, where a selected number of residual blocks are placed after each convolutional layer. Each residual block consists of two consec-

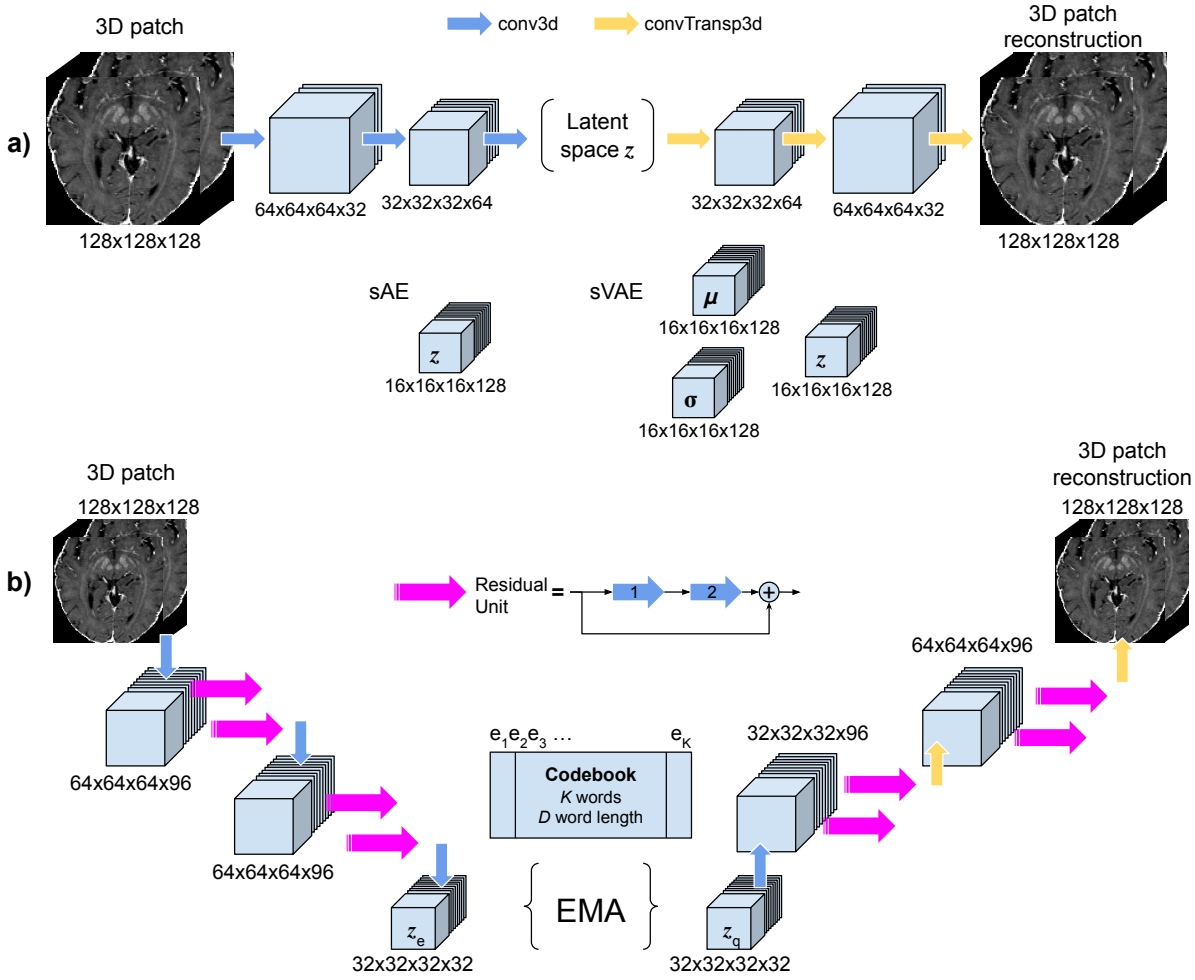

Figure 5: Normative modeling architectures. **a.** General architecture for the spatial autoencoder (sAE) and the spatial variational autoencoder (sVAE), only the latent space changes according to the type of model, z is the latent space embedding, μ and σ correspond to the normal variables. **b.** Vector quantized variational autoencoder (VQ-VAE) implemented architecture, the latent space embedding is vector quantized using a codebook of 32 words (K) of length 256 (D). The codebook is learnt along with the model's parameters using the algorithm exponential moving averages (EMA).

utive convolutional layers, with the output of the second layer being summed with the initial input. In our implementation, we used two convolutional layers in the encoder-decoder, with each layer followed by two residual blocks (Fig. 5b).

In order to augment our dataset, we adopted a patch-based approach for implementing our normative framework. This involved randomly cropping 3D patches from each volume, thereby introducing an additional parameter to consider. We chose a patch size of 128x128x128 to capture sufficient spatial information. To create the train and validation subsets, we split the HC set with a ratio of 70% for training and 30% for validation. For each training subject, we obtained nine patches from their respective volumes. During the inference phase, when testing a new image, we divided it into overlapping sub-volumes and fed each sub-volume to the model. The final reconstructed volume was then aggregated from all the sub-volumes, using a Hann window function to handle the overlapping regions and ensuring a smooth reconstruction.

For each of the model architectures, we used specific loss functions. In the simple sAE we used the L1 loss (Eq. 3). For the spatial VAE (sVAE), we used a loss function (Eq. 4) composed of the L1 norm as the reconstruction error and the Kullback-Leibler (KL) divergence to constraint the encoder to distribute all encodings around the center of the latent space (i.e. $\mu = 0$ and $\sigma = 1$). We weighted the KL term to favor the reconstruction term with a 0.9 ratio.

$$\mathcal{L} = \| x - \hat{x} \|_1 \tag{3}$$

$$\mathcal{L} = \lambda \parallel x - \hat{x} \parallel_1 + (1-\lambda)\left[-\frac{1}{2}\sum_{j=1}^{J}\left(1+\log(\sigma_j)^2-(\mu_j)^2-(\sigma_j)^2\right)\right] \quad (4)$$

Regarding the VQ-VAE training, we also used the L1 norm as reconstruction loss and the exponential moving averages (EMA) equation was used to learn the embedding space (i.e. learn the codebook parameters of the quantizer). With EMA, the embedding vectors e_i of the codebook are moved towards the encoder outputs $z_e(x)$. For the quantization loss details please refer to Oord et al. (2017). In the end, the final loss function comprised a sum of the reconstruction loss and the quantizer loss $\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{quant}$.

To assess the performance of our approach, we obtained reconstructed images and their corresponding RE maps from both the HC validation set and a sub-sample of PD subjects. The sub-sample of PD subjects was chosen to replicate the original imbalance ratio and match the gender distribution and average group age of the HC validation set. For the RE maps, we employed various measures of deviation, including the L1 norm, L2 norm, mean squared error (MSE), and the structural similarity index measure (SSIM) (Wang et al., 2004). The SSIM has been widely used in vision problems as it provides a better evaluation of perceptual image quality and structural similarity. Similar to our analysis in Section 3.2.2, we conducted group average statistics to examine between-group differences. Independent-samples t-tests and one-way ANOVA (F statistic) tests were performed on the mean and median RE values per ROI.

Finally, to evaluate the discriminant ability of each significant ROI, we established two thresholds. The first one, called abnormality threshold (a.t.), is set to detect abnormal voxels, serving for classification at voxel-level. We evaluated the a.t. as an extreme quantile value in the HC validation set error distribution. As noted by Muñoz-Ramírez et al. (2022), reconstruction errors can arise from various sources, such as data noise, loss of spatial information from the model, unaccounted variability in healthy controls, and actual anomalies caused by PD. Hence, selecting an extreme quantile (e.g., 98%) would classify only 2% of voxels in the control population as abnormal due to factors unrelated to PD. On the other hand, choosing a less restrictive quantile (e.g., 80%) would indicate that the model failed to accurately capture the distribution of controls, leading to the inclusion of genuine abnormalities within that threshold. Therefore, the a.t. can be considered a confidence threshold for the successful detection of abnormal voxels by the models.

Once the voxels in each ROI are thresholded based on the selected a.t., the proportion of anomalous voxels is determined, allowing the selection of the second threshold to evaluate the PD versus HC classification performance at the ROI level. Finally, receiver operating characteristic (ROC) curves are generated to assess the discriminating power of each ROI, and metrics such as the area under the curve (AUROC) and geometric mean (g-mean) are computed to quantify the classification performance.

At last, to further validate the classification performance of the models, we performed 5-fold cross-validation (CV). In it, to reduce the computational cost, we trained only sAE and SVAE models with 5 different gender-stratified folds, such that all 23 HC subjects are split into non-overlapping validation sets. Then, L1 RE maps are computed from each of the HC validation sets and their corresponding PD sub-sample, and the abnormality thresholds and final classification performance are evaluated. To simplify the analysis of the CV, we obtained per qMRI map the ten ROIs with the highest median AUROC (i.e. considering PD25 and Synthseg atlas labels together) that were shared between sAE and SVAE results, at a.t. of 98 quantile.

4. Results

4.1. Binary classification

In this section, we present the results of our experiments in a sequential manner, allowing readers to follow the logical progression of our arguments throughout the experiments.

To evaluate the models' performance we utilized the area under the receiver operating characteristic curve (AUROC, or ROC-AUC) and the F1-score. We chose these metrics because they are more appropriate for imbalanced datasets compared to accuracy. The ROC curve plots the true positive rate ($TPR = TP/(TP + FN)$) or sensitivity/recall against the false positive rate ($FPR = FP/(FP + TN)$) or 1 - specificity at varying decision thresholds. With AUROC we measure the model's capability of distinguishing between classes and it ranges from 0 to 1. An AUROC of 0.5 indicates no separation capacity, above 0.5 indicates good separability, and below 0.5 indicates the model predicts the inverse class. The F1-score (Eq. 5) gives more weight to the positive class (i.e., PD) by not considering true negatives (TN). It is worth noting that in our sample, a scenario where the model incorrectly predicts all subjects as PD would yield a high F1-score (e.g., 0.78) due to the larger number of positive cases (Fig. 6).

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (5)$$

Table 1 presents the classification results for all experiments per qMRI map, considering the three levels of pre-training and the three datasets used. The majority of experiments produced results similar to those depicted in Figure 6. However, some experiments demonstrated better performance in terms of high AUROC

Confusion matrix		Predicted	
		PD	HC
Target	PD	9	0
	HC	5	0

= F1-score 0.783

Figure 6: Confusion matrix of the case where all subjects are predicted with the positive label PD and the high F1 score can be misleading.

and F1-score, such as those involving the MTsat and R1 maps, utilizing complete scans and the PPMI pre-training level. For these experiments, we conducted XAI analyses on selected subjects to gain further insights into the models' predictions and evaluate whether they had learned any shortcuts for the classification task (Fig. 7). The attribution heatmaps clearly reveal that the models learned to focus on information outside the brain, specifically in the neck and skull regions, respectively for the PD and HC examples.

We then focused on the R2* map experiment using skull-stripped volumes and the PPMI pre-training level, which exhibited good performance in terms of AUROC and an improved F1 score. To gain insights into the model's decision patterns through the attribution heatmaps in a group analysis, we computed group average statistics for the subjects with the highest prediction scores. Specifically, we obtained the mean values for both OS and IG normalized attribution heatmaps for each ROI label in both the Synthseg parcellation and PD25 atlas. The ROIs that demonstrated p-values below our chosen confidence threshold ($\alpha < 0.05$) for both tests were considered the most significant (Fig. 8). The detailed results for all ROIs can be found in the Appendix A.13.

Among the most significant ROIs, in the IG attribution heatmaps we can identify some nuclei from the brainstem region for both atlases, highlighting that the pallidum (i.e. synthseg) and the globus pallidus interna and externa (i.e. PD25) are ROIs that significantly overlap and thus refer to the same region in the brain. On the other hand, in the OS attribution heatmaps, only the cerebral white matter macro-region and the lateral ventricles displayed a significant average difference between the groups.

Finally, in order to validate the performance of the R2* experiment, we performed 5-fold cross-validation(Table 2), clearly revealing that the model overfitted to that data split.

Table 3 presents the results of the SSL experiments. It is evident that across all maps, the models performed consistently, incorrectly predicting all subjects as PD and exhibiting low AUROC scores.

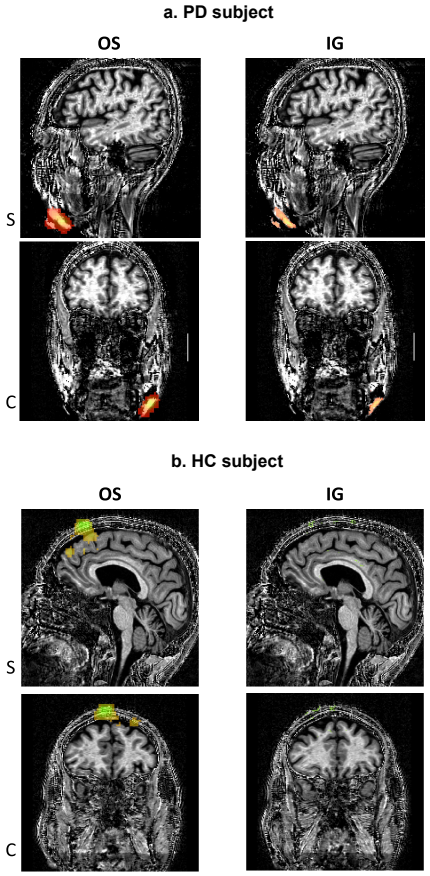

Figure 7: Occlusion sensitivity (OS) and integrated gradients (IG) XAI maps overlayed on two subjects' MTsat maps. Sagittal (S) and coronal (C) views were selected for better visualization. The heatmaps were obtained using the model trained with complete scans, and using the weights from the PPMI pre-train level. The heatmaps were thresholded to display positive attribution values and scaled for proper color intensity.

4.2. Normative modeling

Figure 9 presents an example of qualitative results for a R2* map, showcasing the reconstructed output for each model architecture and the different types of RE maps. Upon visual inspection, it is evident that the reconstructed outputs appear blurred for all models (Fig. 9a). In Figure 9b, we observe that lower values (close to 0) in the L1, L2, and MSE maps indicate fewer deviations from the normal brain, while in the case of SSIM, a higher value signifies greater similarity to the normal brain as it assesses structural similarity. For examples of MTsat, R1, and PD*, please refer to the Appendix A.14.

For each combination of qMRI map, model type, and RE type, we conducted group average statistics per ROI to assess the performance of the normative modeling approach. In the PD group, we anticipated observing an increase in mean or median error for L1, L2, and MSE, and a decrease in similarity according to SSIM. Figure 10 displays the ROIs from all experiments that exhib-

(a) MTsat map

Dataset	Pre-training Level	AUROC	F1
Complete scans	None	0.811224	0.842105
	MedicalNet	0.913265	0.941176
	PPMI	**0.97449**	**0.941176**
Skull-stripped	None	0.678571	0.782609
	MedicalNet	0.69898	0.782609
	PPMI	0.739796	0.782609
Brainstem Region	None	0.678571	0.782609
	MedicalNet	0.709184	0.782609
	PPMI	0.739796	0.782609

(b) PD* map

Dataset	Pre-training Level	AUROC	F1
Complete scans	None	0.770408	0.782609
	MedicalNet	0.785714	0.8
	PPMI	0.714286	0.666667
Skull-stripped	None	0.739796	0.782609
	MedicalNet	0.760204	0.782609
	PPMI	**0.811224**	**0.782609**
Brainstem Region	None	0.709184	0.782609
	MedicalNet	0.668367	0.782609
	PPMI	0.655612	0.782609

(c) R1 map

Dataset	Pre-training Level	AUROC	F1
Complete scans	None	0.94898	0.888889
	MedicalNet	0.938776	0.888889
	PPMI	**0.933673**	**0.947368**
Skull-stripped	None	0.872449	0.782609
	MedicalNet	0.770408	0.782609
	PPMI	0.811224	0.782609
Brainstem Region	None	0.637755	0.782609
	MedicalNet	0.668367	0.782609
	PPMI	0.596939	0.782609

(d) R2*

Dataset	Pre-training Level	AUROC	F1
Complete scans	None	0.668367	0.782609
	MedicalNet	0.760204	0.782609
	PPMI	0.663265	0.727273
Skull-stripped	None	0.770408	0.782609
	MedicalNet	0.831633	0.782609
	PPMI	**0.94898**	**0.888889**
Brainstem Region	None	0.80102	0.782609
	MedicalNet	0.831633	0.782609
	PPMI	0.811225	0.782609

Table 1: Binary classification results for the validation set, per qMRI map. The best experiment's results per qMRI map are shown in bold.

Fold	AUROC	F1
1	0.918367	0.833333
2	0.595939	0.761905
3	0.729592	0.782609
4	0.529592	0.782609
5	0.69898	0.782609

Table 2: 5-fold cross-validation results for the R2* map, using skull-stripped volumes and PPMI level of pre-training.

Map type	AUROC	F1
MTsat	0.760204	0.782609
PD*	0.719388	0.782609
R1	0.760204	0.782609
R2*	0.760204	0.782609

Table 3: SSL pre-training classification results for the validation set, per qMRI map.

ited significant p-values (i.e., $\alpha < 0.05$) for both statistical tests. We can see that each qMRI map showed at least one statistically significant ROI, with several subcortical nuclei being identified, including the right SNc in the R2* map. However, the R1 map only highlighted the left cerebral cortex as a relevant ROI. Furthermore, as expected, the error-based maps exhibited higher error values in the PD group, whereas the similarity-based map (Fig. 10a and d, right) unexpectedly showed higher values for the PD group. In cases where multiple RE map types and statistics (mean or median) yielded statistically significant results, we only report one per ROI.

For each of the significant ROIs identified by group average statistics, we evaluated the impact of the a.t. on the final performance evaluation and selected the extreme quantile that yielded the best result. For that selected a.t. we plotted the ROC curve and the G-mean (Eq. 6) and associated abnormality percentage (i.e. the second threshold that determines the optimal ROI-level classification) (Fig. 11). The highest classification results was achieved by the left globus pallidus interna (GPi) in the MTsat map (AUROC = 0.84, G-mean = 0.82) and the left subthalamic nucleus (STN) in the PD* map (AUROC = 0.84, G-mean = 0.85).

$$G - Mean = \sqrt{TPR * (1 - FPR)} \tag{6}$$

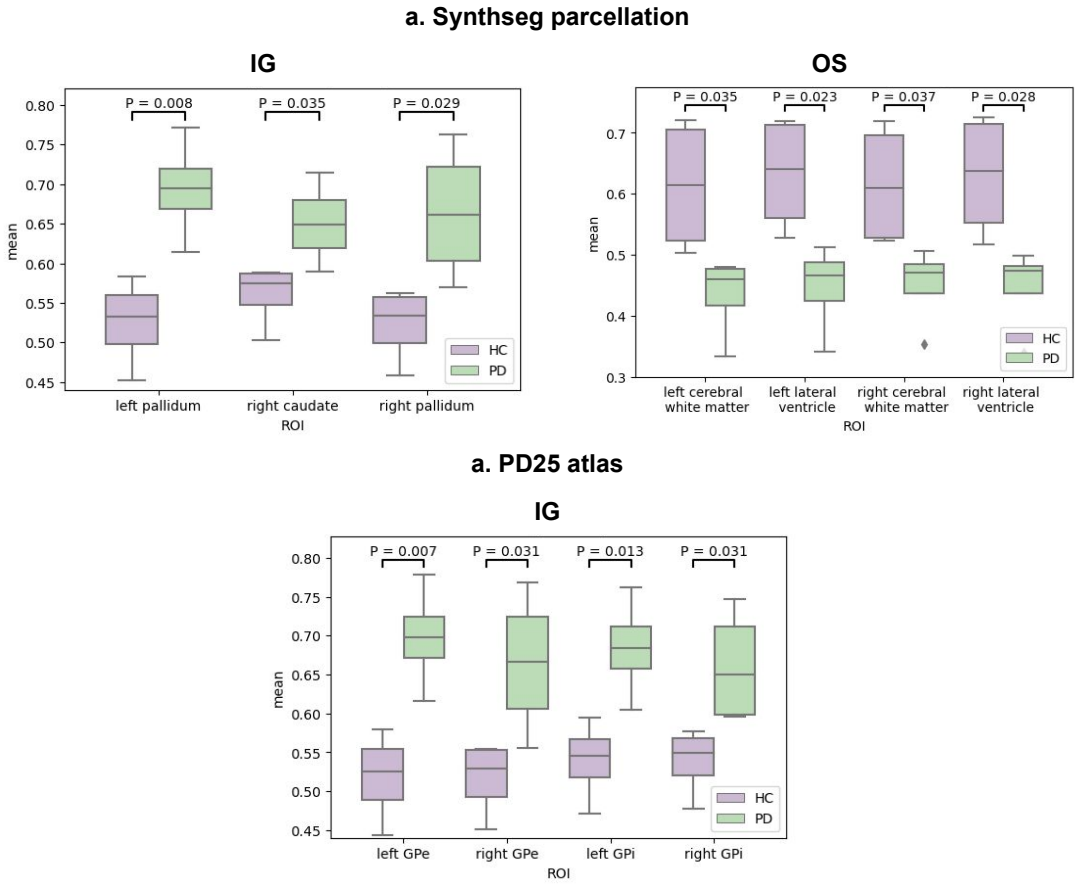

Figure 8: Statistically significant average difference at ROI level for the R2* experiment (PPMI pre-training level, using skull-stripped volumes). **a.** Integrated gradients (IG) and occlusion sensitivity (OS) results for Sythseg ROIs, **b.** IG results for PD25 atlas ROIs, OS had no significant results. Above each pair, the p-value associated to the ANOVA test is displayed, the p-value associated to the t-test was always 0.0285. Each group sampled contained the 4 subjects with higher predicted score. The XAI maps were normalized from 0 to 1 before computing the group statistics.

In Figure 12, we see the CV results per qMRI map, for the top 10 ROIs with the highest median AUROC. We can clearly identify that in all qMRI maps there were many ROIs with highly variant AUROC scores across the different folds, thus reflecting quantitatively how easy it is to over- or under-fit the tested models to very view samples. Furthermore, we can also see the sAE and SVAE models perform very differently across ROIs and qMRI maps, thus not identifying any trend of good or bad performance against each other. From the figure, we can point out the following ROIs showing a good trade-off between a high median AUROC score and small variance: for the R2* map, the left substantia nigra (lSNc) with sVAE, and the right red nucleus (rRN) with sAE; for the R1 map, the left pallidum (lPal) with sAE, and the lSNc with sVAE; for the MTsat map, the left caudate (lCau) and the rRN with sAE, and the lSNc with sVAE; and for the PD* map, the left accumbens area (lAA) and the right globus pallidus externa (rGPe) with sVAE (see appendix for tables with results).

5. Discussion

We did not obtain satisfactory results for the binary classification strategy. However, by employing XAI techniques and conducting proper model validation, we gained valuable insights during the results analysis. Upon examining the XAI attribution heatmaps (Fig. 7), we might infer that the shortcut learning was due to structural information in the form of confounds (e.g. anatomical head variations that only one sample group showed), but because the qMRI maps showed very disrupted patterns outside the brain, we believe the model focused on learning noise. These findings highlight the importance of XAI in validating deep learning models' performance. Nonetheless, interpreting XAI attribution maps can be challenging, especially when higher attributions are found within the brain. It is important to note that inferring novel disease-related neurodegeneration without prior research would be difficult without specific hypotheses, as we have for PD and the SNc and LC ROIs. Nevertheless, by carefully examining the XAI maps, we confirmed that masking the volumes effectively eliminated regions where the model

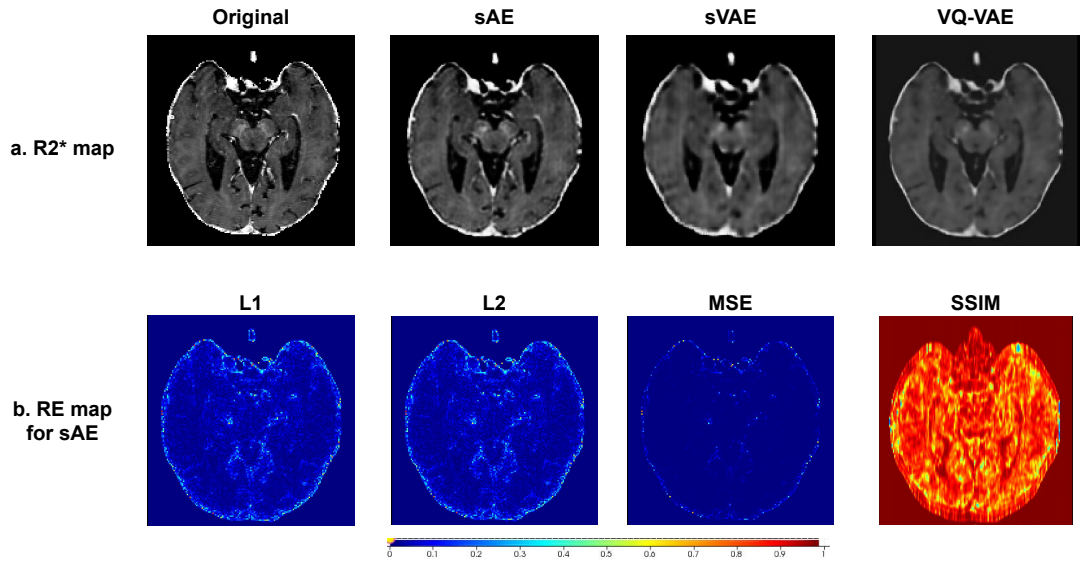

Figure 9: Qualitative results example for R2*. a. Reconstructed output for spatial autoencoder (sAE), spatial variational autoencoder (sVAE), and vector-quantized variational autoencoder (VQ-VAE). b. L1, L2, MSE, and SSSIM reconstruction error (RE) maps for sAE model.

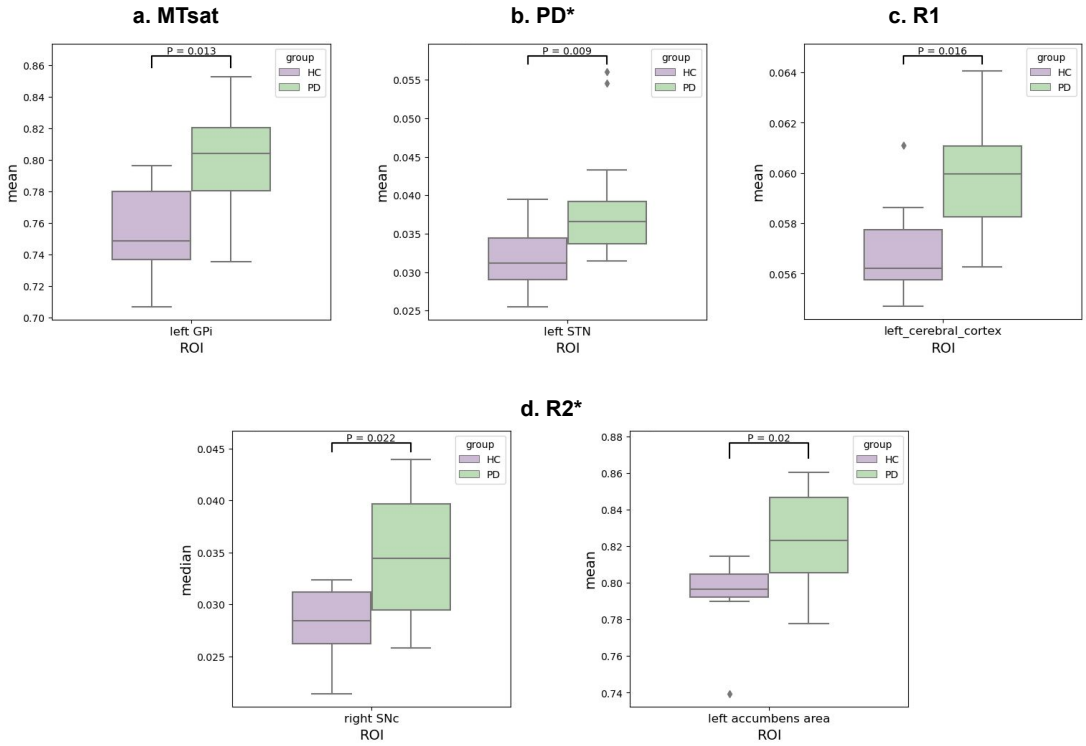

Figure 10: Statistically significant difference at ROI level for different normative modeling experiments. **a.** MTsat map, mean group differences using spatial autoencoder (sAE) and the SSIM RE map. **b.** PD* map, mean group differences using vector-quantized variational autoencoder (VQ-VAE) and the L1 RE map. **c.** R1 map, mean group differences using spatial variational autoencoder (sVAE) and the L1 RE map. **d.** R2* map, median group differences for sAE and L1 RE map (left), and mean group differences for sAE and SSIM RE map (right). Abbreviations: GPi - globus pallidus interna, STN - subthalamic nucleus, SNc - substantia nigra.

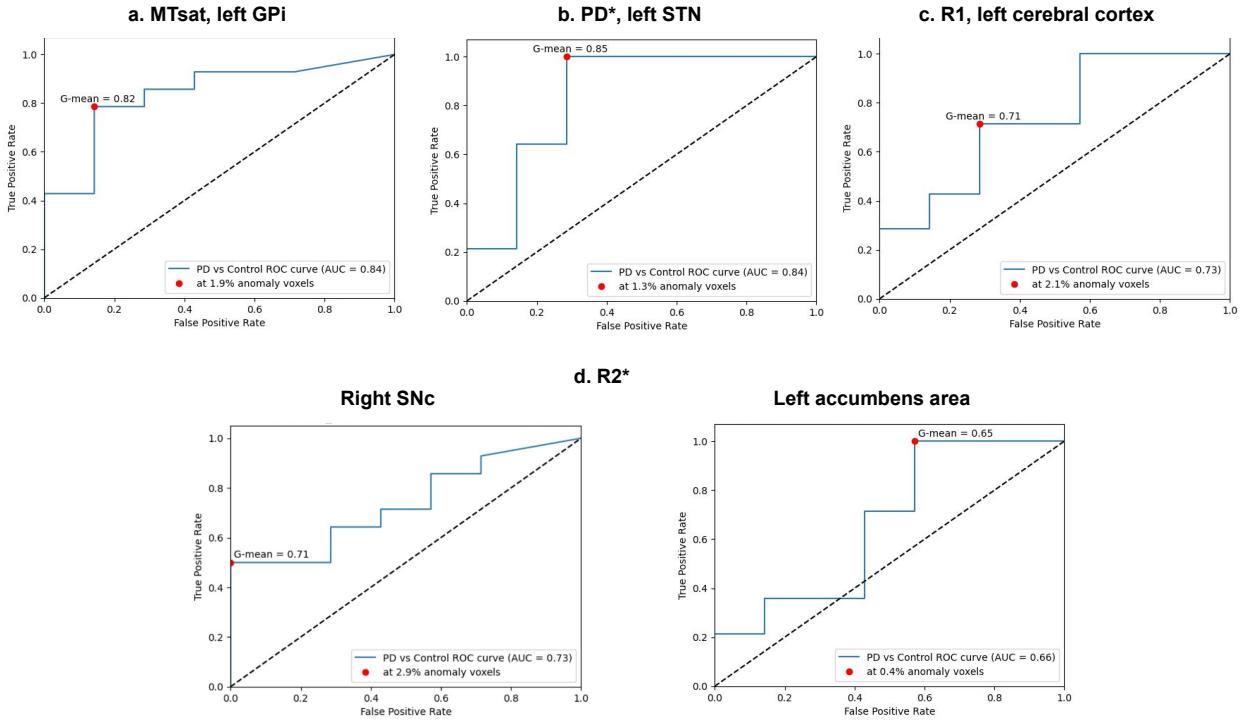

Figure 11: Classification results for normative modeling. Shown are the statistically significant ROIs and their associated ROC curve after selecting the appropriate a.t. In each ROC curve the AUROC is displayed, as well as the G-mean and corresponding abnormality percentage. **a.** Left GPi for MTsat map, at 99 quantile. **b.** Left STN for PD* map, at 98 quantile. **c.** Left cerebral cortex for R1 map, at 98 quantile. **d.** For R2* map, right SNc at 99 quantile (left) , and left accumbens area at 98 quantile (right).

exhibited shortcut learning, thus reinforcing the need to gain a deeper understanding of our data to interpret the model's decision process.

In our analysis of group average statistics for the best R2* experiment (using skull-stripped scans and PPMI pre-training level), although it was latter shown with 5-fold CV that the model overfitted to that data split, we wanted to better understand the model decision and perhaps reveal similarly any shortcut learning evidence. However, adding the previously stated considerations, it is particularly difficult to infer explanations, mainly because there are other constraints to deal with when interpreting XAI attribution maps. For instance, ablation-based algorithms like OS, where some features are dropped and the change in predictions is noted, lead to unrealistic inputs and potentially misleading interpretations when features interact when changing the size of the occluded region (Sundararajan et al., 2017). This might explain the inclusion of the right and left lateral ventricles as relevant ROIs in the OS heatmaps (Fig. 8). Additionally, gradient-based XAI algorithms like Integrated Gradients (IG) can be easily manipulated by applying imperceptible perturbations to the input, making it difficult to interpret the resulting map as a reliable explanation or to use it for assessing our general hypothesis (Dombrowski et al., 2019).

We found that our pre-training strategies to address data scarcity did not yield satisfactory results for our problem. While in some experiments, such as PD* skull-stripped volumes (Table 1b) showed an increase in AUROC accordingly to the pre-training level, there was no clear pattern indicating consistent improvement across all qMRI maps and experimental settings. We might attribute this failure to two things: domain shift and data's high dimensionality. Although transfer learning in the medical domain has shown promising results for certain problems (Chen et al., 2019), domain adaptation is still an evolving field. In our case, the two pre-training stages were performed using imaging modalities different to the qMRI maps, and this domain shift between datasets prevented a proper transfer of learning. Additionally, we still require large amounts of data and our 3D strategy might not have been the most suitable for our dataset size. For our classification task, the limited number of samples and the high-dimensional nature of imaging data, combined with a single label per image, may have restricted the model's ability to make sense of the data as a whole. Furthermore, the poor results in our SSL experiments may be attributed to the choice of transformations for the augmented views, which failed to help the model learn the relevant features for the downstream task. Moreover, our explorative experiments were scarce and a patch-based framework could be explored to increase the dataset size, or even consider trying different proxy tasks, such as the jigsaw puzzle.

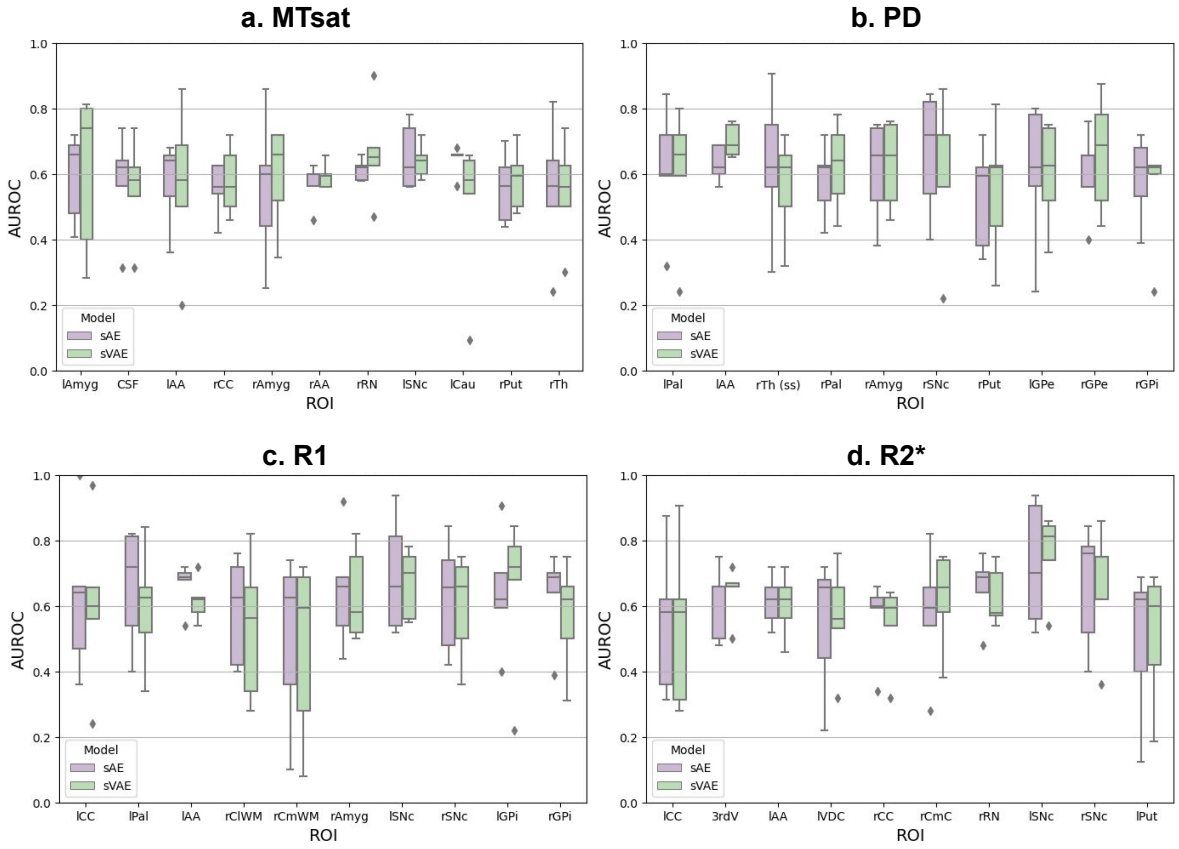

Figure 12: 5-fold CV classification results for normative modeling. Shown are top 10 ROIs with the highest median AUROC in the L1 RE map, for both sAE and sVAE models, computed at 98 quantile as a.t. Abbrevations: l/rAmyg - left/right amygdala, CSF - cerebrospinal fluid, l/rAA - left/right accumbens area, l/rCC - left/right cerebral cortex, rRN - right red nucleus, l/rSNc - left/right substantia nigra, lCau - left caudate, rPut - right putamen, rTh - right thalamus, rTH (ss) - right thalamus from Synthseg atlas, l/rPal - left/right pallidum, l/rGPi/e - left/right globus pallidus interna/externa, rClWM - right cerebral white matter, rCmWM - right cerebellum white matter, 3rdV - 3rd ventricle, lVDC - left ventral diencephalon, rCmC - right cerebellum cortex.

To determine which qMRI map was most suitable for our classification task, the only slight indication was of the brainstem region experiments, and that of the R2* map that showed higher AUROC scores compared to the other maps (Fig. 1d). However, since the model still predicted all subjects as PD (i.e., F1 score of 0.7826), we cannot definitively state that the R2* map contains better or more discriminatory information.

Regarding our normative modeling approach, we believe it was a better strategy for our problem for two main reasons. First, assessing a neurodegenerative disease such as PD as a continuum or a degree of deviation of normality better suits its diagnostic framework. Second, by generating RE maps we obtain explanation maps that allowed us to identify spatial patterns of anomalies, which was a primary goal of our project. Although the choice of error or similarity metrics to generate the RE map is crucial, it significantly reduces the limitations inherent in XAI algorithms, as we can directly interpret the error as a measure of abnormality.

Nevertheless, we still faced some limitations. The data scarcity problem still arised for this approach despite implementing a patch-based strategy to increase our dataset size. Our HC set was smaller than our PD set leading to a disproportionately smaller HC dataset, unlike most existing normative models (see section 2) that were trained with thousands of images. This directly affected our implementation when it came to choose the a.t., as it relates to the degree of confidence for the model to accurately learn the HC distribution. We can reasonably expect that the models failed to capture the full variability of healthy controls and likely overfit to patches from only 14 subjects. To properly validate our approach, it would have been necessary to sample the HC set and create different training-validation subsets. Unfortunately, due to time constraints, we could not perform this validation.

Regarding our reconstruction outputs (Fig. 9), it is evident that our model architectures struggled to accurately reconstruct high-frequency features, resulting in blurred images. Despite our efforts to preserve spatial information by tuning the size of the latent space (z)

and employing shallow fully convolutional networks, the sharpness of the reconstructions was limited by the chosen loss function. The use of L1 as a loss function inevitably drove the model to learn that a blurred image minimizes the error quickly. Furthermore, we attribute the degree of blurring to the constraints imposed by the models on the latent space distribution, whether it be following a multivariate normal distribution (sVAE) or being discrete (VQ-VAE), compared to the less restrictive and non-generative sAE.

In our analysis of group average statistics from the single train-validation split, we found that when using L1, L2, or MSE, the error difference between the PD and control groups was higher than that within the control group, supporting the argument that the PD group deviates from the controls (10. However, we obtained contradictory results for the SSIM maps, where the similarity value should have been higher for the HC group compared to the PD group. This discrepancy may be attributed to the structural component of the SSIM, which is highly sensitive to edge and contour information—factors strongly affected by the blurry nature of the reconstructed outputs (Renieblas et al., 2017).

Finally, the classification results obtained from the single train-val split indicate that while the right SNc in the R2* map displayed a significant group difference, it did not exhibit sufficient discriminant ability to effectively distinguish between PD patients and HC in our test set, in contrast to other ROIs that demonstrated superior performance. This relatively poor outcome might be attributed to an under-fitted model. Therefore, through cross-validation, we can draw a more comprehensive conclusion about the overall performance, favoring the left SNc, which yielded the highest median AUROC amongst all other ROIs and qMRI maps, performing well on all folds but one. We believe that the improved results observed in the SNc of the R2* map can be attributed to the lack of high-frequency reconstruction capabilities in our models, making it more challenging to reconstruct subtle differences from the other qMRI maps, and we knew a priori that the R2* maps contain the highest contrast differences in the SNc.

6. Conclusions

In this thesis project, our goal was to explore the potential of deep learning (DL) models in uncovering novel insights into Parkinson's Disease (PD) neurodegeneration. To achieve this, we employed explainable artificial intelligence (XAI) algorithms to enhance the transparency of complex model decisions and identify relevant regions of interest. We initially pursued the traditional binary classification strategy, but encountered challenges in obtaining satisfactory results. However, this approach provided valuable insights, including the identification of shortcut learning, model validation and overfitting assessment, and the understanding of transfer learning capabilities and limitations.

In our second strategy, normative modeling, we achieved better-suited models for studying the disease and obtained intrinsic explainable reconstruction error maps that led to more interpretable conclusions. However, the results were modest due to the limitations of our generative models to adequately reconstruct important high-frequency information. And more importantly, to properly model the distribution of normal brains, a bigger HC dataset is required, as varying results can be highlighted from the cross-validation.

Our intention was to leverage the high spatial resolution of MRI scans by employing 3D models. However, the small number of samples in our dataset suggests that implementing a 2D or 2.5D model would have been more appropriate. We also aimed to utilize pre-trained models and publicly available datasets, but we faced domain shift limitations to effectively transfer the learned knowledge to our specific classification task.

As future work, it would be beneficial to explore multi-modal strategies, such as combining the four quantitative MRI (qMRI) maps at 3T or integrating additional imaging data, such as 7T NM- and iron-sensitive images, as well as clinical data like PD scale ratings. However, careful consideration must be given to address the challenge of the curse of dimensionality and ensure proper interpretation and explanation of the model's decisions.

Moreover, it would be particularly interesting and clinically relevant to investigate multi-label or multi-class classification approaches, as PD encompasses a continuum of multiple motor and non-motor symptoms.

Acknowledgments

I would like to express my gratitude to Hartwig Siebner for providing me with the opportunity to carry out my thesis internship at the Danish Research Centre for Magnetic Resonance (DRCMR) and for his unwavering support in promoting transparency in research and fostering a passion for MRI. I am deeply thankful to David Meder for his valuable insights and expertise in elucidating the various pathological processes associated with Parkinson's Disease, which greatly contributed to the relevance of my project. I am indebted to the entire DRCMR staff for their generous assistance in answering my inquiries regarding MRI, data processing, and Parkinson's Disease, as their guidance was instrumental in the successful completion of this research endeavor. Special thanks go to José Bernal for his invaluable assistance in visualizing a more effective deep learning approach and to Vladyslav Zalevskyi for his unwavering availability, patient guidance, and for sharing this enriching experience with me. Lastly, I extend my appreciation to all the researchers who have made their research, code, and methodologies openly available, as

well as to the PPMI group for graciously granting me access to their invaluable imaging data.

References

Arribarat, G., Péran, P., 2020. Quantitative MRI markers in parkinson's disease and parkinsonian syndromes. Current Opinion in Neurology 33, 222–229. doi:10.1097/wco.0000000000000796.

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage 54, 2033–2044. URL: https://doi.org/10.1016/j.neuroimage.2010.09.025, doi:10.1016/j.neuroimage.2010.09.025.

Bae, Y.J., Kim, J.M., Sohn, C.H., Choi, J.H., Choi, B.S., Song, Y.S., Nam, Y., Cho, S.J., Jeon, B., Kim, J.H., 2021. Imaging the substantia nigra in parkinson disease and other parkinsonian syndromes. Radiology 300, 260–278. doi:10.1148/radiol.2021203341. pMID: 34100679.

Baur, C., Wiestler, B., Albarqouni, S., Navab, N., 2019. Deep autoencoding models for unsupervised anomaly segmentation in brain MR images, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, pp. 161–169. URL: https://doi.org/10.1007/978-3-030-11723-8_16, doi:10.1007/978-3-030-11723-8_16.

Billot, B., Greve, D.N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A.V., Iglesias, J.E., 2023. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. Medical Image Analysis 86, 102789. URL: https://www.sciencedirect.com/science/article/pii/S1361841523000506, doi:https://doi.org/10.1016/j.media.2023.102789.

Biondetti, E., Gaurav, R., Yahia-Cherif, L., Mangone, G., Pyatigorskaya, N., Valabrègue, R., Ewenczyk, C., Hutchison, M., François, C., Arnulf, I., Corvol, J.C., Vidailhet, M., Lehéricy, S., 2020. Spatiotemporal changes in substantia nigra neuromelanin content in Parkinson's disease. Brain 143, 2757–2770. URL: https://academic.oup.com/brain/article/143/9/2757/5898381, doi:10.1093/brain/awaa216.

Biondetti, E., Santin, M.D., Valabrègue, R., Mangone, G., Gaurav, R., Pyatigorskaya, N., Hutchison, M., Yahia-Cherif, L., Villain, N., Habert, M.O., Arnulf, I., Leu-Semenescu, S., Dodet, P., Vila, M., Corvol, J.C., Vidailhet, M., Lehéricy, S., 2021. The spatiotemporal changes in dopamine, neuromelanin and iron characterizing Parkinson's disease. Brain 144, 3114–3125. URL: https://academic.oup.com/brain/article/144/10/3114/6274641, doi:10.1093/brain/awab191.

Camacho, M., Wilms, M., Mouches, P., Almgren, H., Souza, R., Camicioli, R., Ismail, Z., Monchi, O., Forkert, N.D., 2023. Explainable classification of parkinson's disease using deep learning trained on a large multi-center database of t1-weighted mri datasets. NeuroImage: Clinical 38, 103405. URL: https://www.sciencedirect.com/science/article/pii/S2213158223000943, doi:https://doi.org/10.1016/j.nicl.2023.103405.

Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Darestani, M.Z., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B.S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P.F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L.A.D., Roth, H.R., Xu, D., Bericat, D., Floca, R., Zhou, S.K., Shuaib, H., Farahani, K., Maier-Hein, K.H., Aylward, S., Dogra, P., Ourselin, S., Feng, A., 2022. Monai: An open-source framework for deep learning in healthcare.

Chaddad, A., Peng, J., Xu, J., Bouridane, A., 2023. Survey of explainable AI techniques in healthcare. Sensors 23, 634. URL: https://doi.org/10.3390/s23020634, doi:10.3390/s23020634.

Chaki, J., Woźniak, M., 2023. Deep learning for neurodegenerative disorder (2016 to 2022): A systematic review. Biomedical Signal Processing and Control 80, 104223. URL: https://linkinghub.elsevier.com/retrieve/pii/S1746809422006772, doi:10.1016/j.bspc.2022.104223.

Chen, S., Ma, K., Zheng, Y., 2019. Med3d: Transfer learning for 3d medical image analysis. doi:https://doi.org/10.48550/arXiv.1904.00625.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. doi:https://doi.org/10.48550/arXiv.2002.05709.

DeMaagd, G., Philip, A., 2015. Parkinson's disease and its management: Part 1: Disease entity, risk factors, pathophysiology, clinical presentation, and diagnosis. P T 40, 504–532.

Doersch, C., Gupta, A., Efros, A.A., 2015. Unsupervised visual representation learning by context prediction, in: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1422–1430. doi:10.1109/ICCV.2015.167.

Dombrowski, A.K., Alber, M., Anders, C.J., Ackermann, M., Müller, K.R., Kessel, P., 2019. Explanations can be manipulated and geometry is to blame. URL: http://arxiv.org/abs/1906.07983. arXiv:1906.07983 [cs, stat].

Ernst, R.R., Anderson, W.A., 2004. Application of Fourier Transform Spectroscopy to Magnetic Resonance. Review of Scientific Instruments 37, 93–102. URL: https://doi.org/10.1063/1.1719961, doi:10.1063/1.1719961.

Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A., 2020. Shortcut learning in deep neural networks. Nature Machine Intelligence 2, 665–673. URL: https://doi.org/10.1038/s42256-020-00257-z, doi:10.1038/s42256-020-00257-z.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. doi:https://doi.org/10.48550/arXiv.1512.03385.

Hoopes, A., Mora, J.S., Dalca, A.V., Fischl, B., Hoffmann, M., 2022. Synthstrip: skull-stripping for any brain image. NeuroImage 260, 119474. URL: https://www.sciencedirect.com/science/article/pii/S1053811922005900, doi:https://doi.org/10.1016/j.neuroimage.2022.119474.

Huang, L., Ye, X., Yang, M., Pan, L., hua Zheng, S., 2023. MNC-net: Multi-task graph structure learning based on node clustering for early parkinson's disease diagnosis. Computers in Biology and Medicine 152, 106308. URL: https://doi.org/10.1016/j.compbiomed.2022.106308, doi:10.1016/j.compbiomed.2022.106308.

Hénaff, O.J., Srinivas, A., Fauw, J.D., Razavi, A., Doersch, C., Eslami, S.M.A., van den Oord, A., 2020. Data-efficient image recognition with contrastive predictive coding. doi:https://doi.org/10.48550/arXiv.1905.09272.

Kim, H.E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M.E., Ganslandt, T., 2022. Transfer learning for medical image classification: a literature review. BMC Medical Imaging 22. URL: https://doi.org/10.1186/s12880-022-00793-7, doi:10.1186/s12880-022-00793-7.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., Reblitz-Richardson, O., 2020. Captum: A unified and generic model interpretability library for pytorch.

Madelung, C.F., Meder, D., Fuglsang, S.A., Marques, M.M., Boer, V.O., Madsen, K.H., Petersen, E.T., Hejl, A., Løkkegaard, A., Siebner, H.R., 2022. Locus Coeruleus Shows a Spatial Pattern of Structural Disintegration in Parkinson's Disease. Movement Disorders 37, 479–489. URL: https://onlinelibrary.wiley.com/doi/10.1002/mds.28945, doi:10.1002/mds.28945.

Marek, K., Chowdhury, S., Siderowf, A., Lasch, S., Coffey, C.S., Caspell-Garcia, C., Simuni, T., Jennings, D., Tanner, C.M., Trojanowski, J.Q., Shaw, L.M., Seibyl, J., Schuff, N., Singleton, A., Kieburtz, K., Toga, A.W., Mollenhauer, B., Galasko, D., Chahine, L.M., Weintraub, D., Foroud, T., Tosun-Turgut, D., Poston, K., Arnedo, V., Frasier, M., Sherer, T., Bressman, S., Merchant, M., Poewe, W., Kopil, C., Naito, A., Dorsey, R., Casaceli, C., Daegele,

N., Albani, J., Uribe, L., Foster, E., Long, J., Seedorff, N., Crawford, K., Smith, D., Casalin, P., Malferrari, G., Halter, C., Heathers, L., Russell, D., Factor, S., Hogarth, P., Amara, A., Hauser, R., Jankovic, J., Stern, M., Hu, S.C., Todd, G., Saunders-Pullman, R., Richard, I., Saint-Hilaire, H., Seppi, K., Shill, H., Fernandez, H., Trenkwalder, C., Oertel, W., Berg, D., Brockman, K., Wurster, I., Rosenthal, L., Tai, Y., Pavese, N., Barone, P., Isaacson, S., Espay, A., Rowe, D., Brandabur, M., Tetrud, J., Liang, G., Iranzo, A., Tolosa, E., Marder, K., Sanchez, M., Stefanis, L., Marti, M., Martinez, J., Corvol, J.C., Assly, O., Brillman, S., Giladi, N., Smejdir, D., Pelaggi, J., Kausar, F., Rees, L., Sommerfield, B., Cresswell, M., Blair, C., Williams, K., Zimmerman, G., Guthrie, S., Rawlins, A., Donharl, L., Hunter, C., Tran, B., Darin, A., Venkov, H., Thomas, C.A., James, R., Heim, B., Deritis, P., Sprenger, F., Raymond, D., Willeke, D., Obradov, Z., Mule, J., Monahan, N., Gauss, K., Fontaine, D., Szpak, D., McCoy, A., Dunlop, B., Payne, L., Ainscough, S., Carvajal, L., Silverstein, R., Espay, K., Ranola, M., Rezola, E., Santana, H., Stamelou, M., Garrido, A., Carvalho, S., Kristiansen, G., Specketer, K., Mirlman, A., Facheris, M., Soares, H., Mintun, A., Cedarbaum, J., Taylor, P., Jennings, D., Slieker, L., McBride, B., Watson, C., Montagut, E., Sheikh, Z., Bingol, B., Forrat, R., Sardi, P., Fischer, T., Reith, D., Egebjerg, J., Larsen, L., Breysse, N., Meulien, D., Saba, B., Kiyasova, V., Min, C., McAvoy, T., Umek, R., Iredale, P., Edgerton, J., Santi, D., Czech, C., Boess, F., Sevigny, J., Kremer, T., Grachev, I., Merchant, K., Avbersek, A., Muglia, P., Stewart, A., Prashad, R., and, J.T., 2018. The parkinson's progression markers initiative (PPMI) – establishing a PD biomarker cohort. Annals of Clinical and Translational Neurology 5, 1460–1477. URL: https://doi.org/10.1002/acn3.644, doi:10.1002/acn3.644.

Marquand, A.F., Kia, S.M., Zabihi, M., Wolfers, T., Buitelaar, J.K., Beckmann, C.F., 2019. Conceptualizing mental disorders as deviations from normative functioning. Molecular Psychiatry 24, 1415–1424. URL: https://doi.org/10.1038/s41380-019-0441-1, doi:10.1038/s41380-019-0441-1.

Meder, D., Herz, D.M., Rowe, J.B., Lehéricy, S., Siebner, H.R., 2019. The role of dopamine in the brain - lessons learned from Parkinson's disease. NeuroImage 190, 79–93. URL: https://linkinghub.elsevier.com/retrieve/pii/S1053811918320925, doi:10.1016/j.neuroimage.2018.11.021.

Muñoz-Ramírez, V., Kmetzsch, V., Forbes, F., Meoni, S., Moro, E., Dojat, M., 2022. Subtle anomaly detection: Application to brain MRI analysis of de novo Parkinsonian patients. Artificial Intelligence in Medicine 125, 102251. URL: https://linkinghub.elsevier.com/retrieve/pii/S0933365722000161, doi:10.1016/j.artmed.2022.102251.

Noroozi, M., Favaro, P., 2017. Unsupervised learning of visual representations by solving jigsaw puzzles. doi:https://doi.org/10.48550/arXiv.1603.09246.

van den Oord, A., Li, Y., Vinyals, O., 2019. Representation learning with contrastive predictive coding. doi:https://doi.org/10.48550/arXiv.1807.03748.

Oord, A.v.d., Vinyals, O., Kavukcuoglu, K., 2017. Neural discrete representation learning. URL: https://arxiv.org/abs/1711.00937, doi:10.48550/ARXIV.1711.00937.

Pinaya, W.H.L., Scarpazza, C., Garcia-Dias, R., Vieira, S., Baecker, L., F da Costa, P., Redolfi, A., Frisoni, G.B., Pievani, M., Calhoun, V.D., Sato, J.R., Mechelli, A., 2021a. Using normative modelling to detect disease progression in mild cognitive impairment and Alzheimer's disease in a cross-sectional multicohort study. Scientific Reports 11, 15746. URL: https://www.nature.com/articles/s41598-021-95098-0, doi:10.1038/s41598-021-95098-0.

Pinaya, W.H.L., Tudosiu, P.D., Gray, R., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J., 2021b. Unsupervised Brain Anomaly Detection and Segmentation with Transformers. URL: http://arxiv.org/abs/2102.11650. arXiv:2102.11650 [cs, eess, q-bio].

Poston, K.L., Ua Cruadhlaoich, M.A.I., Santoso, L.F., Bernstein, J.D., Liu, T., Wang, Y., Rutt, B., Kerchner, G.A., Zeineh, M.M., 2020. Substantia Nigra Volume Dissociates Bradykinesia and Rigidity from Tremor in Parkinson's Disease: A 7 Tesla Imaging Study. Journal of Parkinson's Disease 10, 591–604. doi:10.3233/JPD-191890.

Renieblas, G.P., Nogués, A.T., González, A.M., Gómez-Leon, N., del Castillo, E.G., 2017. Structural similarity index family for image quality assessment in radiological images. Journal of Medical Imaging 4, 035501. URL: https://doi.org/10.1117/1.jmi.4.3.035501, doi:10.1117/1.jmi.4.3.035501.

Rizek, P., Kumar, N., Jog, M.S., 2016. An update on the diagnosis and treatment of parkinson disease. Canadian Medical Association Journal 188, 1157–1165. URL: https://doi.org/10.1503/cmaj.151179, doi:10.1503/cmaj.151179.

Rutherford, S., Kia, S.M., Wolfers, T., Fraza, C., Zabihi, M., Dinga, R., Berthet, P., Worker, A., Verdi, S., Ruhe, H.G., Beckmann, C.F., Marquand, A.F., 2022. The normative modeling framework for computational psychiatry. Nature Protocols 17, 1711–1734. URL: https://www.nature.com/articles/s41596-022-00696-5, doi:10.1038/s41596-022-00696-5.

Shinde, S., Prasad, S., Saboo, Y., Kaushick, R., Saini, J., Pal, P.K., Ingalhalikar, M., 2019. Predictive markers for parkinson's disease using deep neural nets on neuromelanin sensitive MRI. NeuroImage: Clinical 22, 101748. URL: https://doi.org/10.1016/j.nicl.2019.101748, doi:10.1016/j.nicl.2019.101748.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M., 2017. Smoothgrad: removing noise by adding noise. doi:https://doi.org/10.48550/arxiv.1706.03825.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks. URL: https://arxiv.org/abs/1703.01365, doi:10.48550/ARXIV.1703.01365.

Tabelow, K., Balteau, E., Ashburner, J., Callaghan, M.F., Draganski, B., Helms, G., Kherif, F., Leutritz, T., Lutti, A., Phillips, C., Reimer, E., Ruthotto, L., Seif, M., Weiskopf, N., Ziegler, G., Mohammadi, S., 2019. hMRI – A toolbox for quantitative MRI in neuroscience and clinical research. NeuroImage 194, 191–210. URL: https://linkinghub.elsevier.com/retrieve/pii/S1053811919300291, doi:10.1016/j.neuroimage.2019.01.029.

Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C., 2020. 3d self-supervised methods for medical imaging, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 18158–18172. URL: https://proceedings.neurips.cc/paper/2020/file/d2dc6368837861b42020ee72b0896182-Paper.pdf.

Trujillo, P., Summers, P.E., Ferrari, E., Zucca, F.A., Sturini, M., Mainardi, L.T., Cerutti, S., Smith, A.K., Smith, S.A., Zecca, L., Costa, A., 2017. Contrast mechanisms associated with neuromelanin-MRI: Neuromelanin-MRI Contrast. Magnetic Resonance in Medicine 78, 1790–1800. URL: https://onlinelibrary.wiley.com/doi/10.1002/mrm.26584, doi:10.1002/mrm.26584.

Tschuchnig, M.E., Gadermayr, M., 2021. Anomaly detection in medical imaging – a mini review URL: https://arxiv.org/abs/2108.11986, doi:10.48550/ARXIV.2108.11986.

Tudosiu, P.D., Pinaya, W.H.L., Graham, M.S., Borges, P., Fernandez, V., Yang, D., Appleyard, J., Novati, G., Mehra, D., Vella, M., Nachev, P., Ourselin, S., Cardoso, J., 2022. Morphology-preserving autoregressive 3d generative modelling of the brain. URL: https://arxiv.org/abs/2209.03177, doi:10.48550/ARXIV.2209.03177.

Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E., 2004. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13, 600–612. URL: https://doi.org/10.1109/tip.2003.819861, doi:10.1109/tip.2003.819861.

Weiskopf, N., Suckling, J., Williams, G., Correia, M.M., Inkster, B., Tait, R., Ooi, C., Bullmore, E.T., Lutti, A., 2013. Quantitative multi-parameter mapping of R1, PD*, MT, and R2* at 3T: a multi-center validation. Frontiers in Neuroscience 7. URL: http://journal.frontiersin.

org/article/10.3389/fnins.2013.00095/abstract, doi:10.3389/fnins.2013.00095.

Wenger, E., Polk, S.E., Kleemeyer, M.M., Weiskopf, N., Bodammer, N.C., Lindenberger, U., Brandmaier, A.M., 2021. Reliability of quantitative multiparameter maps is high for MT and PD but attenuated for R1 and R2* in healthy young adults. preprint. Neuroscience. URL: http://biorxiv.org/lookup/doi/10.1101/2021.11.10.467254, doi:10.1101/2021.11.10.467254.

Xiao, Y., Fonov, V., Bériault, S., Subaie, F.A., Chakravarty, M.M., Sadikot, A.F., Pike, G.B., Collins, D.L., 2014. Multi-contrast unbiased MRI atlas of a parkinson's disease population. International Journal of Computer Assisted Radiology and Surgery 10, 329–341. URL: https://doi.org/10.1007/s11548-014-1068-y, doi:10.1007/s11548-014-1068-y.

Zeiler, M.D., Fergus, R., 2013. Visualizing and understanding convolutional networks. URL: https://arxiv.org/abs/1311.2901, doi:10.48550/ARXIV.1311.2901.

Zucca, F.A., Segura-Aguilar, J., Ferrari, E., Muñoz, P., Paris, I., Sulzer, D., Sarna, T., Casella, L., Zecca, L., 2017. Interactions of iron, dopamine and neuromelanin pathways in brain aging and Parkinson's disease. Progress in Neurobiology 155, 96–119. URL: https://linkinghub.elsevier.com/retrieve/pii/S030100821500101X, doi:10.1016/j.pneurobio.2015.09.012.

Appendix A. Extra figures

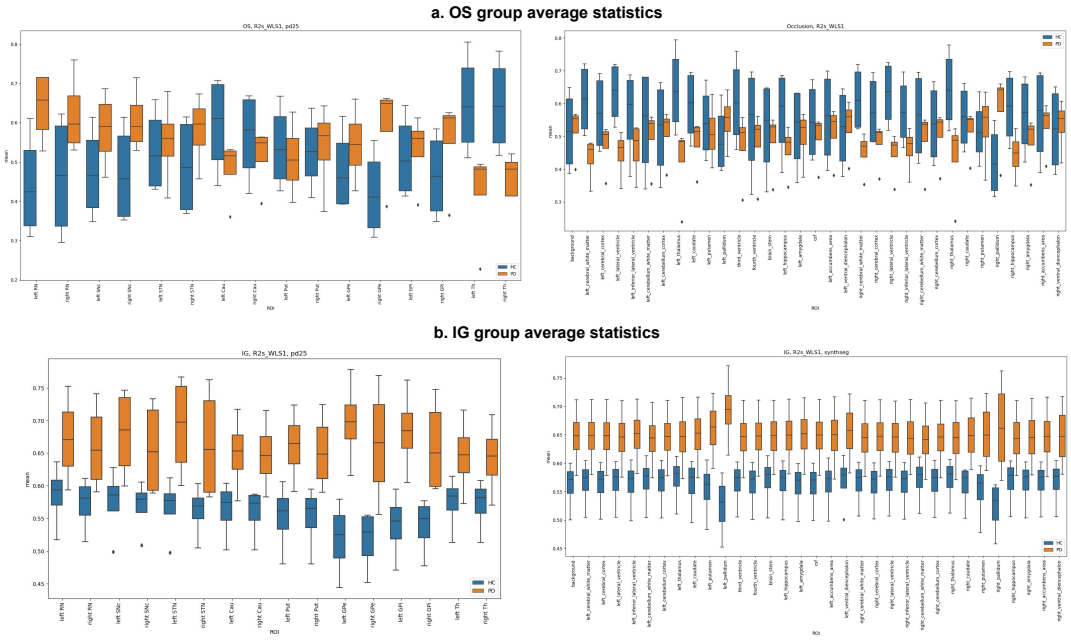

Figure A.13: Group average statistics for R2* experiment, using normalized OS and IG maps

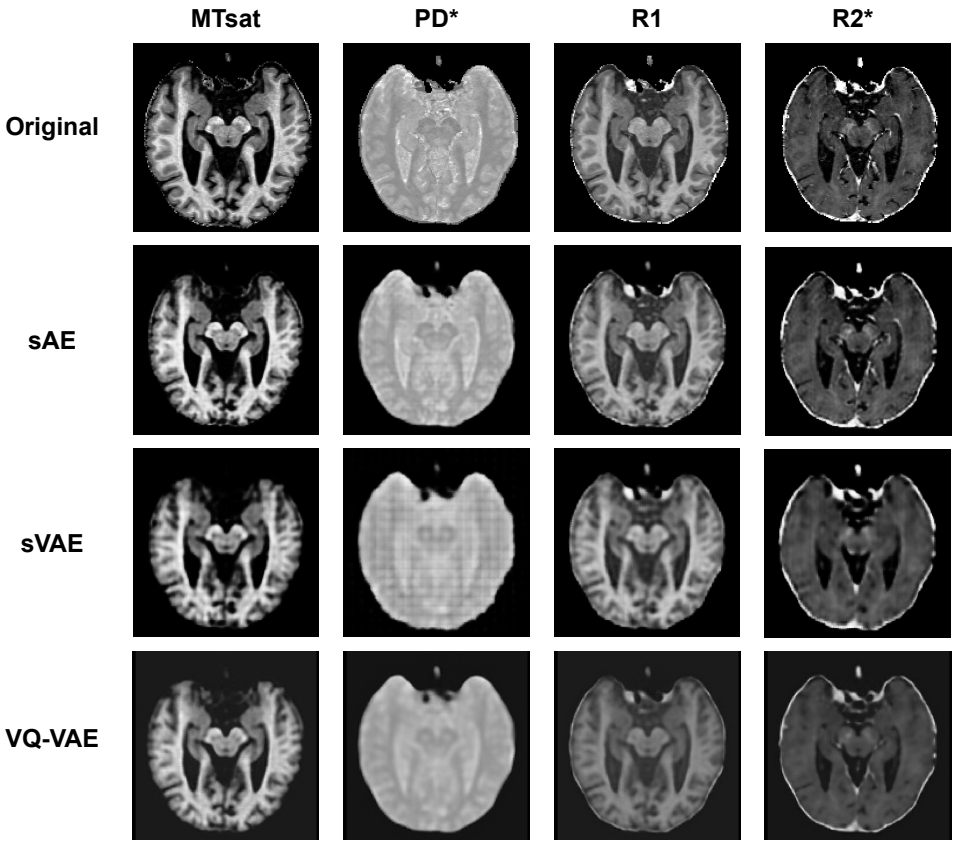

Figure A.14: Reconstruction examples for all qMRI maps and the three different model architectures.

PD25 atlas		Synthseg labels			
Label	Nuclei	Label	ROI	Label	ROI
1	Left red nucleus	2	Left cerebral white matter	26	Left accumbens area
3	Left substantia nigra	3	Left cerebral cortex	28	Left ventral DC
5	Left subthalamic nucleus	4	Left lateral ventricle	41	Right cerebral white matter
7	Left caudate	5	Left inferior lateral ventricle	42	Right cerebral cortex
9	Left putamen	7	Left cerebellum white matter	43	Right lateral ventricle
11	Left globus pallidus externa	8	Left cerebellum cortex	44	Right inferior lateral ventricle
13	Left globus pallidus interna	10	Left thalamus	46	Right cerebellum white matter
15	Left thalamus	11	Left caudate	47	Right cerebellum cortex
2	Right red nucleus	12	Left putamen	49	Right thalamus
4	Right substantia nigra	13	Left pallidum	50	Right caudate
6	Right subthalamic nucleus	14	3rd ventricle	51	Right putamen
8	Right caudate	15	4th ventricle	52	Right pallidum
10	Right putamen	16	Brain-stem	53	Right hippocampus
12	Right globus pallidus externa	17	Left hippocampus	54	Right amygdala
14	Right globus pallidus interna	18	Left amygdala	58	Right accumbens area
16	Right thalamus	24	CSF	60	Right ventral DC

Table A.4: Atlas labels

ROI	sAE		sVAE	
	Median (IQR)	Mean (STD)	Median (IQR)	Mean (STD)
CSF	0.62 (0.078)	0.58 (0.160)	0.58 (0.089)	0.56 (0.157)
lAmyg	0.66 (0.208)	0.59 (0.139)	**0.74 (0.400)**	0.61 (0.248)
lCau	**0.66 (0.004)***	0.64 (0.046)	0.58 (0.100)	0.50 (0.233)
lSNc	0.62 (0.178)	**0.65 (0.102)**	0.64 (0.056)	0.64 (0.054)*
lAA	0.64 (0.125)	0.57 (0.132)	0.58 (0.188)	0.57 (0.245)
rAmyg	0.60 (0.185)	0.56 (0.227)	0.66 (0.199)	0.59 (0.161)
rCC	0.56 (0.085)	0.55 (0.084)	0.56 (0.156)	0.58 (0.108)
rPut	0.56 (0.160)	0.56 (0.110)	0.59 (0.125)	0.58 (0.098)
rRN	0.62 (0.045)	0.61 (0.034)*	0.65 (0.055)	**0.66 (0.155)**
rTh	0.56 (0.140)	0.55 (0.212)	0.56 (0.125)	0.55 (0.163)
rAA	0.60 (0.038)	0.57 (0.065)	0.59 (0.040)*	0.59 (0.039)

Table A.5: MTsat CV results. Results in bold are the highest value in each column. Results with * have the lowest IQR or STD. Abbreviations: IQR - interquantile range, STD - standard deviation.

ROI	sAE		sVAE	
	Median (IQR)	Mean (STD)	Median (IQR)	Mean (STD)
3rdV	0.66 (0.160)	0.61 (0.117)	0.66 (0.010)*	0.64 (0.083)*
lCC	0.58 (0.260)	0.55 (0.226)	0.58 (0.308)	0.54 (0.256)
lPut	0.62 (0.240)	0.49 (0.230)	0.60 (0.240)	0.51 (0.209)
lSNc	0.70 (0.346)	**0.72 (0.192)**	**0.81 (0.104)**	**0.76 (0.131)**
lVDC	0.66 (0.240)	0.54 (0.211)	0.56 (0.125)	0.57 (0.164)
lAA	0.62 (0.094)	0.62 (0.078)*	0.62 (0.094)	0.60 (0.099)
rCC	0.60 (0.031)*	0.56 (0.128)	0.59 (0.085)	0.54 (0.131)
rCmC	0.59 (0.116)	0.59 (0.197)	0.66 (0.160)	0.62 (0.152)
rRN	0.69 (0.063)	0.65 (0.106)	0.58 (0.130)	0.63 (0.092)
rSNc	**0.76 (0.261)**	0.66 (0.191)	0.75 (0.130)	0.67 (0.192)

Table A.6: R2* CV results. Results in bold are the highest value in each column. Results with * have the lowest IQR or STD. Abbreviations: IQR - interquantile range, STD - standard deviation.

ROI	sAE		sVAE	
	Median (IQR)	Mean (STD)	Median (IQR)	Mean (STD)
lCC	0.64 (0.191)	0.63 (0.243)	0.60 (0.096)	0.61 (0.260)
lGPi	0.62 (0.106)	0.64 (0.183)	**0.72 (0.100)**	0.65 (0.248)
lPal	**0.72 (0.273)**	0.66 (0.183)	0.63 (0.136)	0.60 (0.184)
lSNc	0.66 (0.273)	**0.69 (0.179)**	0.70 (0.190)	**0.67 (0.107)**
lAA	0.69 (0.020)*	0.67 (0.072)*	0.62 (0.045)*	0.62 (0.067)*
rAmyg	0.66 (0.148)	0.65 (0.181)	0.58 (0.230)	0.63 (0.143)
rClWM	0.63 (0.299)	0.59 (0.167)	0.56 (0.316)	0.53 (0.223)
rCmWM	0.63 (0.328)	0.50 (0.268)	0.59 (0.408)	0.47 (0.280)
rGPi	0.69 (0.060)	0.63 (0.142)	0.62 (0.160)	0.57 (0.170)
rSNc	0.66 (0.260)	0.63 (0.177)	0.66 (0.219)	0.60 (0.164)

Table A.7: R1 CV results. Results in bold are the highest value in each column. Results with * have the lowest IQR or STD. Abbreviations: IQR - interquantile range, STD - standard deviation.

ROI	sAE		sVAE	
	Median (IQR)	Mean (STD)	Median (IQR)	Mean (STD)
lGPe	0.62 (0.219)	0.60 (0.226)	0.63 (0.220)	0.60 (0.163)
lPal	0.60 (0.126)	0.62 (0.194)	0.66 (0.125)	0.60 (0.216)
lAA	0.62 (0.088)*	0.63 (0.056)*	0.69 (0.090)	**0.70 (0.051)***
rAmyg	0.66 (0.220)	0.61 (0.159)	0.66 (0.230)	0.63 (0.135)
rGPe	**0.66 (0.096)**	0.61 (0.135)	0.69 (0.260)	0.66 (0.180)
rGPi	0.62 (0.149)	0.59 (0.131)	0.62 (0.025)*	0.54 (0.169)
rPal	0.62 (0.105)	0.58 (0.114)	0.64 (0.179)	0.62 (0.137)
rPut	0.59 (0.240)	0.53 (0.163)	0.62 (0.185)	0.55 (0.210)
rSNc	0.72 (0.280)	**0.67 (0.190)**	**0.72 (0.159)**	0.62 (0.245)
rTh (ss)	0.62 (0.190)	0.63 (0.226)	0.62 (0.156)	0.56 (0.158)

Table A.8: PD* CV results. Results in bold are the highest value in each column. Results with * have the lowest IQR or STD. Abbreviations: IQR - interquantile range, STD - standard deviation.