# Machine Learning Based Analysis of Gene Expression and Lipid Composition of the Schizophrenia Brain

Mikhail Zybin

## ABSTRACT

The molecular mechanisms of schizophrenia are still poorly understood. In this research, we apply machine learning techniques with the purpose of discovering what genes, lipids, and brain regions are relevant to this disease. We analyzed the dataset of gene expression and lipid composition post-mortem measurements in multiple brain regions in four healthy individuals and four individuals diagnosed with schizophrenia. We visualized the dataset using PCA (Principal Component Analysis). We applied random forest, logistic regression, and SVM (Support Vector Machine) machine learning classification algorithms to the dataset, treating the disease as a label. We used two approaches: the brain as an object and the region as an object.

The number of people is our dataset is small — this is the limitation of such datasets, because it is hard to obtain biological samples, and at the same time is is easy to make many measurements from one sample. We address this problem by data augmentation. We estimate the distributions of molecule measurements and generate the synthetic dataset. Also, we find and incorporate an external dataset similar to ours.

We were able to estimate the importances of genes, lipids, and regions. Namely, we obtained the Gini importance according to the random forest algorithm and the permutational importances for logistic regression and SVM. The importances of the brain regions were computed as the accuracy of diagnosis prediction for such a region or as the sum of the importances of the molecules in the region. We found out that some molecules and regions distinguish the healthy and schizophrenic brains more than others. Since the molecular changes in schizophrenic brains are relatively subtle, the results obtained from different algorithms are not exactly the same.

Keywords: schizophrenia, data augmentation, distribution estimation, machine learning, dimensionality reduction, transcriptomics, lipidomics

Research advisor:
Name: Philipp Khaitovich
Degree, title: PhD, Professor

# Анализ при помощи машинного обучения экспрессии генов и содержания липидов в мозге при шизофрении

Михаил Зыбин

## Аннотация

Молекулярные механизмы шизофрении до сих пор плохо изучены. В данной работе мы применяем методы машинного обучения с целью выяснить, какие гены, липиды и области мозга имеют отношение к этому заболеванию. Мы проанализировали набор данных посмертных измерений экспрессии генов и содержания липидов в различных регионах мозга у четырех людей с диагнозом шизофрения и у четырех людей без этого диагноза. Мы визуализировали набор данных с помощью PCA (Principal Component Analysis). Мы применили алгоритмы классификации — случайный лес, логистическую регрессию и SVM (Support Vector Machine) — к набору данных, рассматривая болезнь как целевую переменную. Мы использовали два подхода при формировании обучающей выборки: мозг как объект и регион как объект.

Число людей в наборе данных невелико — таково ограничение подобного рода наборов, поскольку биологические образцы добыть сложно, а измерений на основании одного образца можно сделать много. Чтобы решить проблему малого числа объектов, мы оцениваем распределение измерений молекул и генерируем синтетические данные. Также используется внешний набор данных, схожий с нашим.

Мы смогли оценить значимость генов, липидов и регионов. А именно, мы получили значимость Джини по алгоритму случайного леса и пермутационную значимость для логистической регрессии и SVM. Значимость регионов мозга рассчитывалась как точность предсказания диагноза для данного региона или как сумма значимости молекул в регионе. Мы обнаружили, что некоторые молекулы и регионы различают здоровый и мозг с шизофренией больше, чем другие.

# Contents

# List of Figures

# List of Tables

# Chapter 1
# Introduction

**Relevance** The molecular mechanisms of schizophrenia are still poorly understood. The results of this research can help other scientists perform more targeted experiments to discover the mechanisms of schizophrenia.

**Main purpose of the research** In this thesis, we apply statistical and machine learning (ML) methods to schizophrenia (SZ) research. The main goal is to acquire insights into the molecular level of schizophrenia in different anatomical regions of the brain.

This thesis is devoted to discovering which genes, which lipids in which brain regions are associated with schizophrenia. We identify the most relevant regions, genes, and lipids from the available data, so that potential future experiments on the molecular mechanisms of SZ could rely only on a small subset of regions, genes, and lipids.

**Scientific novelty** The novelty of this research is that we combine transcriptomics and lipidomics data across multiple brain regions, and the number of brain regions is especially large. The general limitation of such datasets is the lack of samples to study. However, we overcome it by using data augmentation.

**Practical value** Potential future experiments on the molecular mechanisms of SZ would rely only on a small subset of regions, genes, and lipids; therefore, the experiments would be more cost-effective. And this, in turn, can result in the faster development of novel diagnostic and treatment approaches.

**Statements for Defense** We check that it is possible to classify based on gene expression and lipid composition data whether the brain or the region of the brain comes from a person diagnosed with schizophrenia. Also, it is possible to identify the list of the most relevant regions, genes, and lipids that influence the decision of the ML algorithm.

**Literature review**

**On Schizophrenia in General** Schizophrenia is a psychiatric disorder with a convoluted etiology and a lifetime prevalence of 0.84%. It is thought that an interplay between genetic, epigenetic, and environmental risk factors is involved in SZ etiology.

Symptoms of SZ include positive, negative, and cognitive symptoms. The positive symptoms comprise delusions and hallucinations; the negative symptoms are loss of typical affective functions; and the most prominent cognitive symptoms of SZ are deficits in attention and executive functioning [13].

It is unlikely, and overly simplistic, that single-cause mechanisms or simple (linear) relationships between small sets of biomarkers could explain or even classify the different forms of SZ. SZ is perhaps best understood as an umbrella term, comprising several distinct disorders with partially overlapping phenomenology and neural correlates. Here is the list of data that has been

used in various research projects: fMRI, structural MRI, EEG, genomics data, and transcriptomics data [3].

**Genetic data** In the paper [5] the authors used genomic (SNPs) and other (gene expression, enhancer H3K27ac activation levels, cell fraction estimates, and co-expression module mean expression) data from the prefrontal cortex. They have created a comprehensive online resource for the adult brain across 1866 individuals, containing various types of data.

Another possible approach is described here [2]. In this paper, the authors used the polygenic scores for various conditions comorbid with schizophrenia to predict the diagnosis. They applied LASSO logistic regression and a deep neural network. Unfortunately, we do not have access to GWAS data for the people in our dataset, and we are unaware of their comorbid conditions.

We found 5 papers where the transcriptomic dataset of a similar structure was created: [14], [15], [1], [12], [17]. The first four of them use microarray technology [4], while the last one uses RNA-seq [21]. Since our dataset is obtained using RNA-seq, we can augment it only with the dataset from study [17], because the two technologies are incompatible.

**Lipids** There is a crucial role for brain lipids in depression, schizophrenia, and drug addiction, because brain lipids determine the localization and function of proteins in the cell membrane of neurons and may also act as neurotransmitters or other signaling molecules [18]. We would like to mention that there exist studies of lipid profiles in schizophrenia using blood samples, such as this [19]. Blood lipidome changes might be a sign of systematic lipidome dysregulation in the body, which might also be noticeable in the brain.

Similar to the fatty acid differences observed in peripheral tissues, several studies have reported alterations in polyunsaturated fatty acids (PUFAs) and PUFA-related mediators in certain brain regions postmortem. Let us elaborate on them.

In the first investigation [9], the authors considered whether there were any changes in PUFAs in the entorhinal cortices of patients with schizophrenia and several other psychiatric diseases compared with unaffected controls. They found no significant differences in the major PUFAs. However, they found 8.7% decrease in docosatetraenoic acid in those with schizophrenia, compared with controls. Docosahexaenoic acid (DHA) is an omega-3 fatty acid that is a primary structural component of the human brain, cerebral cortex, skin, and retina.

In another study [10] fatty acids in the phospholipids of the prefrontal cortex (BA8) were evaluated for patients with schizophrenia, bipolar disorder, or major depressive disorder and compared with unaffected controls. The authors found no significant differences in the levels of PUFAs or other fatty acids in the prefrontal cortex between patients and controls.

# Chapter 2
# Problem statement

In this research, we aimed at identifying what genes and what lipids in what brain regions are most relevant for schizophrenia. We understand relevance here as the importance of certain feature for the ML algorithm that would classify brains or regions as having or not having schizophrenia. We wanted to see whether certain genes or lipids contain most of the necessary signal that distinguishes healthy brains from SZ brains.

**Definition 2.1** *A Brodmann area is a region of the cerebral cortex in the human or other primate brain defined by its cytoarchitecture, or histological structure and organization of cells.*

**Definition 2.2** *Lipids can be defined as hydrophobic or amphipathic small molecules that may originate entirely or in part by carbanion-based condensations of thioesters (fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, saccharolipids, and polyketides) and/or by carbocation-based condensations of isoprene units (prenol lipids and sterol lipids) [6].*

Brodmann areas have been discussed, debated, refined, and renamed exhaustively for nearly a century and remain the most widely known and frequently cited cytoarchitectural organization of the human cortex [20].

The Brodmann classification divides the cortex into approximately 52 areas, numbered sequentially, although some regions have been subsequently subdivided and others are only present in non-human primates.

Let us describe in detail what dataset we are working with. This dataset was created at Skoltech. The brains of four individuals diagnosed with paranoid schizophrenia and four individuals without this diagnosis have been collected postmortem. Bulk RNA-seq for 14179 genes was performed in 35 different regions of each brain. High-performance liquid chromatography coupled with mass spectrometry [16] for 347 lipids was performed on 75 different regions in each brain. Thus, for each brain, we have 75 regions with a lipidomics dataset, and of these 35, we also have a transcriptomics dataset.

Exactly 4 HC and 4 SZ measurements are not for all 75 regions. These regions are Brodmann areas, or certain divisions of them. Also, there are omissions and duplicates. Only 50 regions have exactly 4+4 measurements for lipids. In our research, we omitted 25 regions that do not satisfy this.

Here is the list of all 50 regions for which the full set (4 SZ + 4 healthy) of observations is present:

Cingulate Anterior (BA24), Cingulate Posterior (BA31), Prefrontal Medial (BA10m), Orbitofrontal (BA11), Anterior Middle Temporal (BA37-aMT), Amygdala, Insular Posterior Cortex, Nucleus Accumbens, Putamen, Deep Nuclei - Dentate Nucleus, Piriform Cortex, Pulvinar Thalamus, 2ary/3ary Visual Posterior (BA18/19p), Substantia Nigra, Globus Pallidus, Hippocampus, CA1, Medial Dorsal Thalamus, Orbitofrontal (BA47), Ventral Lateral Thalamus, Internal Capsule, Corpus Callosum Anterior, Corpus Callosum Posterior, Cerebellar White Matter, Cingulate Posterior (BA23a), Dorsolateral Prefrontal (BA46), Cerebellar Grey Matter, Cingulate Subcallosal (BA25), 1ary Visual Posterior (BA17p), 1ary Motor (BA4), 1ary Auditory (BA41/42), 1ary Visual Anterior (BA17a), 1ary Somatosensory (BA3/1/2), Posterior Middle Temporal (BA37-pMT), Anterior Supramarginal (BA40a), Premotor Medial (BA6m), Insular Anterior Cortex, Posterior

Inferior Temporal (BA20p), 2ary Auditory, language (BA22a), Angular (BA39), 2ary Auditory, Wernicke (BA22p), Temporopolar (BA38), Posterior Superior Parietal (BA7p), Anterior Prefrontal (BA10), Premotor Anterior Lateral (BA6a), Supramarginal Posterior (BA40p), Ventrolateral Prefrontal, Broca (BA44), Anterior Inferior Temporal (BA20a), Dorsolateral Prefrontal (BA9), Anterior Superior Parietal (BA5/BA7a), Frontopolar (BA10fp), FEF Medial (BA8m), 2ary Auditory Anterior (BA21a).

Here is the table of the lipid classes that were present in the dataset 2.1.

The name of the lipid molecule is formed by its class, the number of carbon atoms, and the number of double chemical bonds. For example, PE 38:7 means phosphatidylethanolamine with 38 carbon atoms and 7 double chemical bonds.

Table 2.1: Lipid Classes

| Abbreviation | Full name |
|---|---|
| Phosphatidylcholines | PC |
| Plasmenyl phosphatidylethanolamines | PE P |
| Ceramides | Cer |
| Phosphatidylethanolamines | PE |
| Free fatty acids | FA |
| Sphingomyelins | SM |
| Hexosylceramides | HexCer |
| Triacylglicerides | TG |
| Plasmanyl phosphatidylcholines | PC 0 |
| Plasmenyl phosphatidylcholines | PC P |
| Lipophosphatidylethanolamines | LPE |
| Lipophosphatidylcholines | LPC |
| Diacylglicerides | DG |
| Sulfohexosylceramides | SulfoHexCer |
| Phosphatidylglycerols | PG |
| Plasmanyl phosphatidylethanolamines | PE 0 |
| Cholesterol esthers | CE |
| Phosphatidylinositols | PI |
| Acylcarnitines | CAR |
| Phosphatidylserines | PS |
| Cholesterol | Cholesterol |

# Chapter 3
# Methodology

## 3.1 Data Visualization

Since the number of subjects (people) is significantly smaller than the number of features (gene expressions and lipid levels), we use dimensionality reduction methods such as PCA [7] for visually observing how well different classes are separated. PCA is good for checking if the batch effect is present. Beyond PCA, we used nonlinear methods for obtaining visualizations that better depict the internal structure of the dataset.

## 3.2 Data Normalization

For different samples to be comparable, the data has to be normalized properly so that nonbiological factors have a minimal effect on the signal. Here we do the following: for each brain, we make the sum of all observations for each molecule equal to zero by subtracting the mean value. We chose this kind of normalization because HC individuals were processed separately from SZ individuals. Normalization over the whole brain potentially removes the batch effect that could arise from that.

## 3.3 Data Augmentation

Notice that for each region and each gene, we have 4 healthy data points and 4 SZ points (same for lipids). We consider these points to come from some probability distributions. We assume that they are normal because they are bell-shaped and because we used the D'Agostino and Pearson's statistical test for normality with Bonferroni correction and no lipids showed a p-value lower than 0,05. We generate synthetic data based on these distributions.

Machine learning algorithms (and especially deeplearning) usually require many training samples; therefore, we focus on data augmentation by generating synthetic data and obtaining external data similar to ours.

We found external data for a transcriptomic dataset in the refine.bio database [8]. We searched using the following keywords: schizophrenia, brain regions, post-mortem, Brodmann area, and psychosis. We found five datasets with the same structure as ours (i.e., gene expression measured post-mortem for some brain regions in schizophrenic and healthy individuals). However, only one of them uses the same technology that was used in our dataset: RNA-seq. We merge this external dataset with our dataset and perform batch effect removal.

## 3.4 Training Dataset Preparation

Several approaches are possible when creating the training dataset for the model. We are using them both.

Firstly, we can treat each brain as an object, and it will correspond to one row in the training dataset. All measures in different regions are combined into one vector. Then each feature is the

level of gene expression or lipid within the specific region. Here, the model is answering the question, ”Is this a schizophrenic brain?” Within this approach, we can judge whether the region is relevant to schizophrenia based on the sum of the features that belong to it.

The motivation for treating the brains as objects is that we do not know exactly which parts of the brain are mostly affected by the illness, and so we need all of the available information about the brain for the classification. However, in this approach, the connection between the same molecules in different regions is lost, because they become different columns in the training table.

Secondly, we can treat each region as an object, and it will correspond to one row in the training dataset. In this case, the model is answering a question, ”Does this region belong to a schizophrenic brain?” Now we can judge whether the region is relevant to schizophrenia based on how well different regions get classified. We expect that for some regions, the accuracy of prediction is better than for others.

The motivation for treating the regions as objects is that here we have fewer features, and the connection between the measurements of the same molecule in different regions is not lost (they become one column in the table). However, here we lose information about the location of the region, and that some regions come from the same brain.

We try both approaches to see how close and stable the results they produce are.

## 3.5 Importance Estimation

In this research, we perform a classification using ML and then estimate the impact that different features have. This is what we mean by the relevance of the molecule to schizophrenia. One can estimate the importances of features in machine learning in several ways.

The relevance of the region to schizophrenia can be computed in two ways. Firstly, if we treat each region as an object, then we can compute the classification accuracy separately for each type of region. We assume that the higher the accuracy, the higher the relevance, because if it is easy for the model to classify the region, then it is significantly different between healthy and schizophrenic patients. Secondly, if we treat each brain as an object, we can sum up all the molecule importances that correspond to the given type of region and obtain the importance of the region.

### 3.5.1 Random Forest

If we use the random forest algorithm [11] for the classification problem, we can compute the impurity-based feature importances. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as Gini importance.

### 3.5.2 Permutational Importances

Secondly, there exists a permutation importance approach for feature evaluation. It is agnostic to what classification algorithm is used. Here, a baseline metric, defined by scoring, is evaluated on a validation dataset. Further, a feature column from the validation set is permuted, and the metric is evaluated again. The permutation importance is defined as the difference between the baseline metric and the metric from permuting the feature column. For better estimation, the permutation is performed multiple times, and the mean value is taken.

As a baseline algorithm, we use logistic regression and support vector classification with a linear kernel.

# Chapter 4
# Numerical experiments

## 4.1 The Tools

In this research, we use the Python programming language and such classical data analysis libraries and tools as NumPy, Pandas, scikit-learn, matplotlib and Jupyter Notebook. We use the R package limma for batch effect removal. The computations were performed on the Skoltech Arkuda supercomputer. The code is available on github.

## 4.2 Data Augmentation

We assume that the distributions of all lipids and genes in all regions are normal and independent of each other. We estimate the mean values and the variances of them. Based on these estimations and assumptions, we generate $N = 10000$ individuals for lipids for both classes (HC and SZ), resulting in $20000$ individuals. In the case of treating regions as objects, this means $320$ feature columns and $20000 \times 50 = 100000$ rows. In the case of treating brains as objects, this means $319 \times 50 = 15900$ feature columns and $20000$ rows.

As for the genes, we generate $N = 1000$ individuals for genes for both classes, resulting in $2000$ individuals. In the case of treating regions as objects, this means $14177$ feature columns and $2000 \times 35 = 70000$ rows. In the case of treating brains as objects, this means $14177 \times 35 = 496195$ feature columns and $20000$ rows. Here we generated fewer individuals due to the limitations of the computational resources (otherwise the dataset would not fit into the memory).

## 4.3 Illustrations

First, we have done a dimensionality reduction using PCA for each brain region for genetic data, considering genes as features and people as objects. Here we provide a scatter plot for the cerebellar white matter as an example 4.1. For all regions, including this one, people with schizophrenia and healthy people are linearly separable.

See this illustration of all regions in one plot (region as object) 4.2.

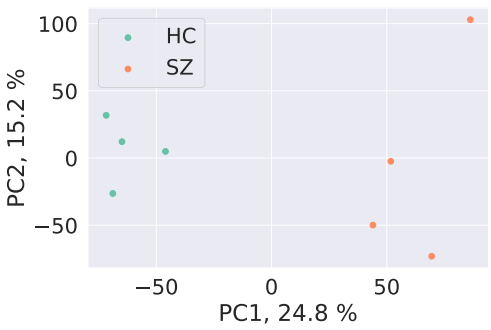


Figure 4.1: PCA for transcriptomics data (brain as object)

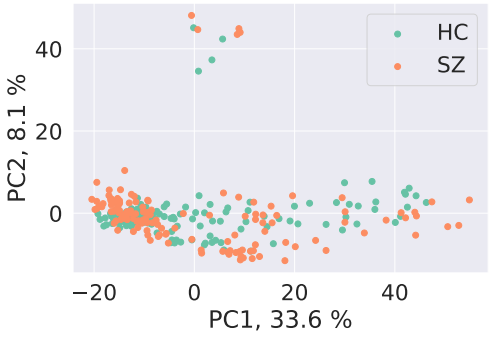


Figure 4.2: PCA for transcriptomics data (region as object)

Let us describe in detail the external dataset we used [17]. The authors performed RNA sequencing on tissue from the anterior cingulate cortex, dorsolateral prefrontal cortex, and nucleus accumbens from three groups of 24 patients, each diagnosed with schizophrenia, bipolar disorder, or major depressive disorder, and from 24 control subjects. The most significant disease-related differences were in the anterior cingulate cortex of schizophrenia samples compared to controls. Transcriptional changes were assessed in an independent cohort, revealing the transcription factor EGR1 as significantly down-regulated in both cohorts and as a potential regulator of broader transcription changes observed in schizophrenia patients. Additionally, broad down-regulation of genes specific to neurons and concordant up-regulation of genes specific to astrocytes was observed in schizophrenia and bipolar disorder patients relative to controls.

Here we provide the visualization of our data combined with the external data for Nucleus Accumbens (4.3 and 4.4). This data is used for classification and importance estimation.

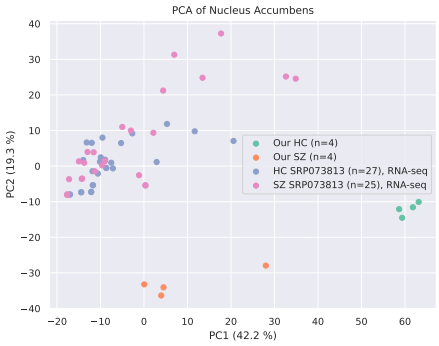


Figure 4.3: PCA for Nucleus Accumbens, transcriptomic data, before BE removal

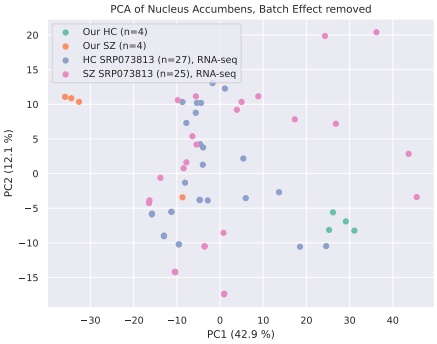


Figure 4.4: PCA for Nucleus Accumbens, transcriptomic data, BE removed

We have drawn the distributions of the lipids based on another dataset provided by Skoltech with lipid measurements in 12 healthy and 12 schizophrenic individuals; the number of regions is 4, and the number of molecules is 445. We believe it is possible to assume that the distributions are normal because we performed the Shapiro-Wilk test for normality with Bonferroni correction and p-values smaller than 0.05 did not appear anywhere (see 4.5).

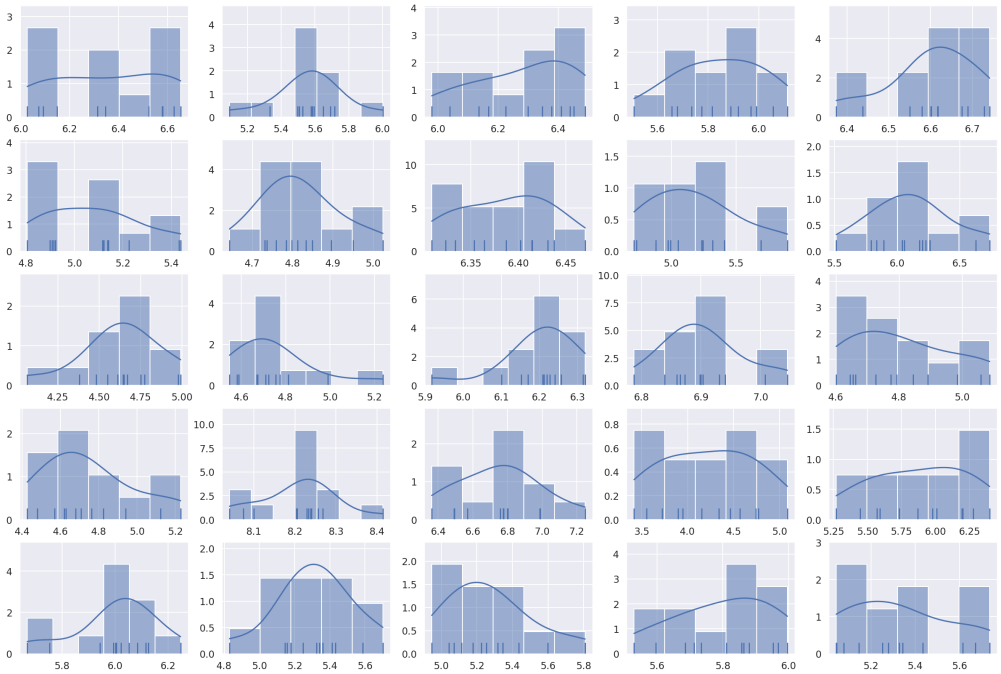

Figure 4.5: Distributions of lipid content

We have performed the data normalization procedure for all the necessary datasets: a given lipidomics dataset, a given transcriptomics dataset, a synthetically generated lipidomics dataset, and a synthetically generated transcriptomics dataset. Namely, for each brain and each lipid, we calculated the mean value over the regions and subtracted it so that the overall sum is equal to zero. See the normalized lipid profiles over 50 regions 4.6.

See the lipidomics dataset, visualized in two approaches: region as object and brain as object 4.7 4.8.


Figure 4.6: Fatty Acid profiles in one of the brains

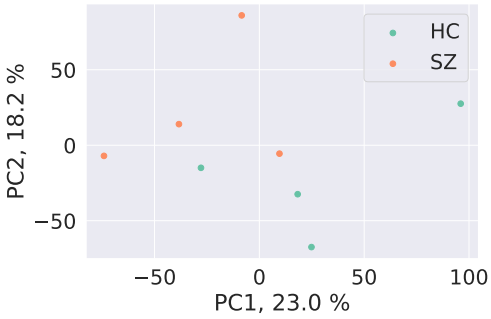


Figure 4.7: PCA for brains as objects, real lipidomics data

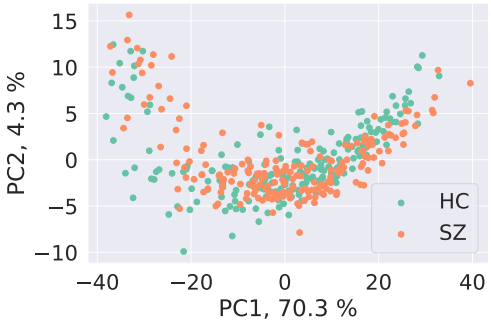


Figure 4.8: PCA for regions as objects, real lipidomics data


Figure 4.9: PCA for regions as objects, generated lipidomics data

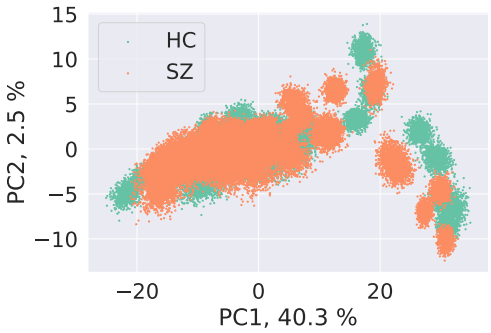


Figure 4.10: Isomap for regions as objects, generated lipidomics data

We also generated the synthetic data while treating each brain as an object 4.11.

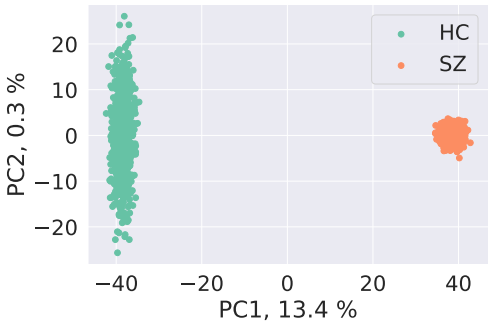


Figure 4.11: PCA for brains as objects, generated lipidomics data

Mean 5-fold CV score for the synthetic dataset based on lipids where we treat regions as objects is **0.942** using random forest algorithm.

## 4.4 Tables of Importances

### 4.4.1 Random Forest

When we apply the random forest algorithm to generated lipidomics data, treating each brain as an object, we obtain the importances of each lipid in each region and the importances of each region as a sum of lipids within it. Of 15900 features, for 15492 the importance is zero. Here we provide top-10 lipids 4.7 and top-10 regions 4.2.

When we apply the random forest algorithm to generated lipidomics data, treating each **region** as an object, we obtain the importances of each lipid, and the importances of each region as an accuracy. Here we provide them 4.3. As for the accuracy, for most regions out of 50, it is equal to 1.0. This means that this value is not suitable for drawing strong conclusions. Here is the plot of the distribution of the accuracies 4.12.

When we apply the random forest algorithm to generated transcriptomic data, treating each brain as an object, we obtain the importances of each gene in each region and the importances of each region as a sum of genes within it. Of 496195 features, 496093 have zero importance (which is 99.9%). See the tables 4.4, 4.5:

Table 4.1: Random Forest importances of each lipid in each region, brain as object

| Brain region | Lipid | importance |
|---|---|---|
| 1ary Visual Posterior (BA17p) | PC 29:0 | 0.031105 |
| Globus Pallidus | HexCer t41:2 | 0.030107 |
| 1ary Motor (BA4) | PC 38:5 | 0.030059 |
| Globus Pallidus | PE 35:2 | 0.028034 |
| Dorsolateral Prefrontal (BA9) | PC_O 40:4 | 0.027058 |
| Cingulate Anterior (BA24) | Cer d42:2 | 0.026909 |
| Corpus Callosum Posterior | SM d36:1 | 0.026368 |
| Anterior Inferior Temporal (BA20a) | PC_P 38:2 | 0.025281 |
| 2ary Auditory, Wernicke (BA22p) | Cer d36:2 | 0.022108 |
| Nucleus Accumbens | CE 18:2 | 0.021994 |

Table 4.2: Random Forest importances of each region, brain as object

| Brain region | Importance |
|---|---|
| 2ary Auditory, Wernicke (BA22p) | 0.158412 |
| 1ary Visual Posterior (BA17p) | 0.131750 |
| Insular Posterior Cortex | 0.093228 |
| Globus Pallidus | 0.078987 |
| Cingulate Anterior (BA24) | 0.060240 |
| Internal Capsule | 0.055661 |
| Angular (BA39) | 0.049371 |
| Corpus Callosum Posterior | 0.042819 |
| Dorsolateral Prefrontal (BA9) | 0.041895 |
| Posterior Inferior Temporal (BA20p) | 0.038030 |

Table 4.3: Random Forest importances of each lipid in each region, region as object

| Lipid | Importance |
|---|---|
| LPE 22:4 | 0.075944 |
| LPE 18:1 | 0.052488 |
| LPE 20:2 | 0.037824 |
| LPE 20:4 | 0.027195 |
| PC_P 38:2 | 0.019500 |
| DG 38:1 | 0.017551 |
| LPE 20:1 | 0.017130 |
| PC_P 38:6 | 0.016511 |
| TG 52:2 | 0.015552 |
| TG 50:3 | 0.015059 |

### 4.4.2 Permutational Importances

When we apply logistic regression or random forest to the lipidomics data, treating brains as objects, all the permutational importances are zero. When we apply logistic regression to the lipidomics data, treating regions as objects, all the permutational importances are also zero. But when we apply SVM to the lipidomics data, treating regions as objects, the permutational importances are not zero: see 4.6.

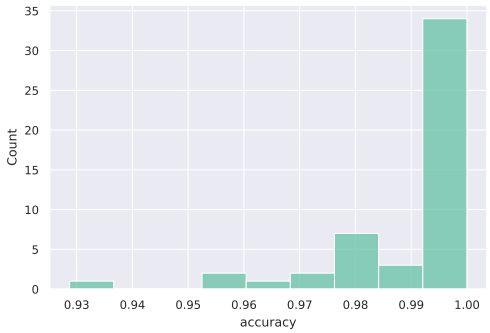


Figure 4.12: Distribution of the prediction accuracy for the regions

Table 4.4: Random Forest importances of each gene in each region, brain as object, value distribution

| Importance | Number of Features with that importance |
|---|---|
| 0.0 | 496093 |
| between 0.0 and 0.1 | 16 |
| 0.1 | 80 |
| 0.2 | 6 |

Table 4.5: Random Forest importances of each region, brain as object

| Brain region | Importance |
|---|---|
| 1ary Auditory (BA41/42) | 0.149265 |
| 1ary Somatosensory (BA3/1/2) | 0.110000 |
| Cingulate Posterior (BA31) | 0.110000 |
| Substantia Nigra | 0.080000 |
| Cerebellar White Matter | 0.068574 |
| Nucleus Accumbens | 0.050000 |
| Cingulate Anterior (BA24) | 0.050000 |
| Orbitofrontal (BA11) | 0.040000 |
| 2ary Auditory, Wernicke (BA22p) | 0.040000 |
| Putamen | 0.031179 |

Table 4.6: SVM permutational importances of each lipid, region as object

| Lipid | Feature Importance |
|---|---|
| PE 38:7 | 0.008274 |
| PC 40:8 | 0.006758 |
| PC_O 38:6 | 0.005336 |
| PE 34:3 | 0.005198 |
| Cer m36:1 | 0.005176 |
| PE_O 40:4 | 0.005148 |
| PC 46:2 | 0.004512 |
| PE 36:6 | 0.004136 |
| PC 40:1 | 0.003942 |
| LPC 20:3 | 0.003708 |

In order to examine the stability of the results, we launched the same experiments twice. Here we provide a scatter plot of the random forest importances obtained from 2 launches for regions 4.13, 4.14.

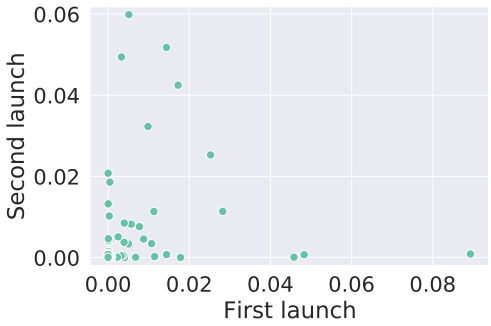


Figure 4.13: Lipidomics data, brain treated as object, importances of regions (transformed by $\log 10(x+1)$)

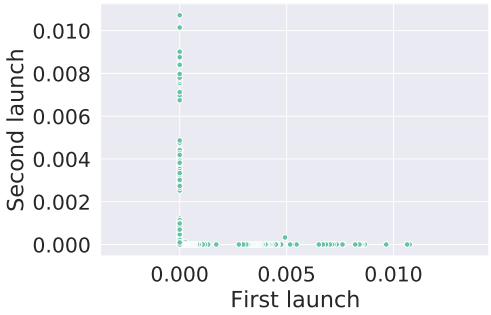


Figure 4.14: Lipidomics data, brain treated as object, importances of molecules in regions (transformed by $\log 10(x+1)$)

Table 4.7: Random Forest importances of each lipid in each region, brain as object

| Molecule class | What is an object | Method | Summary |
|---|---|---|---|
| Lipids | Brain | Random Forest | 97.4% is zero |
| Lipids | Brain | Permutations | Both logreg and SVM give all zeros |
| Lipids | Region | Random Forest | all nonzeros |
| Lipids | Region | Permutations | logreg gives all zeros, SVM gives nonzeros |
| Genes | Brain | Random Forest | 99.9% is zero |
| Genes | Brain | Permutations | Takes too long to compute |
| Genes | Region | Random Forest | 36.9% nonzeros |
| Genes | Region | Permutations | Takes too long to compute |

# Chapter 5
# Discussion and Conclusion

We have augmented the transcriptomic dataset with an external dataset and removed the batch effect. We have augmented our lipidomics dataset using the generated data with estimated empirical distributions. Using the generated data, we have identified the most important lipids, regions, and genes using feature importances estimation by the random forest algorithm, and the permutation approach. The two ways of treating the data—region as object and brain as object—were used.

Using our methods (random forest and permutation on logistic regression and SVM), we obtained the lists of importances, which means the goal of this research was accomplished. The main limitation of the datasets that we are using is the lack of samples, which was overcome by data augmentation via estimation of the distributions and sampling from them. In this research, we used some of the well-known ML algorithms. However, it is possible that some other algorithms that we have not checked would give better results.

Molecular changes in the schizophrenia brain are complex and subtle, which is not the case for many other brain disorders (for example, in the case of Alzheimer's or Parkinson's, the changes are dramatic and noticeable). This can explain why the results we obtain are not completely stable and coherent. As far as we know, importance estimation of lipids, regions, and genes using ML classification is a novel approach.

# Acknowledgements

We would like to thank Maria Osetrova, research engineer at Skoltech Center of Life Sciences, for providing the dataset and explaining its structure.

We would like to thank Skoltech for providing computational resources on the Arkuda supercomputer.

# Bibliography


[1] Barnes, M. R., Huxley-Jones, J., Maycox, P. R., Lennon, M., Thornber, A., Kelly, F., Bates, S., Taylor, A., Reid, J., Jones, N., et al. Transcription and pathway analysis of the superior temporal cortex and anterior prefrontal cortex in schizophrenia. *Journal of neuroscience research 89*, 8 (2011), 1218–1227.

[2] Chen, J., Wu, J.-s., Mize, T., Shui, D., and Chen, X. Prediction of schizophrenia diagnosis by integration of genetically correlated conditions and traits. *Journal of Neuroimmune Pharmacology 13*, 4 (2018), 532–540.

[3] Cortes-Briones, J. A., Tapia-Rivas, N. I., D'Souza, D. C., and Estevez, P. A. Going deep into schizophrenia with artificial intelligence. *Schizophrenia Research 245* (2022), 122–140. Computational Approaches to Understanding Psychosis.

[4] Dufva, M. Introduction to microarray technology. *DNA Microarrays for Biomedical Research: Methods and Protocols* (2009), 1–22.

[5] et al, D. W. Comprehensive functional genomic resource and integrative model for the human brain. *Science 362*, 6420 (2018), eaat8464.

[6] Fahy, E., Subramaniam, S., Murphy, R. C., Nishijima, M., Raetz, C. R., Shimizu, T., Spener, F., van Meer, G., Wakelam, M. J., and Dennis, E. A. Update of the lipid maps comprehensive classification system for lipids1. *Journal of lipid research 50* (2009), S9–S14.

[7] F.R.S., K. P. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2*, 11 (1901), 559–572.

[8] Greene, C. S., Hu, D., Jones, R. W., Liu, S., Mejia, D. S., Patro, R., Piccolo, S. R., Romero, A. R., Sarkar, H., Savonen, C. L., et al. refine. bio: A resource of uniformly processed publicly available gene expression datasets. *Google Scholar*.

[9] Hamazaki, K., Hamazaki, T., and Inadera, H. Abnormalities in the fatty acid composition of the postmortem entorhinal cortex of patients with schizophrenia, bipolar disorder, and major depressive disorder. *Psychiatry Research 210*, 1 (2013), 346–350.

[10] Hamazaki, K., Maekawa, M., Toyota, T., Dean, B., Hamazaki, T., and Yoshikawa, T. Fatty acid composition of the postmortem prefrontal cortex of patients with schizophrenia, bipolar disorder, and major depressive disorder. *Psychiatry Research 227*, 2 (2015), 353–359.

[11] Ho, T. K. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (1995), vol. 1, IEEE, pp. 278–282.

[12] Lanz, T. A., Joshi, J. J., Reinhart, V., Johnson, K., Grantham II, L. E., and Volfson, D. Step levels are unchanged in pre-frontal cortex and associative striatum in post-mortem human brain samples from subjects with schizophrenia, bipolar disorder and major depressive disorder. *PloS one 10*, 3 (2015), e0121744.

[13] Maas, D. A., Martens, M. B., Priovoulos, N., Zuure, W. A., Homberg, J. R., Nait-Oumesmar, B., and Martens, G. J. Key role for lipids in cognitive symptoms of schizophrenia. *Translational psychiatry 10*, 1 (2020), 1–12.

[14] Maycox, P. R., Kelly, F., Taylor, A., Bates, S., Reid, J., Logendra, R., Barnes, M. R., Larminie, C., Jones, N., Lennon, M., et al. Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function. *Molecular psychiatry 14*, 12 (2009), 1083–1094.

[15] Narayan, S., Tang, B., Head, S. R., Gilmartin, T. J., Sutcliffe, J. G., Dean, B., and Thomas, E. A. Molecular profiles of schizophrenia in the cns at different stages of illness. *Brain research 1239* (2008), 235–248.

[16] Pitt, J. J. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *The Clinical Biochemist Reviews 30*, 1 (2009), 19.

[17] Ramaker, R. C., Bowling, K. M., Lasseigne, B. N., Hagenauer, M. H., Hardigan, A. A., Davis, N. S., Gertz, J., Cartagena, P. M., Walsh, D. M., Vawter, M. P., et al. Post-mortem molecular profiling of three psychiatric disorders. *Genome medicine 9* (2017), 1–12.

[18] Schneider, M., Levant, B., Reichel, M., Gulbins, E., Kornhuber, J., and Müller, C. P. Lipids in psychiatric disorders and preventive medicine. *Neuroscience and Biobehavioral Reviews 76* (2017), 336–362. SI:IBNS-2015.

[19] Solberg, D. K., Bentsen, H., Refsum, H., and Andreassen, O. A. Lipid profiles in schizophrenia associated with clinical traits: a five year follow-up study. *BMC psychiatry 16*, 1 (2016), 1–9.

[20] Strotzer, M. One century of brain mapping using brodmann areas. *Clinical Neuroradiology 19*, 3 (2009), 179.

[21] Wang, Z., Gerstein, M., and Snyder, M. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics 10*, 1 (2009), 57–63.