ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Программная инженерия»

УДК 004.942, 515.1

СОГЛАСОВАНО
Руководитель ВКР,
доцент департамента больших данных и информационного поиска факультета компьютерных наук,
канд. ф.-м. наук

__________________ В.Л. Чернышев
«29» мая 2023 г.

УТВЕРЖДАЮ
Академический руководитель образовательной программы «Программная инженерия»,

профессор департамента программной инженерии, канд. техн. наук

__________________ В. В. Шилов

« 31 » мая 2023 г.

Выпускная квалификационная работа
(академическая)

на тему: Топологический анализ мультимодальных медицинских данных
(Topological Analysis of Multimodal Medical Data)

по направлению подготовки 09.03.04 «Программная инженерия»

ВЫПОЛНИЛА
студентка группы БПИ193
образовательной программы
09.03.04 «Программная инженерия»
К.А. Шилова____________________
И.О. Фамилия

К. Шилова 28.05.2023
Подпись, Дата

Москва 2023

Abstract

The purpose of this work was to implement Topological Data Analysis (TDA) on functional Magnetic Resonance Imaging (fMRI) and gene expression data. The mentioned types of medical data could be represented as networks, which describe interactions between brain regions and genes respectively. The main idea is to use these networks as simplicial complexes, which could be studied with the persistent homologies tool. The results presented in this work illustrate that topological features of the networks, which are based on data associated with Alzheimer's disease, markedly differ from control samples. It is also shown that dimensionality reduction methods applied to gene expression data preserve the difference and make working with data more convenient. It is worth mentioning that the ADNI (Alzheimer's Disease Neuroimaging Initiative) dataset is used to get fMRI data, which are presented as imaging files, and gene expression profiles. Additionally, the CNI challenge (2019) and the OASIS-3 datasets are used to get preprocessed time series fMRI data.

The work contains 29 pages, 24 figures, 6 tables, 45 sources.

Keywords – TDA (Topological Data Analysis), persistent homology, GRN (Gene Regulatory Network), gene expression, fMRI (functional Magnetic Resonance Imaging), Alzheimer's disease.

Main terms and acronyms

Alzheimer's disease (AD) is a neurodegenerative disease, the most common type of dementia.

Late mild cognitive impairment (LMCI) is the stage between the expected decline in memory, which is related to aging, and the abnormal decline of dementia.

Functional magnetic resonance imaging (fMRI) is the way to measure brain activity by detecting changes associated with blood flow. The blood flow depends on brain regions usage, that is, a more intense flow means that a brain region is in use.

Gene expression is the process of using gene information in producing proteins or non-coding RNA, which affect a phenotype.

High-order Gene Regulatory Network (GRN) is a network, which consists of nodes represented by genes, and edges, the weights of which are the strengths of interaction between genes.

Topological Data Analysis (TDA) is the way of analysis of datasets using techniques from topology. In this work, the persistent homologies are a key tool for the analysis.

Contents

Introduction

This work focuses on the topological data analysis (TDA) applied to medical data, which are associated with patients with Alzheimer's disease and healthy individuals. Overall, TDA applied to fMRI data is quite an effective and developing approach. Computational resources are effectively used to explore dynamic processes of brain networks. As a result, network neuroscience continues to develop because it is very correlated with modern trends in data analysis. Brain structures represented as networks are strongly influenced by methods of graph construction, such as thresholding and edge definition. Consequently, there are a wide variety of graph construction approaches, and it could be a foundation for future computational experiments, which will extend the results presented in this work. However, the methods provided in this work are classical, and correlation coefficients between measured brain regions' activity without neglecting any values, which are close to zero, are considered strengths of brain region interaction. Overall, one of the purposes of this work is to use TDA applied to fMRI data to investigate the impact of Alzheimer's disease on brain region interactions.

In addition, neurological diseases are highly influenced by genetic factors. The link between gene expression and Alzheimer's disease is the cause of the obvious importance of genetic data analysis. The interactions between genes based on highly correlated expression values could be represented as a high-order gene regulatory network (GRN). The high-order GRN (hereafter referred to as GRN) is derived from normal samples or samples associated with a disease, and distances between genes are based on correlation coefficients between their expression. The TDA-based approach of GRNs studying recently was applied to cancer-specific genomes, and an essential difference between patients' and controls' data was presented ([1]). Consequently, the second purpose of the work was to reproduce a similar approach, where topological features of a gene expression network are studied using persistent homology theory.

It is considered that genetic data is noisy, and gene expression datasets are also high-dimensional [2]. Thus, this work also provides the outcomes of TDA after implementing dimensionality reduction of the gene expression dataset. A new approach, called the Topological Autoencoders technique, is used to implement topology-preserving dimensionality reduction. Concerning fMRI data, it is not needed to use dimensionality reduction approaches, because the data contains brain activity measurements averaged over large-scale regions.

Finally, another problem to which this work is devoted is the lack of studies, which concern the topological analysis of both mentioned types of medical data. Therefore, it is possible that TDA of brain imaging and genetic data has similar problems and advantages. It might be discovered with a deep analysis of such a joint approach.

Section 1. Literature review

Topological Data Analysis is a widespread approach applied to biosciences. The key advantage of the persistent homologies tool, which is implemented to study biological data, is providing a comprehensive summary of topological features of the studied data. For instance, the paper [3] offers an overview of biological TDA applications, which include the exploration of diseases as well. It is also noted that TDA is an efficient tool for genomics and even evolutionary biology. In general, TDA is widely used for the data of biological nature analysis. This fact could be proved by a significant number of papers in biology and medicine where TDA applications are considered (Fig.1).

Overall, there are a huge number of studies, where the foundations of network neuroscience together with TDA are discussed. In [4], the general advantages of approaches of network neuroscience are shown. In [5], authors focus on the strengths of the persistent homology tool applied to network neuroscience. Overall, [5] shows applications of topology to structural networks, namely, structural connectomes. However, TDA is mostly applied to functional brain networks.

The first type of data, which is studied in this work, is fMRI data. Functional Magnetic Resonance Imaging, or fMRI, is a technique of brain imaging, which identify the brain regions' activity over measuring time. The measurements are based on the blood flow changes. A functional connectome (a network) presents the strengths of brain regions' interaction, and this network is based on raw neuroimaging files. The raw files are used in this work as input files from one of the considered datasets. As a result, some preprocessing steps help to get functional brain activity data in a format of time series for each region. The time series are used to get a network, that is a functional connectome. As shown in [6], the final functional connectome is sensitive to a method of graph construction. However, in this work, standard techniques are used, which could lead to further experiments with the other ones. In general, the difference between topological features of functional brain networks may reflect whether or not there is a neurological disease. It is this approach that makes it possible to analyze AD patients' data in this work. However, TDA could be also used for broader tasks.

The analysis of the functional Magnetic Resonance Imaging data with TDA is presented in [7], [8], and [9]. Additionally, there are also a variety of approaches to analyzing and comparing persistent homologies represented as persistence diagrams or Betti curves. In [8], the

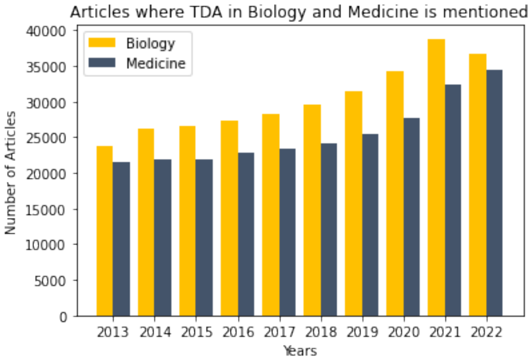


Fig. 1: Number of papers, where TDA in biosciences is mentioned (Google Scholar, November 16, 2022).

cognitive impairments are investigated in terms of distinguishing topological features. The key result related to the persistent homology tool, which [9] presents, includes the fact that the early stage of cognitive impairment disease could be poorly seen by TDA. Early stages affect specific brain regions, and the influence is not propagated on large-scale connections. In this work, early stages together with late ones are considered, and it is illustrated below that the obtained results correspond to the paper [9].

Concerning genetic data, Alzheimer's Disease could be studied with the help of gene regulatory networks [10]. Genes are involved in a complex process of producing proteins, which influence a phenotype. Expression of a gene is a key aspect of phenotype construction, and expression values also could be measured numerically. The access to patients' and controls' expressions data gives the possibility to study metric spaces based on the data of selected risky genes (for example, relevant genes are described in [11]). Thus, in [1], the aforementioned approach (but concerning cancer genomes) leads to significant results. First of all, correlation coefficients between gene expression values and persistence diagrams were analyzed. Following this, the persistence diagrams related to a dataset of patients' genetic information and healthy gene expression profiles were compared. The results from [1] show the difference between GRNs of cancer genomes and normal ones. Therefore, this work will implement a similar approach of using TDA applied to datasets, which are associated with Alzheimer's disease. It should be noted that AD is highly influenced by gene expression and its regulation, which is deeply shown in [12] and [13].

Often genetic data needs additional processing such as dimensionality reduction. There is a significant number of genes and subjects in the gene expression dataset, an intrinsic dimension of which could be reduced by different techniques. First of all, approaches are divided into local and global ones, which tend to preserve local or global structures respectively. For instance, PCA (Principal component analysis) [14] mostly preserves a global structure, while UMAP [15] and t-SNE [16] focus on preserving local neighborhood structure at the cost of destroying the global one. However, there is also a trade-off between mentioned types of approaches. Methods based on topological data analysis could preserve topological similarity between datasets while reducing dimensionality. The most significant TDA-based methods of dimensionality reduction are Topological Autoencoders, also known as TopoAE [17], and Representation Topology Divergence Autoencoders, called RTD-AE [18]. In [18], the comparison of global, local, and TDA-based methods is provided, and TDA-based methods seem to be more effective in general. In this work TopoAE method applied to the gene expression dataset is compared with some local and global techniques. However, the brand-new RTD-AE will be considered as part of the future extension of this work. It has not been implemented yet due to the lack of computational resources.

Concerning joint functional brain networks and genomes analysis, there are remarkable works where both data types are studied. By way of example, [19] investigates single nucleotide polymorphisms (SNPs) and their relation to brain functional data. Schizophrenia (studied in [19]) is also under the impact of genetic factors, and the paper gives information about the link between fMRI data and single nucleotide polymorphisms, which influence the disease-related phenotype. However, there is no network-based approach. In [20] and [21], the associations between neuroimaging phenotypes and SNPs are provided, and these works focus on the explanation of behavioral phenotypes. In addition, [20] and [21] offer the UK Biobank dataset, which includes both neuroimaging and genetic information. By way of another example, [22] (Brain Genomics Superstruct Project) gives some insights into the association between genetic information and behavioral traits of healthy people. Although there are works where both types of medical information are considered, the study which applies a TDA-based approach for multimodal medical data is still to be introduced. This work gives the comparison of TDA results related to both types of biological data from chosen datasets. In more detail, the description of the datasets alongside the steps of the research are outlined in the next section.

Section 2. Methodology

The main steps of persistent homologies analysis are given in Fig.2. Firstly, it is needed to load all relevant data, which contain fMRI and gene expression datasets. Then, networks, which reflected the data, must be built and processed by the persistent homologies tool. Following this, the difference between patients and a control group might be observed on persistence diagrams. Finally, statistically proven results must be presented together with conclusions. This section focuses on the datasets and methods, which are used in this work. Additionally, it provides the mathematical background of the persistent homologies tool and a brief description of studied networks.

2.1 Datasets description

To investigate fMRI and genetic data together, it could be helpful to use a dataset that contains both types of data for the same sample of subjects. The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [24] provides both neuroimaging and gene expression information. The ADNI dataset also gives LMCI patients' data (with less crucial changes in comparison with AD ones), and this leads to further comparison presented below. This dataset offers the normalized and well-prepared for networks construction gene expression data. However, it provides neuroimaging data, which should be significantly preprocessed.

With regards to gene expression data, expression profiling is implemented with the help of The Affymetrix Human Genome U219 Array, the modern commonly used microarray platform [25]. It is used to get raw expression values extracted from a cell's RNA fragments. Then raw values are preprocessed and normalized using the RMA (Robust Multi-chip Average) method [26]. Firstly, the RMA method focuses on background correction. In the following steps, quantile normalization and logarithmic scale transformation are performed. The quantile normalization equalizes the distribution of probe intensities among microarrays, while the log transformation step stabilizes a variance of intensities. It is followed by summarizing and averaging the probe intensities corresponding to the same gene. Genetic data from the ADNI dataset were additionally verified, using a comparison of results with gender-specific gene expression values and with SNPs (single nucleotide polymorphisms). Questionable subjects were removed from the dataset.

As for fMRI data, raw files with imaging data are presented as subject-related arrays of three-dimensional slices in a DCM format (brain volumes). These arrays could be considered as time-varying brain states, while the time of measurement repetitions is 3000 ms. Overall, there are 140 or 280 timesteps for each subject. A type of neuroimaging data is resting-state fMRI, which means that brain activity was measured in the absence of specific sensory or cognitive tasks. Overall, there are 46 subjects, namely, 37 controls and 9 patients. Because of the format of raw neuroimaging files, the construction of a graph from fMRI data needs

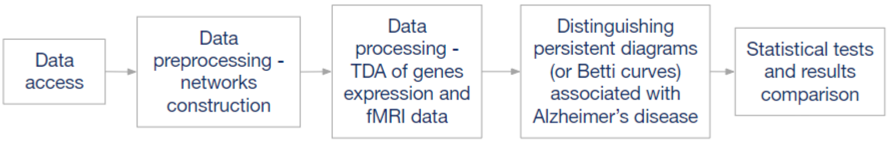

Data access
Data preprocessing - networks construction
Data processing - TDA of genes expression and fMRI data
Distinguishing persistent diagrams (or Betti curves) associated with Alzheimer's disease
Statistical tests and results comparison

Fig. 2: The pipeline of fMRI and gene expression data analysis.

some preprocessing steps, such as spatial normalization, and segmentation according to brain regions. The mentioned steps are described in the next subsection (2.2) in detail.

Additionally, there are two extra datasets with more balanced and comprehensive data, called CNI challenge datasets [27] and OASIS-3 dataset [28]. Both of these datasets contain time series data of brain regions' activity. The CNI challenge dataset contains data that is related to ADHD (Attention-deficit/hyperactivity disorder) patients, while the OASIS-3 provides AD patients' data as well as the ADNI dataset.

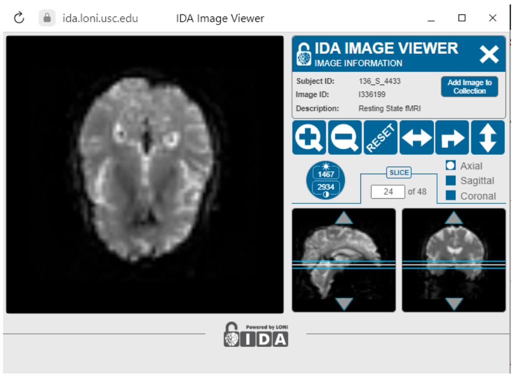
The CNI Challenge data contains 120 patients and 120 subjects, which are divided into training and validation categories. The data are presented as time series tables for each subject in the dataset. Moreover, the CNI Challenge provides segmentation, which corresponds to three frequently used brain atlases: the AAL atlas [29], the Harvard-Oxford atlas [30], the Craddock 200 atlas [31]. The repetition time is equal to 2500 ms, and each table consists of 156 or 128 samples. The CNI Challenge data does not provide any genetic information. Consequently, the CNI data is used with the intention to more comprehensively test a piece of code written for TDA of fMRI data.

Finally, from the OASIS-3 datasets 400 subjects data, equally divided into patients and controls, are used. Data were preprocessed using the fMRIPrep method [32], and final time series tables also provide AAL atlas segmentation [29]. Patients from the OASIS-3 were diagnosed with mild to moderate Alzheimer's disease, which makes the comparison between the ADNI, which includes different stages of AD, and the OASIS-3 datasets possible and even reasonable.

2.2 Data preprocessing

The ADNI dataset provides fMRI data in raw neuroimaging files, which must be preprocessed. These files (Fig.3) are available in a DCM format, which contains a sequence of 3D images (brain volumes). To convert the sequence to one 4D NIFTI file (with brain representations over the measuring time), the 'CONN: functional connectivity toolbox' [33] is used. In addition, the MNI space normalization of the NIFTI files is also made with the CONN toolbox. The mentioned Montreal Neurological Institute normalization template (so-called MNI template) is a widely-used format for neuroimaging files, which makes it possible to read these files by the majority of toolboxes and packages, including the Nilearn package [34], which is heavily used in this work. Following normalization, the SPM12 software [35],[36] is used to make segmentation of regions and to save final fMRI files, which contain data of a time-varying brain regions' activity. Finally, the files are transformed into time series using the Nilearn methods.

Following steps concentrate on connectomes construction. A functional connectome creation is based on the paired correlation measure between the time series of brain regions. The outcome of this step is a network, which illustrates brain regions and the strengths of their cooperation, that is a functional network. Likewise, GRNs are constructed using correlation values between gene expression in patients and healthy individuals. In other words, an edge of a network has a weight that is equal to $1 - |r|$, where r is the correlation coefficient between a pair of genes.



Subject ID: 136_S_4433
Image ID: I336199
Description: Resting State fMRI
24 of 48
● Axial
■ Sagittal
■ Coronal

Fig. 3: Example of a neuroimaging file's view from the ADNI dataset.

2.3 Networks description

To start with, gene expression is the main factor that affects a phenotype. First of all, gene expression has an effect on protein building, while proteins define a phenotype. Therefore, the gene expression data should be taken into consideration while a disease studying. For example, as shown in [1], cancer genomes and their GRNs could be successfully analyzed with the TDA-based approach. As mentioned above, a Gene Regulatory Network (GRN) is a network, which consists of nodes assigned to genes, and edges described the interactive couplings between them. Pairwise correlation coefficients of gene expression values form weights of the networks, which are associated with the patients' and controls' data. To be more precise, a weight of an edge is equal to $1 - |r|$, where r is the correlation coefficient between gene expression. The schematic pipeline of networks construction is given in Fig.4.

Moreover, only the most relevant genes should be considered in the expression values table. Genes affect AD phenotype with varying degrees. Some of them are considered to be high-influencing, while others are less risky concerning AD. Consequently, in this work, only the most relevant genes are selected.

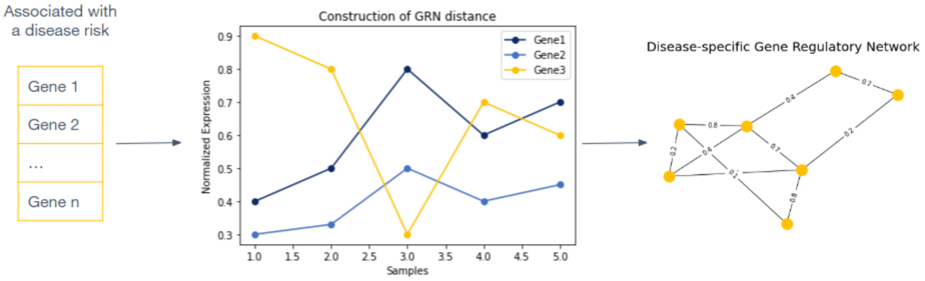
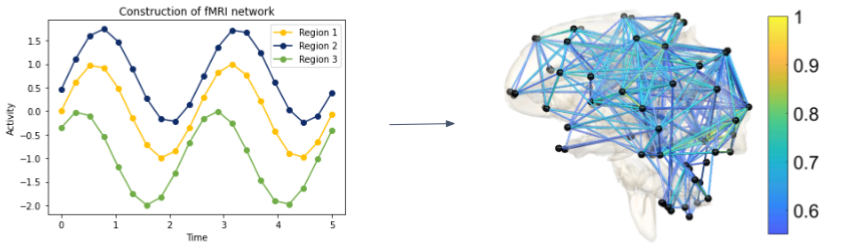


Fig. 4: Scheme of Gene Regulatory Network construction based on gene expression dataset.



Soumya Das, Hernando Ombao, Moo K. Chung,
Topological Data Analysis For Functional Brain Networks

Fig. 5: Scheme of functional connectome construction based on brain regions' activities.

The brain regions' activity could be transformed into time series, and this type of data is an input for a functional brain network creation. The interaction between brain regions is evaluated with correlation coefficients, and the value of $1 - |r|$ (where r is a correlation coefficient) is assigned to a weighted edge of a network. Thus, Fig.5 gives the idea of brain functional network construction. After the network's construction, the simplicial complexes are built from the networks to get persistence diagrams (and also barcode diagrams), which are used to identify the topological difference between patients' and controls' data.

2.4 Persistent Homologies

Overall, persistent homologies theory gives an efficient tool for providing a summary of the topological features of the data. In this section, key definitions are provided.

Definition 1. A simplicial complex K is a set of simplices, where every face of a simplex from K is also in K; and the non-empty intersection of two simplices σ_1 and σ_2 is a face of both σ_1 and σ_2.

Definition 2. A k-simplex is the convex hull of $k + 1$ affinely independent points.

The advantage of using persistent homologies is the dynamic result that is accomplished by a filtration.

Definition 3. A filtration is a sequence of topological spaces $X_0 \subseteq X_1 \subseteq ... \subseteq X_n$. A Vietoris-Rips filtration is the most widely used filtration for real-data analysis. Vietoris-

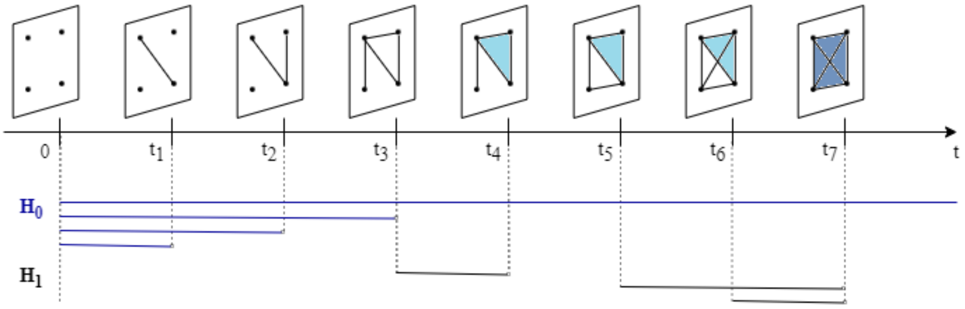


Fig. 6: An example of a filtration and a barcode.

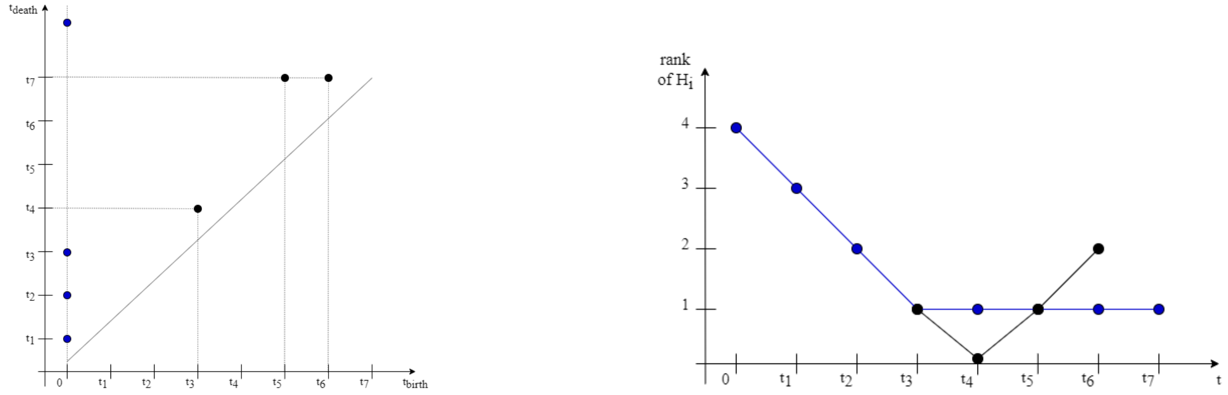


Fig. 7: a) An example of a persistence diagram. b) An example of a Betti curve.

Rips complex $VR_\epsilon(X)$ at scale ϵ (filtration value) is defined as: $VR_\epsilon(X) = \{\sigma \subseteq X | d(x,y) \leq 2\epsilon, \forall x,y \subseteq \sigma\}$.

Definition 4. The p^{th}-persistent homology of X is the pair $(\{H_p(X_i)\}_{1\leq i\leq n}, \{f_{ij}\}_{1\leq i\leq j\leq n})$, where $f_{ij} : H_p(X_i) \to H_p(X_j)$ is a linear map induced by the inclusion map $X_i \to X_j$.

Definition 5. Let $\alpha \in H_p(X_i, \epsilon_i)$ be a nontrivial homology class. α is born at ϵ_i if it is not in the image of $f_{i-1,i} : H_p(X, \epsilon_{i-1}) \to H_p(X, \epsilon_i)$. α dies at ϵ_j if $j > i$ is the smallest index for which $f_{ij} = 0$. The lifetime of α is half-open interval $[i, j)$.

Definition 6. For $p > 0$, the p-Wasserstein distance between two persistence diagrams D_1 and D_2 is defined as $W_{p,q}(D_1, D_2) = inf_{\gamma:D_1\to D_2}(\Sigma_{x\in D_1}||z - \gamma(x)||_q^p)^{1/p}$, where $||.||_q$ denotes the q-norm, $1 \leq q \leq \infty$ and γ ranges over all bijections between D_1 and D_2.

A persistence diagram contains points, the coordinates of which represent the birth and death value of a topological feature. Similarly, a barcode diagram contains the lines which represent lifetime intervals of topological features. In addition, Betti curves contain ranks of H_i (i^{th} homology group) in correspondence with each filtration value. Examples of the graphs are Fig.6 and Fig.7.

The insights into some computational and theoretical approaches applied to persistent homologies are given in [37] and [38]. Firstly, [37] includes simple examples of the process of computing persistent intervals. There is also an in-depth description of the mathematical background, which is a basis for studying persistent homologies. Additionally, [38] also gives a comprehensive theoretical background as well as computational methods.

2.5 Data processing

The aforementioned persistence diagrams might be created and visualized with a commonly used tool, which is called the Gudhi (Geometry Understanding in Higher Dimensions) package [39]. It is worth mentioning that the Gudhi package is well-optimized to calculate persistent intervals from metric spaces. Moreover, the package contains persistence diagrams vectorization methods, which are used for statistical tests and classification tasks.

Overall, the genetic data is processed as shown in Fig.8. Firstly, the relevant genes, which are explicit markers for AD, are selected from the ADNI dataset. Relevant genes are defined in accordance with recent findings from [40]. Although males and females have slightly different degrees of genes' influence on AD, in this work, gender is not taken into consideration. In general, the set of genes is similar considering both genders. Consequently, the difference in the genes' effect on the disease progress should not have a significant contribution to the results.

Following the extraction of relevant genes, dimensionality reduction is implemented. In more detail, three techniques were chosen: Topological autoencoders [17], UMAP [15], and Basic autoencoders. As it was mentioned above, these approaches compared in [18], and topological autoencoders illustrated better results. From the obtained data, original or reduced, two correlation matrices are constructed. One of them describes patients' data. and the other one is related to control data. The dimension of the matrices is equal to $N \times N$, where N is the number of selected genes. Then, the Vietoris-Rips complex is constructed from the matrices with the help of the Gudhi package, which is also used to calculate persis-

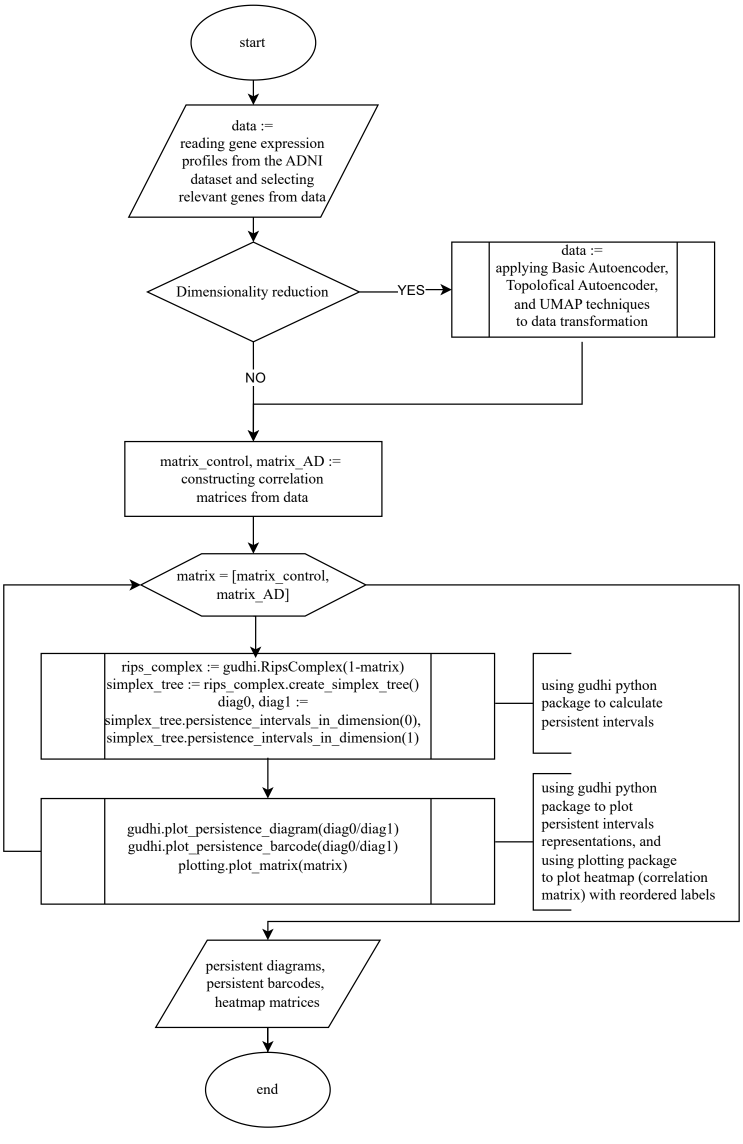


Fig. 8: Flowchart diagram of gene expression data processing.

tence intervals. Finally, as an outcome, the persistence diagrams and barcodes are drawn. The diagrams are also used for statistical tests, which are described below.

Concerning fMRI data, each subject is processed by solution presented in Fig.9. Overall, there are three datasets, which are considered in this task: the ADNI, the CNI challenge and the OASIS-3 datasets. For each subject in the chosen dataset, a time series table could be loaded from a file. As well as genetic data, the table is processed by the Gudhi library to get a Vietoris-Rips complex and persistence view (persistence intervals). In addition, the classification algorithm is implemented using *sklearn.model_ selection* module and its *GridSearchCV()* class from the Scikit-learn package [41]. Also, the simple neural network with PyTorch package [42] is implemented to solve the classification task. There are several vectorization techniques and classifiers. Vectorization methods, which turn a persistence diagram into a vector, are described as following:

- *Persistence landscape* [43]. Firstly, a diagram is rotated by $-\frac{\pi}{4}$. Following this, triangular functions are built on each point of the diagram, and the k^{th} is the k^{th} largest

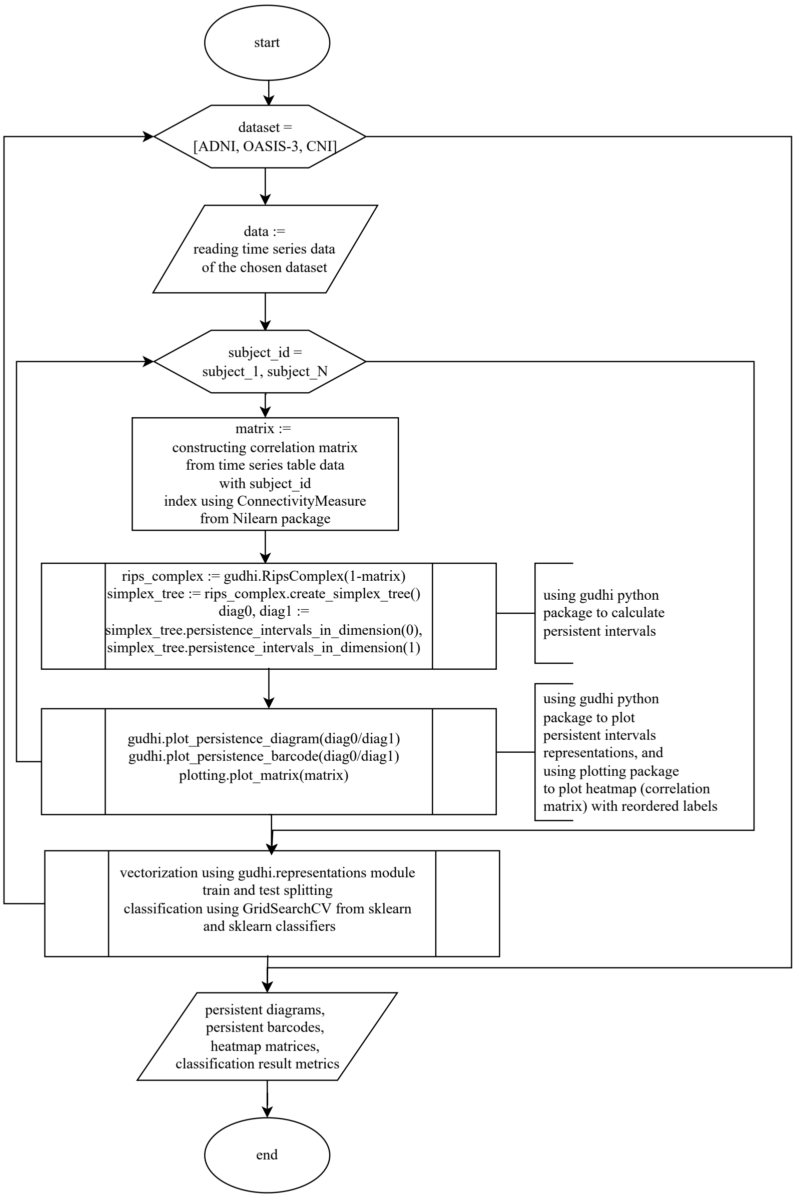

start
dataset =
[ADNI, OASIS-3, CNI]
data :=
reading time series data
of the chosen dataset
subject_id =
subject_1, subject_N
matrix :=
constructing correlation matrix
from time series table data
with subject_id
index using ConnectivityMeasure
from Nilearn package
rips_complex := gudhi.RipsComplex(1-matrix)
simplex_tree := rips_complex.create_simplex_tree()
diag0, diag1 :=
simplex_tree.persistence_intervals_in_dimension(0),
simplex_tree.persistence_intervals_in_dimension(1)
using gudhi python
package to calculate
persistent intervals
gudhi.plot_persistence_diagram(diag0/diag1)
gudhi.plot_persistence_barcode(diag0/diag1)
plotting.plot_matrix(matrix)
using gudhi python
package to plot
persistent intervals
representations, and
using plotting package
to plot heatmap (correlation
matrix) with reordered labels
vectorization using gudhi.representations module
train and test splitting
classification using GridSearchCV from sklearn
and sklearn classifiers
persistent diagrams,
persistent barcodes,
heatmap matrices,
classification result metrics
end

Fig. 9: Flowchart diagram of fMRI data processing.

value among built triangular functions. The vectors' elements correspond to uniformly sampled points of the rotated diagonal (x-axis).

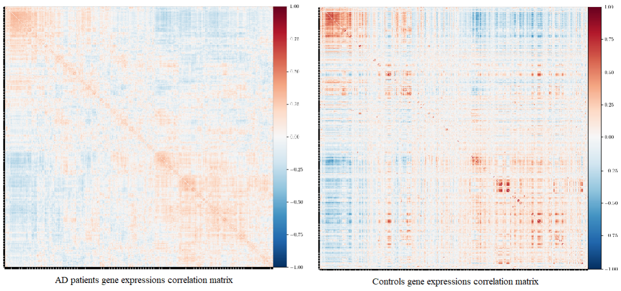
- *Silhouette* [44]. This method is quite similar to the previous one (Landscape), however, the weighted average of built triangular functions is taken instead of k^{th} largest value. The weights correspond to the distance between the selected point and the diagonal, which is rotated as in the previous case.

Then, the persistence diagrams are classified into two categories using classification models and the GridSearchCV method from the Scikit-learn package as well as PyTorch package.

To estimate the similarity between matrices and persistence intervals, the Kolmogorov–Smirnov test [45] is frequently used. The Kolmogorov–Smirnov test, or KS test, compares two datasets, or a dataset and a theoretical distribution, by measuring the maximum vertical distance between the cumulative distribution functions. In this work, concerning the genetic data, the test is used to compare weights distributions from AD patients' and controls' correlation matrices. In the case of the fMRI data, the distributions are obtained by concatenation of subject-related weights. In addition, the test is also used to estimate the similarity of persistence diagrams, which are preliminarily vectorized.

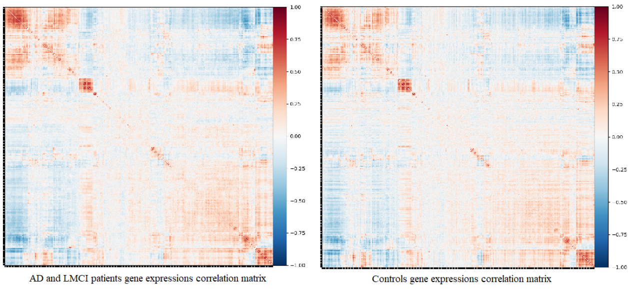
Section 3. Gene expression data analysis

To explore the difference between patients' and controls' gene expression profiles, the ADNI dataset was used. The ADNI gene expressions data contains the subject's IDs, symbols of genes, and a corresponding expression value for each subject and gene. By matching an ID of a subject with information about the subject's category (control or patient), it is possible to construct two tables with correlation values between pairs of genes. Overall, there are 16 subjects related to the AD category and 193 subjects from a control group. Considering the fact that the dataset is not well-balanced, it is also possible to take into account such a category as 'LMCI' (late mild cognitive impairment). In this case, there are 187 subjects with some impairments.


AD patients gene expressions correlation matrix

Controls gene expressions correlation matrix

Fig. 10: Gene expression correlation matrices. AD patients and contol group.


AD and LMCI patients gene expressions correlation matrix

Controls gene expressions correlation matrix

Fig. 11: Gene expression correlation matrices. AD and LCMI patients and contol group.

To select the most relevant genes, the paper [40] is valuable. Preliminarily, 200 top genes from [40] were matched with the ADNI data. Hence, correlation tables are based on these important genes, which are related to the development of Alzheimer's disease. Finally, there are 635 expression rows in the constructed table, given different loki of genetic markers.

As it was mentioned above, following matrix creation, the correlation values between gene expressions among each subject are calculated. The correlation matrix with a size of 635 × 635 (Fig.10 and Fig.11) is an input for topological analysis. As a result, persistence diagrams and barcodes are illustrated in Fig.12 and Fig.13. Concerning AD patients and controls comparison, categories are quite different, because (1) there is more contrast correlation matrix in the controls dataset, (2) the topological features' lifetimes are much longer in patients dataset. Overall, there is an obvious importance of one-dimensional holes and two-dimensional holes in AD patients' dataset, because it is shown that their lifetimes are longer during the filtration than in the controls' dataset.

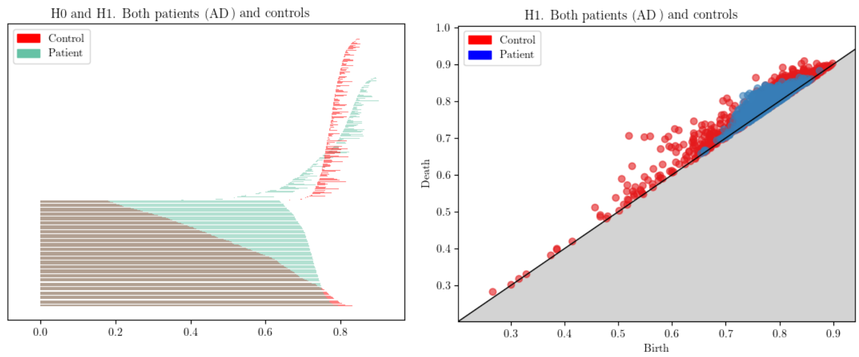


Fig. 12: Barcode diagram and persistence diagram. Comparison of patients and controls.

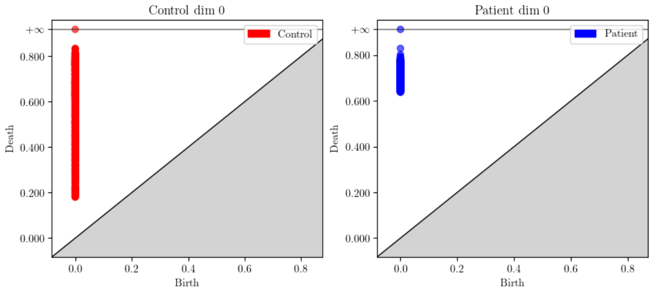


Fig. 13: Persistence diagrams, H0. Comparison of patients and controls.

With regards to AD and LMCI patients, there is a less explicit difference in comparison with the control group, however, the key features, which are mentioned above, are preserved. The effect is expected because LMCI is the stage between normal cases and AD ones. However, it tends to be close to a normal aging process, which does not influenced by significant genetic changes. In general, a similar effect could be observed on barcode and persistence diagrams. The mentioned diagrams, which are related to the LMCI dataset, are shown in Fig.14 and Fig.15. They demonstrate the similarity between controls' persistence intervals (0 and 1 dimensions) and LMCI patients' ones.

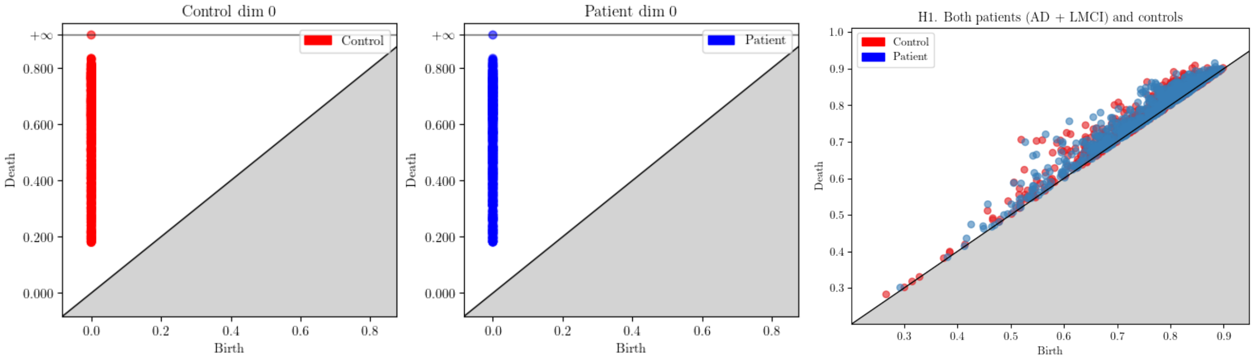


Fig. 14: a) and b) Controls and LMCI patients data presented as 0-dimensional persistence diagrams. c) Controls and LMCI patients data presented as 1-dimensional persistence diagrams (controls are red, patients are blue).

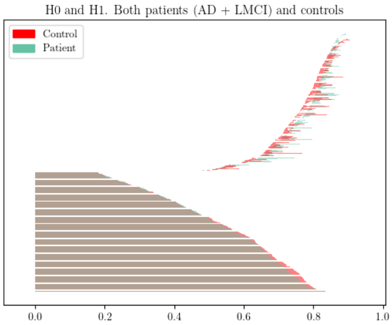


Fig. 15: Controls and LMCI patients data presented as 0- and 1-dimensional persistence barcodes (controls are red, patients are blue).

As it is shown in Fig.16, Fig.17, and Fig.18, dimensionality reduction techniques do not brake general patterns, and the results of topoAE and UMAP implementation are quite better in terms of showing the difference between patients' and controls' data. In general, it must be stated, that computations of reduced data are faster, and results are similar.

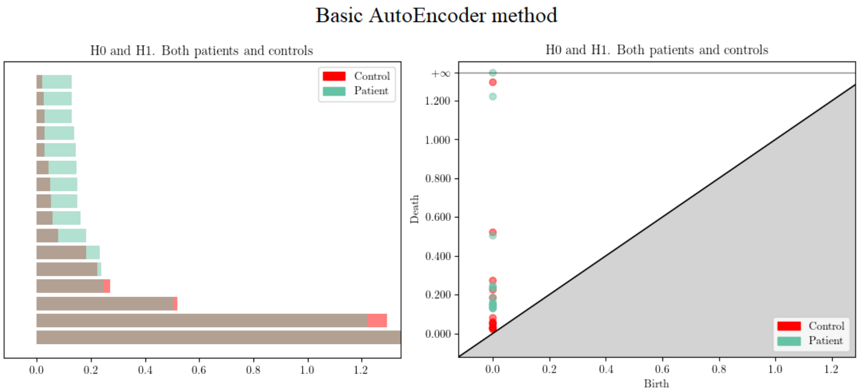


Fig. 16: Result of dimensionality reduction using Basic AutoEncoder method (AD patients).

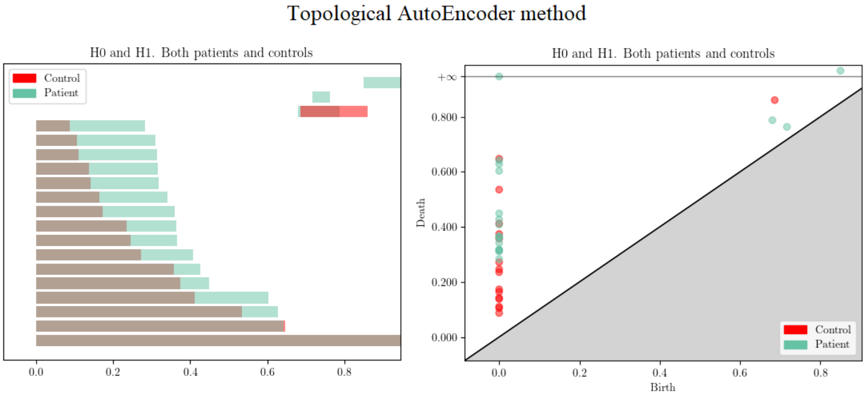


Fig. 17: Result of dimensionality reduction using Topological AutoEncoder method (AD patients).

As for the Kolmogorov-Smirnov test, the results are consistent with the visual difference between topological features of genetic data related to patients and controls. Firstly, the

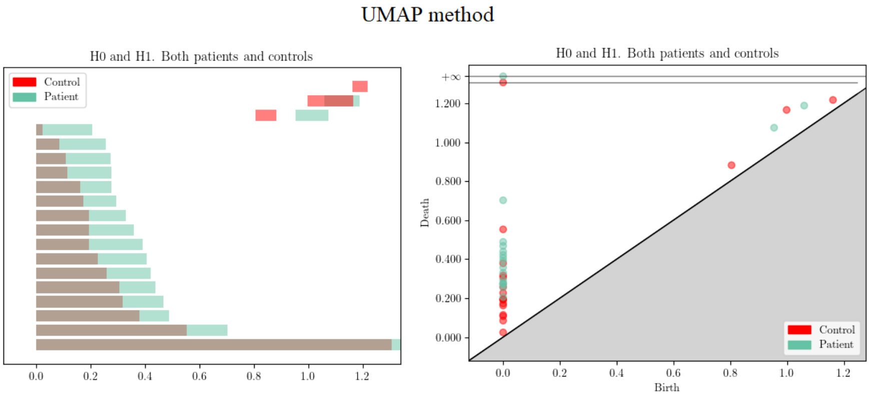


Fig. 18: Result of dimensionality reduction using UMAP method (AD patients).

AD and controls dataset							
dim 0				**dim 1**			
original	basicAE	topoAE	umap	original	basicAE	topoAE	umap
0.0000***	0.0030***	0.0030***	0.0112*	0.0000***	-	0.0097***	0.0196*

Table 1: P-values for the Kolmogorov-Smirnov test. AD patients' and controls' persistence diagrams comparison (gene expression dataset).

vectorized persistence diagrams are compared. The null hypothesis states that the two given distributions are similar. The alternative hypothesis is that the distributions are not identical. The Kolmogorov-Smirnov test statistic is calculated as the maximum absolute difference between the two CDFs (cumulative distribution functions). Consequently, the test statistic is calculated as $max|F(x) - G(x)|$, where $F(x)$ and $G(x)$ are the empirical distribution functions of the given samples. Let significance levels be 0.01, 0.05, and 0.1. Table 1 and Table 2 represent p-values for the Kolmogorov-Smirnov test, and asterisks mean that the null hypothesis could be rejected on 0.01, 0.05, and 0.1 significance levels respectively to (***), (**), and (*) symbols. From Table 1 and Table 2, it is possible to conclude the following:

1. The null hypothesis is rejected in the cases of AD and controls datasets comparison. The dimensionality reduction methods lessen the difference, but it is still significant.

2. The null hypothesis could not be rejected in the cases of LMCI and control group comparison. There is no significant difference between datasets, and p-values are close to 1. However, the lower p-value could be observed in the case of original data. Hence, the dimensionality reduction methods decrease the difference between LMCI patients' and controls' data.

LMCI and controls dataset							
dim 0				**dim 1**			
original	basicAE	topoAE	umap	original	basicAE	topoAE	umap
0.1416	0.9998	0.9998	0.9998	0.0546*	-	0.9937	0.9999

Table 2: P-values for the Kolmogorov-Smirnov test. LMCI patients' and controls' persistence diagrams comparison (gene expression dataset).

The results for genetic data of LMCI patients transformed by dimensionality reduction methods are presented in Fig.19, Fig.20, and Fig.21. It could be seen that patients' data is close to the controls' one. Confirming the mentioned result, Table 2 shows that dimensionality reduction methods even make patients' and controls' gene expression profiles more similar than in the original dataset.

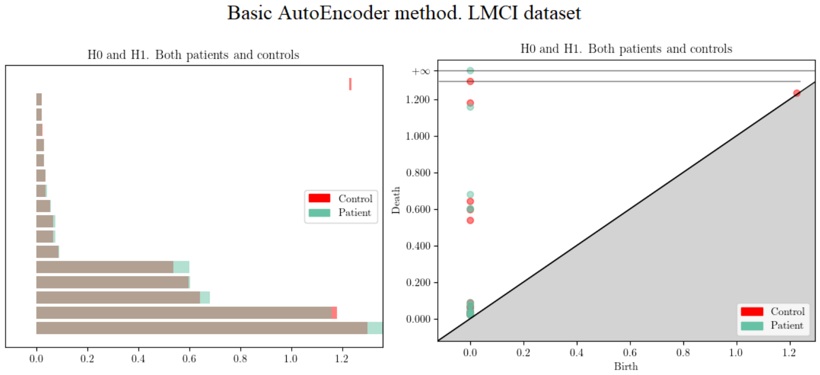


Fig. 19: Barcode and persistence diagrams for genetic data transformed by Basic Autoencoder dimensionality reduction method (LMCI patients).

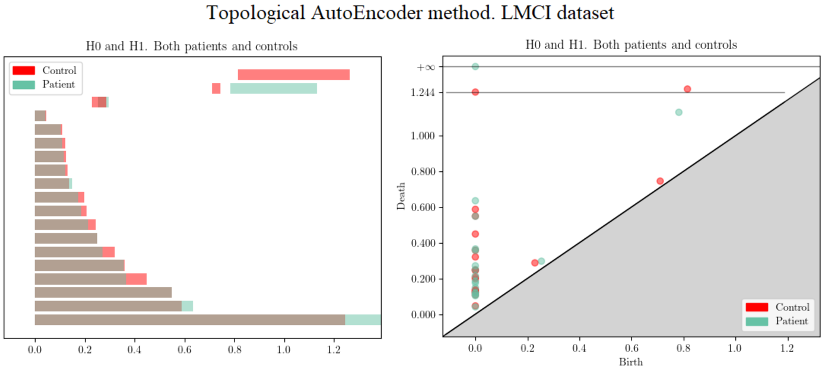


Fig. 20: Barcode and persistence diagrams for genetic data transformed by Topological Autoencoder dimensionality reduction method (LMCI patients).

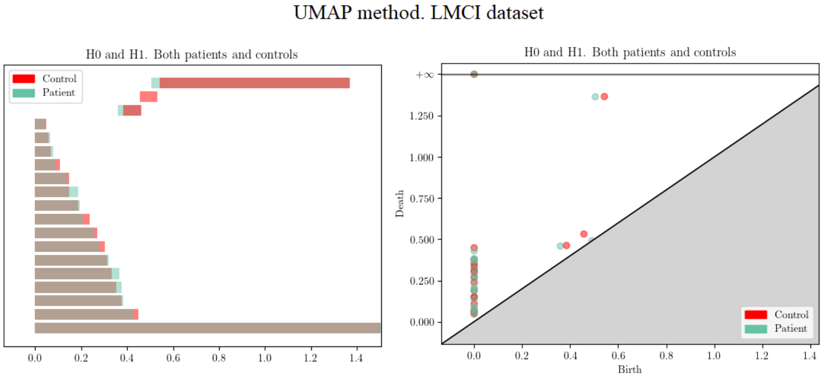


Fig. 21: Barcode and persistence diagrams for genetic data transformed by UMAP dimensionality reduction method (LMCI patients).

Following, the weights distributions from correlation matrices of patients' and controls' data are compared by the Kolmogorov-Smirnov test. In general, the null hypothesis, about the similarity of the given distributions, could not be rejected, and there is no significant difference between controls' and patients' weights. Only in one case, the hypothesis is rejected: there is a difference between AD and control datasets, considering original data. The results could be explained by the fact that AD patients' data present the late stages of AD, while LMCI data are similar to controls in terms of weight distributions, because of the earlier stages of dementia. The mentioned information is also reflected in Table 3 and Table 4.

AD and controls dataset			
original	basicAE	topoAE	umap
0.0678*	0.9977	0.9999	0.9999

Table 3: P-values for the Kolmogorov-Smirnov test. AD patients' and controls' weights distributions from correlation matrices comparison (gene expression dataset).

LMCI and controls dataset			
original	basicAE	topoAE	umap
0.9999	0.9999	0.9667	0.9977

Table 4: P-values for the Kolmogorov-Smirnov test. LMCI patients' and controls' weights distributions from correlation matrices comparison (gene expression dataset).

To summarize the findings related to gene expression data analysis, it should be mentioned that the obvious difference between AD patients and a control group from the ADNI dataset is observed. First of all, the controls' genes are more correlated in terms of expression values, and this fact might mean the unstable disruption of the work of risky genes in AD cases. In other words, patients' genes are expressed in different ways, and disruption might be caused by both too high expression value and too low one. Secondly, the persistence diagram of AD patients shows that connected components (0-dim persistent homologies) are mostly formed by low correlated genes, while controls' data illustrate that genes with high and moderate interaction values are involved in connected components as well. This means that patients' network lacks diverse correlation values between genes. As for the 1-dimensional persistence diagram of AD patients, it shows that healthy gene interactions form holes with a variety of genes (low and high interacting), while in the case of patients, all points on the diagram are shifted to one single cluster.

Finally, it should be noted that dimensionality reduction techniques preserve the results making the patients' dataset more balanced and closer to the controls' one in terms of correlation values. However, LMCI patients are almost identical to a control group in cases of original and reduced data. Moreover, dimensionality reduction makes LMCI and control datasets indistinguishable.

Section 4. Functional MRI data analysis

As a commonly used solution, the AAL atlas [29] regions are taken as a basis for time series tables. The correlation matrices for each dataset and each subject are obtained from the given (or constructed from raw files) time series tables. An example of a correlation matrix for a subject is shown in Fig.22. Each dataset contains patients' and controls' data, and two summarized persistence diagrams are created for the comparison. One of the diagrams contains all patients' data, while the other diagram contains only controls. A similar approach is implemented for weights distributions: there are two histograms for each dataset with patients and controls respectively.

To start with, Fig.23 illustrates the difference between controls' and patients' weights from correlation matrices for each dataset. The ADNI dataset is distinguishable from the CNI and the OASIS-3 data because there is a significant difference between patients and controls. A similar result is shown in Table 5, where p-values for the Kolmogorov-Smirnov test are written. In the case of the ADNI data, the null hypothesis about significant similarity between data is rejected, while in the other cases, the null hypothesis could not be rejected (there is no explicit difference between patients and controls on a reasonable significance level).

Dataset	ADNI	CNI	OASIS-3
P-value	0.0000	0.9999	0.7166

Table 5: P-values for the Kolmogorov-Smirnov test. AD patients' and controls' weights distributions from correlation matrices comparison (fMRI data from three datasets).

The OASIS-3 dataset includes patients with early stages of dementia, and the CNI dataset includes ADHD patients (Attention-deficit/hyperactivity disorder). Consequently, the results are reasonable, and the ADNI dataset reflects significant changes in brain regions' cooperation. As stated in [9], the well-correlated regions in not significant impairments could compensate for local changes, which are not propagated through the whole brain structure. Consequently, it is expected to see almost identically correlated regions for patients and a control group in the OASIS-3 dataset.

Barcode diagrams are quite convenient to illustrate persistent homologies related to the

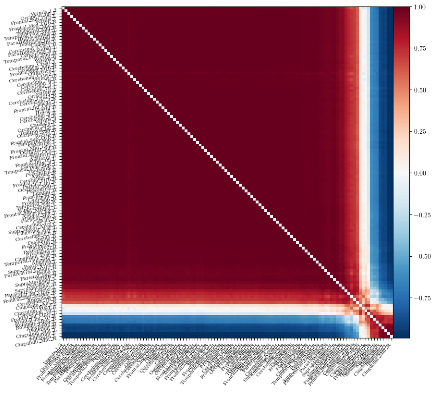


Fig. 22: Example of a correlation matrix of AAL brain regions.

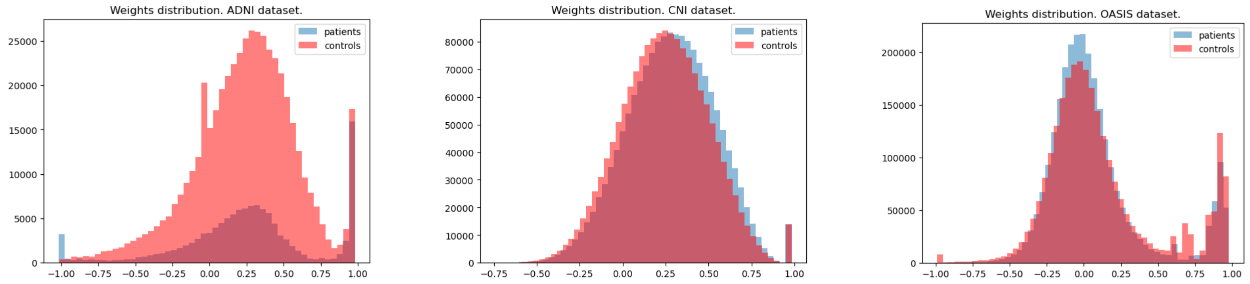


Fig. 23: a) ADNI dataset weights distributions. b) CNI challenge dataset weights distributions. c) OASIS-3 dataset weights distributions.

mentioned datasets. In Fig.24 it is possible to see that controls' topological features have longer lifetimes than patients' ones. This statement is applicable to all three datasets, which leads us to the finding of the general pattern. Overall, the one-dimensional and two-dimensional holes are formed by close regions (in terms of cooperation), and this fact might describe the loss of cooperation strength between weakly-interacting regions concerning a disease. Table 6 demonstrates the significant difference between patients and controls in three given datasets (according to Kolmogorov-Smirnov statistics).

dim 0			**dim 1**		
ADNI	CNI	OASIS-3	ADNI	CNI	OASIS-3
6.16e-21	6.61e-13	3.07e-128	7.61e-28	6.76e-08	2.82e-31

Table 6: P-values for the Kolmogorov-Smirnov test. AD patients' and controls' persistence diagrams comparison (fMRI data from three datasets).

The fMRI data is also used to classify subjects into two categories: patients and controls. Unfortunately, the first dataset (ADNI) contains only 37 controls and merely 9 patients, so it is not a representative case to classify the dataset into two categories, because the result (accuracy or f1-score) depends on a random permutation for training and testing datasets. However, to solve the classification task, the CNI and the OASIS-3 datasets are better, since the datasets contain more observations (400 subjects are used from the OASIS-3, and 240 subjects are from the CNI challenge dataset).

First of all, persistence diagrams of 0 and 1 dimensions are vectorized. Then, the simple models from the Scikit-learn package are implemented. Among Scikit-learn classification

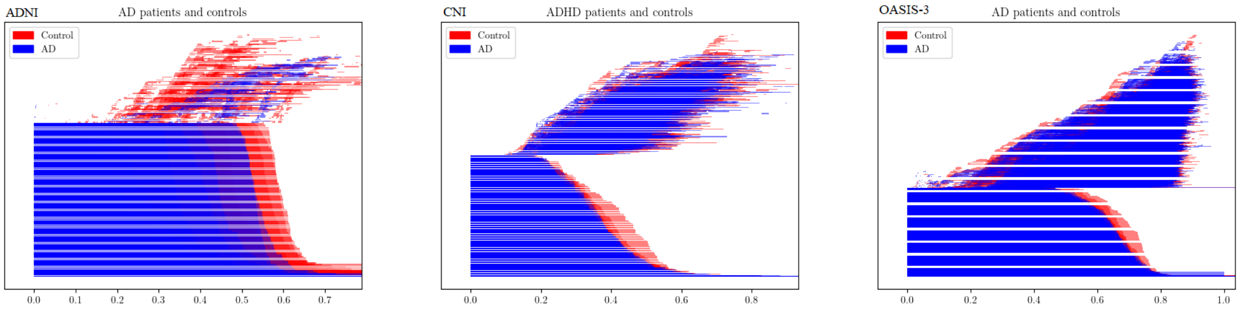


Fig. 24: a) ADNI dataset barcode diagram. b) CNI challenge dataset barcode diagram. c) OASIS-3 dataset barcode diagram.

models, the *GradientBoostingClassifier* gives the best results for the classification of the OASIS-3 and the CNI datasets, and the accuracy is about 60% in both cases. However, to improve the result, a neural network with four linear layers and three *ReLU* activation functions is implemented. After each linear layer, the Dropout regularization technique is used to deal with overfitting. After the final layer, a sigmoid activation function is used to return a value between 0 and 1, making it suitable for binary classification tasks. The model is trained using binary cross-entropy loss (BCELoss) and stochastic gradient descent (SGD) optimizer. For the OASIS-3 dataset the learning process is implemented without the Dropout technique.

The CHI challenge dataset is classified with 91% and 60% values for accuracy values on training and test samples respectively. Meanwhile, the classification for subjects from the OASIS-3 dataset gives an accuracy of 92% on the training sample and 82% on the test one. It is possible to conclude, that vectorized persistence diagrams could be used to classify fMRI data, however, the results are not as good as expected.

To sum up, the main result, which is presented in this section, is the comparison of weight distribution and persistence diagrams. Despite the fact that weights (correlation values between regions' activity) could be indistinguishable in controls' and patients' data, the topological features show the difference even in the early stages of AD as well as in ADHD cases. This means that the topology of a functional network changes even when the strengths of interaction between brain regions remain the same. ADHD patients have strengths of interaction that are almost identical to controls, while the topological view is quite different. The OASIS-3 dataset demonstrates more distinguishable weights (but the difference is still insignificant), whereas the persistence diagrams of patients and controls are different with a close to zero p-value.

Section 5. Conclusion

It should be mentioned that some general patterns, which are related to the difference between normal and abnormal medical data, were found in both data types. Also, both types of medical data lead to some similar difficulties while processing. Mainly, it is related to the lack of data and especially the lack of well-balanced data. However, even considering existing datasets, it is possible to find the difference between patients and controls, which is based on persistence diagrams.

Concerning gene expression profiles data, a significant difference between AD patients' and controls' gene expression values was observed. It could be proven by the Kolmogorov-Smirnov test. There are also reasonable observations related to the LMCI patients because there is no explicit difference between their genetic data and the controls' ones. Considering changes caused by LMCI as less marked than ones caused by AD, the mentioned results are well-based. It is worth noting that patients' correlation coefficients between genes could be considered distinguishable according to the Kolmogorov-Smirnov test, however, only on the high level of significance (0.07). Dimensionality reduction methods do not change the results of the analysis of persistence diagrams while decreasing the difference between patients' and controls' correlation matrices.

As far as fMRI data are concerned, there is a remarkable difference between normal cases and diseases in terms of persistence diagrams. In accordance with the Kolmogorov-Smirnov test, which compared vectorized 0-dimensional and 1-dimensional diagrams, in three given datasets patients' data differ from controls' data with near-zero p-values. In contrast, the correlation coefficients between brain regions could be considered identical in the CNI and OASIS-3 datasets (excepting for the ADNI). The effect could be explained with the following: the mentioned datasets reflect insignificant impairments, which are caused by local changes in brain structure, so it has no impact on regions' cooperation. Consequently, the obtained results are quite reasonable.

Moreover, the CNI and the OASIS-3 datasets were classified into two categories (patients and controls) with an approach based on neural networks and classifiers from the Scikit-learn package. Overall, the classification scores for the neural network with a few linear layers and the Dropout technique are quite better than for *GradientBoostingClassifier* and other classifiers from the Scikit-learn. The OASIS-3 dataset was classified with the best accuracy values of 82% on the test sample.

There is also a common observation for the comparison of patients and controls for both types of data: the correlation coefficients between genes and brain regions activity have almost identical distributions, while the difference between persistence views is quite more explicit. This fact supports the existing evidence of the importance of TDA applied to biological data. Topological features distinguish categories well in both types of medical data.

As for future steps, they might include experiments with network construction methods, and also the implementation of the newest dimensionality reduction technique, called RTD-AE (Representation Topology Divergence AutoEncoders). Concerning network construction, the different threshold-independent methods could be implemented as an improvement of the standard used approach. In this way, the effect of some minor brain region interactions could be smoothed, and this might lead to the improvement of the existing results.

References

[1] H. Masoomy, B. Askari, S. Tajik, A. K. Rizi, G.R. Jafari, Topological analysis of interaction patterns in cancer-specific gene regulatory network: persistent homology approach, Sci Rep 11, 16414, 2021.

[2] S. Simmons, J. Peng, J. Bienkowska, B. Berger, Discovering What Dimensionality Reduction Really Tells Us About RNA-Seq Data. J Comput Biol. 8, 715-28, Aug. 2015.

[3] P. G. Camara, Topological methods for genomics: present and future directions, Curr Opin Syst Biol, 1: 95–101, Feb. 2017.

[4] D. S. Bassett, O. Sporns, Network neuroscience, Nat Neurosci 20, 353–364, 2017.

[5] A. E. Sizemore, J. E. Phillips-Cremins, R. Ghrist, D. S. Bassett, The importance of the whole: Topological data analysis for the network neuroscientist, Network Neuroscience; 3 (3): 656–673, 2019.

[6] M. N. Hallquist, F. G. Hillary, Graph theory approaches to functional network organization in brain disorders: A critique for a brave new small-world, Network Neuroscience; 3 (1): 1–26 ,2018.

[7] S. Das, H. Ombao, M. K. Chung, Topological Data Analysis for Functional Brain Networks, arXiv:2210.09092v1, Oct. 2022.

[8] S. Das, D. V. Anand, M. K. Chung, Topological Data Analysis of Human Brain Networks. Through Order Statistics, arXiv:2204.02527v3, Oct. 2022.

[9] J. Wang, R. Khosrowabadi K. K. Ng, Z. Hong, et al., Alterations in Brain Network Topology and Structural-Functional Connectome Coupling Relate to Cognitive Impairment, Front. Aging Neurosci. 10:404, 2018.

[10] S. Cussat-Blanc, K. Harrington, W. Banzhaf, Artificial Gene Regulatory Networks – A Review, Artificial Life 24(5):1-33, Jan. 2019.

[11] C. Bellenguez, F. Küçükali, I.E. Jansen, et al., New insights into the genetic etiology of Alzheimer's disease and related dementias, Nat Genet 54, 412–436, 2022.

[12] X.-D. Wang, S. Liu, H. Lu, Y. Guan, H. Wu, Y. Ji, Analysis of Shared Genetic Regulatory Networks for Alzheimer's Disease and Epilepsy, Biomed Res Int. 14;2021:6692974. eCollection, Oct. 2021.

[13] W. Gao, W. Kong, S. Wang, G. Wen, Y. Yu, Biomarker Genes Discovery of Alzheimer's Disease by Multi-Omics-Based Gene Regulatory Network Construction of Microglia, Brain Sci, 5;12(9):1196, Sep. 2022.

[14] K. Pearson . 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):559–572, 1901.

[15] L. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction, arXiv:1802.03426, 2018.

[16] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE. Journal of machine learning research, 9(11), 2008.

[17] M. Moor, M. Horn, B. Rieck, K. Borgwardt, Topological Autoencoders, arXiv:1906.00722v5, 2021.

[18] I. Trofimov, D. Cherniavskii, E. Tulchinskii, N. Balabin, E. Burnaev, S. Barannikov, Learning Topology-Preserving Data Representations, arXiv:2302.00136v2, 2023.

[19] J. Liu, G. Pearlson, A. Windemuth, G. Ruano, N. I. Perrone-Bizzozero, V. Calhoun, Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA, Hum Brain Mapp, 30(1): 241–255, Jan. 2009.

[20] L.T. Elliott, K. Sharp, F. Alfaro-Almagro, et al., Genome-wide association studies of brain imaging phenotypes in UK Biobank. Nature 562, 210–216, 2018.

[21] S.M. Smith, G. Douaud, W. Chen, et al., An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. Nat Neurosci 24, 737–745, 2021.

[22] A. Holmes, M. Hollinshead, T. O'Keefe, et al., Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures. Sci Data 2, 150031, 2015.

[23] UK Biobank Multimodal Dataset, S.M. Smith, G. Douaud, W. Chen, et al., An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. Nat Neurosci 24, 737–745, 2021, Oxford Brain Imaging Genetics Server - BIG40 [Online]. Available: https://open.win.ox.ac.uk/ukbiobank/big40/, Accessed: Feb. 23, 2023.

[24] The Alzheimer's Disease Neuroimaging Initiative (ADNI) (since 2004) Dataset [Online]. Available: https://adni.loni.usc.edu/about/, Accessed: Feb. 20, 2023.

[25] Affymetrix GeneChip Human Gene 2.0 ST Array (U219) [Online]. Available: https://www.affymetrix.com/. Accessed: May 1, 2023.

[26] R.A. Irizarry, B. Hobbs, F. Collin, et al., Exploration, normalization, and summaries of high-density oligonucleotide array probe level data. Biostatistics, 4(2):249-264, 2003.

[27] M.D. Schirmer, et al., Neuropsychiatric disease classification using functional connectomics-results of the connectomics in neuroimaging transfer learning challenge. Medical image analysis 70: 101972, 2021.

[28] D.S. Marcus, T.H. Wang, J. Parker, et al., Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. Journal of Cognitive Neuroscience, 19(9):1498-1507, 2007.

[29] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, et al., Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. Neuroimage 15:273–289, 2002.

[30] R.S. Desikan, F. Segonne, B. Fischl, et al., An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage 31:968–980, 2006.

[31] R.C. Craddock, G.A. James, P.E. Holtzheimer III, et al., A whole brain fmri atlas generated via spatially constrained spectral clustering. Human brain mapping 33:1914–1928, 2012.

[32] O. Esteban, C.J. Markiewicz, R.W. Blair, et al., fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat Methods 16:111–116, 2019.

[33] S. Whitfield-Gabrieli, A. Nieto-Castanon, Conn: A functional connectivity toolbox for correlated and anticorrelated brain networks. Brain connectivity, 2(3), 125-141, 2012.

[34] NiLearn (RRID:SCR001362) [Online]. Available: http://nilearn.github.io, Accessed: Feb. 14, 2023.

[35] W. Penny, K. Friston, J. Ashburner, S. Kiebel, T. Nichols, Statistical Parametric Mapping: The Analysis of Functional Brain Images, 1st Edition: Oct. 6, 2006.

[36] SPM12 [Online]. Available: https://www.fil.ion.ucl.ac.uk/spm/, Accessed: March 1, 2023.

[37] N. Otter, M.A. Porter, U. Tillmann, et al., A roadmap for the computation of persistent homology. EPJ Data Sci. 6, 17, 2017.

[38] P. Leung, Y. Cao, A. Monod, k-Means Clustering for Persistent Homology, arXiv:2210.10003v1, Oct. 2022.

[39] M. Clément, J.-D. Boissonnat, M. Glisse, M. Yvinec, The Gudhi Library: Simplicial Complexes and Persistent Homology, Conference: International Congress on Mathematical Software, 2014.

[40] M.A. Hill, S.C. Gammie, Alzheimer's disease large-scale gene expression portrait identifies exercise as the top theoretical treatment. Sci Rep 12, 17189, 2022.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-learn: Machine Learning in Python. JMLR 12(85):2825-2830, 2011.

[42] PyTorch [Online]. Available: https://pytorch.org/docs/stable/index.html, Accessed: May 20, 2023.

[43] P. Bubenik, Statistical topological data analysis using persistence landscapes. Journal of Machine Learning Research (16):77-102, 2015.

[44] F. Chazal et al., Stochastic Convergence of Persistence Landscapes and Silhouettes, arXiv:1312.0308, 2013.

[45] N.V. Smirnov, Estimate of deviation between empirical distribution functions in two independent samples. (Russian). Bull. Moscow Univ. 2(2):3–16, 1939.