Imperial College London

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

FINAL REPORT

Predicting depression recovery with brain scan data during psychedelic therapy

Supervisor:
Dr Pedro A.M. Mediano

Author:
Lewis James Ng

Second Marker:
Dr Konstantinos Gkoutzis

Submitted in partial fulfillment of the requirements for the MEng degree in Computing of Imperial College London

June 2023

Abstract

Major depressive disorder, often referred to as depression, is an increasingly common mental illness with significant negative impacts on a person's well-being. Current treatments for depression have a poor long-term prognosis, which necessitates the exploration of alternative options. A recent study suggested that psilocybin, a psychedelic compound, may produce improved outcomes compared to selective serotonin reuptake inhibitors (SSRIs), the clinical standard of treatment. To evaluate the efficacy of psilocybin versus SSRIs for an individual, this work proposes a machine learning-based approach utilizing functional magnetic resonance imaging (fMRI) data and Beck's Depression Inventory (BDI) scores to predict post-treatment BDI scores. Trial data provided by the Centre for Psychedelic Research includes pre- and post-trial fMRI scans and BDIs of qualified participants. By constructing a high-performance predictor, potential treatment options can be evaluated for individual patients. The proposed model demonstrates promising results, with significant predictive performance (Pearson r = 0.757, p = 6.40e-9, R^2 = 0.559 (all 3sf)). The model leverages informed feature engineering and graphical encoding techniques to capture robust patterns in the data, surpassing visual encoding approaches. Although the final model shows limitations in predicting high post-treatment BDI scores for psilocybin patients due to limited data points, it exhibits the potential to learn more accurate regressions with expanded datasets. Overall, this research contributes to individualized depression treatment by providing a predictive tool to aid in clinical decision-making and attempting to learn the underlying anatomical relationships that contribute to treatment effectiveness.

Acknowledgments

I would like to thank my supervisor Dr Pedro A.M. Mediano for offering direction and support during my project and allowing me to explore unconventional approaches throughout.

I would also like to thank my family and friends, whose support has been invaluable during my time at university.

Contents

Chapter 1

Introduction

Major depressive disorder is a severe mental illness that affects a significant portion of the population and can significantly reduce the quality of life for those affected. Depression also greatly increases the risk of engaging in self-harming and suicidal behaviour, with a reported rate of 10.7 people per ten thousand committing suicide in the UK in 2021 [1]. Additionally, individuals with depression can experience mental illness-related stigma from their friends and colleagues, which can lead to discrimination and social isolation.

Currently, we lack viable tools for combatting depression, with the current clinical standard of treatment proving effective for only 66% of patients [2]. Additionally, 50% of individuals who see an improvement experience a recurrence of their symptoms. It has also been shown that a greater number of depressive episodes correlates with a higher rate of recurrence [3]. This signals a necessity to explore alternative treatment options which have a better long-term prognosis. Unfortunately, this is coupled with the fact that depression can be caused by a combination of many diverse risk factors [4]. Some common examples include personal issues, socioeconomic and environmental factors, family history, and comorbidity of other mental illnesses. The diversity of causes often results in many different presentations of the disease, and therefore we require individualised treatment approaches to tackle the symptoms an individual is facing.

Recently, a team at the Centre for Psychedelic Research conducted a study which suggested that psilocybin can reduce depressive symptoms and outperform standard SSRI-based treatments. This improvement was shown in the short-term prognosis (6-week period) as well as the long-term prognosis (over 6 months) [5]. A key observation taken from this study is that although the psilocybin arm had better average post-trial depression metrics, there was still an overlap between the SSRI and psilocybin arms. This suggests that for some patients, psilocybin may be a less effective treatment option than SSRIs.

Considering this fact, we propose that by utilizing a combination of functional magnetic resonance imaging (fMRI), Beck's Depression Inventory score (BDI, a standard depression severity metric) and relevant background data of a patient, we can apply machine learning techniques to predict their post-treatment BDI if they were to undergo either psilocybin or SSRI treatment. In this way, we can determine whether pursuing the treatment is a viable option for an individual. The Centre for Psychedelic Research at Imperial College London has graciously provided us with the trial data from the study [5], in which participants (N=42 qualified) were scanned (using fMRI) at baseline pre-treatment, and subsequently at post-treatment to compare brain network and functional differences. We will primarily make use of the baseline fMRI and BDI, which we will use in a variety of models in an attempt to create a high-performance predictor with the potential to be used in a clinical diagnostic capacity. The construction of a high-performance predictor will also aid in the analysis of the trial data and may help determine specific predictive features that contribute to a larger BDI decrease under psilocybin or SSRI treatment.

Chapter 2

Background

2.1 Psychiatry and Related Concepts

2.1.1 Major Depressive Disorder

Major depressive disorder (commonly referred to as depression) is a mental illness characterised by consistent feelings of unhappiness or anhedonia, which is often described as the lowered capability of experiencing joy in one's daily life. Depression can become very severe and lead individuals to withdraw from social activities, struggle to complete daily tasks and engage in self-isolatory behaviour. All of these behaviours can strengthen symptoms of comorbidities (conditions often diagnosed alongside depression) such as anxiety and substance abuse disorders [6]. Additionally, depression is one of the leading risk factors for suicide and its comorbidities greatly increase the chance of engaging in suicidal behaviour [7].

We currently have a very rudimentary understanding of the pathogenesis of depression, partially due to the underfunding of vital research [8]. Uncertain probable causes can create some difficulty in understanding the underlying mechanisms of effective treatment. For some time, the scientific community attributed the therapeutic effects of anti-depressant medication to an increase in mono-aminergic transmission, which is the synaptic transfer of monoamine neurotransmitters such as serotonin and dopamine. [9]. Modern antidepressant medication can specifically target the mono-aminergic transmission of several neurotransmitters responsible for mood regulation but still, these medications have an underwhelming efficacy of roughly 30-40% [10]. These theories have been largely questioned, with the work by Moncrieff et al. [11] performing a large-scale meta-analysis on the link between serotonin concentration and activity with depression and finding poor evidence to corroborate any significant correlation.

While studying the effects of anti-depressant medication, researchers found that although the rate of mono-aminergic transmission greatly increased as soon as treatment started, patients would only begin to exhibit an alleviation of depressive symptoms approximately 4 weeks [10] after starting the medication. This is an indication that the increase in the mono-aminergic transmission is not the key element in reducing depressive symptoms, and instead, the concept of promoting neuroplasticity has become increasingly prevalent in recent research [9].

2.1.2 Beck's Depression Inventory (BDI)

Beck's Depression Inventory (often abbreviated to BDI) is a comprehensive questionnaire specifically designed to assess the severity of a person's depression and guide treatment choices. The questionnaire consists of 21 questions, each assigned a score ranging from 0 to 3, resulting in a possible minimum score of 0 and a possible maximum score of 63.

It is worth noting that the format of the test is not on a Likert scale. While the responses in figure

0	I am not particularly discouraged about the future.
1	I feel discouraged about the future.
2	I feel I have nothing to look forward to.
3	I feel the future is hopeless and that things cannot improve.

Figure 2.1: Question 2 from Beck's Depression Inventory [12]

2.1 provide more concrete options than a Likert scale, patients may still deliberate between several options. The distinction between adjacent responses is ambiguous and so it is ultimately a subjective assessment. BDI serves as a commonly used proxy measure in the absence of any purely objective indicators.

2.1.3 The Physiology of Depression

Depression is highly correlated with abnormal patterns of connectivity and over-active regions, especially within certain systems such as the default mode network (DMN), executive network (EN), and salience network (SN) [13] (see figure 2.2). These three systems together are commonly referred to as the triple-network model.

The DMN is associated with introspective thought and is generally active when a person is not actively interacting with their environment, for example when daydreaming. The EN is associated with multi-stage problem-solving and extended attention. Impaired executive function experienced by many depressed individuals could be attributed to dysfunction in this system [14]. Finally, the SN is associated with higher-order functions such as social awareness and consciousness through gathering and processing sensory information.

The work of Wang et al. [13] suggests a 'downward spiral' model - which is commonly used to refer to a progressive worsening of depressive symptoms that promote each other. Firstly the SN has a negative pre-conceived notion which is reinforced by some external stimulus, this causes the DMN to dominate over the EN which leads to ineffective activity in the EN. This collectively produces a poor mental state which is run back through the cognitively impaired DMN and causes a cycle that tends to amplify negative stimuli and ignore positive stimuli.

2.1.4 Psychedelics

Psychedelics are a subset of hallucinogenic drugs which are primarily used to achieve otherwise unreachable states of consciousness. They have recently gained traction in clinical trials to determine their therapeutic potential for a variety of psychiatric disorders. The use of psychedelics has extended throughout history, with many ancient cultures incorporating the usage of entheogenic substances into their spiritual practices [15]. The study of psychedelics has been hampered for some time, notably since the prohibition of many psychoactive substances in the 'Misuse of Drugs Act' in 1971. Psilocybin, alongside other psychedelics, is still scheduled as a class A controlled substance which means it is illegal to possess any quantity. Consequently, the use of psychedelics is highly stigmatized, with many still fearful of adverse effects. It is essential to recognize and carefully evaluate the potential side effects of therapeutic drugs, taking into account the benefits they offer, without being influenced by any biases or stigma against the use of drugs.

Psychedelics have been observed to promote neuroplasticity [9], which can be described as the flexibility of connections between neurons. Although promoting neuroplasticity can potentially be beneficial in overcoming dysfunctional patterns, the brain can equally transition to a more dysfunctional state - which leads us to the conclusion that solely promoting neuroplasticity is insufficient in reducing depressive symptoms, and in itself does not have an anti-depressant action [9].

Recent works have suggested that depression may be attributed to dysfunctional connectivity within this triple-network model and the aim of treatment should be to move these systems into healthier

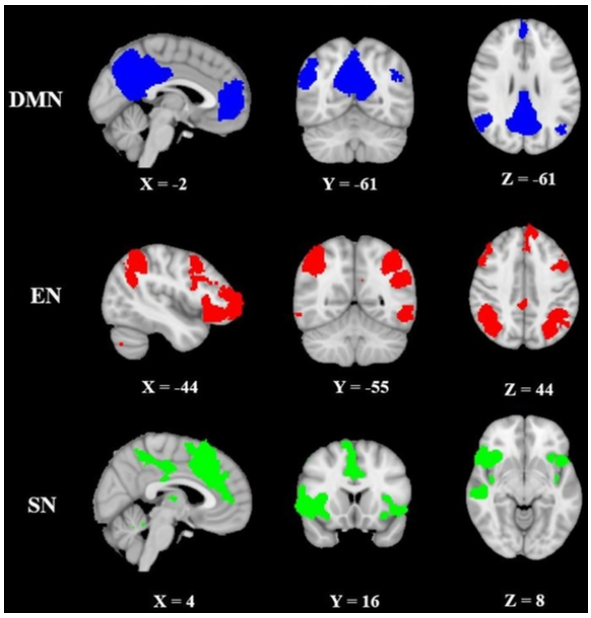


Figure 2.2: fMRI scans denoting the locations of the DMN, EN, and SN, cross-sections in all 3 spatial dimensions, image from `https://onlinelibrary.wiley.com/doi/full/10.1002/hbm.25771` Accessed 21/1/2023

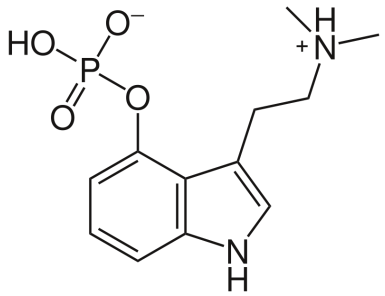


Figure 2.3: Chemical structure of psilocybin, image from `https://en.wikipedia.org/wiki/Psilocybin` Accessed 21/1/2023

patterns of function. A plausible mechanism of treatment suggested by Carhart-Harris RL et al. [16], is that while the brain circuitry is malleable from increased neuroplasticity, psychotherapy has a much greater ability to coax the brain out of negative cycles, and into healthier patterns of function.

In a separate line of enquiry, it has been shown that psychedelics can increase network integration, alongside a reduction in regional over-activity, leading to more varied states with higher connectivity. Regional over-activity was especially curbed in the DMN, which is highly implicated in depressive symptomatology [5]. In addition, regions that are particularly dense in 5-HT_{2a} had a significant increase in functional inter-connectivity [5][17].

2.2 Neuro-computational Concepts

2.2.1 Functional Magnetic Resonance Imaging

Functional Magnetic Resonance Imaging (fMRI) is a method of scanning brains that measures small changes in cerebral blood flow. Neuronal activity can be shown to increase when blood flow to that region increases, which establishes blood flow as a proxy metric for neuronal activity [18]. fMRI produces a 4D tensor: 3 spatial dimensions and 1 temporal dimension (see figure 2.4). This means

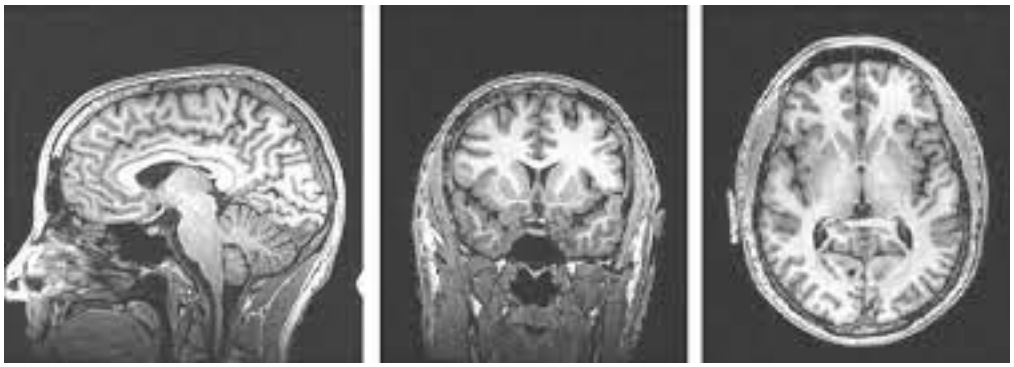


Figure 2.4: 1-time slice of fMRI, showing cross sections of all 3 spatial dimensions, image from `http://fmri.ucsd.edu/Howto/3T/structure.html` Accessed 21/1/2023

fMRI describes contiguous 3D renderings of neuronal activity through time (see figure).

Nilearn [19] is a python package built on top of sci-kit learn [20] that implements all of the operations on fMRI which are described in the following sections. It implements the methods in a way that is abstracted away from the actual process but provides a powerful and fast way to construct an fMRI processing pipeline. All of the code written uses these abstractions to increase readability and changeability.

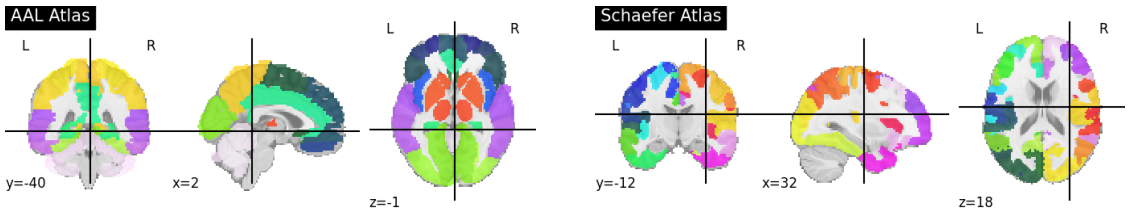
2.2.2 Confounds

MRI machines have an innate thermal noise component in addition to noise added by the subject - this needs to be pre-processed to compute a clean fMRI. There isn't a consensus on the best way to denoise fMRI, only that some methods remove more of the noise but also as a result can damage more of the underlying signal. The innate noise is specific to the MRI machine and is generally denoised experimentally during the calibration of the machine. We describe subject-related noise fluctuations using variables called confounds. They are commonly described using frame displacement metrics for the corresponding subject and sometimes motion metrics between frames. This is with the motivation to translate the frame into a standard space (lies straight and about the origin) ready to be fit by the atlas. The machines are notably very loud and often cause patients to be restless and move unnecessarily. This can lead to a large frame displacement which will damage the underlying signal greatly when transformed. Generally, when the damage is significant, the data has to be discarded.

Pre-processing can vary greatly between datasets which can lead to large differences in the amount of retained volumetric data. In this work, we make use of another auxiliary dataset, namely HCP 1200 subjects [21]. In the HCP pre-processing pipeline, a surface projection is applied to the fMRI which has the potential to remove a substantial portion of the volumetric data. It is important to be aware of this disparity and acknowledge its potential impact on both the visual and graphical distribution of the FCs.

2.2.3 Atlases and Parcellation

An atlas defines the regions of interest (ROIs) in the brain that the fMRI should be mapped onto. Atlases can be categorised as probabilistic, where voxels (equivalent to a pixel in 3D) can belong to more than one region, or deterministic, where all voxels only belong to one region. Various atlases exist which incorporate different structures of the brain. An atlas used in this work is the Schaefer atlas [22] (see figure 2.5) which solely maps cortical regions. The cortex is the outermost layer of the brain which is associated with high-order mental function [23]. An issue with using a purely cortical atlas is that we disregard a significant portion of the data pertaining to the subcortical regions. This omission may result in the loss of valuable information related to important structures within the sub-cortex such as the amygdala which is associated with emotion processing. For this reason,



((a)) 3D rendering of AAL atlas

((b)) 3D rendering of Schaefer atlas

Figure 2.5: Notice how the Schaefer atlas neglects the subcortical regions as opposed to the AAL which appears to include a much greater volume of the brain, both images computed using Nilearn[19]

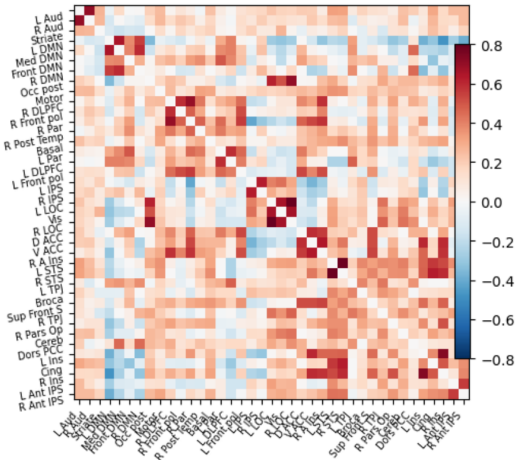


Figure 2.6: An example of a functional connectome computed using Pearson correlation coefficients and 39 ROIs using publicly available data [26]

there are also mixed cortical-subcortical atlases, notably the Automatic Anatomy Atlas (AAL) [24] (see figure 2.5), which we also make use of in this work. Both of these atlases are deterministic.

Parcellation is the process by which we divide the fMRI signal into values corresponding to each RoI defined by an atlas. To parcellate the fMRI, we first transform the fMRI using the confounds into standard space. Next, we generate a mask which maps voxels in a standard space map to an ROI by fitting the atlas to the space. Then, for each ROI, we compute the average value by taking the arithmetic mean of all the voxels within that particular ROI. This procedure yields a multi-variate time series, represented as a 2D tensor. This time series captures the average voxel intensity at the time "t" within each ROI which provides an informative representation of regional activity.

2.2.4 Connectivity Measures

There are multiple approaches to convert the time series into a connectivity matrix. Pearson correlation coefficients are often computed, resulting in a symmetrical 2D tensor, called a functional connectome (FC) [25]. In this matrix, the value at [i, j] is the Pearson correlation between the time series of RoI_i and RoI_j. This FC matrix can be interpreted as a graph, where each ROI corresponds to a node, and the correlation coefficient serves as the edge weight connecting the ROIs (see figure 2.6). This graph form provides a suitable representation to leverage graph neural networks, as explored in section 2.4.5.

2.2.5 Lempel Ziv Complexity

The Lempel-Ziv complexity is a measure that quantifies the number of unique substrings when reading a binary string from left to right. This provides an intuitive measure of repetitiveness or entropy

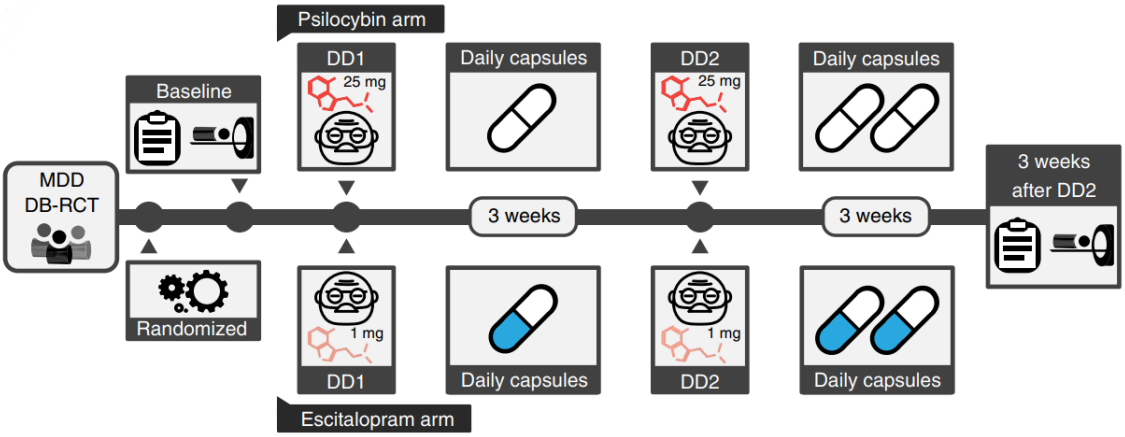


Figure 2.7: Image taken from psilocybin NatMed [5] explaining the trial setup.

present in a binary sequence. Lempel Ziv complexity has been suggested as a measure to assess depression in several studies [27][28]. Given a binary string X, the measure's maximum value is roughly proportional to:

$$\frac{\text{len}(X)}{\log_2(\text{len}(X))}$$

In our calculations, we divide the Lempel Ziv complexity by this value to loosely bound our values from 0 to 1.

2.3 Psilocybin Trial Procedure

The Centre for Psychedelic Research conducted a psilocybin trial to assess its therapeutic value for depression. A double-blind randomized controlled trial was conducted (see figure 2.7) where subjects were split into 2 arms, namely the psilocybin (treatment arm) and the escitalopram (control arm).

At the beginning of the trial, all patients underwent baseline fMRI scans and completed a BDI questionnaire to assess their initial depression severity. The treatment arm received two rounds of psilocybin treatment, with placebo pills administered for three weeks following each treatment. On the other hand, the control arm received a placebo treatment but was given escitalopram (a state-of-the-art SSRI) for three weeks after each treatment. Following the treatment period, both arms underwent follow-up fMRI scans and BDI measurements.

The Centre for Psychedelic Research generously shared the pre- and post-treatment measurements, for our analysis. In this work, our objective is to explore the predictive power of the pre-trial metrics in determining the post-trial BDI scores.

2.4 Machine Learning Concepts

2.4.1 Curse of Dimensionality

In regression problems, the curse of dimensionality refers to the exponential growth of space with respect to the dimensionality of the input. This implies that with each additional dimension, we require an exponentially larger number of samples to maintain the same coverage ratio of the sample space (see figure 2.8).

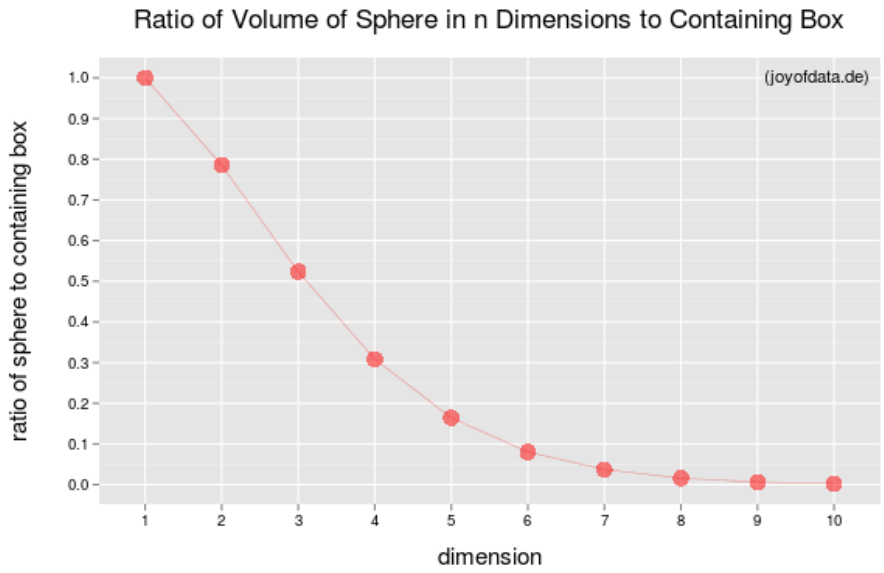


Figure 2.8: Figure illustrates the diminishing ratio of an n-dimensional sphere inside an n-dimensional box. The corners of the n-box sample space dominate the area and the n-sphere covers less and less of the hyper-space, image from `https://www.joyofdata.de/blog/curse-dimensionality/` Accessed 18/06/2023

This becomes an issue when data is scarce, leading to a very poor sample space coverage of the hyperspace (higher dimensional space). Attempting to learn from a small subset of the hyperspace leads to large regions of the hyperspace being neglected and results in a complex and over-fitted boundary that poorly generalises to unseen data. It becomes imperative to either source more data or reduce the dimensionality of the input.

2.4.2 Pre-training, Fine-tuning and Mixed Class Training

Pre-training is a technique employed to enhance the sample space coverage of a high-dimensional dataset by leveraging an auxiliary dataset with a similar structure, typically of a larger size. The auxiliary dataset should possess a distribution that is roughly equivalent to the main dataset. The objective is to capture the underlying features that represent the data distribution to achieve a more generalised regression model.

In this work, we leverage pre-training to improve the generality of our VAE and VGAE, which are both non-linear dimensionality reduction architectures. We explore two separate methods on built top of this: fine-tuning, and mixed-class training.

Fine-tuning is a method that first trains on the auxiliary dataset and then subsequently trains on the actual dataset. This method aims to learn more general features initially with the auxiliary dataset and then fine-tune specific features on the actual dataset, which may help overcome distribution differences in the datasets. By contrast, mixed-class training learns from both datasets simultaneously. This method aims to achieve a more balanced performance between the datasets which may help mitigate over-fitting to achieve a more robust model.

In leveraging these techniques, we aim to boost the performance and generality of our VAE and VGAE, allowing them to effectively encode high-dimensional data and capture robust patterns.

2.4.3 Decision Trees

Decision trees are a form of supervised learning in which the data is continuously partitioned on specific value criteria, for example, flowers being split into colour, then size, then petal length etc.

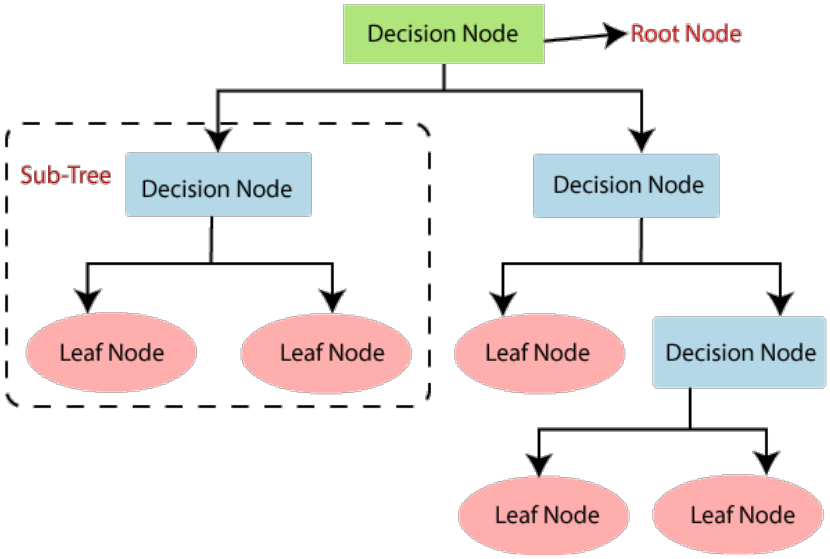


Figure 2.9: Structure of a decision tree, image from `https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm` Accessed 21/1/2023

A prediction from a decision tree involves passing the input down the tree until a leaf node is hit in which case the prediction is the label associated with the leaf. Decision nodes are often conditioned on minimising entropy (essentially splitting the data as evenly as possible). This process helps to evenly distribute the data across different branches, resulting in a more balanced and effective partitioning. By minimizing entropy, the depth of the decision tree can also be reduced, making it more generalised and efficient.

2.4.4 Random Forest Regression

Random Forest Regression is a form of supervised learning that employs ensemble learning - a technique which creates many different sub-networks and aggregates their results into a final prediction.

The algorithm takes a set of random subset from the training data and then creates a decision tree - this provides one prediction. However, this prediction is biased towards the samples that we drew from the training data.

The key aspect of random forest is that it simultaneously constructs hundreds of other trees at the same time which means we have hundreds of predictions which are all biased towards the samples from which the respective tree was created. Crucially, as the amount of trees increases, the bias of the average prediction diminishes (see figure 2.10 for example). Random forest regression is often used as a baseline for evaluation as there are relatively few hyper-parameters when compared to complex neural network architectures. The random forest can establish a lower performance bound with minimal effort, allowing for a direct comparison with more complex models.

2.4.5 Neural Network Architectures

Multi-layer Perceptrons

Multi-layer perceptrons (MLP) are powerful neural networks that offer us a way of processing data and detecting patterns that are infeasible for humans. MLPs are feed-forward neural networks with at least three layers: one input layer, one or more hidden layers and one output layer.

Crucially they also must have a non-linear activation between layers, otherwise, these multiple non-activated layers have the same effect as one layer. An MLP can have an output layer of arbitrary length, which can represent various types of data, such as probabilities, classes, value prediction, etc.

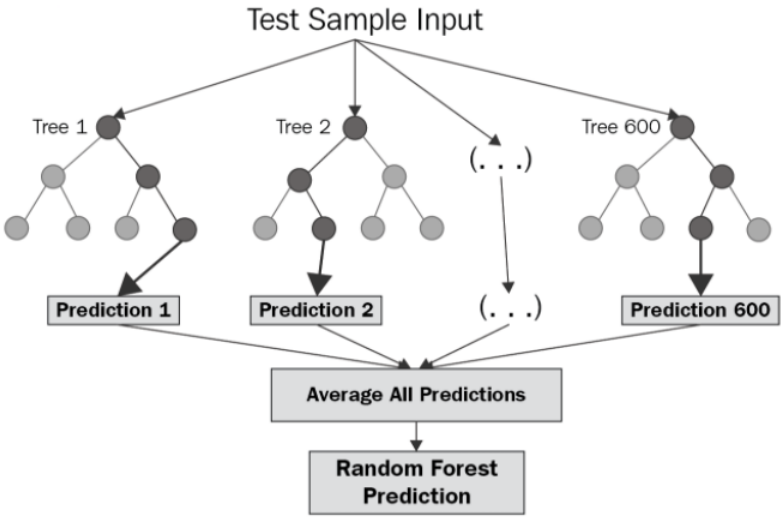
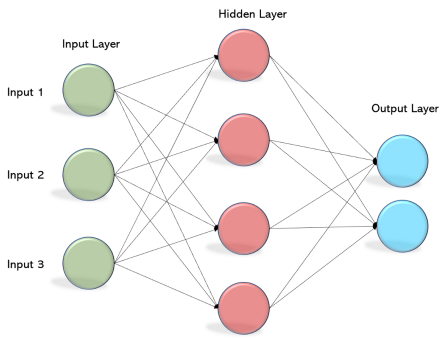
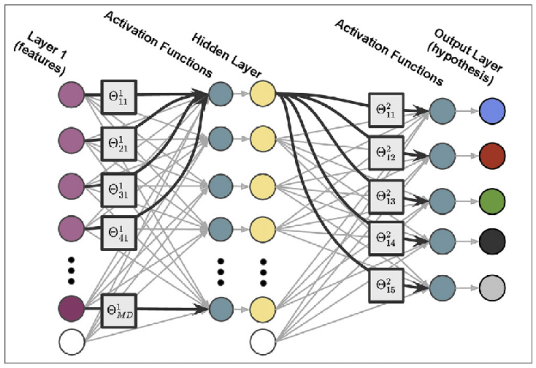


Figure 2.10: A visualisation of the random forest regression algorithm, image from `https://levelup.gitconnected.com/random-forest-regression-209c0f354c84` Accessed 21/1/2023



((a)) A basic multi-layer perceptron, image from `https://jameskle.com/writes/rec-sys-part-5` Accessed 21/1/2023



((b)) A more complex MLP with a more than one node in the output layer - suited for a classification task, image from `https://www.researchgate.net/publication/339423202_Volcano_video_data_characterized_and_classified_using_computer_vision_and_machine_learning_algorithms` Accessed 21/1/2023

Figure 2.11

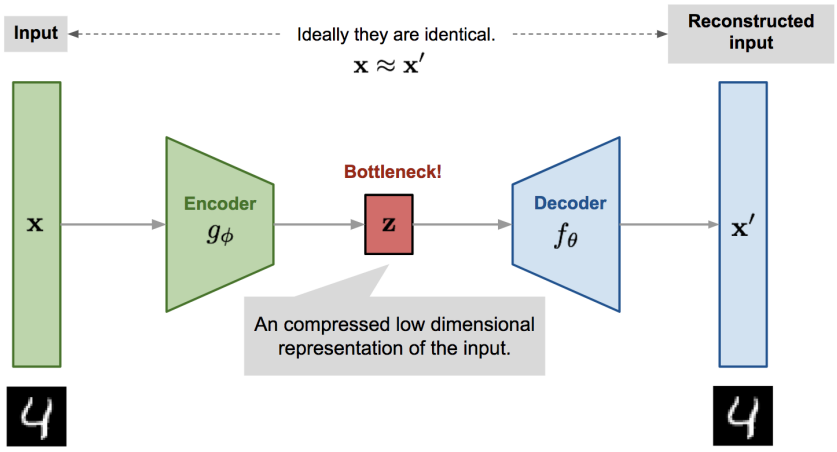


Figure 2.12: General auto-encoder architecture, image from `https://lilianweng.github.io/posts/2018-08-12-vae/` Accessed 21/1/2023

MLPs learn by computing a task-specific loss function. By treating the entire network as a function, the objective is to minimize the loss function using stochastic gradient descent (SGD). The goal is to adjust the network's weights and biases to reach the point of minimum loss, where the MLP's output closely matches the expected labels. Backpropagation is widely used, where derivatives of the loss function are calculated with respect to the individual weights and biases in the network. This enables the network to keep updating its parameters to improve its predictions.

Some examples of MLPs are shown in Figure 2.11. The left figure represents a basic MLP with one hidden layer, while the right figure showcases a more complex MLP with multiple nodes in the output layer, which is suitable for classification tasks. These diagrams depict the flow of information through the network, where each node in one layer is connected to every node in the subsequent layer.

Auto-Encoders

In complex large-scale datasets, it can often be difficult to directly engineer features for ML algorithms. Auto-encoders are a form of unsupervised learning which attempts to learn a lower-dimension representation of the input. The architecture of an autoencoder consists of several hidden layers, known as the encoder, which progressively reduce the dimensionality of the input data until reaching a bottleneck layer where the data is represented by a latent vector. The decoder then takes the latent vector and attempts to reconstruct the input data (see Figure 2.12).

The loss function is then generally defined as the similarity between the input and the decoded image - often the L_2 norm of the difference matrix between input and output. By minimizing this loss, the autoencoder learns to reconstruct the input data accurately. The latent vector representation generated by the autoencoder can be seen as a non-linear version of Principal Component Analysis (PCA), as it summarises the essential features and relationships from the data in a lower-dimensional latent vector.

These trained autoencoder layers can then be used as input to a neural network. Instead of directly working with the high-dimensionality raw input, the network operates on the latent vector extracted by the autoencoder. This approach offers several advantages, including faster training and improved accuracy. By using the latent vectors, the network can isolate the most informative features of the data, leading to more efficient and accurate performance [29].

Graph Neural Networks

Graphs are non-euclidean data structures which capture information about points of interest (nodes) and the links between them (edges). Graphs can be represented in a variety of ways, but for machine learning, they are often given as feature and adjacency matrices.

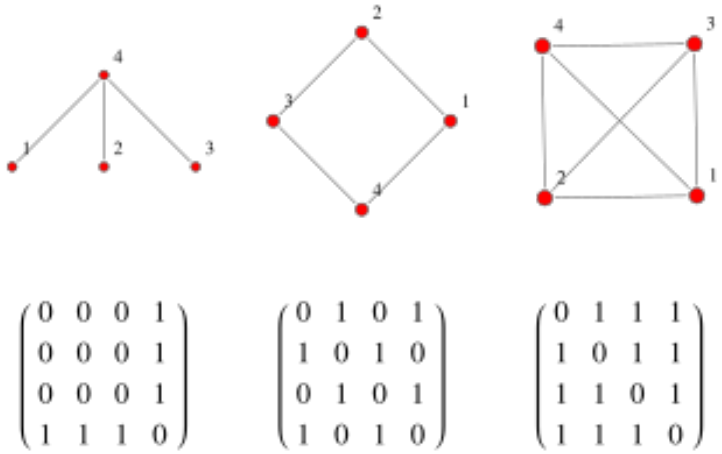


Figure 2.13: Graphs and their equivalent adjacency matrix, image from `https://mathworld.wolfram.com/AdjacencyMatrix.html` Accessed 21/1/2023

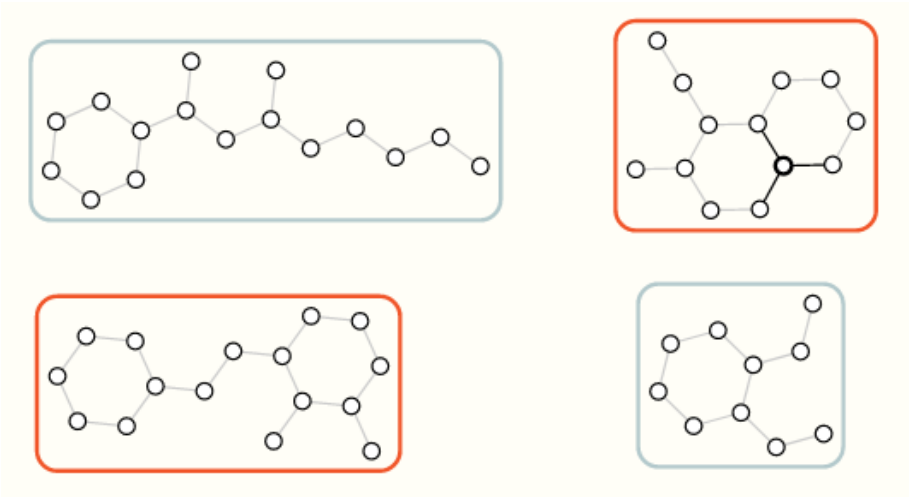


Figure 2.14: Illustration of a global level task, graphs have been classified into 2 subsets: red and grey, image from `https://distill.pub/2021/gnn-intro/` Accessed 21/1/2023

The adjacency matrix offers a compact representation of the graph structure, where each entry indicates the presence or absence of an edge between two nodes (see Figure 2.13). Additionally, adjacency matrices can have float value elements to indicate the weight of edges.

Conventional neural network architectures, designed for tasks such as image processing, struggle to effectively integrate the structural relationships encoded by graphs. To address this, graph neural networks (GNNs) were developed. GNNs are specifically designed to be able to work on graph-structured data and integrate the spatial relationships within graphs into the model.

In conventional GNNs, a key component is the MetaLayer, which consists of multiple multilayer perceptrons which operate on the associated nodes, edges, and graph-level variables of a graph element. GNN layers often train through message passing between graph components, such as nodes and edges, and the aggregation of these messages. By leveraging message passing and aggregation, GNNs can incorporate spatial relationships.

Chapter 3

Ethical Issues

This work involves a secondary use reanalysis using personal medical data collected in the psilocybin trial at the Centre for Psychedelic Research [5]. This data used to train the models is pseudonymised medical data, which can be linked back to the individual given a key. To ensure the privacy and security of this sensitive data, we have stored it on a secure Imperial password-protected server. Upon completion of the project, we will delete all personal data, in addition to agreeing that we will not attempt to obtain the key for the data.

In addition to the psilocybin trial data, we make use of the '1200 Extensively Processed fMRI Adult Resting State' dataset from the Human Connectome Project [21] which contains fMRI data of 1200 participants. HCP has the right to distribute such material as an open dataset, and since we do not analyse or present any of the restricted data categories, the information we analyse or present is de-identified.

Considering the sample size of our study, it is important to critically assess the statistical significance of our models. Incorrect predictions by our models could potentially lead to different treatment outcomes for patients and cause unnecessary suffering. In an average case, patients can be subjected to unnecessary suffering and misery if our tool gives them a worse clinical outcome. There is a high prevalence of suicidality among individuals with depression, meaning in the worst-case scenario if our tool prescribes the wrong treatment, patients may commit suicide. As a result, it will be of utmost importance to conduct a reanalysis of the proposed method as soon as more data becomes available to improve the robustness and reliability of our results.

Our work also contributes to the analysis of brains of depressed individuals which may have implications towards the development of depression classifiers. This could potentially create discrimination in the form of depression screenings for job candidates. Due to the stigma associated with mental illness, this could negatively impact candidates who would otherwise be highly suited for their position. To address this issue, it is necessary to ensure our work is in line with anti-discrimination laws.

To summarise, our contribution is not inherently problematic, but it illustrates a need to address issues related to privacy, discrimination, and ethical considerations. By sticking to established protocols and encouraging responsible analysis of the findings, we aim to curb any potential risks and ensure that our research is used responsibly.

Chapter 4

Results

4.1 Dataset statistics

To motivate the direction of the work, we started by calculating some rudimentary distribution statistics on the trial data (we will often refer to the data from this dataset as psilocybin data). The distribution of psilocybin and escitalopram samples is relatively balanced, with a split of 22:20. However, it's important to note that the dataset itself consists of only 42 total samples, which is very scarce considering that the dimensionality of the fMRI data is very high.

We chose to use the Mean Absolute Error (MAE) loss for our predictor due to its straightforward interpretation. When a predictor has an MAE loss of L, it means that, on average, our predictions deviate from the actual values by L in either direction. We calculated the standard deviation of the predicted BDI (see subsection 2.1.2) values in the psilocybin dataset as 10.1 (3sf). The standard deviation is a pertinent metric to compare against the Mean Absolute Error (MAE) of a predictor. This is because when the average post-trial BDI is uniformly predicted, the average MAE loss will equal the standard deviation. Therefore, as a rigorous baseline, any valuable model we develop should yield an MAE below 10.1.

In addition to MAE, we also want to assess the goodness of fit for our regressions. Therefore, we plan to calculate the Pearson correlation coefficient (Pearson r) and the coefficient of determination (R-squared, or R^2). These measures will provide us with an insight into the correlation strength and explained variance respectively.

We calculated the lower and upper quartiles as 4 and 16.5 respectively, with a median of 10. The

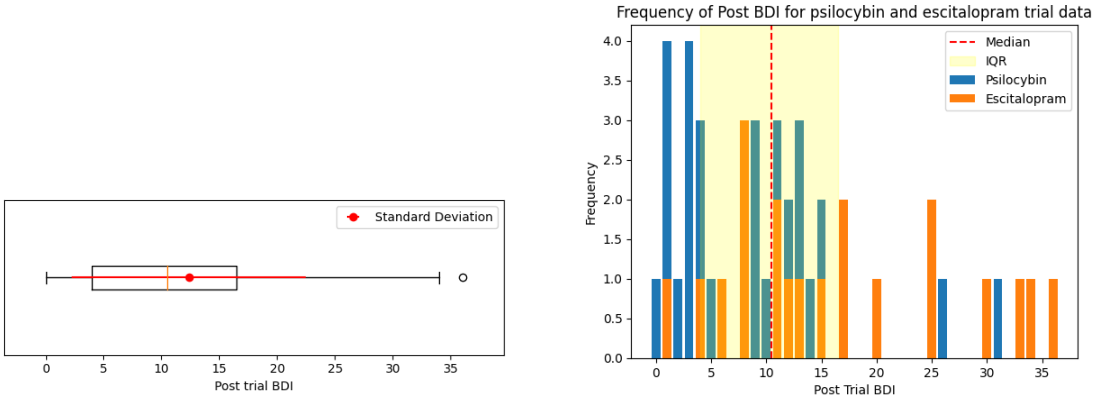


Figure 4.1: A box plot and stacked frequency bar chart illustrating the distribution of post-trial BDIs.

frequency bar chart (see figure 4.1) clearly illustrates a hard downward skew to our data, with the upper quartile resting below the midpoint of the range. With only 25% of the data samples covering the area in the top half of our range we recognized that it may prove difficult to learn relationships for data in this range.

There is only one sample from the psilocybin arm where BDI increased post-trial, so it is infeasible to learn features that would cause a BDI increase for psilocybin samples. We also see that there are 4 out of 20 samples where the BDI increased for the escitalopram arm, so we are unlikely to fit any general features that would predict the BDI increase for escitalopram samples. For these reasons, we focus on a regression-based approach, where given the fMRI and pre-trial BDI, we output the predicted post-trial BDI as opposed to only predicting a binary improvement.

4.2 Processing Functional Connectomes (FCs)

fMRI (see section 2.4) data is extremely high-dimensional with each data sample containing 3 high-resolution spatial dimensions in addition to a time component, resulting in a dimensionality of approximately 10^7 for an average MRI scan with 100 frames. Attempting to learn naively from such a high-dimensional input will fit an extremely rough and overfitted boundary to the function space, making it highly unlikely to generalise to unseen data. Therefore, it becomes essential to condense the information to any reasonable regression performance. An established method for processing fMRI is the computation of a time series. This requires parcellating (see section 2.2.3) the fMRI with an atlas. Given a time series, we can compute the functional connectome (FC, see section 2.2.4), which gives an intuitive measure of the pairwise interaction between regions of interest (ROIs). This computation has already massively reduced the dimensionality to N_{RoI}^2. To further reduce the dimensionality, most of the models created will take an upper triangular FC input (top-right triangle of the FC), which exploits the symmetry of the FC and further reduces the input dimensionality by over 50%.

To establish baseline models, time series were processed using two separate atlases: the Schaefer 100 RoIs (2mm resolution, 7 yeo networks) and the AAL 116 RoIs atlas. Schaefer is a purely cortical atlas, whereas AAL is a mixed cortical and subcortical atlas. Regressions will be computed using FCs from both atlases, to compare the extent of critical information preservation between parcellations.

It is important to note that direct comparison between the two atlases is not straightforward due to the misalignment of cortical regions between Schaefer and AAL. Any observed differences in information can potentially be attributed to both the additional subcortical information and the dissimilar cortical regions covered by the atlases.

4.3 Baseline Regressors

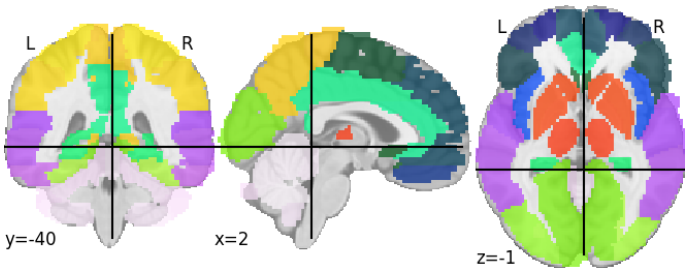
4.3.1 Linear Regressors and Ensemble Methods

At this stage, it was unknown whether the data held any useful relationships, so we aimed to create basic models to discover whether we could improve upon the standard deviation. We utilized three different kinds of regression models:

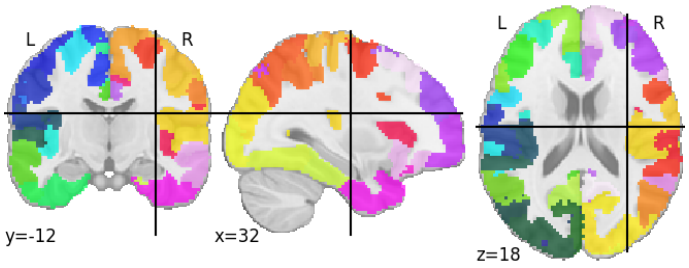
1. Random forest regression - a standard ensemble method that uses decision trees as the unitary model.
2. Ridge regression - linear regression with L2 regularization.
3. Lasso regression - linear regression with L1 regularization.

For each kind of model, two model configurations were considered:

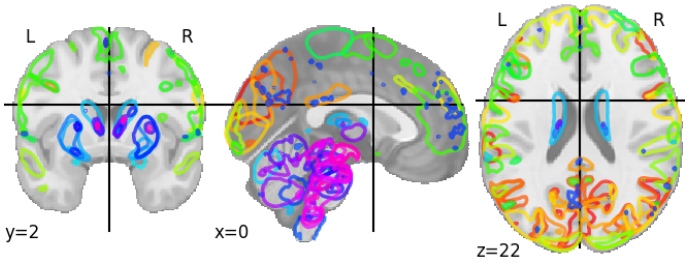
1. Input consists of pre-trial BDI and a one-hot encoding of whether they were given escitalopram or psilocybin (referred to as the 'drug flag' hereafter).
2. Input consists of pre-trial BDI, drug flag, and an upper triangular Schaefer FC.



((a)) 3D rendering of AAL 116 RoI atlas

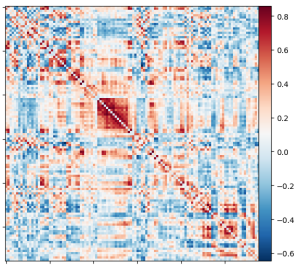


((b)) 3D rendering of Schaefer 100 RoI atlas

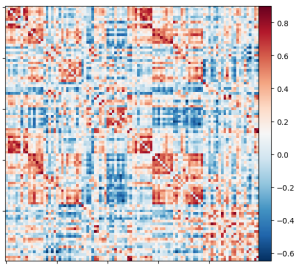


((c)) 3D rendering of ICA 100 RoI atlas

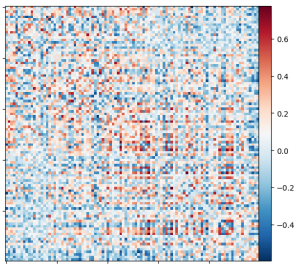
Figure 4.2: Renderings of the atlases used in this work: note that the ICA 100 RoI map is non-anatomical and defines highly complex and asymmetric regions.



((a)) An FC parcellated with AAL (116 RoIs)

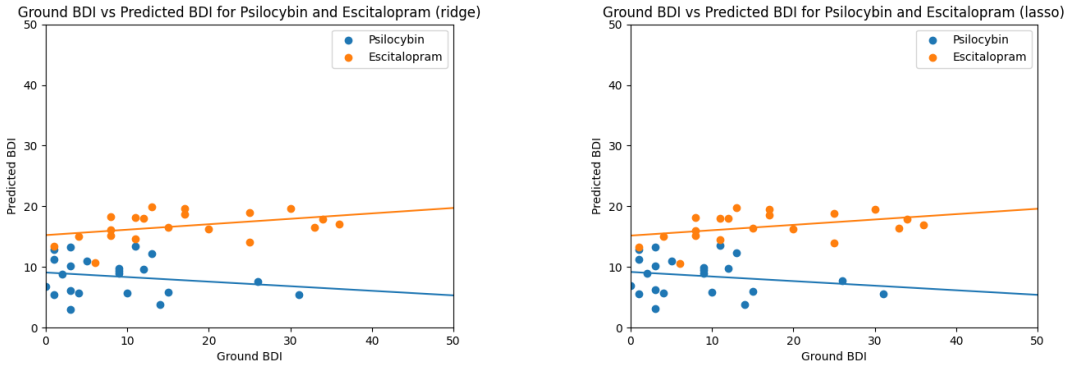


((b)) An FC parcellated with Schaefer (100 RoIs)



((c)) An FC parcellated with ICA (100 RoIs)

Figure 4.3: FCs parcellated with each atlas. Note the clear visual difference: AAL and Schaefer FCs appear more ordered than the ICA FC



((a)) MAE = 7.35, $R^2 < 0$ **((b))** MAE = 7.35, $R^2 < 0$

Figure 4.4: Graphs showing the two best performing models, which were trained without FC input (see all graphs in appendices 6.1)

3. Input consists of pre-trial BDI, drug flag, and an upper triangular AAL FC.

For random forest regression, the amount of trees was varied but it caused no significant improvement beyond the default 100 trees. Additionally, any reduction in the tree depth only resulted in model deterioration. We hypothesize that the poor ensemble performance is due to the sample skew (see figure 4.1) in addition to dimensionality issues with the FC.

The model with the best performance was ridge/lasso without FC input, achieving an MAE of 7.35. Notably, the FC input deteriorated performance for all models. This indicates that the dimensionality of the input is too large to infer any useful relationship. Upon calculation of R^2 scores, all of the values are sub-zero meaning that the model is learning a degenerate solution that is capturing some superficial relationship in the data.

4.3.2 Simple MLP

Having obsoleted simpler linear regressors, we moved on to the construction of MLPs (see section 2.4.5) Two configurations of simple fully connected neural net regressors were tested:

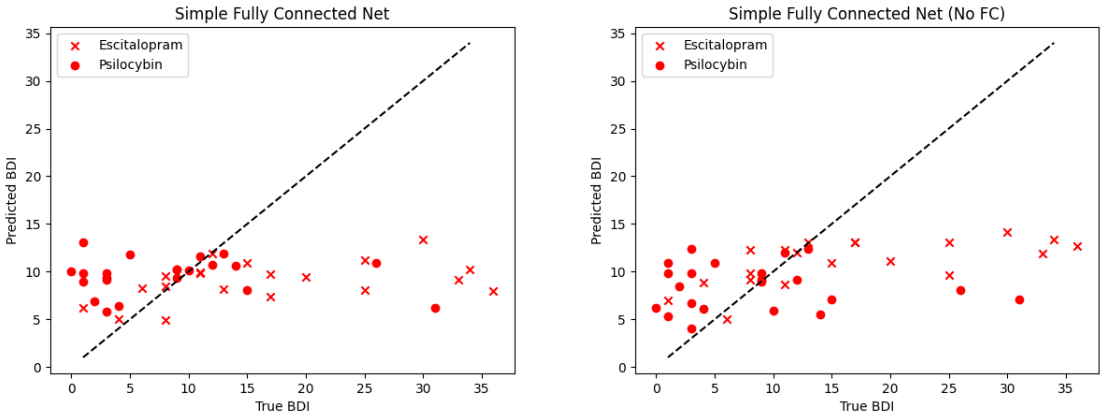
1. Input consists of pre-trial BDI and drug flag
2. Input consists of pre-trial BDI and drug flag in addition to upper triangular FC

The initial evaluation of the MLP was conducted using a single train, validation, and test split, which initially performed well. However, when the split was changed to a different permutation, the performance of the model deteriorated aggressively. This illustrated the sensitivity of the model to the split of training and evaluation data.

To address this issue and obtain a more complete evaluation, a K-fold cross-validation (CV) regime with K=5 was implemented. This method involves splitting the dataset into five folds, where the model is trained on four folds and evaluated on the remaining fold. This process is repeated K times, with each fold serving as the evaluation set once. By averaging the performance across multiple splits, we aim to reduce the bias introduced by a single test set and provide a more comprehensive view of the model's performance.

This approach serves as a fixed point of comparison for all the subsequent regression models in this work, allowing for a fair and consistent evaluation across different models.

An improvement in the Mean Absolute Error (MAE) from 7.35 to 6.77 is observed in the model with FC input (see figure 4.5). This improvement suggests that the multi-layer perceptron (MLP) is capable of capturing non-linear relationships in the data, leading to slightly better predictions. However, it is



((a)) MAE = 7.29, $R^2 < 0$ **((b))** MAE = 6.77, $R^2 < 0$

Figure 4.5: Graphs of regression calculated by 5 fold CV for both configurations.

important to note that both models still have a sub-zero R^2 value, indicating a degenerate solution and poor fit to the data.

Interestingly, the model without FC input outperformed the model with FC input, further highlighting the dimensionality issues. The model with FC input fails to converge to the same solution as the model without FC input due to the implementation of L2 regularization. The regularization tries to balance the weights of the input features, causing the redundant FC input to be considered, leading to suboptimal performance.

The degradation in model performance with the inclusion of FC input highlights the futility of including high-dimensionality FC input in any type of regressor given the limited sample size.

4.4 Linear Dimensionality Reduction

After recognising the extent of the dimensionality issues, we searched for a method of mitigation. As a starting point, we experimented with simple linear reduction methods such as Partial Least Squares (PLS) and Principal Component Analysis (PCA) in an attempt to establish a baseline for their performance.

For PLS, we applied the technique to the Schaefer FCs. The PLS model was trained on an 80:20 train test split of the dataset. The results were disappointing, with sub-zero R^2 scores across all tested numbers of components (1 to 9). The best R^2 score obtained was still sub-zero (see figure 4.6).

Next, PCA was performed on the Schaefer FCs (see figure 4.6). The experiments revealed that the unexplained variance ratio (UVR) on the training set reached an elbow at 4 components, with a UVR of 0.05. When the test set was transformed using these components, the UVR is significantly higher at 0.78. The t-SNE visualization of the linear embeddings showed no discernible patterns or clusters in the data. The train and test data points both appeared randomly distributed. The poor performance of these linear reduction techniques can likely be attributed to the suspected nonlinear nature of the underlying function boundary.

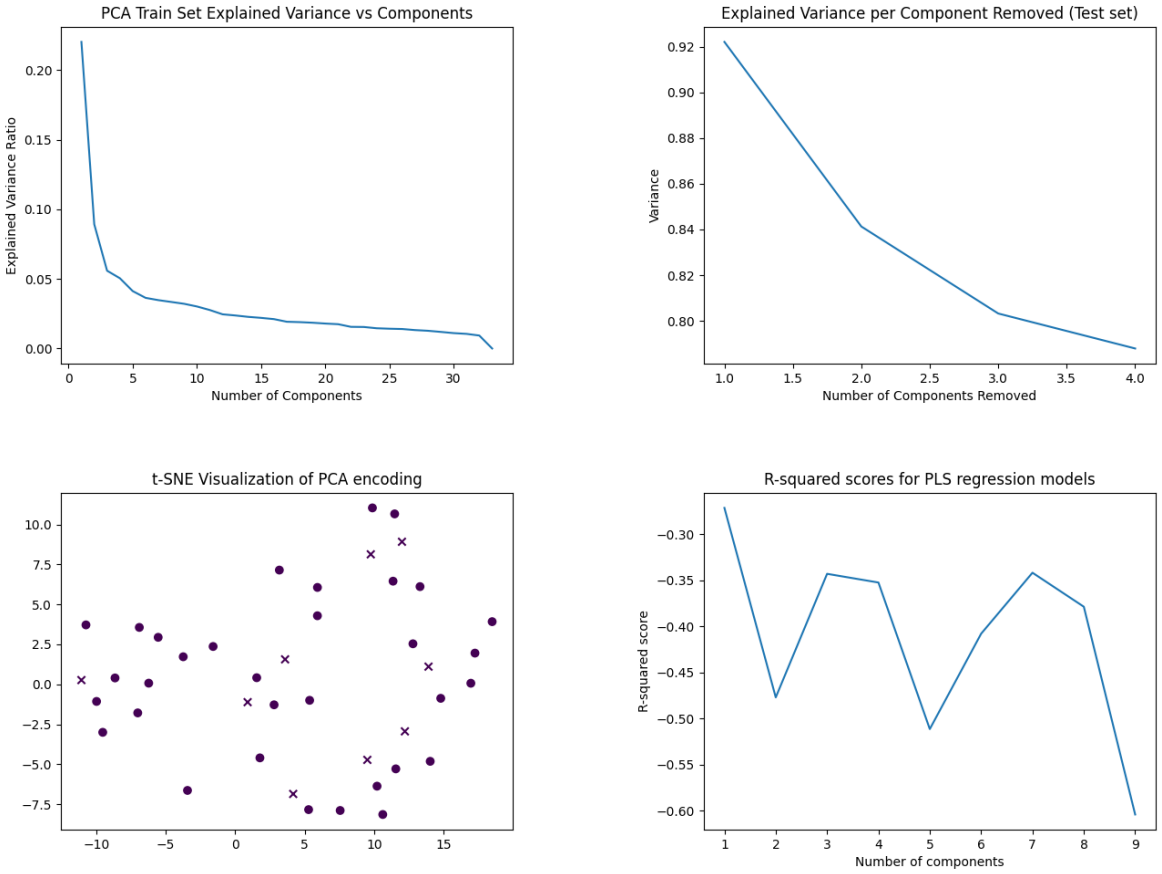


Figure 4.6: Results of PCA and PLS experimentation. Scree plots for train and test sets in addition to a t-SNE representation of the latent transformations for PCA. R^2 against the number of components for PLS test set transformations.

4.5 Constructing a Variational Autoencoder (VAE) (Non-linear Dimensionality Reduction)

4.5.1 Training the VAE

To further investigate the cause of the poor performance, we experimented with non-linear dimensionality reduction. We decided to construct a variational autoencoder (VAE, see section 2.12) to produce a latent encoding that can then be fed to an MLP. Intuitively, if we can compress the information to a smaller encoding and then recreate it, then the encoding must be a losslessly compressed version of the original FC. This would result in a greatly reduced input dimensionality for our predictor.

In addition, we had access to a large (N=1003) auxiliary dataset (HCP [21]) which contained fMRI data with a similar structure to our psilocybin dataset that we leveraged for pre-training 2.4.2 (we will highlight concerns and potential future improvements regarding this approach later in our analysis).

To access the HCP data, we needed to carefully consider the practicality of obtaining the entire dataset. The HCP fMRI data is of notably high resolution, and the temporal dimension of the data is approximately ten times larger than our psilocybin dataset. Each pre-processed sample occupies a significant amount of storage space, roughly around 40GB. To download the entire HCP dataset, we would require approximately 40 TB of storage space.

Given the impracticality of downloading and processing such a large dataset, we decided to download the pre-computed time-series data provided by HCP. These pre-computed time series were obtained using an Independent Component Analysis (ICA) 100 RoI atlas. We then parcellated our psilocybin fMRI data with the same atlas into equivalent time series and processed FCs for both the HCP ICA time series and the psilocybin ICA time series.

To train our VAE, we used an 80-20 training-validation split. The reasoning behind this split was that the visual distribution of the data was not overly noisy, allowing us to obtain fairly consistent results with a single validation set. Considering that the magnitude of reconstruction loss lacks a meaningful interpretation, we decided to use a Mean Squared Error (MSE) loss, which is often used to calculate differences between images. Results of training are shown in Table 6.1.

Our VAE training configurations are as follows:

1. Train using HCP ICA FCs first, and then fine-tune using psilocybin ICA FCs
2. Train using psilocybin ICA FCs
3. Train using psilocybin Schaefer FCs
4. Train using psilocybin AAL FCs

For each configuration listed, we considered two sub-cases:

1. Psilocybin dataset only contains pre-trial FCs (referred to as ‘before’)
2. Psilocybin dataset contains both pre-trial and post-trial FCs (referred to as ‘combined’)

The fine_tune_before configuration achieved the lowest reconstruction loss on the psilocybin data. This indicates that the pre-training process had a positive effect on the VAE’s ability to reconstruct the data.

An interesting observation is that for the VAEs that were not pre-trained, the combined dataset outperformed the before dataset in terms of reconstruction loss. This can be attributed to the fact that the additional samples in the combined dataset (sample size 42 vs. 84) result in more proportionally significant coverage of the function space, compared to the pre-training dataset (sample size 1045 vs. 1087).

It’s important to note that the validation set for the combined VAE contains combined data, which may contribute to better overall reconstruction performance for the combined dataset, but potentially worse reconstruction for only the pre-trial dataset.

Table 4.1: VAE reconstruction losses for configurations detailed above (full dropout testing available in appendices 6.1)

Configuration	Val Loss (MSE 2dp)
hcp	28.84
psilo_ica_before	145.55
psilo_ica_combined	131.77
psilo_schaefer_before	154.88
psilo_schaefer_combined	131.14
psilo_aal_before	194.22
psilo_aal_combined	165.14
fine_tune_combined	100.57
fine_tune_before	103.41

Comparing different atlases, the AAL atlas had the worst reconstruction loss. This can be attributed to the fact that it has more RoIs, but the same number of samples and the same sized latent dimension. On the other hand, the Schaefer atlas showed a markedly higher reconstruction loss than the ICA atlas. This suggests that the ICA FCs are easier to compress using the VAE, potentially because the ICA atlas is constructed by minimizing statistical dependence between RoIs, making their relationships more readily learnable by the VAE.

It could also be the case that the ICA atlas captures a different level of informational variation compared to the Schaefer atlas, making it easier for the VAE to compress and reconstruct the latent representations.

We generated a t-SNE representation of the latent vectors created by the schaefer_before VAE, as shown in Figure 4.7. This was to try and identify any patterns or clustering in the latent space.

Firstly we observed no significant difference in the distribution of training and validation data, which implies our VAE is appropriately fitted to the data.

However, we found limited evidence of clustering over all perplexities. This implies that the VAE struggles to encode any strong commonality between samples. The relationship may be too complex for the VAE to effectively capture. Alternatively, the samples may be inherently dissimilar in the original hyperspace.

4.5.2 Fine Tuning vs Mixed Class Training Results

In addition to our previous experiments, we explored the use of a mixed-class training regime where we combined the pre-training data with the psilocybin data. We also implemented class weights that were inversely proportional to class representations.

As expected, when aiming for balanced performance across two datasets, the performance for each dataset tends to be worse. In our case, we found that the fine-tuned VAE produced more faithful representations compared to the mixed class VAE, as demonstrated in Figure 4.8 where we show two randomly selected FCs.

4.6 Building an MLP using latent VAE embeddings

To integrate the latent vectors generated by the VAE into our MLP model, we trained several configurations with different VAE supports. The configurations we considered are as follows:

1. Latent vectors are generated using the ICA-trained VAE (referred to as 'ica')
2. Latent vectors are generated using the Schaefer-trained VAE (referred to as 'schaefer')
3. Latent vectors are generated using the AAL-trained VAE (referred to as 'aal')

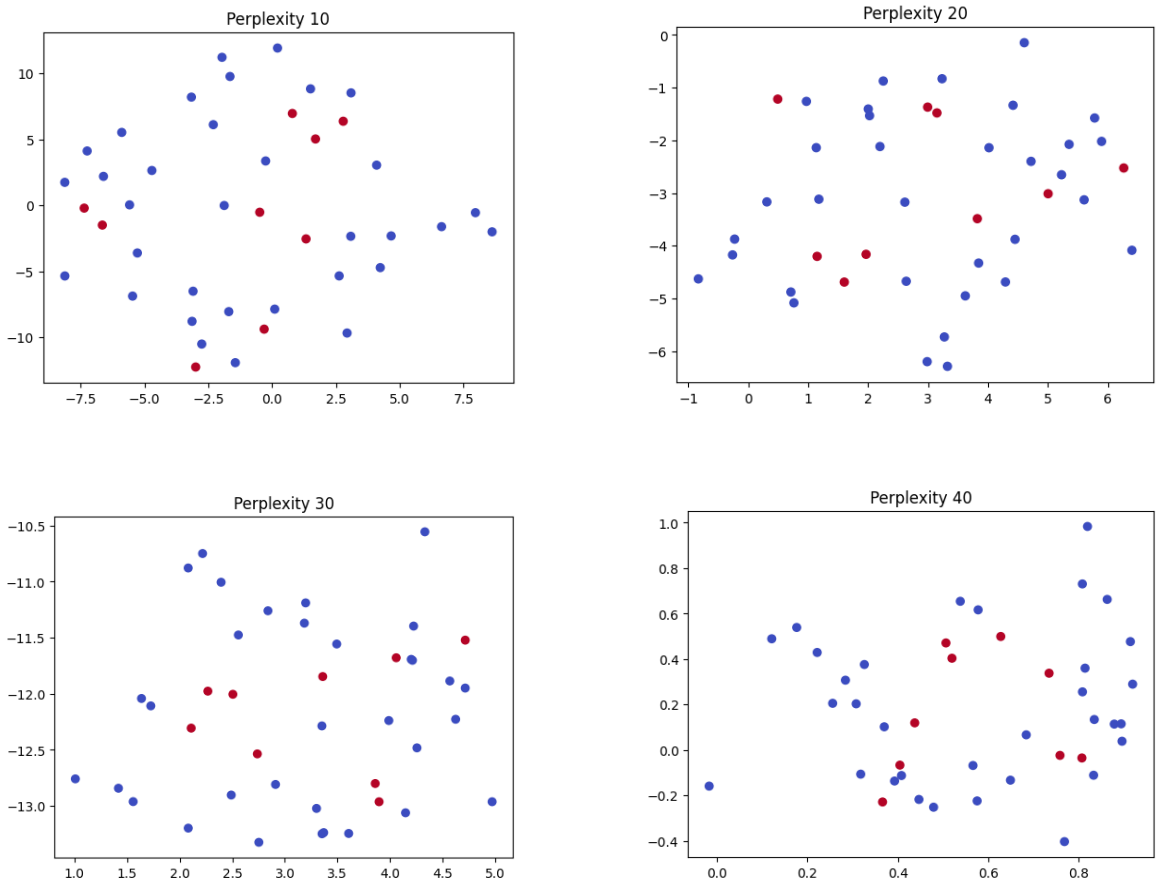


Figure 4.7: t-SNE representations of latent vectors created by the VAE trained with psilocybin Schaefer FCs. Blue points are training data and red points are validation data.

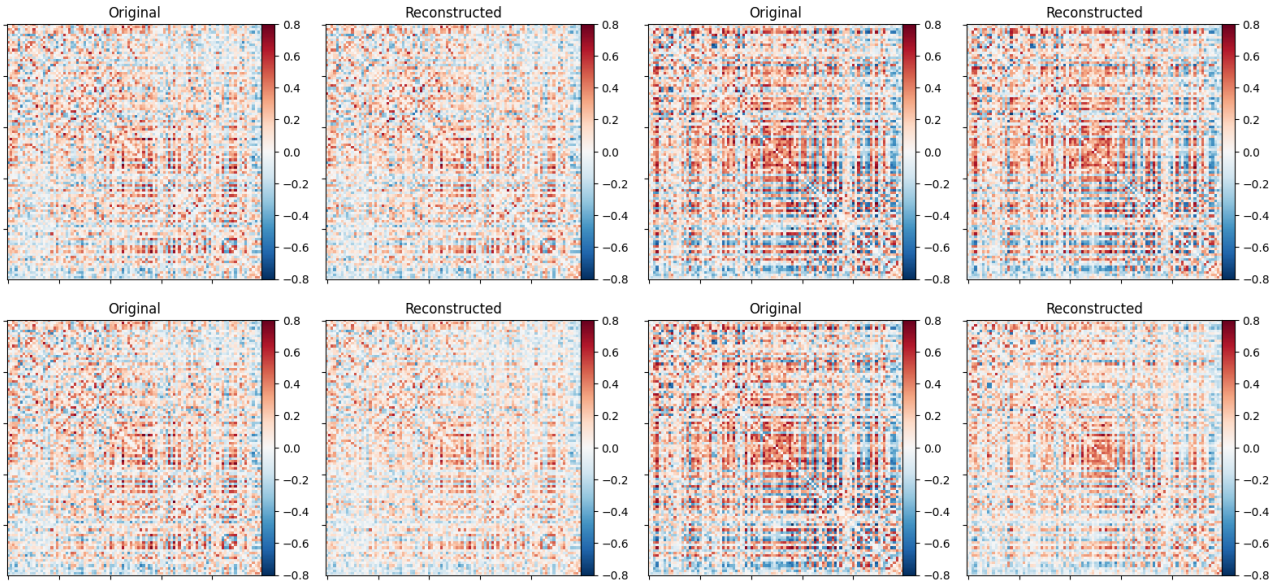


Figure 4.8: Reconstructions of the validation set after passing through the fine-tuned VAE (top) and the mixed class VAE (bottom). Notice the visibly better construction with the fine-tuned VAE.

Table 4.2: MAE losses and goodness of fit measures for the best-performing dropout for each model configuration.

Config	Dropout	MAE	Pearson r	p-value	R^2
fine_tune_before	0.35	6.99	0.319	3.94E-02	0.0848
fine_tune_combined	0.25	6.97	0.383	1.23E-02	0.132
ica_before	0.1	5.37	0.704	1.95E-07	0.478
ica_combined	0.2	5.67	0.613	1.59E-05	0.356
schaefer_before	0.25	5.59	0.640	5.10E-06	0.400
schaefer_combined	0	5.56	0.619	1.26E-05	0.379
aal_before	0.45	5.81	0.448	2.94E-03	0.162
aal_combined	0.15	6.01	0.536	2.54E-04	0.260

4. Latent vectors are generated using the ICA pre-trained + fine-tuned VAE (referred to as 'fine_tune')

For each configuration listed, we considered two sub-cases:

1. VAE was trained using a pre-trial dataset (referred to as 'before')
2. VAE was trained using the pre- and post-trial dataset (referred to as 'after')

The training results of the different configurations for a range of dropouts are summarised in table 4.2. Let's first compare the performance of the different atlases. The ICA atlas consistently outperforms the other atlases (R^2 = 0.478, Pearson r = 0.704). This indicates that the ICA atlas provides the most informative and meaningful latent representations for the MLP regression task.

We see that the performance of the atlases tends to improve as the reconstruction loss of the corresponding VAE decreases. This suggests that a more accurate and faithful encoding of the data by the VAE leads to better performance in the subsequent MLP regression. The VAE's ability to create a meaningful latent vector allows the MLP to extract more useful information and make more accurate predictions.

Comparing the performance of the combined VAE and before VAE, we find that the use of the combined VAE has a negligible to slightly negative effect on the regression performance. This implies that the visual distribution differences between the pre- and post-trial data are significant enough to outweigh the benefits of increased sample space coverage provided by the combined VAE.

However, despite the overall success of the VAE-based approach, the pre-training regime proved to be ineffective with the worst performance among all configurations. This is likely due to the large difference in visual distribution between the HCP and psilocybin datasets, which mitigates the benefit of pre-training.

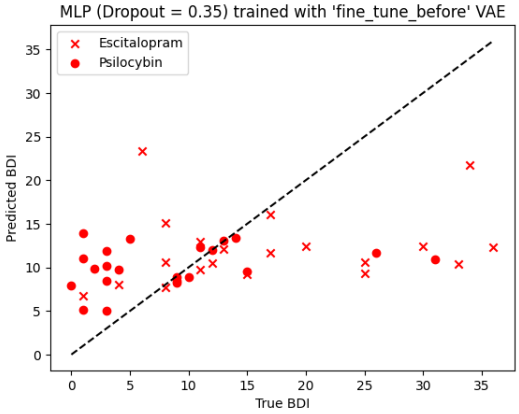
Given the potential for incorporating graphical relationships in the FC, we believe that a VGAE approach could provide even greater predictive power. This motivates us to explore the VGAE approach in the subsequent section.

4.7 Constructing a Variational Graph Autoencoder (VGAE)

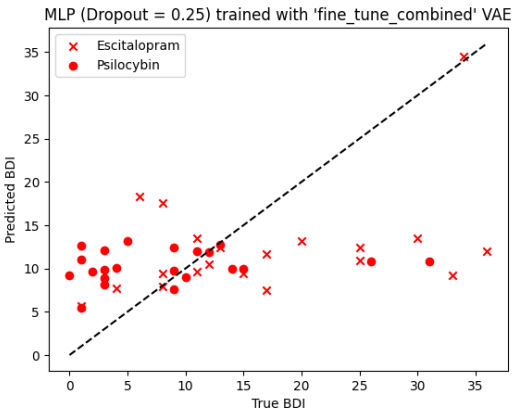
4.7.1 Architecture Iterations

To exploit the graphical nature of the FC, we constructed a VGAE as described in sections 2.4.5 and 2.12. The VGAE incorporates a graphical encoder that takes advantage of the inherent graph structure in the FC.

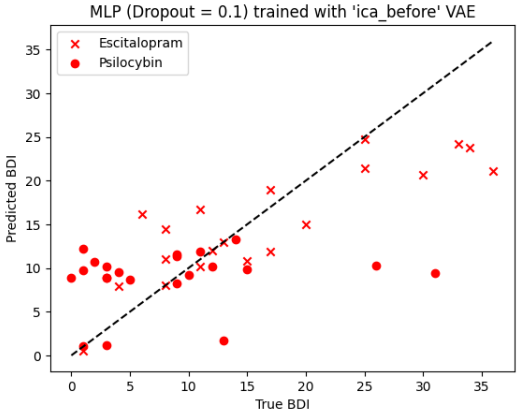
To enhance the information available for each RoI, we calculated additional features based on their respective time series. In this work, we calculated Lempel Ziv values (see section 2.2.5), which provide an entropy measure for each RoI. By using these values as node features, we capture additional information that would otherwise be lost during the FC computation.



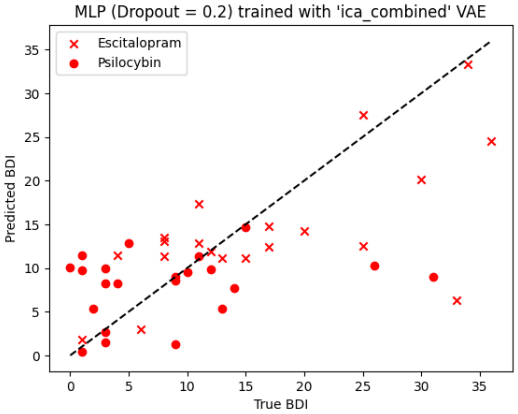
((a)) MAE = 6.99, Pearson r = 0.319 (p = 3.94e-2, R^2 = 0.0848 (all 3sf)



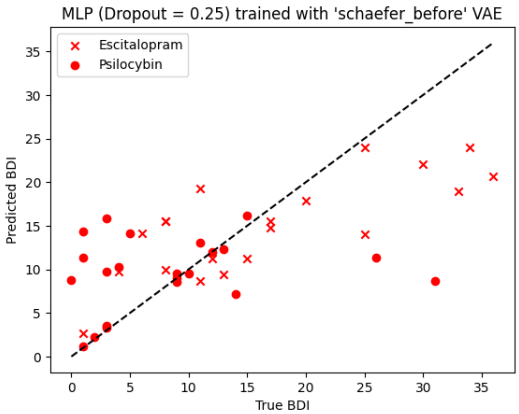
((b)) MAE = 6.97, Pearson r = 0.383 (p = 1.23e-2, R^2 = 0.132, (all 3sf)



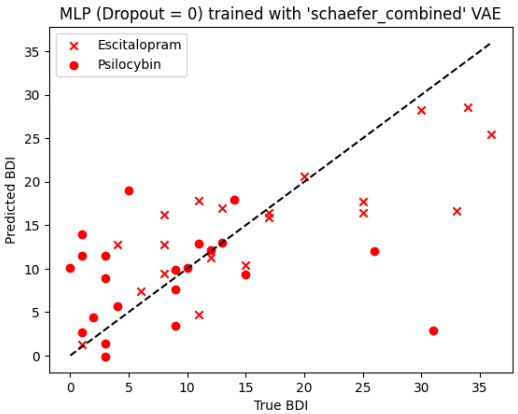
((c)) MAE = 5.37, Pearson r = 0.704 (p = 1.95e-7, R^2 = 0.478, (all 3sf)



((d)) MAE = 5.67, Pearson r = 0.613 (p = 1.59e-5, R^2 = 0.356, (all 3sf)

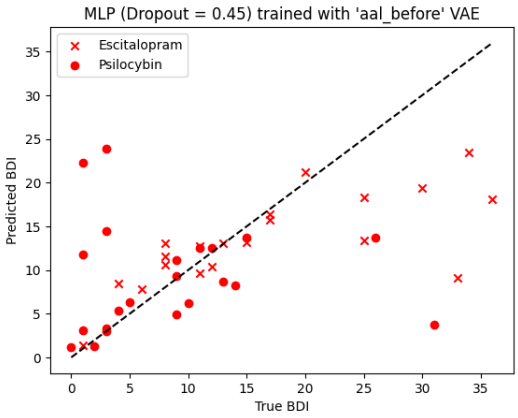


((e)) MAE = 5.59, Pearson r = 0.640 (p = 5.10e-6, R^2 = 0.400, (all 3sf)

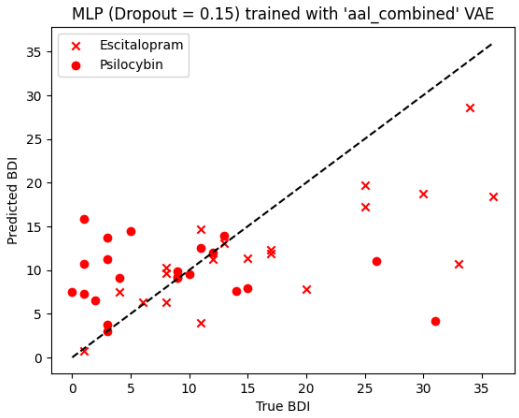


((f)) MAE = 5.56, Pearson r = 0.619 (p = 1.26e-5, R^2 = 0.379, (all 3sf)

Figure 4.9



((g)) MAE = 5.81, Pearson r = 0.448 (p = 2.94e-3, R^2 = 0.162, (all 3sf)



((h)) MAE = 6.01, Pearson r = 0.536 (p = 2.54e-4, R^2 = 0.260, (all 3sf)

Figure 4.9: Predicted BDI vs truth BDI graphs for MLPs trained on VAE latent vector and baselines, only showing the best-performing dropout for each VAE configuration. X = Y is plotted as a dotted line.

The VGAE model then compresses these node features into a latent representation, which captures the essential information of the FC graph in a lower-dimensional space. This latent representation enables the model to learn meaningful and informative representations of the FC, which can then be fed to an MLP.

Upon analyzing the Lempel Ziv structure of the HCP and psilocybin data, we observed a notable difference in their distributions, as depicted in figure 4.10. This difference in distribution is very apparent and could not be solely attributed to underlying variations in the time series data between the two datasets. We hypothesize that the disparity in distribution is created by differences in the preprocessing methods used in each dataset (see section 2.2.3).

In our trial architecture, we used MetaLayers to construct the encoder and decoder components of the VGAE. MetaLayers are specialized layers that incorporate various graph neural network (GNN) techniques such as propagation and convolution, as described in section 2.4.5. However, during the implementation, we encountered limitations in the reconstruction capabilities of MetaLayers. The convolutional message passing inherent in MetaLayers was not well-suited for reconstructing individual values within a graph structure. Consequently, we decided to switch to a conventional decoder architecture, employing simple fully connected layers instead.

To allow for a direct comparison of the benefits derived from the encoding architecture, we used the same latent dimension and decoder structure as in the VAE. By maintaining consistency in the latent dimension and decoder, any observed improvements in performance can be specifically attributed to the differences in the encoding architecture. This approach enables us to isolate and evaluate the impact of the encoder component in enhancing the model's capabilities.

After addressing the limitations of MetaLayers in the decoding process, we observed that even after the initial forward pass through the MetaLayer, our latent dimension exhibited a highly uniform distribution. As a result, we began to question the effectiveness of the MetaLayer in the context of compression.

4.7.2 Final Architecture for VGAE

To address the challenge of over-propagation, we opted for a fixed-edge node MLP architecture. This approach takes advantage of the fully connected nature of the FC, where each RoI has connections to every other RoI. By leveraging this characteristic, we can design an MLP that processes each node's input, including its region number and node features, along with the features of all the edges connected to that node.

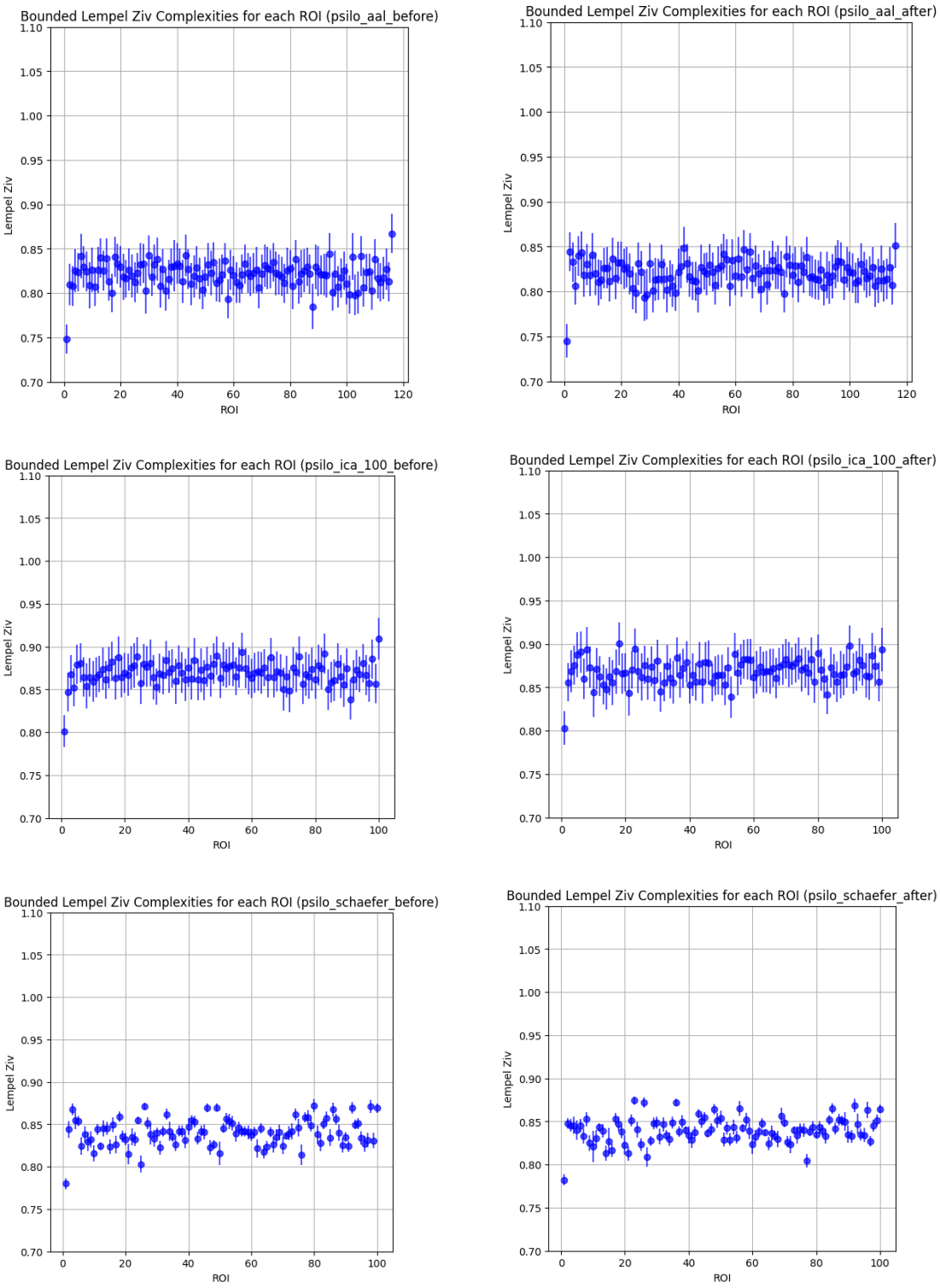

Bounded Lempel Ziv Complexities for each ROI (psilo_aal_before)
Bounded Lempel Ziv Complexities for each ROI (psilo_aal_after)
Bounded Lempel Ziv Complexities for each ROI (psilo_ica_100_before)
Bounded Lempel Ziv Complexities for each ROI (psilo_ica_100_after)
Bounded Lempel Ziv Complexities for each ROI (psilo_schaefer_before)
Bounded Lempel Ziv Complexities for each ROI (psilo_schaefer_after)
Lempel Ziv
ROI

Figure 4.10

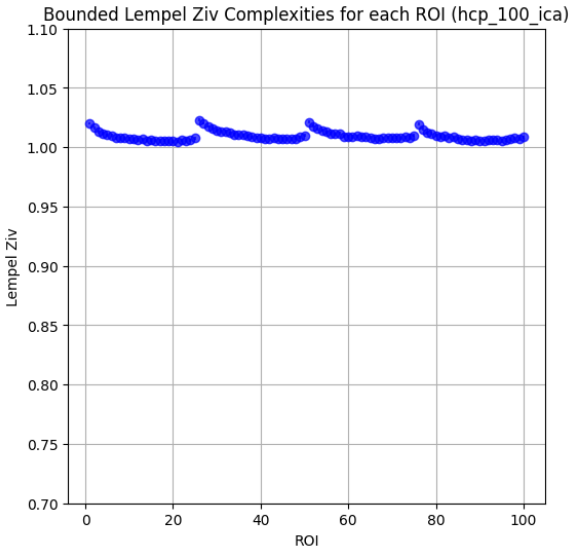


Figure 4.10: Bounded Lempel Ziv complexities computed per atlas.

The fixed-edge node MLP architecture enables us to learn a unique node representation for each RoI. By incorporating the features of the connected edges, we can capture relational patterns between the regions. This approach allows us to effectively encode the graph's structure and enhance the representation learning process.

During the experimentation phase, we initially considered inputting only the edges of the functional connectivity (FC) graph while excluding the diagonal (self-connections). However, upon further testing, we observed that including the entire graph as input led to improved performance. This improvement can be attributed to the spatial consistency in the input representation.

By including the entire graph, we preserve the spatial alignment between the neurons in the input layer and the corresponding edge weights connecting regions. This alignment ensures that the model can readily capture the relationship between the region identifier and the input location of each edge weight. This means the model does not need to learn the deeper relationship between these factors, resulting in improved performance.

The initial implementation of the fixed-edge node MLP involved using the node feature layer as the bottleneck. However, this approach encountered a couple of issues. Firstly, the node feature layer remained relatively large as it is always proportional to the number of RoIs. This larger size limited the effectiveness of the bottleneck, as it did not sufficiently compress the information within the node features. Furthermore, the model lacked a variational loss component, specifically the KL loss. The absence of this loss term resulted in poor conditioning of the latent representation of the model.

To regulate the node feature encodings, we implemented an infomax loss, which aims to construct encodings that provide evidence that a node belongs in its corresponding graph. However, during training, we observed that the infomax loss was static and made little contribution to the overall training process.

We hypothesize that the fully connected nature of the graph made it difficult to significantly differentiate between nodes and determine whether they belonged to specific graphs. The lack of distinct structural variation within the graph may have hindered the effectiveness of the infomax loss.

Based on our experimentation, we decided to incorporate a linear VAE in our architecture. After generating node descriptors, we flatten the descriptors and input them to a linear VAE. This approach allows us to compress the learned non-euclidean relationships within the graph into a latent vector of an appropriate dimension.

By introducing the linear VAE, we gain the advantage of conditioning the latent space through the

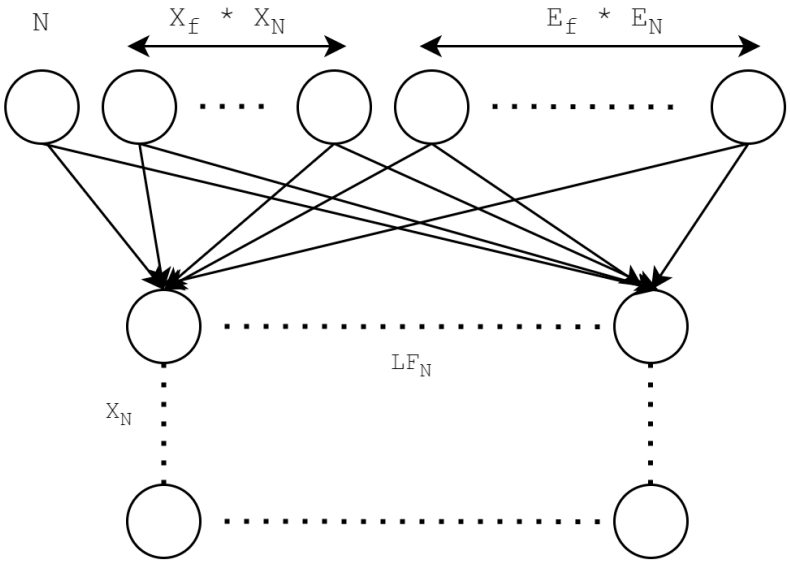


Figure 4.11: Stage 1

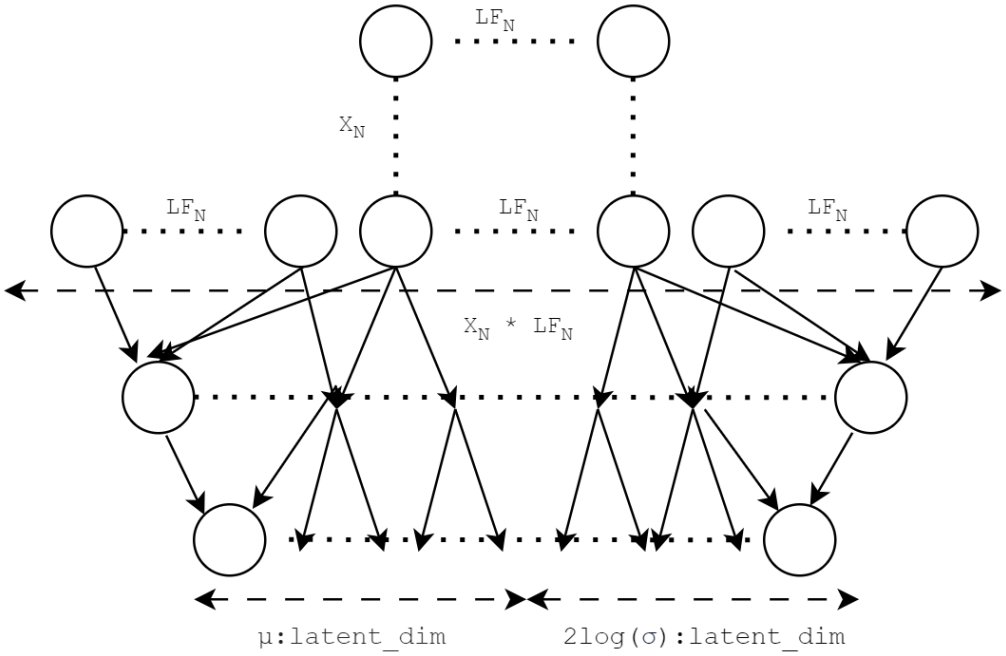


Figure 4.12: Diagram showing the 2 stages of the VGAE encoder. Stage 1 generates node descriptors of arbitrary length LF_N. This stage may implement several hidden layers. Stage 2 flattens the node descriptors to size $X_N * LF_N$, where X_N is the number of nodes and LF_N is the size of the node descriptors. This flattened vector is then input to a VAE to summarize global information about the learned graphical relationships. The VAE returns a variational encoding in μ and 2log(σ), both of size latent_dim which is a hyper-parameter.

Table 4.3: MSE reconstruction losses (2dp) over validation set for each VGAE configuration. HCP featureless failed to train and is omitted

VGAE type	HCP	ICA	Schaefer	AAL	Fine-tuned
Featureless		154.26	167.79	214.85	
Feature	32.66	152.71	158.15	203.11	150.9
VAE Val (comparison)	29.14	137.98	163.87	197.55	99.06

minimization of the KL divergence. This conditioning helps to ensure that the function boundary we learn is smoother and more representative of the underlying relationships within the data.

The combined approach of the fixed-edge node MLP and the linear VAE allows us to capture the graph's structural information and generate a low-dimensional vector with the VAE. The final architecture which we implemented is detailed in figure 4.12

4.7.3 Training the VGAE

Our VGAE training configurations are as follows:

1. First train using HCP ICA FCs and then fine-tune using psilocybin ICA FCs
2. Train only using psilocybin ICA FCs
3. Train only using psilocybin Schaefer FCs
4. Train only using psilocybin AAL FCs

For every configuration listed, we considered two additional cases:

1. FC is input alongside baseline metrics (referred to as 'featureless VGAE')
2. FC is input alongside baseline metrics in addition to node features (referred to as 'feature VGAE')

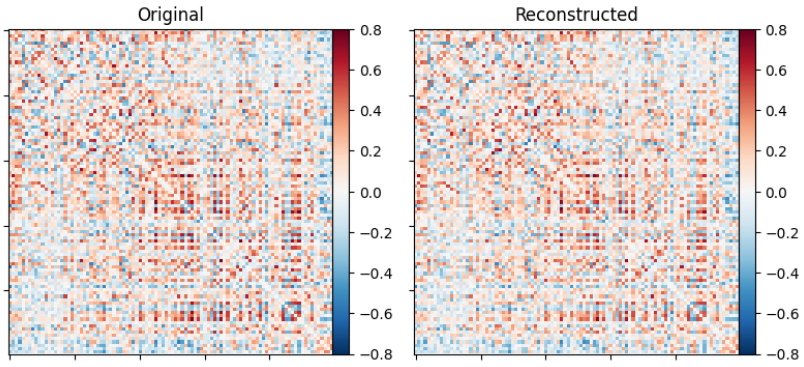
We chose to train the VGAE-MLP models using only pre-trial FCs, as the VAE training did not show significant improvement when using a combined dataset. This also ensures that their performance is solely based on the predictive patterns observed in the pre-trial dataset, eliminating any potential influence from post-trial information. Training for the VGAE was performed using an 80:20 training-validation split. The results of training are shown in table 4.3.

Training for the featureless VGAE using the HCP data was unsuccessful due to issues with exploding gradients. After some experimentation, the issue could not be solved trivially and so it has been purposely omitted from the table. The node feature encodings in some cases likely need further conditioning to converge properly.

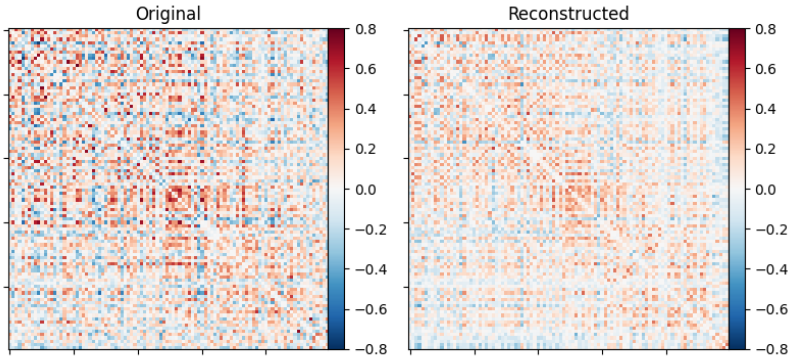
We observe that the reconstruction losses of the VGAE are slightly worse than those of the VAE. This could be attributed to the complexity of capturing the full graphical relationship of the FC using the node MLP. A simpler approach like the VAE may be able to heuristically capture some aspects of this relationship. Another factor to consider is that encoding the graphical relationship may require a larger latent vector or node descriptor.

It is important to note that a slightly worse reconstruction loss may not necessarily indicate a worse latent representation. If the latent vector captures a more meaningful graphical relationship, it may require more effort to accurately decode it. However, our VGAE uses the same decoder as the VAE. Therefore, it is possible to have a worse reconstruction loss and yet have a better latent transformation, as we will see in section 4.8.

The performance order by atlas remains consistent with that of the VAE. However, it is interesting to note that the feature VGAEs achieve lower reconstruction losses than the featureless VGAEs across all atlases. This finding is unexpected, considering that more information is being compressed into the bottleneck. We hypothesize that the inclusion of node features serves as a regularizer for the graphical structure, encouraging a deeper and more meaningful representation of the underlying graphical



((a)) Reconstruction using the fine-tuned VAE



((b)) Reconstruction using the fine-tuned VGAE

Figure 4.13: Comparison of reconstructions between architectures. Notice the significant difference in reconstruction quality.

relationship. Consequently, the feature VGAEs can capture the essential aspects of the graph more effectively, leading to improved reconstruction performance.

The t-SNE representations of the fine-tuned VGAE's latent vectors (see figure 4.14) show an even distribution between the training and validation sets, indicating an appropriate fit. However, similar to the t-SNE results for the VAE's latent space, there is a lack of clear clustering or discernible patterns. This observation further supports the idea that the samples within their original hyperspace may be disparate.

4.8 Building an MLP using latent VGAE embeddings

We trained multiple configurations of MLP using the latent vectors generated by different VGAEs. The configurations are as follows:

1. Latent vectors are generated using the ICA-trained VGAE (referred to as 'ica')
2. Latent vectors are generated using the Schaefer-trained VGAE (referred to as 'schaefer')
3. Latent vectors are generated using the AAL-trained VGAE (referred to as 'aal')
4. Latent vectors are generated using the ICA pre-trained + fine-tuned VGAE (referred to as 'fine_tune')

For each configuration listed (apart from fine_tune due to the non-convergence issue we encountered), we considered two sub-cases:

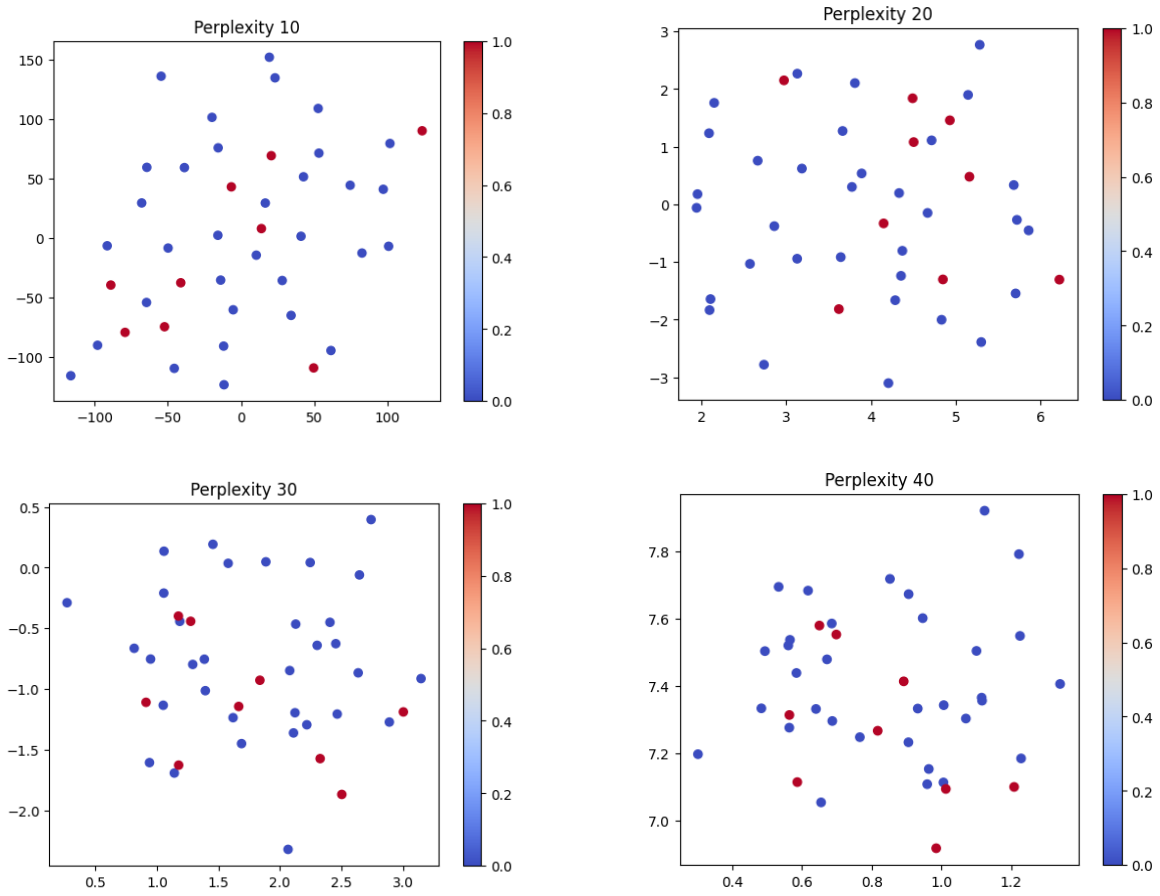


Figure 4.14: t-SNE representations of latent vectors created by the fine-tuned feature VGAE trained with psilocybin ICA FCs. Blue points are training data and red points are validation data.

Table 4.4: Table showing training results of 5 fold CV with MLP built on VGAE for each configuration. 'feature' denotes a configuration trained with node features and 'featureless' denotes a configuration trained without node features. Only listing the best-performing dropout (highest R^2) for each configuration (see full tables in appendices 6.2, 6.3). The best performance is achieved by the model highlighted in blue.

VGAE type	Dropout	MAE	Psilo MAE	Esc MAE	R^2	Pearson r	p-value
ica_feature	0.1	5.81	6.14	5.45	0.354	0.617	1.36E-05
ica_featureless	0.05	6.05	5.62	6.52	0.290	0.570	8.12E-05
aal_feature	0.15	5.24	5.30	5.18	0.472	0.704	2.02E-07
aal_featureless	0	5.82	5.69	5.97	0.315	0.585	4.69E-05
schaefer_feature	0	4.89	5.22	4.52	0.504	0.718	8.71E-08
schaefer_featureless	0.15	6.09	5.81	6.39	0.305	0.590	3.94E-05
fine_tune_feature	0.2	4.43	4.91	3.90	0.559	0.757	6.40E-09

1. VGAE was trained using node features ('feature')
2. VGAE was trained without node features ('featureless')

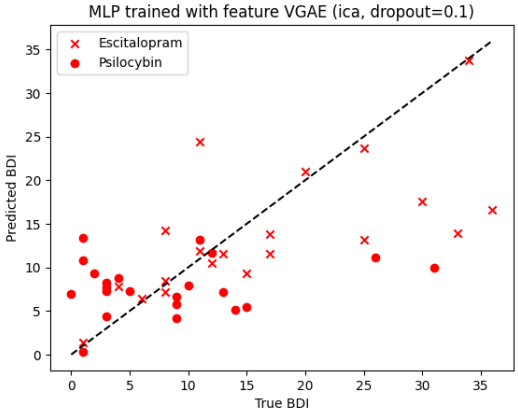
The results of training the VGAE-MLPs are summarised in Table 4.4. We can observe the significant impact of pre-training with a VGAE compared to a VAE. The configuration utilizing the pre-trained VGAE achieves the best performance among all model configurations, with a highly significant Pearson r value of 0.757 (p=6.40e-9). The significant improvement in the pre-training method between VAE and VGAE implies that large differences in visual distribution don't necessitate a large difference in graphical distribution. The VAE pre-training likely failed because the visual relationship that it learns to encode from the HCP data was not supported by the visual distribution of psilocybin FCs. The VGAE however, encodes a graphical relationship of HCP that is supported by the graphical distribution of the psilocybin FCs, and so the sample coverage ratio increases which enhances the model performance.

While the models demonstrate good performance in predicting post-trial BDI scores in the lower range, they encounter difficulties in accurately predicting scores for the two psilocybin samples in the upper half of the range. In contrast, the escitalopram regression shows a reasonable regression across the entire range. This observation suggests that with a larger number of psilocybin samples within the upper range, the model would likely be able to effectively capture the patterns and improve predictions for the upper range of the psilocybin samples as well.

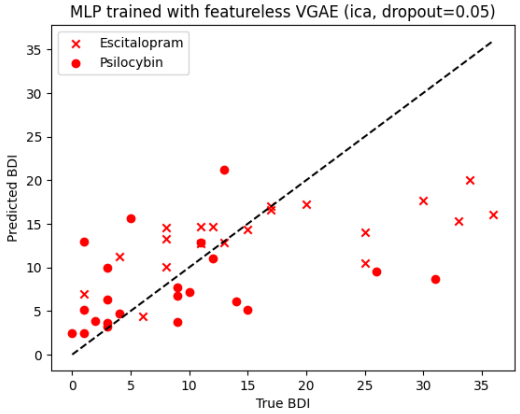
We observe that the two anatomical atlases outperform the non-anatomical atlas in terms of performance. This can be attributed to the fact that the graphical relationships between anatomical regions are more likely to be closely related to actual anatomical mechanisms. Interestingly, the ICA atlas is the only configuration that shows a deterioration in performance with the graphical approach. This suggests that the non-anatomical regions in the ICA atlas may have weaker or less clear graphical relationships compared to the anatomical regions.

These results are highly promising, especially because the best performance was achieved by fine-tuning with the atlas that had the worst performance without pre-training. The Schaefer feature VGAE already shows significant performance with an MAE of 4.89, approaching the performance of the fine-tuned feature VGAE with an MAE of 4.43. This suggests that further fine-tuning with the other atlases would likely yield even better regression. More interestingly, conducting input relevance detection on a fine-tuned anatomical VGAE would reveal the graphical relationships between anatomical regions, rather than the ICA regions. These relationships can be further analysed and interpreted by researchers, contributing to a deeper understanding of the underlying mechanisms involved in the recovery process.

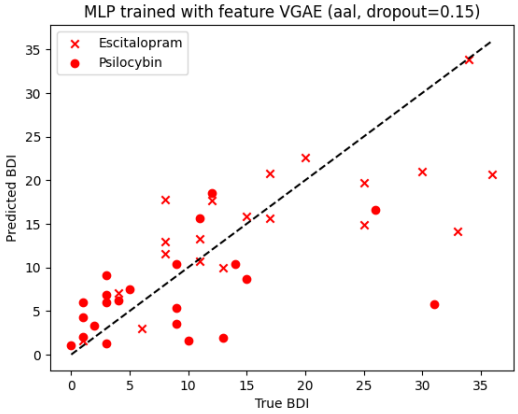
It is important to acknowledge that the encoder architecture used in the VGAE is an ad hoc design, specifically tailored for this particular task of fully connected graph encoding. While the current architecture has shown promising results, it is crucial to approach it with a critical mindset and recognize that alternative architectural choices or parameter configurations may exist that could potentially lead to even better performance.



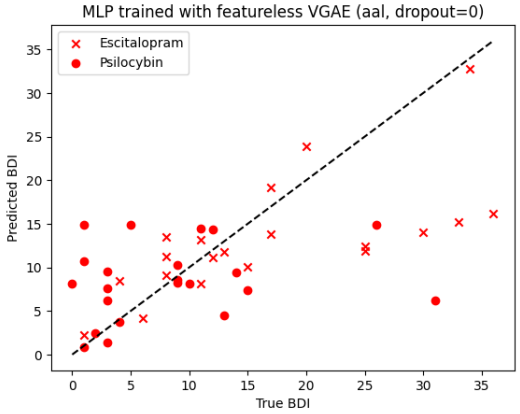
(a) MAE = 5.81, Pearson r = 0.617 (p = 1.36e-5), R^2 = 0.354, all to 3sf.



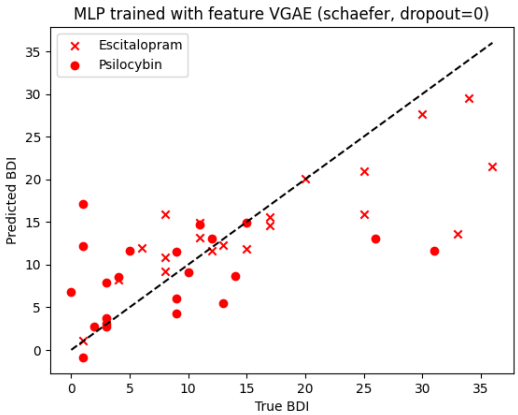
(b) MAE = 6.05, Pearson r = 0.570 (p = 8.12e-5), R^2 = 0.290, all to 3sf.



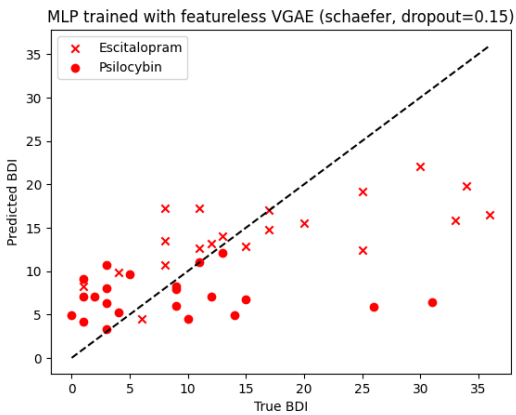
(c) MAE = 5.24, Pearson r = 0.704 (p = 2.02e-7), R^2 = 0.472, all to 3sf.



(d) MAE = 5.82, Pearson r = 0.585 (p = 4.69e-5), R^2 = 0.315, all to 3sf.

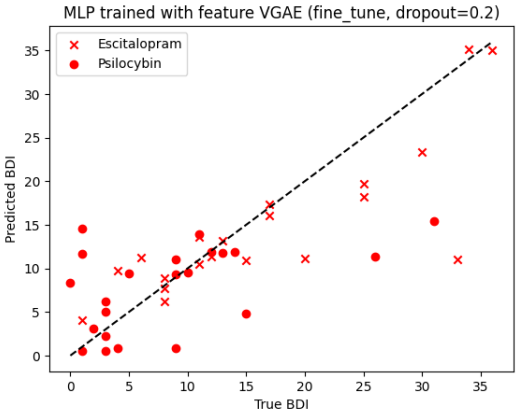


(e) MAE = 4.89, Pearson r = 0.718 (p = 8.71e-8), R^2 = 0.504, all to 3sf.



(f) MAE = 6.09, Pearson r = 0.590 (p = 3.94e-5), R^2 = 0.305, all to 3sf.

Figure 4.15



(c) MAE = 4.43, Pearson r = 0.757 (p = 6.40e-9), R^2 = 0.559, all to 3sf, outperforms all other configurations by all metrics.

Figure 4.15: Predicted BDI vs truth BDI graphs for MLPs trained on VGAE latent vector and baselines, only showing the best-performing dropout for each VAE configuration. X = Y is plotted as a dotted line.

Chapter 5

Conclusions

5.1 Summary of Results

The regression achieved by the best-performing model aligns with the findings from the initial psilocybin trial at the Centre for Psychedelic Research [5], where correlations were established between post-trial modularity and BDI change after 6 months. Notably, our model exclusively trains using pre-trial data, demonstrating its predictive potential with significant statistical measures (Pearson r = 0.757, p = 6.40e-9, R^2 = 0.559).

The incorporation of engineered node features in our model resulted in a substantial improvement, highlighting the value of informed feature engineering when training models with limited datasets. The ability to learn and encode robust graphical relationships, and condense them into a rich latent encoding, provides a strong foundation for accurate predictions.

Despite the higher reconstruction loss of the VGAE compared to its VAE counterpart, the VGAE's corresponding MLP demonstrates superior predictive performance. This indicates that reconstruction losses alone do not fully reflect the encoding quality and that a higher-quality latent encoding may lead to comparatively worse reconstructions when using a fixed decoder.

It is worth noting that the limited availability of high post-treatment BDI values for psilocybin patients hinders accurate predictions for the upper range. However, the fine-tuned feature MLP shows promising capability in accurately predicting escitalopram BDIs at the upper end of the range, even with a relatively small number of data points. This suggests that with an expanded dataset encompassing more points within the range, the model may achieve similar regression accuracy for psilocybin data.

5.2 Future Work

Although the ad hoc graphical encoder showcased impressive performance, it does exhibit convergence issues for certain datasets. Exploring methods to condition the node feature vectors without compromising the quality of the latent representation would yield a general-purpose structural encoder for fully connected graph encoding.

Further exploration of architecture parameters is a promising avenue for future research. While the current model performs well, many architectural decisions have not been exhaustively tested. For instance, the latent dimension may be insufficient to adequately capture the graphical relationships for atlases with a greater number of Regions of Interest (ROIs). Benchmarking the optimal latent dimension against the number of RoIs for a fixed atlas, facilitated by atlases like Schaefer with various parcellations, would be valuable. Additionally, considering alternative or deeper decoder structures may help reduce the reconstruction loss and improve the latent representation further.

In this study, we utilized Lempel Ziv as a node feature, but other informed measures such as synergy or redundancy could enhance the richness of the latent representation. Additionally, the inclusion of additional baseline metrics as input to the predictors could be a straightforward development.

The VAE pre-training approach failed with the HCP dataset due to potential disparities in visual distribution, likely caused by differences in pre-processing. It would be beneficial to investigate the effect of unifying the pre-processing or parcellation with a different atlas on the effectiveness of VAE and VGAE pre-training. Additionally, exploring different pre-training strategies with Schaefer or AAL FCs may lead to improved regression results due to stronger anatomical correlation.

As an additional avenue of exploration, performing automatic relevance determination on the VGAE models could help identify the input features that are predictive of the post-trial BDI. This analysis would provide insights into the graphical relationships between RoIs that are correlated with changes in BDI. Shifting to an anatomical atlas would yield mechanistically informative RoIs as input, aiding in informing further trials and uncovering the causal mechanisms of depression, which is the ultimate goal.

By addressing these future research directions, we can advance our understanding of graph-based methods for FC analysis, refine the encoding architecture, and uncover valuable insights into the underlying mechanisms of depression.

Bibliography

[1] for National Statistics (ONS) O. Suicides in England and Wales: 2021 registrations; 2022. Accessed: 21/1/2023. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/suicidesintheunitedkingdom/2021registrations.

[2] Rush AJ, Trivedi MH, Wisniewski SR, Nierenberg AA, Stewart JW, Warden D, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR* D report. American Journal of Psychiatry. 2006;163(11):1905-17. Accessed: 21/1/2023.

[3] Burcusa SL, Iacono WG. Risk for recurrence in depression. Science Direct. 2007;27(8):959-85. Accessed: 21/1/2023.

[4] Ferenchick EK, Ramanuj P, Pincus HA. Depression in primary care: part 1—screening and diagnosis. BMJ. 2019;365. Accessed: 21/1/2023.

[5] Daws RE, Timmermann C, Giribaldi B, Sexton JD, Wall MB, Erritzoe D, et al. Increased global integration in the brain after psilocybin therapy for depression. Nature medicine. 2022;28(4):844-51. Accessed: 21/1/2023.

[6] Association AP. Diagnostic and statistical manual of mental disorders (5th edition). 2022. Accessed: 21/1/2023.

[7] Bertolote JM, Fleischmann A. Suicide and psychiatric diagnosis: a worldwide perspective. World psychiatry. 2002;1(3):181. Accessed: 21/1/2023.

[8] Wykes T, Haro JM, Belli SR, Obradors-Tarragó C, Arango C, Ayuso-Mateos JL, et al. Mental health research priorities for Europe. The Lancet Psychiatry. 2015;2(11):1036-42. Accessed: 21/1/2023.

[9] Harmer CJ, Duman RS, Cowen PJ. How do antidepressants work? New perspectives for refining future treatment approaches. The Lancet Psychiatry. 2017;4(5):409-18. Accessed: 21/1/2023.

[10] Artin H, Zisook S, Ramanathan D. How do serotonergic psychedelics treat depression: The potential role of neuroplasticity. World Journal of Psychiatry. 2021;11(6):201. Accessed: 21/1/2023.

[11] Moncrieff J, Cooper RE, Stockmann T, Amendola S, Hengartner MP, Horowitz MA. The serotonin theory of depression: a systematic umbrella review of the evidence. Molecular psychiatry. 2022:1-14. Accessed: 21/1/2023.

[12] Beck AT, Steer RA, Brown G. Beck depression inventory–II. Psychological assessment. 1996.

[13] Wang X, Öngür D, Auerbach RP, Yao S. Cognitive vulnerability to major depression: view from the intrinsic network and cross-network interactions. Harvard review of psychiatry. 2016;24(3):188. Accessed: 21/1/2023.

[14] Stordal KI, Lundervold AJ, Egeland J, Mykletun A, Asbjørnsen A, Landrø NI, et al. Impairment across executive functions in recurrent major depression. Nordic Journal of Psychiatry. 2004;58(1):41-7. PMID: 14985153. Available from: https://doi.org/10.1080/08039480310000789.

[15] Richardson G. Ayahuasca Use Throughout Time: A Literature Review. 2020. Accessed: 21/1/2023.

[16] Carhart-Harris RL, Goodwin GM. The therapeutic potential of psychedelic drugs: past, present, and future. Neuropsychopharmacology. 2017;42(11):2105-13. Accessed: 21/1/2023.

[17] Vohryzek J, Cabral J, Lord LD, Fernandes HM, Roseman L, Nutt DJ, et al. Brain dynamics predictive of response to psilocybin for treatment-resistant depression. bioRxiv. 2022:2022-06. Accessed: 21/1/2023.

[18] Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A. Neurophysiological investigation of the basis of the fMRI signal. nature. 2001;412(6843):150-7. Accessed: 21/1/2023.

[19] Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Frontiers in neuroinformatics. 2014:14. Accessed: 21/1/2023.

[20] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011;12(Oct):2825-30. Accessed: 21/1/2023.

[21] Project HC. 1200 Subjects Data Release; 2017. "Accessed: 21/1/2023". https://www.humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release.

[22] Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo XN, Holmes AJ, et al. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cerebral cortex. 2018;28(9):3095-114.

[23] Fuster JM. The prefrontal cortex—an update: time is of the essence. Neuron. 2001;30(2):319-33. Accessed: 21/1/2023.

[24] Rolls ET, Joliot M, Tzourio-Mazoyer N. Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. Neuroimage. 2015;122:1-5. Accessed: 21/1/2023.

[25] Contreras JA, Goñi J, Risacher SL, Sporns O, Saykin AJ. The structural and functional connectome and prediction of risk for cognitive impairment in older adults. Current behavioral neuroscience reports. 2015;2:234-45.

[26] Richardson H, Lisandrelli G, Riobueno-Naylor A, Saxe R. "MRI data of 3-12 year old children and adults during viewing of a short animated film". OpenNeuro; 2018.

[27] Kalev K, Bachmann M, Orgo L, Lass J, Hinrikus H. Lempel-Ziv and multiscale Lempel-Ziv complexity in depression. In: 2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE; 2015. p. 4158-61.

[28] Bachmann M, Kalev K, Suhhova A, Lass J, Hinrikus H. Lempel Ziv complexity of EEG in depression. In: 6th European Conference of the International Federation for Medical and Biological Engineering: MBEC 2014, 7-11 September 2014, Dubrovnik, Croatia. Springer; 2015. p. 58-61.

[29] Xia L, Zhang X, Li B. Improving Deep Learning Accuracy with Noisy Autoencoders Embedded Perturbative Layers. In: ICIC; 2016. Accessed: 21/1/2023.

Chapter 6

Appendices

Table 6.1: VAE validation losses over all dropouts and all configurations.

Configuration	Dropout	Validation Loss (2dp)
hcp	0.00	28.84
hcp	0.05	30.77
hcp	0.10	32.49
hcp	0.15	33.13
hcp	0.20	34.10
hcp	0.25	35.28
hcp	0.30	36.14
hcp	0.35	36.86
hcp	0.40	37.02
hcp	0.45	37.53
hcp	0.50	37.82
psilo_ica_before	0.00	145.55
psilo_ica_before	0.05	150.43
psilo_ica_before	0.10	149.37
psilo_ica_before	0.15	147.18
psilo_ica_before	0.20	148.13
psilo_ica_before	0.25	150.64
psilo_ica_before	0.30	151.74
psilo_ica_before	0.35	152.97
psilo_ica_before	0.40	153.44
psilo_ica_before	0.45	153.74
psilo_ica_before	0.50	158.21
psilo_ica_combined	0.00	131.77
psilo_ica_combined	0.05	135.45
psilo_ica_combined	0.10	135.47
psilo_ica_combined	0.15	134.95
psilo_ica_combined	0.20	136.25
psilo_ica_combined	0.25	138.63
psilo_ica_combined	0.30	141.30
psilo_ica_combined	0.35	143.10
psilo_ica_combined	0.40	145.44
psilo_ica_combined	0.45	154.00
psilo_ica_combined	0.50	156.21
psilo_schaefer_before	0.00	154.88
psilo_schaefer_before	0.05	156.20

Continued on next page

Table 6.1 – Continued from previous page

Configuration	Dropout	Validation Loss
psilo_schaefer_before	0.10	159.19
psilo_schaefer_before	0.15	158.92
psilo_schaefer_before	0.20	162.14
psilo_schaefer_before	0.25	162.13
psilo_schaefer_before	0.30	165.35
psilo_schaefer_before	0.35	164.06
psilo_schaefer_before	0.40	171.13
psilo_schaefer_before	0.45	189.09
psilo_schaefer_before	0.50	196.91
psilo_schaefer_combined	0.00	131.14
psilo_schaefer_combined	0.05	133.81
psilo_schaefer_combined	0.10	138.43
psilo_schaefer_combined	0.15	140.98
psilo_schaefer_combined	0.20	143.05
psilo_schaefer_combined	0.25	149.46
psilo_schaefer_combined	0.30	152.11
psilo_schaefer_combined	0.35	156.53
psilo_schaefer_combined	0.40	156.91
psilo_schaefer_combined	0.45	158.55
psilo_schaefer_combined	0.50	159.43
psilo_aal_before	0.00	194.22
psilo_aal_before	0.05	197.52
psilo_aal_before	0.10	200.29
psilo_aal_before	0.15	204.16
psilo_aal_before	0.20	200.26
psilo_aal_before	0.25	202.00
psilo_aal_before	0.30	205.47
psilo_aal_before	0.35	218.69
psilo_aal_before	0.40	222.45
psilo_aal_before	0.45	231.13
psilo_aal_before	0.50	235.63
psilo_aal_combined	0.00	165.14
psilo_aal_combined	0.05	174.90
psilo_aal_combined	0.10	178.26
psilo_aal_combined	0.15	176.15
psilo_aal_combined	0.20	177.76
psilo_aal_combined	0.25	180.83
psilo_aal_combined	0.30	183.60
psilo_aal_combined	0.35	193.57
psilo_aal_combined	0.40	198.08
psilo_aal_combined	0.45	196.94
psilo_aal_combined	0.50	198.27
fine_tune_combined	0.00	103.41
fine_tune_combined	0.05	127.91
fine_tune_combined	0.10	126.13
fine_tune_combined	0.15	119.11
fine_tune_combined	0.20	111.83
fine_tune_combined	0.25	112.84
fine_tune_combined	0.30	111.29
fine_tune_combined	0.35	107.70
fine_tune_combined	0.40	115.01
fine_tune_combined	0.45	111.33

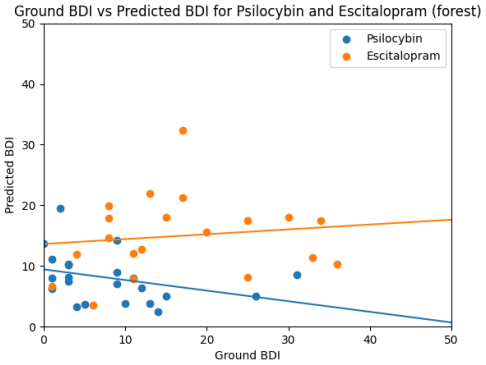
Continued on next page

Table 6.1 – Continued from previous page

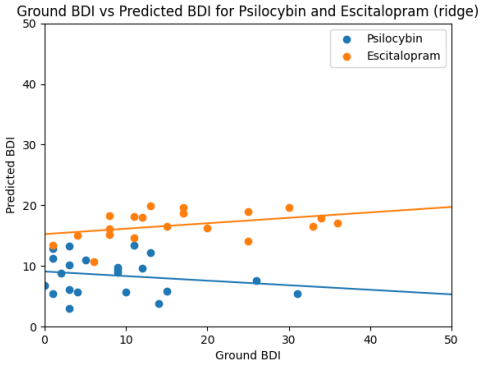
Configuration	Dropout	Validation Loss
fine_tune_combined	0.50	114.62
fine_tune_before	0.00	100.57
fine_tune_before	0.05	125.71
fine_tune_before	0.10	124.61
fine_tune_before	0.15	120.57
fine_tune_before	0.20	119.03
fine_tune_before	0.25	122.34
fine_tune_before	0.30	110.06
fine_tune_before	0.35	112.53
fine_tune_before	0.40	113.50
fine_tune_before	0.45	112.17
fine_tune_before	0.50	118.15

Table 6.2: Feature VGAE training results. A model with the best R^2 is highlighted in blue for each configuration.

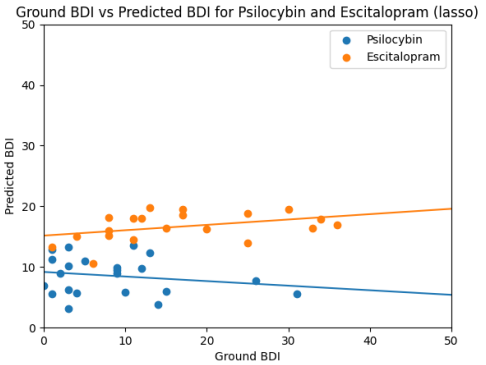
Atlas	Dropout	MAE	Psilo MAE	Escitalopram MAE	R^2	Pearson r	p-value
ica	0.00	5.96	6.17	5.73	0.301	0.600	2.70e-5
ica	0.05	5.74	5.56	5.94	0.329	0.604	2.28e-5
ica	0.10	5.81	6.14	5.45	0.354	0.617	1.36e-5
ica	0.15	5.92	5.84	6.01	0.261	0.554	1.40e-4
ica	0.20	5.67	6.47	4.78	0.344	0.616	1.41e-5
ica	0.25	5.85	5.78	5.93	0.334	0.604	2.28e-5
ica	0.30	6.11	6.44	5.74	0.250	0.553	1.48e-4
ica	0.35	5.90	5.93	5.88	0.328	0.607	2.00e-5
ica	0.40	5.87	6.06	5.67	0.269	0.553	1.44e-4
ica	0.45	5.94	6.09	5.78	0.304	0.609	1.88e-5
ica	0.50	6.14	6.29	5.97	0.248	0.552	1.50e-4
aal	0.00	5.50	5.49	5.51	0.407	0.646	3.83e-6
aal	0.05	5.29	5.42	5.14	0.400	0.660	1.96e-6
aal	0.10	5.44	5.34	5.55	0.401	0.661	1.87e-6
aal	0.15	5.24	5.30	5.18	0.472	0.704	2.03e-7
aal	0.20	5.30	5.44	5.15	0.457	0.708	1.53e-7
aal	0.25	5.26	5.40	5.10	0.405	0.662	1.82e-6
aal	0.30	5.20	4.87	5.57	0.460	0.692	3.96e-7
aal	0.35	5.31	5.24	5.38	0.428	0.666	1.46e-6
aal	0.40	5.46	5.44	5.48	0.385	0.638	5.50e-6
aal	0.45	5.48	5.72	5.22	0.373	0.653	2.74e-6
aal	0.50	5.46	5.54	5.37	0.345	0.622	1.12e-5
schaefer	0.00	4.89	5.22	4.52	0.504	0.718	8.71e-8
schaefer	0.05	5.32	5.48	5.16	0.459	0.690	4.40e-7
schaefer	0.10	5.20	5.16	5.23	0.470	0.693	3.65e-7
schaefer	0.15	5.27	5.62	4.88	0.466	0.692	3.87e-7
schaefer	0.20	5.05	5.32	4.76	0.489	0.708	1.55e-7
schaefer	0.25	5.36	5.79	4.88	0.430	0.659	2.08e-6
schaefer	0.30	5.07	5.26	4.85	0.429	0.668	1.37e-6
schaefer	0.35	5.09	5.64	4.49	0.428	0.668	1.32e-6
schaefer	0.40	5.16	5.46	4.83	0.451	0.680	7.29e-7
schaefer	0.45	5.32	5.49	5.13	0.413	0.662	1.78e-6
schaefer	0.50	5.46	5.91	4.96	0.409	0.649	3.29e-6
fine_tune	0.00	4.75	4.98	4.50	0.493	0.710	1.43e-7
fine_tune	0.05	4.98	5.59	4.31	0.468	0.695	3.36e-7
fine_tune	0.10	4.83	5.25	4.37	0.478	0.699	2.64e-7
fine_tune	0.15	4.84	5.46	4.16	0.492	0.713	1.18e-7
fine_tune	0.20	4.43	4.91	3.90	0.559	0.757	6.40e-9
fine_tune	0.25	4.58	5.35	3.74	0.511	0.723	6.30e-8
fine_tune	0.30	4.73	5.51	3.88	0.494	0.713	1.16e-7
fine_tune	0.35	4.74	5.42	3.99	0.494	0.711	1.30e-7
fine_tune	0.40	4.64	5.37	3.84	0.516	0.729	4.44e-8
fine_tune	0.45	4.85	5.25	4.41	0.446	0.687	5.15e-7
fine_tune	0.50	4.68	5.04	4.28	0.503	0.724	6.04e-8



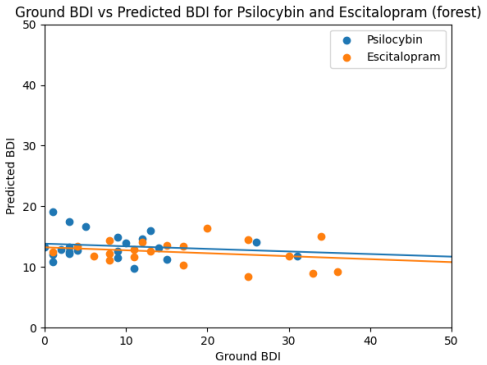
((a)) No FC input for random forest regression, MAE = 8.61, $R^2 < 0$



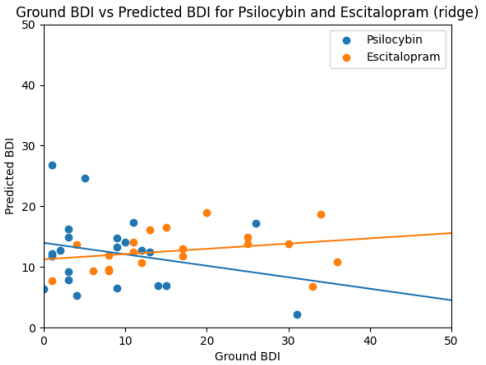
((b)) No FC input for ridge regression, MAE = 7.35, $R^2 < 0$



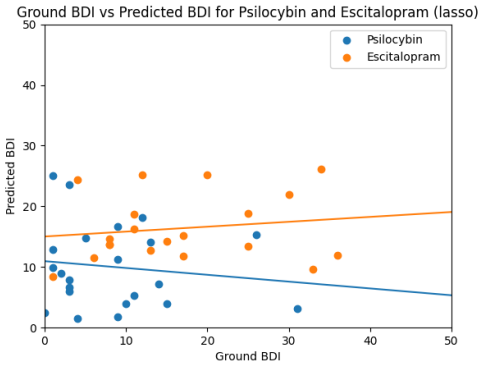
((c)) No FC input for lasso regression, MAE = 7.35, $R^2 < 0$



((d)) AAL FC input for random forest regression, MAE = 8.60, $R^2 < 0$

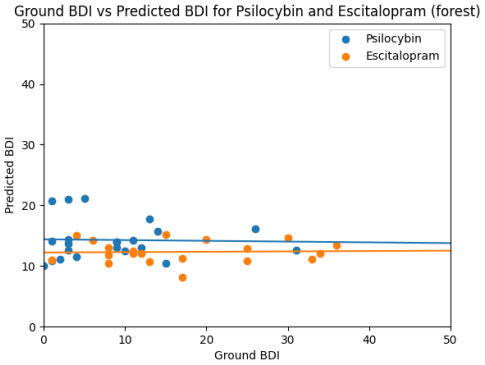


((e)) AAL FC input for ridge regression, MAE = 8.35, $R^2 < 0$

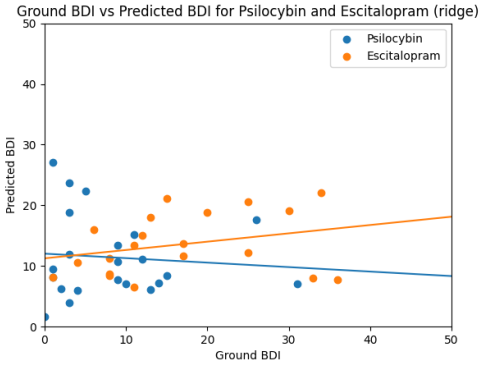


((f)) AAL FC input for lasso regression, MAE = 8.63, $R^2 < 0$

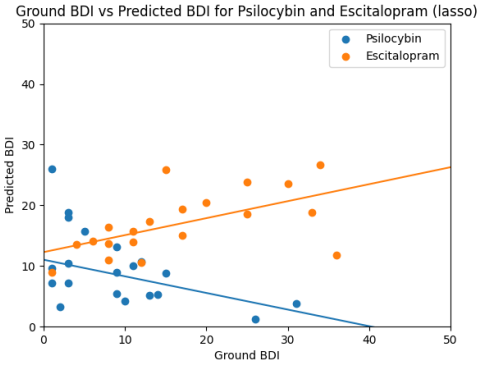
Figure 6.1



((g)) Schaefer FC input for random forest regression, MAE = 8.78, $R^2 < 0$



((h)) Schaefer FC input for ridge regression, MAE = 7.94, $R^2 < 0$



((i)) Schaefer FC input for lasso regression, MAE = 7.66, $R^2 < 0$

Figure 6.1: Graphs showing Ridge, Lasso and Random Forest regressions for Schaefer and AAL atlases as well as no FC input.

Table 6.3: Featureless VGAE training results. A model with the best R^2 is highlighted in blue for each configuration.

Atlas	Dropout	MAE	Psilo MAE	Escitalopram MAE	R^2	Pearson r	p-value
ica	0	6.07	5.74	6.43	0.247	0.544	1.94e-4
ica	0.05	6.05	5.62	6.52	0.290	0.570	8.12e-5
ica	0.1	5.97	5.27	6.74	0.206	0.517	4.54e-4
ica	0.15	5.81	5.02	6.67	0.276	0.558	1.22e-4
ica	0.2	6.00	5.19	6.90	0.208	0.497	8.20e-4
ica	0.25	5.96	5.20	6.80	0.213	0.512	5.26e-4
ica	0.3	6.02	5.33	6.77	0.250	0.534	2.72e-4
ica	0.35	5.90	4.88	7.04	0.197	0.512	5.35e-4
ica	0.4	5.83	5.21	6.52	0.279	0.561	1.12e-4
ica	0.45	6.04	5.21	6.96	0.163	0.481	1.25e-3
ica	0.5	5.90	4.86	7.04	0.181	0.507	6.18e-4
schaefer	0	6.16	5.72	6.65	0.274	0.575	6.92e-5
schaefer	0.05	6.07	5.60	6.59	0.264	0.589	4.00e-5
schaefer	0.1	6.16	5.82	6.53	0.287	0.564	1.00e-4
schaefer	0.15	6.09	5.81	6.39	0.305	0.590	3.94e-5
schaefer	0.2	6.09	5.74	6.47	0.294	0.571	7.74e-5
schaefer	0.25	6.10	5.81	6.41	0.280	0.576	6.61e-5
schaefer	0.3	6.09	5.76	6.46	0.268	0.565	9.72e-5
schaefer	0.35	6.11	5.78	6.48	0.258	0.560	1.14e-4
schaefer	0.4	6.02	5.75	6.30	0.254	0.550	1.61e-4
schaefer	0.45	6.10	5.71	6.53	0.237	0.562	1.06e-4
schaefer	0.5	6.12	5.59	6.71	0.240	0.553	1.45e-4
aal	0	5.82	5.69	5.97	0.315	0.585	4.69e-5
aal	0.05	6.00	6.18	5.80	0.307	0.573	7.34e-5
aal	0.1	6.14	6.09	6.19	0.266	0.540	2.22e-4
aal	0.15	5.97	6.07	5.86	0.268	0.535	2.62e-4
aal	0.2	5.95	5.84	6.07	0.248	0.534	2.68e-4
aal	0.25	5.92	5.51	6.38	0.299	0.577	6.29e-5
aal	0.3	6.09	5.69	6.53	0.230	0.529	3.20e-4
aal	0.35	5.89	5.40	6.44	0.248	0.557	1.29e-4
aal	0.4	5.98	5.87	6.10	0.199	0.515	4.82e-4
aal	0.45	5.83	5.41	6.30	0.281	0.580	5.68e-5
aal	0.5	5.89	5.54	6.28	0.218	0.562	1.07e-4