Technical Report

UCAM-CL-TR-973
ISSN 1476-2986

Number 973

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Data-driven representations in brain science: modelling approaches in gene expression and neuroimaging domains

Tiago M. L. Azevedo

July 2022

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

https://www.cl.cam.ac.uk/

This technical report is based on a dissertation submitted 13 February 2022 by the author for the degree of Doctor of Philosophy to the University of Cambridge, Churchill College.

Some figures in this document are best viewed in colour. If you received a black-and-white copy, please consult the online version if necessary.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

https://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Abstract

Data-driven Representations in Brain Science:

Modelling Approaches in Gene Expression and Neuroimaging Domains

Tiago Manuel Lourenço Azevedo

The assumptions made before modelling real-world data greatly affect performance tasks in machine learning. It is then paramount to find a good data representation in order to successfully develop machine learning models. When no considerable prior assumption exists on the data, values are directly represented in a "flatten", 1-Dimensional vector space. However, it is possible to go one step further and perceive more complex relational patterns: for example, a Graph-Dimensional space is used to illustrate the more structured way to represent data and their relational inductive bias.

This thesis is focused on these two computational data dimensions across two scales of human biology: the micro scale of molecular biology using gene expression data, and the macro scale of neuroscience using neuroimaging data. Different modelling approaches will be explored to understand how one can model and represent high-dimensional brain data across the specific needs in the applied fields of these two scales. Specifically, for Graph-Dimensional data two approaches will be developed. Firstly, specific and shared genetic profiles that can be generalisable to external datasets will be extracted by applying multilayer co-expression networks across 49 human tissues. Then, a novel deep learning model will be introduced to leverage the entirety of resting-state fMRI data (i.e., spatial and temporal dynamics), as opposed to previous approaches in the literature that simplify and condense this type of data, while illustrating its robustness in an external multimodal dataset and explainability capacities. For 1-Dimensional data, an interpretable model will be developed for understanding cognitive factors using multimodal brain data.

Overall, the research adopted in this thesis explores explainable data-driven representations and modelling approaches across the multidisciplinary scientific fields of machine learning, molecular biology, and neuroscience. It also helps highlight the contributions of these fields when modelling the brain and its intra- and inter-dynamics across the human body.

Acknowledgements

This thesis is an example of how lucky and privileged I think I am: it was conducted in a stimulating environment, in a beautiful city, with such good people. I truly had one of the best experiences of my life! I would never have been able to complete this challenging marathon without the invaluable help of many generous and supportive people, either in my personal or professional life, and sometimes in both. I have therefore to acknowledge many of them.

It goes without saying that I am eternally grateful to my family: my parents, my grandparents, my brothers, my uncle, ... You are my foundations and the fundamental reason anything good exists in my life. I am the person I am today because of you.

Among my friends in Portugal, I must thank those who were always with me online, in the UK, or in Portugal. To my best friends Ana Marta, Diogo Almeida, and Eduardo Fernandes, with whom I have spent countless moments through thick and thin, a big thank you. You are the true meaning of friendship as a second family we choose. Thanks are more-than-obviously due to my big and “old” group of friends for all the friendship and many amazing moments spent together: Cris, Aves, Catarina, Gonçalo, Inês, Luísa, and Perdiz. To my amazing friends Soneca, Alexey, and Filipe, with whom I am so lucky still to keep in touch after a very intense degree together: thank you so much for all the time that shaped me, and for your friendship, despite the distance. I also thank the awesome Filipa Mariz, who always reminds me of the wonderful adventure that life is, and Inês Fantástica, who is always ready for a good catch-up. I am grateful to still keep in touch with Joana, Maria, and Rita, and for the good moments we still have! Finally, my eternal gratitude for Dr. Rosaldo Rossetti’s paternal friendship: you are the one who really made me like research, and directly supported me while applying to Cambridge. I am sure I would never have been admitted if it were not for you.

One of the nice perks of studying at Cambridge is the diversity of people met due to the Collegiate system. Here, I must thank those with whom I shared most of my personal and social life, making the PhD path so much easier and funnier. To Bill Moriarty, I owe thanks for your immense patience in my first year: you surely made my new life in a different country much more enjoyable. To Alex Vetterl, Diana Popescu, and Sai Adloor, thank you for sharing daily life and so many good times. To Dushanth Seevaratnam and Victor Kang, I am indebted for the pleasant time spent in our exciting project. To the bluest Mavi thanks are due for all the energy and support when communication in College was still so difficult to me, and for teaching me how we can be driving forces of change at so many levels; you opened my eyes to issues I never thought I would ever consider when pursuing a PhD degree. A big hug to Jocelino Rodrigues and Miguel El Guendy and to the lovely surprise you both were. To all the other people I met at Churchill College, in the MCR and elsewhere, thanks for all the support and for shaping the experience that this PhD was. Finally, a very special thanks to my X-friend Kaspars Karlsons ($X \in \{$🏠,

🍃, 🌲, ♫, 🎮, 🏋, 🚶, 🍺, 🍴}), who was always there from the very beginning of my life at Cambridge, through the middle, until the very end and beyond: there is so much to thank that I had to summarise in these symbols.

Cambridge is known for its brilliant minds, and I am humbled to have met many talented and approachable people. I have learned so much due to these interactions that I say without a doubt they were vital in forming me throughout these recent years. Personal thanks to Jin Zhu, with whom I shared so many struggles and worries from the beginning. I am grateful for the good moments sometimes spent late in the department learning Mandarin and discussing music theory. 非常感谢你！ Great thanks are due to Giovanna Dimitri and Simeon Spasov, who besides teaching me so much, were the reason for some enjoyable travel experiences. Credit is due to Alessandro Di Stefano, Alex Campbell, and Ramon Viñas for such enriching discussions in which I was so lucky to participate and learn from. Also, thanks to Tim Rittman for the early support and for introducing me to the fascinating field of connectomics. Of course, nothing would be possible without the wider support of the department itself and my group, to whom a "thank you" is not enough. I am truly grateful to have had the chance to work with all these people and receive such excellent advice. I must also thank the students whose projects I was lucky to supervise (some of which I briefly mention in this thesis): Alexandru-Catalin Filip, Kamilė Stankevičiūtė, Renjie Zhou, and Roxana Dumitriu. It was a very gratifying experience to see the projects form and grow, while learning so much from you all.

Huge credit is due to all the exceptional people with whom I collaborated outside the research environment of my department, culminating in some fine research. Therefore I want to acknowledge Carlo Maj, Eric Gamazon, Ivan Merelli, Luca Passamonti, Nicola Toschi, Oleg Borisov, Richard Bethlehem, and Valentina Giansanti. Such a multidisciplinary environment allowed me to humbly grow as a researcher, always knowing that there is so much in this world I will never know.

I would like to acknowledge the W. D. Armstrong Trust Fund for supporting three years of my PhD, and both my department and the Cambridge Philosophical Society for supporting part of my fourth year. Likewise, a word of acknowledgement to my examiners and for the time taken to read this thesis and provide feedback to make it stronger.

I want to leave a special thanks to Partha Maji and (Baroness) Bere, who were especially important for me to finish the last stressful stages of my doctorate studies. To Partha, I am thankful for his patient understanding and support, which allowed me to have time without sacrificing open doors in my career. And to Bere, *muchas gracias* for the love, happy encouragement, and tender care, while bringing me whole(some) new meanings to the idea that other worlds are possible.

Lastly, but certainly not least, my deepest gratitude is owed to my supervisor, Pietro Liò, for the endless networking opportunities, tireless support in everything I ever asked, and the once-in-a-lifetime opportunity to pursue a PhD at such a prestigious institution. It is difficult to count all the doors that were opened to me thanks to you: I can only hope your trust in my capacities was justified.

To everyone: maybe I will never truly return all I have received, but I hope that one day I can give back in some way part of that which I was so generously given. As I think Arthur Ashe once said, "From what we get, we can make a living; what we give, however, makes a life".

Contents

Chapter 1

Introduction

1.1 Complexity in Data Representations

Science has been very successful in advancing our understanding of the world by modelling it across many fields while focusing on distinct orders of magnitude. For instance, it is possible to focus on the astronomical scale (i.e. around 10^8 and 10^{24} metres) all the way down to the subatomic (i.e. below 10^{-18} metres). This fascinating range of scales illustrates the inherent complexity involved when trying to model the world. Finding a good data representation is then paramount to successfully modelling some aspect of science using computational methods. In the field of machine learning this is commonly called *representation learning* [31], and its correct use can greatly affect subsequent learning tasks.

This thesis considers two main ways of looking at specific scales when modelling real-world cases, which affects how information is acquired and represented. The most simple way is when one directly perceives simple features or values; for example, in a human scale this could be the height of a person, or even simple colour perceptions (e.g. this table is *brown*, this mug is *yellow*). Technically, these values are represented in a "flatten" representation or, in other words, in a **1-Dimensional** vector space; we do not make any prior assumptions on the data, and just directly perceive what is observed. However, we can go one step further and perceive more complex patterns beyond these more simple, direct ones. For instance, besides the *brown* table and the *yellow* mug, one can see some *relations*: the mug *is on* the table, and the person *is looking* at the mug. A more typical example of *relations* exists in a social network where, beyond the features we have for each person, we also have friendship relations connecting them. I call this representation a **Graph-Dimensional** space to illustrate the more structured way of representing data.

A Graph-Dimensional space is a consequence of considering a *relational* inductive bias[1], where we try to perceive not just the flatten, simple features but also their arbitrary relations. Relational inductive biases are pervasive in the fields of molecular biology and neuroscience. When observing a structural brain scan (see Section 2.3), one can perceive flatten characteristics, such as the thickness of the prefrontal cortex in the right hemisphere, or its grey volume in the left hemisphere. These two regions may be *related* in different ways: do their neurons fire in a similar pattern across a period of time, or in a non-synchronous way? In the same way, if one takes a blood sample and measures its gene expression profile (see Section 2.2), we can see how specific genes are more or less

[1] An inductive bias is a set of prior assumptions when modelling/representing data.

expressed (i.e. flatten representation); if this expression profile is shared with other parts of the body, then the blood is *related* to those body parts.

George Box famously mentioned in some of his work that “all models are wrong, but some are useful” [42]. It is under this vast and complex real world that this thesis fits: in developing “wrong” models that can still be “useful” in some way. I will narrow my attention to two captivating scales in human brain biology: firstly, the micro scale of molecular biology using gene expression data, and then the macro scale of neuroscience using neuroimaging data.

1.2 Research Questions

The previous section illustrated how representation learning can be influenced by using flatten features (i.e. 1-Dimensional representation) or features with related entities in the form of a graph (i.e. Graph-Dimensional representation). This thesis aims to explore different modelling approaches in brain science using 1-D and Graph-D representations, with a focus on the gene expression and neuroimaging domains. Given this context, there are three main research questions I will seek to explore in this thesis (in parenthesis a single name is added for textual identification):

1. *(Representation)* How can we model and represent very complex and high-dimensional brain data according to specific needs in the applied fields of molecular biology and neuroscience?

2. *(Explanations)* Is it possible to provide models to applied researchers that can provide explanations on how decisions are made, even if learning complex non-linear patterns?

3. *(Graph)* How can we integrate graph-based data in order to better understand neurological and genetic mechanisms of the brain?

The historical and recent successes of machine learning in a multitude of fields leveraging different data structures [86, 153] make this computational subfield the obvious candidate to answer these research questions. It is impossible to develop a one-size-fits-all solution; therefore, I will investigate the suitability of deep learning and more traditional machine learning methods for each applied field in *Research Question 1 (Representation).* In the case of molecular biology, gene expression data will be explored, while I will focus on neuroimaging data in the case of neuroscience. Although in the neuroimaging domain analyses are typically focused on the brain alone, in the gene expression domain we have access to data of other parts of the human body in the same format as those of the brain. Therefore, I will leverage this particular characteristic of gene expression data to allow the development of brain models that take advantage of information in other body tissues.

Using machine learning approaches to explore the intricate non-linear patterns of data could lead to well-known issues in the field. In the past, scientists developed cutting-edge models beating benchmarks without much knowledge of what the model was actually learning [53]. Although these black-box models can be very helpful [140], a lack of understanding of what the model is learning can lead to adversarial attacks [8, 99] or failure to generalise in out-of-distribution data [261]. These issues can have severe consequences for humans when they happen in safety-critical environments and justify the considerable importance of *Research Question 2 (Explanations)* (see Section 2.1.7).

Finally, *Research Question 3 (Graph)* is focused on the specific case of using data in Graph-D space, and what additional mechanistic advantages this space brings when compared with using a flatten data representation (see Section 2.1.8).

All in all, given the inherent complexity of the real-world fields of molecular biology and neuroscience, I aim to develop possible answers through different, though complementary, perspectives. Specifically, I will explore data representations (i.e. in 1-D and Graph-D spaces) and machine learning approaches (i.e. supervised and unsupervised).

1.3 Thesis Structure

I structure this thesis as indicated in Figure 1.1 in order to tackle the three research questions posed in the previous section. This thesis is divided into two main parts. The first, which explores Graph-Dimensional data representations, is comprised of two chapters. The second explores 1-Dimensional data representations, and is comprised of one chapter. The first two research questions will be explored in all these three main chapters, whereas the third research question will be expressly explored only in the first two main chapters.

To help answer *Research Question 1 (Representation)* I will be dealing with distinct challenges in each of the main three chapters. In Chapter 3 I will focus on using a multilayer approach to model co-expression networks across 13 brain tissues and 36 other human tissues, aiming to extract specific and shared genetic profiles that can be generalisable to external datasets. Chapter 4 will introduce a novel deep learning model which can leverage the entirety of resting-state fMRI data (see Section 2.3) as opposed to previous approaches in the literature that simplify and condense this type of data. Finally, in Chapter 5 I will focus on how to extract more interpretable information from brain data using feature engineering to better understand cognitive factors.

I will seek to provide specific insights in all three chapters on what each model learns in order to tackle *Research Question 2 (Explanations)* regarding explainable models. Indeed, in Chapter 3 all the code and information on each community across human tissues is provided so anyone can analyse how each genetic profile is shared across body tissues. Chapter 4 will leverage the clusters formed across samples to understand which patterns emerge from the model, and in Chapter 5 I will provide a ranked list of the most informative brain features for each task at hand.

Finally, *Research Question 3 (Graph)* is explored only in chapters 3 and 4. In these chapters, data are represented in a Graph-D space, that is, modelled as graphs. Indeed, in the former Chapter 3, I adopt a multilayer approach to represent the relations of co-expression networks (i.e. graphs), while in the latter Chapter 4 I propose a deep learning architecture which is able to specifically leverage the graph structure of the data.

It is a particularly exciting time to research on these topics [65, 74, 81, 287], given the high quality and well-curated datasets being released to use by researchers. During this thesis development, I had access to almost a thousand neuroimaging scans explored in chapters 4 and 5, and more than 35 thousand neuroimaging scans in Chapter 4 (see Section 2.3.3). The last release of GTEx dataset was publicly released close to the end of this thesis [4]; indeed, I was able to explore this last version in Chapter 3. Further details of all the datasets used in this thesis, as illustrated in Figure 1.1, are provided in each chapter.

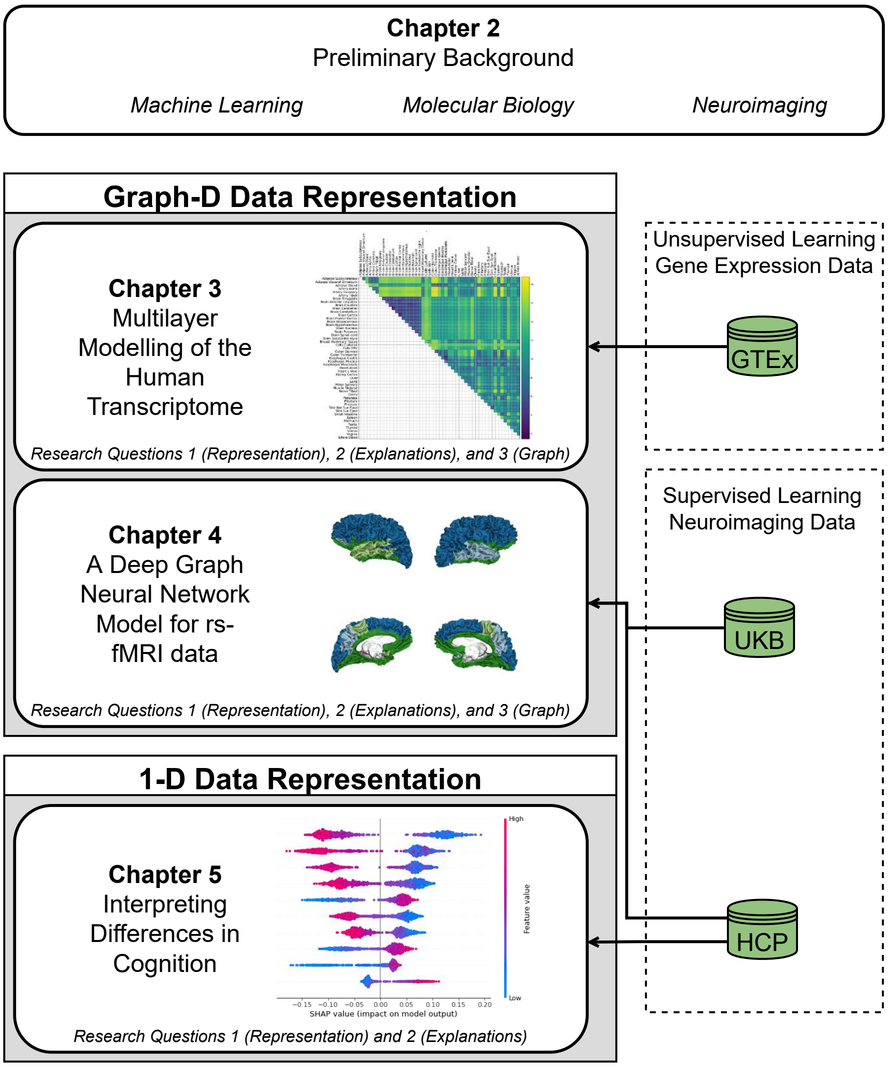

Chapter 2
Preliminary Background
Machine Learning
Molecular Biology
Neuroimaging
Graph-D Data Representation
Chapter 3
Multilayer Modelling of the Human Transcriptome
Research Questions 1 (Representation), 2 (Explanations), and 3 (Graph)
Chapter 4
A Deep Graph Neural Network Model for rs-fMRI data
Research Questions 1 (Representation), 2 (Explanations), and 3 (Graph)
1-D Data Representation
Chapter 5
Interpreting Differences in Cognition
Research Questions 1 (Representation) and 2 (Explanations)
Unsupervised Learning
Gene Expression Data
GTEx
Supervised Learning
Neuroimaging Data
UKB
HCP
High
Feature value
Low
SHAP value (impact on model output)

Figure 1.1: Thesis structure for the main chapters. Different perspectives will be explored and integrated across distinct data representations (i.e. 1-D and Graph-D spaces), machine learning approaches (i.e. supervised and unsupervised), data fields (i.e. molecular biology and neuroscience), and respective main datasets comprised of gene expression and neuroimaging data. The research questions tackled in each chapter are shown, and explained in Section 1.2. Each dataset is presented in detail in its respective chapter. Chapter 2 provides the fundamental concepts from machine learning, molecular biology, and neuroimaging explored in the main chapters.

Further to the three main chapters of this thesis, in Chapter 2 I will outline an overview of the topics explored in this thesis. Specifically, I will summarise fundamental concepts on machine learning training procedures, as well as background concepts in molecular biology and neuroimaging. In Chapter 6 I will highlight the main limitations and contributions of this thesis and offer my thoughts on interesting future directions of research stemming from its limitations.

1.4 Related Publications

The following is a list of relevant publications that convey part of the contributions described in this thesis, in which I was the first author:

[21] Azevedo, T.*, Dimitri, G. M.*, Liò, P., & Gamazon, E. R. (2021). ***Multilayer modelling and analysis of the human transcriptome***. npj Systems Biology and Applications.

[23] Azevedo, T., Campbell, A., Romero-Garcia, R., Passamonti, L., Bethlehem, R. A. I., Liò, P., & Toschi, N. (2022). ***A Deep Graph Neural Network Architecture for Modelling Spatio-temporal Dynamics in resting-state functional MRI Data***. Medical Image Analysis.
Preliminary approaches of this work were also presented at the 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) [20] and at the ICLR 2020 Workshop on AI for Affordable Healthcare (AI4AH) [19].

[18] Azevedo, T., Passamonti, L., Liò, P., & Toschi, N. (2019). ***A Machine Learning Tool for Interpreting Differences in Cognition Using Brain Features***. IFIP Advances in Information and Communication Technology.

In all papers, I contributed to the majority of ideas, text, and analysis. Alex Campbell and Giovanna Dimitri contributed with further data analysis and useful discussions of the right conceptualisation and methodology to follow. Richard A. I. Bethlehem and Rafael Romero-Garcia provided very useful domain-specific insights and preprocessed data from the UK Biobank dataset. Pietro Liò, Eric R. Gamazon, Luca Passamonti, and Nicola Toschi played a critical advisory role in all papers they have co-authored. Eric R. Gamazon and Nicola Toschi further provided access to preprocessed data from GTEx/TCGA and HCP, respectively.

I have also contributed to other papers closely related to the theme of this thesis:

[97] Filip, A.-C., Azevedo, T., Passamonti, L., Toschi, N., & Liò, P. (2020). ***A novel Graph Attention Network Architecture for modeling multimodal brain connectivity***. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).

[254] Stankevičiūtė, K., Azevedo, T., Campbell, A., Bethlehem, R. A. I., & Liò, P. (2020). ***Population Graph GNNs for Brain Age Prediction***. ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+).

The papers by Alexandru-Catalin Filip and Kamilė Stankevičiūtė are abridged versions of their Part II dissertations at the Department of Computer Science and Technology at the University of Cambridge, developed under my supervision. The remaining authors provided the same type of contribution as for the previous papers co-authored with me.

Whenever possible, I aimed to publicly release the source code with the papers, either by helping in shared repositories or by releasing them on my personal GitHub page: `https://github.com/tjiagoM`.

Chapter 2

Preliminary Background

The research questions posed in Section 1.2 centre on machine learning applications in the molecular biology and neuroimaging domains. Accordingly, in this chapter I will provide an introduction to essential machine learning techniques used in this thesis. I will also offer an overview of fundamental molecular biology and neuroimaging concepts used in this thesis.

2.1 Machine Learning Fundamentals

Since the identification in October 2012 [213] of data scientist as the “sexiest [sic] job of the 21st century”, the hype of related fields like machine learning still exists. Indeed, machine learning has been, and still is, an active field of research with clear implications for society. Many try to pinpoint reasons for such successful and still ongoing advancements: artificial intelligence is outpacing Moore’s Law in a very significant way, increased amounts of big data are being publicly released, storage is becoming cheaper every year, and more sophisticated machine learning algorithms are still being proposed every month. Machine learning is entering almost every field of research, making it the most popular computational tool in applied artificial intelligence, despite known concerns on whether some advances are real [144].

Programming languages are maturing and diversifying fast, allowing virtually everyone in the world to develop and try their own ideas. For example, in a report by *Stack Overflow* in 2017 [227], they claim that “Python has a solid claim to being the fastest-growing major programming language”, with a clear growing pattern not only in wealthy but also in lower-income countries. Python is a scripting language with a vast online supportive community and comprehensive documentation for technical and non-technical people alike. This makes it easier for people without a computer science background to use it in their own non-computational subfield. The field of machine learning follows these trends: we see a matured scientific python-based ecosystem with core numeric libraries (e.g. *NumPy* and *SciPy*), advanced interactive environments (e.g. *IPython* and Jupyter notebooks), and domain-specific packages (e.g. *statsmodels* for statistics, *pandas* for tabular data structures, *scikit-learn* for machine learning, and *TensorFlow*/*PyTorch* for deep learning). The main programming language used in all chapters of this thesis was therefore Python, given all these advantages.

A single definition of machine learning is difficult, especially with such closely related fields such as statistics and data science. As the name indicates, however, I consider machine learning to be a field that tries to make machines/computers to learn something. Deep learning is thus a subfield of machine learning, and machine learning is a subfield of artificial intelligence. All these subfields make effective use of statistical techniques. As a distinctive feature, machine learning changes how one thinks about a problem: instead of thinking logically as a software developer, the focus is shifted to think more like a natural scientist. That is, one makes observations and experiments in which the algorithm is not told exactly which rules to follow, but rather to find patterns from the examples given by the scientist. This shift can reduce the programming time to develop a successful machine learning model.

This section will describe fundamental concepts on how to train machine learning models, thus providing the general framework adopted throughout this thesis.

2.1.1 Types of Learning

As previously mentioned, there is no single definition of machine learning, apart from its consensual inclusion of some type of "learning" - as indicated in its name. Even though authors may consider a variety of learning frameworks, in the machine learning field it is generally accepted that when a fixed dataset is present, a machine learning model can broadly be defined as **supervised** or **unsupervised**, depending on how the learning occurs. Indeed, I employ unsupervised learning in Chapter 3, and supervised learning in chapters 4 and 5. There are four important concepts necessary for understanding these two learning types:

- **Feature**: a measurable characteristic of something, which can be generally represented as a vector $\boldsymbol{x} \in \mathbb{R}^N$ with N elements: $x_1, \ldots x_N$.
- **Label**: another measurable characteristic, which instead is predicted by an algorithm, usually represented as $\boldsymbol{y}$.
- **Sample/Example**: a particular instance of data, which can be represented by a tuple $(\boldsymbol{x}, \boldsymbol{y})$ for labelled data, and only $\boldsymbol{x}$ for unlabelled data.
- **Dataset**: a set of samples.

In unsupervised learning the dataset contains unlabelled features, and a model tries to find patterns without any knowledge of the ground truth. In other words, it tries to find the complex probability distribution that generated the features of the dataset. In supervised learning, however, each sample contains a ground-truth label (or labels) which will guide/supervise the learning algorithm; strictly speaking, the model tries to learn how to predict a particular label $\boldsymbol{y}$ from features $\boldsymbol{x}$ by estimating the probability distribution $p(\boldsymbol{y}|\boldsymbol{x})$.

It is worth noting that despite these precise definitions, the two concepts are sometimes mixed, and we can define other types of learning. For example, it is possible to use an unsupervised model in a pipeline together with a supervised model [184]; in this case, there is not a single type of learning involved, but rather a combination of both types. Another popular type of learning is **reinforcement learning**, in which the dataset is dynamic and the machine learning algorithm needs to interact with an environment to receive constant rewards from its actions.

Although supervised, unsupervised, and reinforcement learning are very famous types of learning in the field, they still do not answer what *learning* actually is. To be more precise, Mitchell [192] tries to define *learning* by stating that "a computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". Goodfellow et al. [124] further clarify: "learning is our means of attaining the ability to perform the task. For example, if we want a robot to be able to walk, then walking is the task. We could program the robot to learn to walk, or we could attempt to directly write a program that specifies how to walk manually".

From this very broad definition it is possible to see that supervised, unsupervised, and reinforcement learning depends on the specific type of experience E that is used during the learning process. The way experience E is acquired will then influence what performance measure P is needed. However, exploring this is beyond the scope of this section.

There are many possible tasks T that can be tackled by machine learning models. Two frequent tasks are **classification** and **regression**, and both correspond to supervised learning. In both cases a model tries to learn a target function f_θ, parametrised by θ, such that it can predict $\boldsymbol{y} = f_\theta\left(\boldsymbol{x}\right)$. In the case of classification, the function is such that $f : \mathbb{R}^N \rightarrow \{1, \ldots, k\}$, where k is the number of possible categories (i.e. labels); for example, for a certain image with pixels represented in a feature vector $\boldsymbol{x}$, this function could try to predict whether it contains a dog, a cat, or a bird. In the case of regression the function is such that $f : \mathbb{R}^N \rightarrow \mathbb{R}$: that is, the output is a continuous value; for example, one can try to predict the economic value of a house from its characteristics. Depending on the dataset, some parts of the feature vector could be missing, mandating other models that can expect missing values.

As highlighted by Goodfellow et al. [124], there are some other common tasks which are not directly explored in this thesis. For instance: (1) **machine translation**, where the model receives a sequence of symbols in a certain language and tries to translate it into a sequence of symbols in another language; (2) **anomaly detection**, where the model predicts whether a certain sample is atypical of a situation, as when detecting credit card fraud; (3) **data synthesis**, in which the model learns how to synthesise (i.e. generate) new artificial samples based on those existent in the dataset; (4) **imputation of missing values** to try to predict missing values on $\boldsymbol{x}$ based on the remaining existent values of that vector; (5) **denoising** where the algorithm is provided with a corrupted example $\tilde{\boldsymbol{x}}$ and tries to transform it into a "clean" feature vector $\boldsymbol{x}$. The possibilities are endless, and the reader could certainly find other examples in the literature. Note how all these tasks T could be performed with just one type of learning or a combination of them.

2.1.2 Training Procedure

There are many types of models in the field of machine learning, as it will be illustrated in this thesis. Even though this variability exists, there are ubiquitous concepts when making a model to learn. In the previous section I introduced the concept of "performance measure P"; the way we can correctly measure it, directly affects how the training procedure should be conducted.

To train a model or, in other words, to *fit* a model to a particular dataset, one needs to tune the learnable parameters of that model so that a performance measure P is the best possible. The typical scenario includes training data that are "fed" to a model so that it can learn to perform a certain task. There are then some holdout test data (e.g. a completely different dataset) with which the model is evaluated. I should highlight that this seemingly simple process of training and evaluating the model in different datasets is fundamental to avoid optimistic (i.e. biased) performance measures. Indeed, if one trains and evaluates a model in the same set of samples, we cannot know for sure whether the model has learned or simply memorised the data. Therefore, the interest is to determine whether a model can perform well (i.e. generalise) with unseen data.

If only a single dataset is available, there is the need to artificially split it into training and test sets. Random sampling is a straightforward way to perform such splitting; however, when the dataset is known to contain samples belonging to different categories, randomly sampling might end up with train/test sets with a category distribution distinct from that of the original. For instance, if a dataset contains equal amounts of small, medium, and tall people (i.e. around 33% for each category), randomly sampling could create a training set with different category ratios. In such a case a training set could have 25% / 35% / 40% of small, medium, and tall people, respectively, and a test set could have ratios of 35% / 25% / 40% instead. This means that the two sets of data have not only different category ratios from the original dataset, but also opposite ratios: the training set has fewer small than medium people, and the test set has more small than medium people. By contrast, a **stratified** approach randomly splits a dataset so that the category ratios in the two sets are roughly the same; this will also guarantee that the ratios are representative of the original dataset. This stratification strategy was used throughout this thesis because real-world datasets tend to have imbalanced categories, and therefore this strategy can provide more confident evaluation quantifications.

Now that these different strategies for splitting a dataset into train/test sets are presented, defining a method to train a model using such sets is important. The *holdout validation method* is frequently used and, as described by Raschka [224], it contains four main steps to train and evaluate a machine learning model: (1) divide the dataset into training and test sets which is needed to "work-around for dealing with the imperfections of a non-ideal world, such as limited data and resources, and the inability to collect more data from the generating distribution" [224], (2) pick an appropriate model and fit it to the training data, while manually specifying its hyperparameters, (3) evaluate the model's behaviour with unseen data (i.e. a test set), which is assumed to provide an unbiased estimate of good performance, and (4) use the entire dataset (instead of just the training set) to fit the model again, after proving that the learned model is capable of generalisation; this step is optional and more useful when one needs a final model to be deployed in practice.

2.1.3 Hyperparameter Tuning and Cross-Validation

When one is developing a machine learning model, some settings need to be specified. These are formally called *hyperparameters* or, more informally, knobs, as they usually need to be tweaked while optimising a model. In the case of neural networks, these hyperparameters may be the number of layers or nodes used; in the case of random forests, this may be the number of trees used. I have previously focused on the need of a holdout set (i.e. test set) to obtain a final, unbiased estimate of performance for a specific model. This section will focus on how to include *hyperparameter optimisation* in the training process, sometimes also called *hyperparameter tuning.*

For this extra optimisation task one also needs to use some performance metric, which can be the same as the metric used for the final evaluation on the test set. The objective is then to select a model with the hyperparameters that give the best final performance metric. However, reusing the test set repeatedly to evaluate different hyperparameters could introduce an overly optimistic estimate on the generalisation performance. To avoid this issue it is necessary to introduce a third set, commonly called *validation set*, of roughly the same size, or slightly smaller than the test set. This will allow us to keep a single test set to be used only once for final evaluation measurements.

With this in mind, Raschka [224] extends the previous *holdout validation method* to include hyperparameter optimisation, in what he calls a *three-way holdout method.* I represent its five main steps in Figure 2.1, with modifications introduced for clarity and completeness. Among other details, the possible need of a validation set by the learning algorithm in *step 2* is not included in the original paper, whereas this is the default case in this thesis; also, the original paper defines six steps, whereas here I simplify them to only five.

The five main steps of the *three-way holdout method* are the following:

1. Divide the dataset into training, validation, and test sets. These will be used respectively for model fitting, model selection, and final evaluation of the selected model. This step differs from *step 1* in the *holdout validation method* as it creates a further validation set.

2. Conduct the actual hyperparameter tuning by fitting different models to the training data, each model with a different set of hyperparameters. This step will result in multiple models and performance estimates, and the model with the best performance metric is chosen for the next step. Given sometimes an infinite choice of hyperparameters exists, I will explore in the next section different ways to determine which hyperparameter to choose in this step. Depending on the model, sometimes the validation set might also be needed by the learning algorithm, rather than just used for performance evaluation.

3. If the validation set was only used for evaluation in the previous step, refit the model on both the training and validation sets to avoid a pessimistic bias due to a smaller amount of data. This refit uses the hyperparameters that yielded the best performance metric in the previous step. This step is often ignored in the literature mostly because many models need a separate validation set during the training procedure which cannot be merged with the training data. For instance, in deep learning models, the validation set is needed to know when to stop the learning process.

4. The selected model is evaluated on the independent test set for estimation of generalisation performance. This corresponds to the same *step 3* in the *holdout validation method.*

5. Optionally, use the entire dataset to refit the model on the maximum number of samples available and select that model for deployment. This step will likely yield a model with different learned parameters from the model evaluated in the previous step but, in theory, using more informative data should improve model performance. On this assumption, the performance evaluated in *step 4* is an underestimation of the actual performance of the model in this step; however, as there are no holdout data left, only by monitoring the deployed model in the real world it will be possible

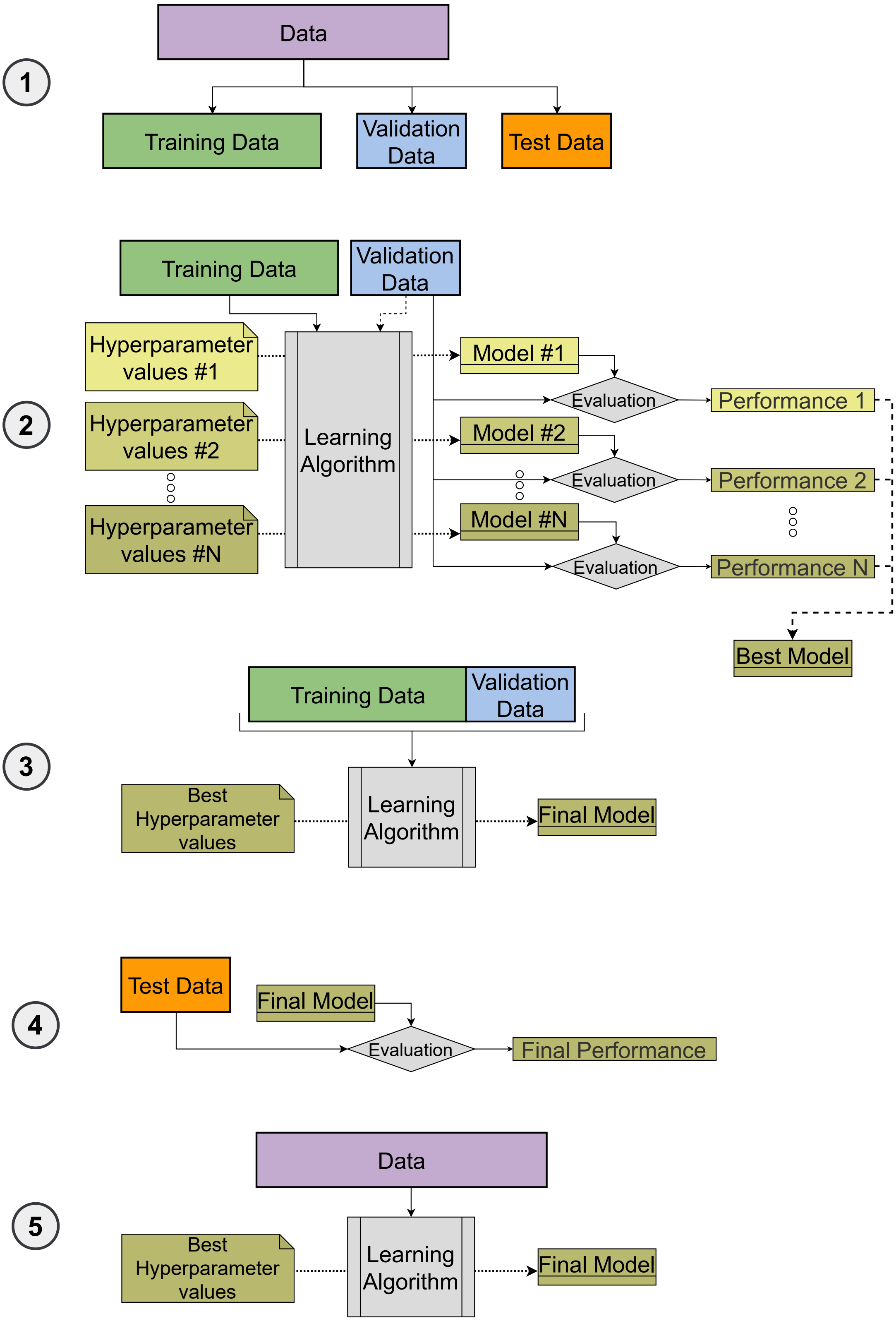


Figure 2.1: The five steps for hyperparameter tuning using the *three-way holdout method*, where each number corresponds to one step: dataset division, hyperparameter tuning, model refit, test set evaluation, and final model refit. Each step is explained in more detail in Section 2.1.3, and adapted from Raschka [224].

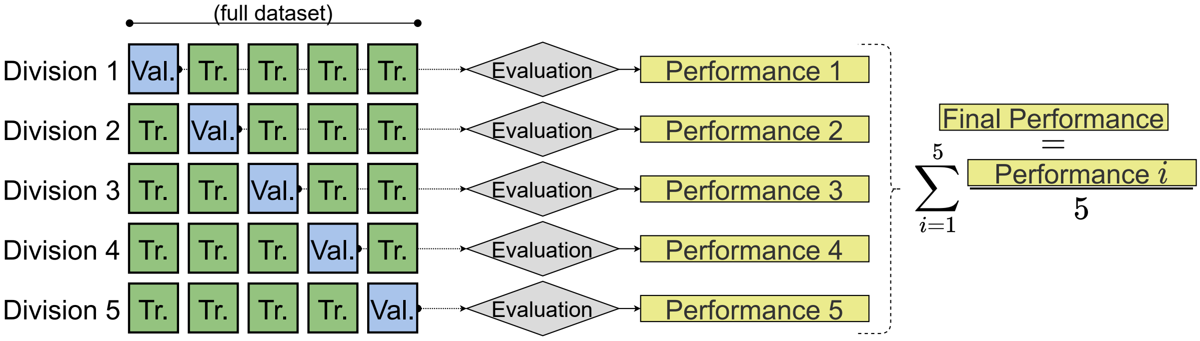


Figure 2.2: Data division into folds in a k-fold cross-validation procedure for $k = 5$. At each division one fold is used as a validation set, while the remaining folds are used for training. Performance is then averaged across all the performances.

to practically evaluate its performance. This step corresponds to the final step in the *holdout validation method*.

Raschka [224] presents the *three-way holdout method* for the particular case of supervised learning by highlighting the need for labels in all the steps. However, I argue that this method is general enough for any type of learning, and have adapted the original steps accordingly in Figure 2.1. Due to their nature, learning types such as the unsupervised one do not contain ground truth labels to evaluate performance directly. Still, some evaluation of the model is needed to validate the choice of hyperparameters or simply to understand how successful the learning process was. The key here is to consider that evaluating the performance of the model could go beyond systematic quantitative metrics: it can include visual checks (e.g. subjectively evaluating how clusters from clustering algorithms look in a 2D embedding), checking the evolution of an objective function as a proxy for performance, or evaluating the accumulated reward in a reinforcement learning task.

One way to frame steps 2 and 3 in the *three-way holdout method* is to think of them as a function f_M that for a given set of possible hyperparameters $\boldsymbol{H}$ and a dataset $\boldsymbol{D}$, outputs the best model and respective best hyperparameters:

$$\text{model}, \text{hyperparameters} = f_M\left(\boldsymbol{H}, \boldsymbol{D}\right). \tag{2.1}$$

In the case of the *three-way holdout method* this function f_M only considers a single training/validation split, but one could use a **cross-validation** procedure instead. This term has loose and varied semantics in the literature, being used not only for hyperparameter tuning and model selection but also to report a more unbiased performance estimation.

A type of cross-validation is the k-fold cross-validation, where one divides a dataset into k non-overlapping folds. At each division, one fold is used as a validation set, and the remaining $k - 1$ folds are used as a training set. Figure 2.2 illustrates the process for a 5-fold cross-validation, in which performance is averaged across the 5 non-overlapping validation sets.

The k-fold cross-validation procedure can be directly applied inside f_M for the same type of input and output. For each set of hyperparameter values, the learning algorithm trains a model k times for each different split of training/validation sets. This means that the performance reported for a model with a certain set of hyperparameter values will now be an average of these different folds. At the end this process avoids an indirect overly optimistic selection of hyperparameter values for a specific validation set; instead, it finds the best hyperparameters that generalise better for different data splits rather than for a single validation set.

It is common in machine learning literature to use cross-validation to avoid a performance evaluation in a single test set. In smaller datasets, it is not clear whether a single test set could have a *lucky* set of samples that produce an overly optimistic performance measure. Therefore, cross-validation can be used to define k non-overlapping test sets, and repeat the *holdout validation method* or the *three-way holdout method* for each k test set. The final performance measure estimation is averaged across these k test folds.

Using cross-validation for hyperparameter tuning or providing an averaged final performance brings a noticeable overhead in computational costs, as more models need to be trained and evaluated. Accordingly, this is usually conducted only for smaller datasets where there is an increased likelihood of producing overly optimistic evaluation metrics.

As it is possible to see, despite the common concepts needed for training and evaluating a machine learning model, the details still need to be adapted for each particular need. Indeed, in this thesis each chapter has slightly different training procedures. For instance, Chapter 4 has used a straightforward *three-way holdout method* with cross-validation for final performance evaluation. On the other hand, in Chapter 5, due to the small and very high-dimensional dataset, I have used a variant called nested cross-validation, where cross-validation is used both for hyperparameter tuning and final performance evaluation.

2.1.4 Hyperparameter Tuning Methods

As it was mentioned in the previous section, each distinct set of hyperparameters will require a model to be fit to the training data, thus resulting in multiple performance estimates. In this section I will describe which methods are typically used to conduct hyperparameter tuning, given it is not practical - and most of the times not possible - to try every combination. It should be highlighted that there are other methods in the literature with a certain level of complexity (e.g. genetic algorithms) but that are less used in practice for common hyperparameter tuning procedures to fine-tune a single model.

Manual search consists in manually choosing and experimenting with different hyperparameter combinations. Typically, an initial choice is set based on judgement and experience, and at each iteration a new, subjective choice is made according to the previous results. This loop is repeated for as long as possible until a satisfactory metric is scored. This is a tedious process and not used in practice when the search space (i.e. the set of possible hyperparameter values) is too large.

Grid search is perhaps the most basic hyperparameter tuning method commonly used. As the name indicates, a grid of possible hyperparameter values is chosen, and the model is trained using every combination in that grid. In this sense, the choice of the values in the grid is left to the person implementing the tuning procedure. When a big number of values is used in the grid, this method can cover a thorough amount of the search space and be quite inefficient as a consequence. In practice, most of the times the search space is infinite thus this search is prone to biases when implementing it (e.g. the grid choice might not capture the best combination). This method still contains some level of manual choice like with manual search, but to a much lesser extent given the choices are made *a priori*.

Random search differs from grid search as a discrete set of values is no longer required; instead, each hyperparameter is defined as a statistical distribution from which it can be randomly sampled from. If a hyperparameter contains categorical values, then each value is sampled with equal probability. A notorious advantage of random search when compared to grid search is its flexibility: the number of iterations can be decided based on time or desired number of combinations instead of a fixed grid. Another key advantage of random search is that even if the optimal value of hyperparameters is in a grid, random search will usually find a "close-enough" solution in far fewer iterations, making this method significantly efficient; this happens because grid search can spend considerable time evaluating unpromising regions of the search space. In a famous seminal work by Bergstra and Bengio [33], it is shown that random search can sometimes even find better values than when using grid search, and that "for most datasets only a few of the hyperparameters really matter, but that different hyperparameters are important on different datasets". Figure 2.3 illustrates these issues in a simple scenario; however, the assumption that not all hyperparameters are equally important holds true for most datasets.

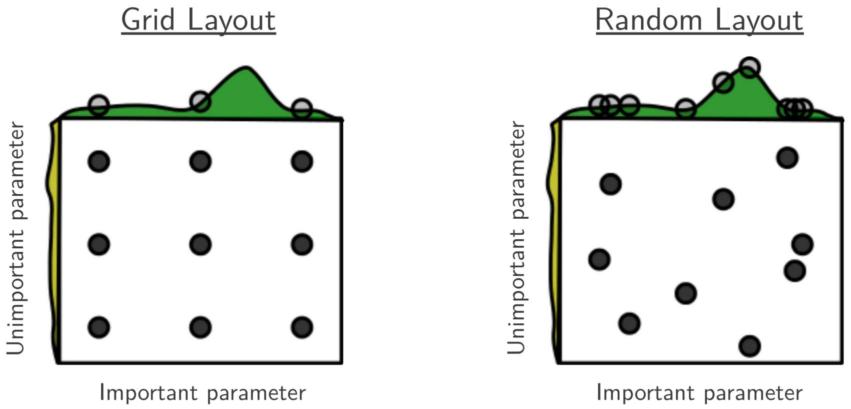


Figure 2.3: Grid and random search space when optimising a function $f(x_1, x_2) = g(x_1) + h(x_2) \approx g(x_1)$, where $g(x_1)$ is shown in green above and $h(x_2)$ in yellow on the left side of each square. With grid search, g is only tested in three distinct places while with random search it is explored with more distinct values. Image taken from Bergstra and Bengio [33].

Bayesian optimisation uses Bayesian theory to optimise objective functions that take a long time to evaluate. This method offers a principled approach to weight the importance of hyperparameter values such that the results of a previous iteration can be used to improve the sampling method for the next choice of values. This process starts by training a model with a specific configuration (i.e. set of hyperparameter values) which will have a score based on some metric. Then, for a probability model $P(\text{score}|\text{configuration})$, one could use Gaussian processes to model the prior probability of model scores across the search space [116, 247][1]. This means that the previously evaluated configuration is used to compute a posterior expectation of the search space, which in turn can be used to sample more informed hyperparameter values. This process is run iteratively until a

[1] Other regression models such as decision trees [145] and even neural networks [248] can also be used, but explanation of either cases is beyond the scope of this thesis.

certain threshold is met. A key constrain of this method is that it belongs to a class of sequential model-based optimisation (SMBO) algorithms, meaning that in order to search a combination of N hyperparameter values, one needs to run them sequentially. In contrast, with random or grid search one can launch those evaluations in parallel in a server, making the search more efficient in these later cases with similar resulting metrics. This constrain was the main reason why Bayesian optimisation was not used in this thesis.

2.1.5 Supervised Learning

In this section I will introduce two supervised learning methods which will be used or adapted in chapters 4 and 5.

2.1.5.1 Support Vector Machine (SVM)

The support vector machine (SVM) [39] is a supervised learning technique that is typically used for classification tasks. In its standard binary classification form, it maps the data into a higher-dimensional space where the two classes can be separated by a hyperplane. The goal of SVM is then to maximise the gap (usually called functional margin) separating the closest pair of data samples from the hyperplane. These closest points are called the support vectors (therefore the name of the method), because they are the data observations that "support" (i.e. determine) the decision boundary between the two classes.

For a training set with samples $x_i \in \mathbb{R}^d, i = \{1, 2, \ldots, N\}$ and corresponding labels $y_i \in \{-1, 1\}$, this training set can be separable in feature space if there is a vector $\boldsymbol{w}$ and scalar b such that[2]:

$$y_i(\boldsymbol{w}^T \phi(x_i) + b) \geq 1, \quad \forall i \in \{1, 2, \ldots, N\} \tag{2.2}$$

where $\phi : \mathbb{R}^d \to \mathbb{R}^F$ denotes a fixed feature-space transformation that maps the d-dimensional inputs to a F-dimensional feature space. By definition, there will be at least one data sample in which the equality $y_i(\boldsymbol{w}^T \phi(x_i) + b) = 1$ holds true, and for those samples whose the equality holds true, they are the *support vectors.*

The optimisation problem requires the minimisation of the following optimisation problem:

$$\underset{\boldsymbol{w},b}{\text{argmin}} \frac{1}{2} \|\boldsymbol{w}\|^2, \tag{2.3}$$

subject to the constrain given in Equation 2.2. In real applications such perfect separation does not exist, thus, to take into account misclassifications, it is necessary to introduce a penalty term (the so-called *slack variable* ξ_i) for each data sample such that

$$\xi_i = \begin{cases} 0, & \text{if } x_i \text{ is correctly classified} \\ \left| y_i - \left(\boldsymbol{w}^T \phi(x_i) + b \right) \right|, & \text{otherwise.} \end{cases} \tag{2.4}$$

[2] It is sometimes seen in the literature $y_i \in \{0, 1\}$, which would require some changes in the optimisation rules, but still producing equivalent results in practice.

This penalty term is therefore a linear representation of the distance of the data sample to the decision boundary. The optimisation problem in Equation 2.3 now turns into:

$$\underset{\boldsymbol{w},b,\boldsymbol{\xi}}{\operatorname{argmin}} \left(\frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{N} \xi_i \right), \tag{2.5}$$

subject to the constrain $y_i(\boldsymbol{w}^T \phi(x_i) + b) \geq 1 - \xi_i$, and where C is a parameter controlling the trade-off for the penalisation of misclassifications.

Higher dimensional transformations can allow SVM to separate data in higher-dimensional space, whose transformation is defined based on the *support vectors*. The decision function for the classification problem of an unlabelled vector x_i can be given by

$$sign \left(\sum_{m \in S} y_m \alpha_m \phi(x_m) \phi(x_i) + b \right), \tag{2.6}$$

where S denotes the indices of the support vectors and α_m are coefficients previously determined by the SVM algorithm. This would mean that it would have to perform operations with the higher dimensional vectors in the transformed feature space, which could lead to impractical computational costs. Another way to tackle this problem is to use the *kernel trick*. The "trick" is that kernel methods are able to represent the data in terms of pairwise similarity comparisons between the data samples in the original d-dimensional space without the need to explicitly apply the transformation $\phi(x_i)$ mentioned above. For a dataset with N samples, we would represent a kernel matrix of size $N \times N$ where each element (i, j) is calculated by a kernel function that is defined as:

$$K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}, \quad s.t. \quad K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j). \tag{2.7}$$

The three most commonly used kernels are:

- **Linear kernel:** $K(x_i, x_j) = x_i^T x_j$, representing a simple dot product. Combining equations 2.6 and 2.7, it follows that with this kernel, $\phi(x_i) = x_i$.

- **Polynomial kernel:** $K(x_i, x_j) = \left(\gamma x_i^T x_j + r\right)^q$, where γ, r, and q are kernel parameters in which q represents the order of the polynomial and $\gamma > 0$.

- **Radial basis kernel (RBF):** $K(x_i, x_j) = \exp(-\gamma |x_i - x_j|^2)$, where $\gamma > 0$ is a kernel parameter, and $|.|$ denotes the Euclidean distance. This is sometimes called the Gaussian kernel.

All in all, the *kernel trick* allows the SVM optimisation to find an optimal higher-dimensional hyperplane without the need to directly find the function $\phi()$. For more details on SVM and on how this is calculated see, for example, chapter 7 of Bishop [36]'s book.

2.1.5.2 Extreme Gradient Boosting (XGBoost)

The eXtreme Gradient Boosting (XGBoost) [57] is an open-source optimised gradient boosting library. It was originally developed in C++, but soon it received implementations in many languages such as Python and R, while starting to be widely popular given its successes in top Kaggle competitions[3].

[3] https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions

A key characteristic of XGBoost is, as the name indicates, *boosting. Boosting* allows to build a single model by iteratively adding individual ones (so-called *weak learners*), in practice making it an ensemble model. The idea is therefore to learn from past mistakes by focusing the learning on the difficult cases or, in other words, focusing on correcting the mistakes from previous iterations. This is different, for example, from a standard ensemble model in which if all models are trained in separate, all of them might make the same mistake.

Closely following the notation from the original paper [57], let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ represent a dataset with pairs of features $\boldsymbol{x}_i \in \mathbb{R}^m$ (with m features), and labels $y_i \in \mathbb{R}$ if a regression task in present, or $y_i \in \mathbb{R}^c$ for a classification task with c categories. An XGBoost model is built using an ensemble of "Classification And Regression Trees" (CARTs) [43] which are represented as adapted decision trees where each leaf (i.e., end node) will contain a continuous score based on the decisions of the internal nodes leading to each leaf.

Formally, for a tree ensemble model, the output for an input $\boldsymbol{x}_i$ is calculated as the (additive) summation of the output of K functions (i.e., CARTs):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i), \tag{2.8}$$

where each function $f_k(\boldsymbol{x})$ corresponds to a CART in the ensemble model. Specifically, $f_k(\boldsymbol{x}) = w_{q(\boldsymbol{x})}$, where w_j is the continuous score on the j-th leaf at the respective tree, and q maps an input to the corresponding leaf index such that, for a tree with T leaves, we have $q : \mathbb{R}^m \to T$ and $w \in \mathbb{R}^T$. In this sense, q represents the decision rules in a tree mapping the input to the corresponding score.

The following regularised objective is minimised to learn the set of trees in the model:

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{2.9}$$

where l is a differentiable loss function and $\Omega(f)$ is the regularisation term that penalises complex trees. This regularisation term is defined for XGBoost as $\gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$, where λ and γ are hyperparameters.

Given we are in the presence of gradient boosting, the tree ensemble model is optimised in an additive fashion; therefore, in order to optimise Equation 2.9, some modifications need to be introduced as learning all trees at once is intractable. This additive strategy is illustrated in the following iterative way, where at each step t we fix what has been learned so far, and add one tree at a time:

$$\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(\boldsymbol{x}_i) = \hat{y}_i^{(0)} + f_1(\boldsymbol{x}_i) \\
&\dots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(\boldsymbol{x}_i) = \hat{y}_i^{(t-1)} + f_t(\boldsymbol{x}_i),
\end{aligned} \tag{2.10}$$

where the first step contains a dummy function that predicts 0 for all inputs. Given this iterative representation, and using the Taylor expansion up to the second order [106], the loss at step t is approximated to the following objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{N} [g_i f_t(\boldsymbol{x}_i) + \frac{1}{2} h_i f_t^2(\boldsymbol{x}_i)] + \Omega(f_t), \tag{2.11}$$

where $g_i = \partial_{\hat{y}_i^{(t-1)}} l \left(y_i, \hat{y}_i^{(t-1)} \right)$ and $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l \left(y_i, \hat{y}_i^{(t-1)} \right)$ are the first and second order gradients of the loss function at that step t, respectively. This definition allows the usage of different loss functions as only g_i and h_i are needed as inputs to the solver.

With Equation 2.11, it is possible to evaluate the quality of a tree at a specific timestep; however, it is still needed to decide how to construct a new tree at each step. Ideally, all tree structures q would be enumerated and evaluated but obviously that is not tractable. The solution is a greedy algorithm that starts from a single leaf and splits it into two leaves iteratively. Let $I_j = \{i | q(\boldsymbol{x}_i) = j\}$ be the set of indices of data points assigned to leaf j. Also let I_L, I_R be the instances sets of the left and right leaves after a split. By letting $I = I_L \cup I_R$, Chen and Guestrin [57] defined the loss reduction after a split (i.e. the gain) as:

$$\text{Gain} = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G_I^2}{H_I + \lambda} \right] - \gamma, \tag{2.12}$$

where $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$. Intuitively, it can be seen that $G_L^2/(H_L + \lambda)$ corresponds to the quality of the left tree structure; using the same intuition in the remaining equation, it is possible to see that if the gain is not big enough (dependent on γ), it is better not to add that split. This process corresponds to the pruning techniques used in decision tree-based models.

There are two further techniques used in XGBoost that are worth mention in this section: shrinkage and subsampling. Shrinkage scales newly added weights by a factor (tunable as a hyperparameter), after each step; in practice, it reduces the influence of each tree hence allowing future trees to improve the final tree ensemble model. Subsampling can come in two facets: (1) subsampling training instances at each step, and (2) subsampling features. Both subsampling techniques can not only prevent overfitting, but also speed-up computation.

Despite following the same principles as gradient boosting machines (GBMs), there are some details that separate XGBoost from other GBM implementations which explain why it has been such a popular model. When it was introduced, XGBoost provided a set of technical features that no other GBM had implemented altogether in the same tool. Some included out-of-core computation (i.e., when data is too large to fit entirely in memory) and sparsity-aware split finding. A key characteristic besides these performance enhancements is that XGBoost used a fundamentally more regularised model formulation. With time, other GBM models incorporated characteristics from XGBoost, as well as vice-versa, but as mentioned before, XGBoost continues to be widely popular.

XGBoost and, in general, GBMs, contain some known limitations. A practical one concerns the many possible hyperparameters that need to be optimised, sometimes compared to some decisions on the architectural decision of neural networks, therefore requiring a large search space. The intractability of enumerating all possible tree structures and therefore the solution of doing one split at a time is an obvious limitation. GBMs, due to their additive strategy, could continue indefinitely to minimise all errors, therefore overemphasising outliers and possibly causing overfitting if not controlled properly. A final major problem with additive trees is their memory usage, which is exactly one of the key issues that XGBoost tried to tackle for a more efficient implementation.

2.1.6 Good Representations

Training the supervised methods from Section 2.1.5 with a training procedure and selection of hyperparameters using the techniques presented in sections 2.1.2-2.1.4 can be seen as performing representation learning. This is the case because these methods are learning representations from input data such that they can learn a target function f_θ, parametrised by θ, such that it can predict $\boldsymbol{y} = f_\theta(\boldsymbol{x})$ (see Section 2.1.1).

Bengio et al. [31] posited that the goal of representation learning is to "to disentangle as many factors as possible, discarding as little information about the data as is practical". It is a challenge to determine what is *possible* and *practical*, but one way to tackle the practicality is to define the purpose of representations as to make subsequent learning tasks easier [124]. Based on this assumption, Dimanov [72] defined six primary requirements for an ideal representation:

- **Expressive**. The expressivity of a representation, also known as capacity, helps in guaranteeing the learnability of models and can be intuitively seen as the number of input regions that parameters can encode. There are different ways to actually make an approximate calculation of expressiveness [31, 221]; for instance, the Rademacher complexity measures the capacity of a hypothesis space by fitting random labels such that a higher complexity means that the model can fit a larger number of random labels and thus it has higher capacity [193].

- **Abstract**. A level of abstraction is needed so the representation knows which information needs to keep (i.e. the most salient factors of variation in the data) and which to ignore (i.e. those variations that are uninformative or invariant to the subsequent task).

- **Disentangling**. Beyond abstraction, a good representation should also be able to separate the underlying factors of variation of a dataset such that each feature in representation space is independent/orthogonal and corresponds to a different explanatory cause.

- **Easy to model**. When each factor can be observed without affecting the other factors (e.g., the representation is disentangled), then the representation could be easier for a human to comprehend and interpret.

- **Compact**. A smaller representation can be more efficiently handled both computationally (e.g. less computations when multiplying vectors) as well as statistically (i.e. less parameters to learn).

- **Robust**. This can concern robustness to perturbations in the inputs (e.g. adversarial examples), as well as out-of-distribution generalisation to unseen data.

It is important to highlight that, although ideal, these requirements might not be all satisfied at the same time. For example, a representation with independent features (e.g. disentangled) might not preserve as much information as possible, and a compact representation might not be expressive enough [72]. As it will be explored in the next section, the increased expressive power of representations (as very deep neural networks are one example of) can come at the cost of their interpretability.

Despite the possible conflicts, these ideal requirements are a good framework to understand the approaches used in this thesis when I explored *Research Question 1 (Representation)*. Indeed, these approaches can be framed under the assumption that the purpose of representations is to make subsequent learning tasks easier [124] when I use specific modelling approaches for the objectives I have in each chapter; specifically, when using multilayer networks for the transcriptomic networks in Chapter 3, graph neural networks for the supervised task using graph-structured data in Chapter 4, and XGBoost for the tabular data used in Chapter 5 in which it was important to have interpretable model decisions.

Finally, some evaluations made in this chapter can be directly framed under the requirements presented here. For instance, compactness of model can be seen in Section 4.4.2 when a more compact representation is able to achieve similar performance, therefore also indicating a good level of abstraction; robustness of models are directly evaluated using cross validation in chapters 4 and 5, and in some parts of Chapter 3; the "easy to model" requirement is also seen in Chapter 5 when interpretability insights are made due to underlying representation achieved by XGBoost.

2.1.7 Explainability and Interpretability

Research Question 2 (Explanations) asks: "is it possible to provide models to applied researchers that can provide explanations on how decisions are made (...) ?". Therefore, this section will motivate this question while providing a more precise view of what an explanation means.

Unfortunately, there is no single consensus in the field regarding what explainability means precisely, and this term is constantly used interchangeably with interpretability [7, 52, 177], despite some efforts in trying to find a common definition [14, 129, 162, 194, 197]. A simple definition of interpretability is given by Doshi-Velez and Kim [76] as "the ability to explain or to present in understandable terms to a human", but as it is possible to see from different works, different authors will look to this term from different perspectives.

In this thesis I will consider the definitions from the recent survey by Arrieta et al. [14], in which explainability "is linked to post-hoc explainability since it covers the techniques used to convert a non-interpretable model into a explainable one", while "the interpretability of the model is something that comes from the design of the model itself". The authors postulate a general definition of the field of explainable artificial intelligence as "given an audience, (...) is one that produces details or reasons to make its functioning clear or easy to understand". This general definition of explainable artificial intelligence then indirectly includes interpretability and explainability as equally important to move the field forward, while arguing that the type of *explanations* might differ depending on the target audience. Notice this difference is present in the different nomenclature used in chapters 4 and 5: while in the former I strive to provide post-hoc explanations to a deep learning model, in the latter I use the term "interpretability" given the explanations come from leveraging characteristics from how XGBoost is designed.

It is worth motivating why looking for these explainable properties is important instead of just looking to the best evaluation metrics for a certain task, even if a single, precise definition of *explanation* does not exist. One practical reason is that, as I briefly mention at the start of this thesis, artificial intelligence systems are increasingly used in safety-critical environments, and therefore it is challenging to rely on systems in which one does not know how they work or make decisions; indeed, the U.S. Defense Advanced Research Projects Agency has a project on explainable artificial intelligence[4]. One particular example to illustrate the importance of explainability is when it was discovered that models learned to predict the class “horse” in the ImageNet dataset not by using features from the horse but by looking to the copyright tag existent in many of those images [169]. On a more general level, this means that explainability can help debug very complex models and discover biases in the datasets that should be accounted for [12, 46, 170]. Ultimately, the field of explainable artificial intelligence could push models to be trusted by more people [72] while being safer [268] and auditable [40].

Together with the approaches to try to define the field of explainable artificial intelligence, it is already possible to see specific progress relevant to the contributions of this thesis. In this context, SHapley Additive exPlanations (SHAP), which will be explained and used in Chapter 5, has been a recurrent choice to try to find post-hoc explanations. For instance, Yap et al. [293] used SHAP to explain a deep learning model predicting tissue type from transcriptome data (see Section 2.2), and Yu et al. [297] integrated SHAP with auto encoders to evaluate the contribution of different genes for various cancer-related classes in the hidden variables of those auto encoders. Despite the apparent focus of the explainable artificial intelligence field on deep learning architectures, it is also possible to see rule-based models to find biologically relevant patterns from gene-gene temporal relations in gene expression data [13]. Explainable artificial intelligence is also widely present in the neuroimaging field, with review/surveys of deep learning-based medical image models already available [207, 264, 267]; likewise, despite the even more recent paper explosion in the field of graph neural networks (see Section 2.1.8), the interplay of these new architectures with explainability is already being explored [295, 298]. The inclusion of explainable artificial intelligence in neuroimaging developments can be done to help physicians understand why someone is being diagnosed with a certain neurodegenerative disease [84] and which structural connections are involved in neurodegenerative diseases [88]; sometimes, those models achieve better evaluation metrics than non-explainable models [263]. Finally, the identification of regions involved in brain ageing is also a promising application of explainable artificial intelligence in neuroimaging [110, 286].

I argue that these reasons fully motivate *Research Question 2 (Explanations)* in the sense that finding *explanations* is a good objective to strive while exploring the other two research questions. However, for completeness, it is important to mention that some authors still disagree with some of these ideas. Some defend that in some cases artificial intelligence models can be clinically validated even when their function is not completely understood [189], and that current explainability methods are unlikely to achieve the goals that the field is looking for so we should look for more rigorous evaluation processes instead [118]. I consider that an underlying cause for these doubts is the belief that explainable/interpretable models cannot perform as well as current state-of-the-art ones [165]; however, growing evidence shows that the development of interpretable models does not need to negatively influence their performance [231]. As the several surveys referenced in this section show, this is a growing field and I lean on the cautious side that this is a needed functionality; therefore, it should be looked for together with rigorous evaluation processes.

[4] `https://www.darpa.mil/program/explainable-artificial-intelligence`

2.1.8 Graph Neural Networks

This section will briefly outline the utility of graph neural networks (GNNs) in the context of this thesis, as they directly motivate how *Research Question 3 (Graph)* was tackled in Chapter 4. This type of neural networks has seen an explosion in number of related publications around 2017 on a multitude of applications where data can be naturally represented as a graph [288, 302]. The field is still growing in part due to very significant successes in both academic [257] and industrial [69] applications. The original GNN is attributed to Gori et al. [127], though earlier approaches in the 1990s exist often involving directed acyclic graphs [105, 250, 251].

GNNs can be seen as a way to apply deep learning methods on graph data, in which the underlying idea is to generate representations that depend on the actual structure of the graph and any information (i.e. features) from the nodes, edges, and the graph as a whole. Making a parallel with CNNs which are effective at extracting features from grid-like data (e.g. 2D and 3D images), and RNNs which are able to learn features from data that are temporally organised as a sequence of steps, GNNs can learn from data that can be depicted in the form of unordered entities and relations such as graphs. This connects with the framework presented in Section 2.1.6 where it was stated that good representations are those that make subsequent learning tasks easier: CNNs, RNNs, and GNNs leverage specific prior information known about the input data in order to have better representations, and therefore better performance on different learning tasks.

There are different ways to frame and explain GNNs (including the formalisation taken in Chapter 4), but in all these different ways there is a notion of a neural message passing mechanism where "messages" are exchanged between nodes and updated using neural networks. These are generically called message-passing neural networks (MPNN) [119]; more formally, if we consider $\mathbf{v}_i^{(l-1)}$ the features of node i in layer $(l-1)$, and $j \in \mathcal{N}(i)$ the neighbour nodes of i connected through an edge with edge features $\mathbf{e}_{j,i}$, an MPNN layer can be defined as:

$$\mathbf{v}_i^{(l)} = \gamma^{(l)} \left(\mathbf{v}_i^{(l-1)}, \rho_{j \in \mathcal{N}(i)}^{e \to v} \left(\varphi^{(l)} \left(\mathbf{v}_i^{(l-1)}, \mathbf{v}_j^{(l-1)}, \mathbf{e}_{j,i} \right) \right) \right), \tag{2.13}$$

where $\rho_{j \in \mathcal{N}(i)}^{e \to v}$ is a differentiable aggregation function, invariant to input permutation and applied across node's neighbours, and γ and φ are also differentiable functions such as multi-layer perceptrons.

In summary, the previous paragraphs illustrate an important advantage of GNNs over regular deep learning models which is the fact that they allow the creation of "good" representations for graph data. This happens because they create features that are invariant to node permutations while still taking into account the connectivity of the underlying graph structure.

As GNNs are still a very intense field of study, it is necessary to note that they contain known limitations. One of the topics still being researched concern the actual expressive power of GNN representations, as some authors pointed out that common GNN architectures are not able to capture certain simple graph structures [25, 240, 289]. The fact that it is a relatively recent field also means that hardware acceleration has not matured yet for these architectures, and implementations of these models are not as fast as what currently exists for CNNs: this is due to a combination of dense and very sparse operations and the need to scale some operations to huge graphs [5, 291]. These known limitations led the results in Chapter 4 to be compared to other GNN models in the field, as well as non deep learning baselines.

2.2 Molecular Biology Fundamentals

Chapter 3 explores the applied field of molecular biology, and more specifically gene expression from transcriptomic data. Therefore, in this section, I will clarify not only the key semantics of a transcriptomic dataset, but also additional concepts that are critical to an understanding of the broader applied field in which transcriptomics exist in the context of this thesis.

The "central dogma" of molecular biology was explicitly stated in 1958 by Francis Crick [63], and posits that the flow of genetic information is unidirectional: through a process called *transcription* the DNA encodes genetic information (nucleic acid) which is passed on to the messenger RNA molecules; proteins are then synthesised from information in this RNA in a process known as *translation.* These proteins ultimately end up influencing how a phenotype is expressed through metabolites[5] and environmental factors. A phenotype is an observable trait in an individual, such as the colour of hair, blood type, or structural features of the brain such as those analysed in Chapter 5's neuroimaging scans.

This "central dogma" was stated around the same time as the discovery of the three-dimensional, double-helical model for the structure of DNA. The latter discovery was ultimately the topic of the Nobel prize in Physiology or Medicine in 1962, famous for not crediting the groundbreaking crystallography work by Rosalind Franklin and Raymond Gosling [104] which provided the vital clue (commonly called "photo 51") to the double helix structure [1, 182]. Since the assertion of the "central dogma" and the discovery of the DNA structure, our knowledge of biological systems has greatly evolved, and we now know that more complex interactions exist among the different components [241]. For instance, it is known that transcription factors are proteins that can initiate and regulate the *transcription* of genes, therefore indicating that this process is not completely unidirectional.

2.2.1 Multi-omics

In molecular biology, it is common to add the suffix "-omics" to imply a comprehensive, or global, assessment of a set of molecules [132]. I now present several "-omics" directly and indirectly related to the topics analysed in this thesis; each one has unique characteristics and provides a different view at the molecular level.

Genomics concerns the study of an organism's full DNA. This term is distinct from the genome, which is an organism's complete set of genetic information. In the early 2000s, thanks to the availability of whole-sequencing data provided by large consortia such as the Human Genome Project, many genome-wide association studies (GWAS[6]) flourished in order to find associations between genes and phenotypes in multiple human

[5] A metabolite is an intermediate or end product of metabolism, and can therefore provide information on cellular activity and physiological status.

[6] GWAS studies are critical but a thorough explanation is beyond the scope of this thesis. Very briefly, thousands of individuals are genotyped at more than a million genetic markers, and statistically significant differences between cases and controls are considered evidence of association.

populations [125]. These advances ultimately helped in understanding highly complex genetic traits in health and disease, leading genomics to be the most mature "omic" field [132]. Despite the remarkable advances in identifying many complex genotype-phenotype associations, there are still known limitations [260], the main one concerning that only correlations are revealed in these studies, and other factors play a critical role in explaining phenotype such as gene regulation and environmental stimuli. One of the reasons for the development of multi-omic analysis is to overcome some of these limitations.

Transcriptomics examines genome-wide RNA levels qualitatively (i.e. which transcripts are present), as well as quantitatively (i.e. how much each transcript is expressed). This term usually refers to all the RNA existent, but may sometimes just refer to the messenger RNA (mRNA) containing necessary information for protein synthesis. The transcriptome (i.e. the set of all RNA transcripts) is therefore an expression of the genome by capturing the genes expressed at the transcription level for a particular condition. At this stage, it is usually possible to generate **gene expression profiles** such as those I have used in this thesis. I should highlight that the transcriptome is susceptible to environmental conditions, highlighting the importance of gene expression analysis to identify genes that exhibit differences in expression between health conditions and even among different body tissues; therefore, this type of data addresses some of the issues still existent in GWAS studies regarding gene regulation and external factors [159]. Besides traditional statistical methods, machine learning and deep learning methods have proved to be efficient at identifying transcriptomic profiles associated with specific phenotypes, considering different input data such as measured RNA-seq data[7] [279], single cell expression [142], and also imputed transcriptomic data [128].

Proteomics is the quantification of cellular levels of proteins encoded by the genome, as well as the study of the diverse properties of proteins in a cell or tissue [214]. As mRNA resulting from transcription can decay quickly and therefore not be translated, looking only at gene expression levels can be misleading. It is at this stage where well-known protein-protein interactions are studied. The transcriptome may include genes that do not encode proteins (e.g. genes with a regulatory function), and as a consequence the study of the proteomics alone might not give a complete picture in molecular biology. It can therefore be seen as more of a complementary addition to the study of such complex environments.

Epigenomics is the study of the complete set of chemical changes to the genetic material of a cell, also known as the epigenome [204]. Even though I do not directly explore this concept in this thesis, it is important to fully comprehend the possible applications of this field. Unlike the static genome, the epigenome can be dynamically affected with long-lasting effects, sometimes even heritable, on genetic and environmental factors. Although a bit controversial, growing evidence supports a role for epigenetic regulation as a key mechanism underlying lifelong regulation of gene expression that mediates stress vulnerability [200] and disease development [54].

Other "-omics". It is worth mentioning that there are other "-omics" which are not as important for the understanding of this thesis, but still worth mention for completeness

[7] RNA-seq is a particular sequencing technique for RNA quantification.

for the interested reader. They include, but are not limited to: *metabolomics*, the study of metabolites which can have a complementary role with transcriptomics and proteomics [139]; *lipidomics*, the analysis of a specific type of metabolites called lipids which can be associated with various diseases [285]; and *microbiomics*, the study of the microorganisms in certain parts of the human body [222].

2.2.2 Networks

Networks are pervasive in the molecular biology field, and therefore in this short section I will give some application examples for a wider understanding of the field which indirectly underpinned the hypothesis explored in Chapter 3. The representation in the form of networks provides a more intuitive perspective of molecular biological systems thus allowing the analysis of different interactions between molecules from the different "-omics" presented in Section 2.2.1.

Examples of common networks include, but are not limited to: (1) gene co-expression networks (the ones I explored in Chapter 3), (2) protein-protein interactions to understand (or predict) how different proteins can interact and be activated under certain conditions [181, 303], (3) metabolic networks which can inform how metabolites are transformed to synthesise other substances (and therefore each edge represents a metabolic reaction) [149], and (4) the interactome which loosely represents the integrated network of all physical/molecular interactions in a cell [51], therefore allowing for a holistic integration of many "-omics" in the same representation.

With this representation of nodes referring to molecules and edges depicting interactions, each node can then be characterised according to topological measures that can suggest biological roles in those networks. Indeed, one can look to power-law networks [300] and explore how the removal of small-degree nodes does not have a substantial effect in the network's properties, as well as looking to cliques/modules and corresponding biological implications [131, 147]. A useful network modelling approach concerns multilayer networks, which are formally defined in Section 3.1.7 when used in the context of Chapter 3. As I will describe, they are able to aggregate different modelling levels in a single network [164] and thus are typically used to integrate different "-omics" [130, 218, 292]; however, in this thesis, I will use this modelling approach to focus on the field of transcriptomics.

2.3 Neuroimaging Fundamentals

Chapters 4 and 5 explore the applied field of neuroimaging. Therefore, this section will explore important concepts for a better understanding of the applied work in the context of this thesis.

2.3.1 Types of Neuroimaging Scans

Neuroimaging is a field concerned with creating visual representations of a brain for clinical and research purposes. There are two main types of neuroimaging: structural and functional imaging. Each type contains many possible techniques, which vary in temporal and spatial resolution, as well as in the type of components that are targeted (e.g. grey matter, blood vessels, tumours). The imaging methods used in this thesis are not invasive, but some other techniques could include the injection of radioactive material into the bloodstream to interact with specific molecules in the brain, called metabolic imaging. These metabolic imaging methods are expensive and not widely available; therefore, gathering sufficient data for meaningful analysis with artificial intelligence methods is impractical.

The most common type of structural imaging is magnetic resonance imaging (MRI). In MRI scanners a uniform magnetic field is first employed, taking then advantage of the magnetic properties of hydrogen which is prevalent in the human body [64, 168, 274]. In short, a radio frequency (RF) pulse is emitted from the scanner at which the hydrogen protons change their alignment and emit energy. Some time after the RF, the emitted signals are measured using Fourier transformations in the frequency domain. After varying the sequence of RF pulses, different types of images are created and two variables of the scan sequence will determine tissue contrast in these images: repetition time (TR) and echo time (TE). TR is the time between successive pulse sequences and TE is the time between the application of a RF pulse and the receipt of the echo signal or, in other words, the time between the excitation pulse and the peak of the signal.

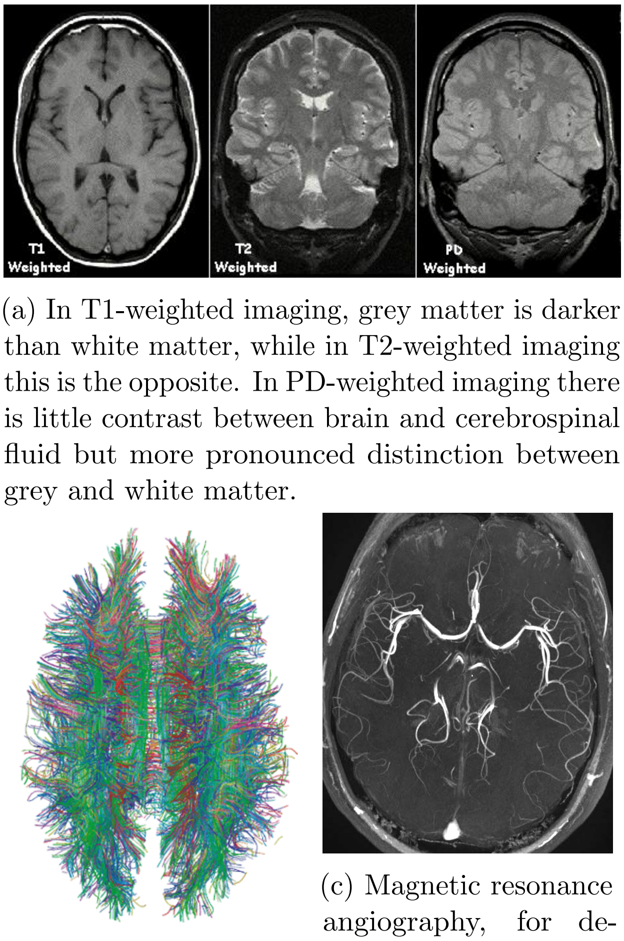
By varying the TR and TE values, different MRI sequences can be produced with distinct contrasts for different body components as shown in Figure 2.4 which illustrates five MRI techniques, each one targeted to detect different components of the brain [64]. For example, T1-weighted scans are produced using short TE and TR times which produce dark cerebrospinal fluid (CSF), light white matter, and grey cortex; these are particularly useful for looking at brain structure. In contrast, T2-weighted scans are produced using longer TE and TR times, producing bright CSF, darker white matter and lighter cortex; these are particularly useful for examining changes in the brain's white matter.

One significant advantage of MRI scanners is that they do not involve ionising radiation in contrast to other techniques such as computer tomography (CT) scans, making it a safe option with little to no hazard of increased cancer risk. Therefore, it is highly versatile and widely used in medical diagnosis, monitoring disease, and general research purposes [66, 68].

Functional neuroimaging concerns the measure of brain activity throughout the brain; such measures help understand relationships between activities in specific brain areas and mental functions of interest in cognition, psychology, and social neuroscience [95]. In this neuroimaging type, it is common to have a temporal resolution, as activity is measured across time, and a spatial resolution that tends to be significantly lower than the static structural neuroimaging techniques. For instance, electroencephalography (EEG) measures voltage fluctuations from electrodes placed along the scalp and magnetoencephalography (MEG) measures magnetic fields fluctuations from magnetometers placed along the scalp; therefore, even though their temporal resolution is extremely high, its spatial resolution is sparse as it mainly concerns positions outside the skull. Other functional neuroimaging techniques exist, such as positron emission tomography (PET) in which radioactive substances are injected into the body to detect specific metabolic processes [95], but this technique is not used in this thesis.

As it is possible to see, a key difference between structural and functional neuroimaging concerns their resolution at different scales [94]. For instance, EEG and MEG focus on high frequency brain activity, but are limited by being able to examine only cortical signals. MRI has a lower frequency of time sampling, but is able to visualise the whole brain and provide structural and volumetric data for further analysis.

A typical type of functional neuroimaging which will be analysed in Chapter 4 is



(a) In T1-weighted imaging, grey matter is darker than white matter, while in T2-weighted imaging this is the opposite. In PD-weighted imaging there is little contrast between brain and cerebrospinal fluid but more pronounced distinction between grey and white matter.

(b) White matter tracts obtained from MRI diffusion tensor imaging. Colours are calculated at post-processing time based on water diffusion direction.

(c) Magnetic resonance angiography, for detecting blood vessels. This image was obtained using particular magnetic pulse dynamics with a short echo time (TE) called time-of-flight (TOF).

Figure 2.4: Examples of five distinct MRI techniques, each highlighting different components of the brain. Images taken from Wikimedia Commons[8].

functional magnetic resonance imaging (fMRI). This imaging type uses the blood-oxygen-level-dependent (BOLD) contrast, with limited spatial resolution typically ranging between 2-4mm for each voxel[9], and temporal resolution ranging between 1-2 seconds. The BOLD contrast can detect changes in oxygen saturation in haemoglobin due to its magnetic properties. These changes serve as a proxy for detecting brain activity as the vascular system changes to respond to the brain's need for glucose [98].

There are mainly two types of fMRI image: task-specific fMRI and resting-state fMRI (rs-fMRI). In task-specific fMRI, a person is asked to perform some cognitive task and a researcher can then explore which brain areas are responsible for the person's response. In rs-fMRI, a person enters the scanner without performing any particular task, and the

[8] `https://commons.wikimedia.org/wiki/File:T1t2PD.jpg`, `https://commons.wikimedia.org/wiki/File:White_Matter_Connections_Obtained_with_MRI_Tractography.png`, `https://commons.wikimedia.org/wiki/File:Mra-mip.jpg`

[9] A voxel corresponds to a pixel in 3D space.

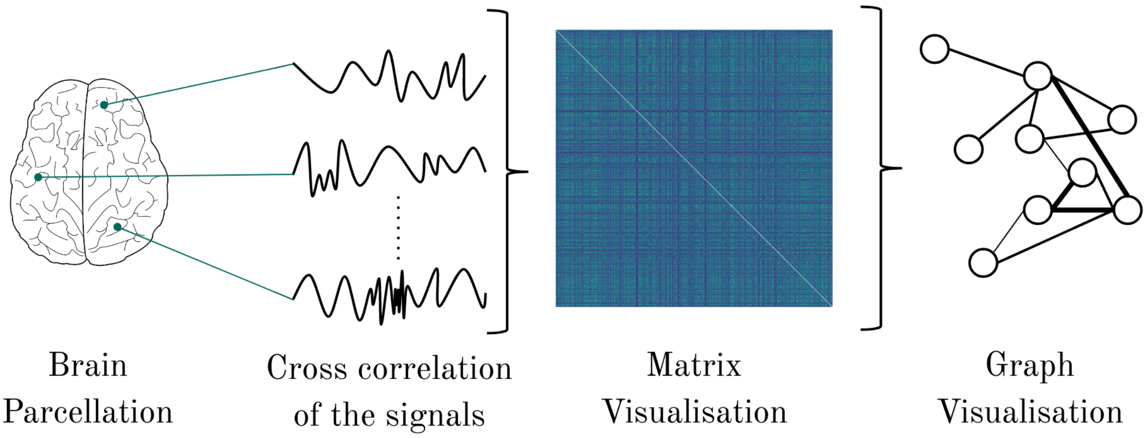


Figure 2.5: High-level steps to generate a graph representation from fMRI time series data. A brain is initially parcellated (i.e. divided) into regions of interest, from which time series are extracted. Correlations are calculated between every pair of brain regions to generate a symmetric matrix and a respective graph representation. Brain parcellation techniques are explored in Section 2.3.2.

objective is to analyse the activations that the brain produces at rest (i.e. in the absence of any external stimulus) [98].

As depicted in Figure 2.5, by correlating the fMRI time series of every pair of brain regions we can obtain a graph visualisation representing how different brain regions are more *functionally* connected over a certain period of time. In this way, functional connectivity can be represented as an association matrix using the correlation strength between every brain region pair. This matrix can be highly informative as a stronger correlation indicates that two brain regions were similarly activated/deactivated during a certain time (i.e. in-sync), thus indicating some form of communication. The study of brain connectivity, sometimes also called "connectomics" has revealed insights in to normal brain function and neuropsychiatric disease [102].

The study and mapping of the neural connections in the brain is a field that has recently attracted great interest in the scientific community. It is known that topology and functional connectivity change with age and form the basis for things such as learning, ageing and disease. Thus, it is possible to better understand these aspects of brain dynamics just by analysing its working connections [238, 256]. These functional brain networks differ between people and may respond uniquely to different external stimuli or treatments. This heterogeneity opens up the possibility for personalised medicine based on the knowledge of brain networks [91, 138].

2.3.2 Processing Neuroimaging Scans

The complex physical mechanisms involved with these neuroimaging scans introduce yet another challenge: the preprocessing steps for downstream analysis. A number of issues need to be addressed in preprocessing to remove noise and permit comparison between individuals. For example, physical movements of a person inside the scanner and blood flowing through the brain can introduce noise. As the brain is not a rigid organ like a bone, it can move and produce motion artifacts in the resulting image, though this is less of an issue than the person moving in the scanner.

Other MRI artifacts are worth mentioning, which after their identification might require a radiologist or another specialist to understand whether a solution needs to involve hardware change, new scanner parameters, better controlling the patient, redoing the scan, use software correction when preprocessing the data, or simply excluding the scan in later analysis. The following artifacts are not an extensive list but illustrate other known MRI artifacts:

- Susceptibility artifact, when a patient has an implant or another medical device that responds to the magnetic field, therefore producing wrong contrasts in the resulting image [237].
- Zipper artifact, in which spurious RF signals entering the shielded room can produce noise in the image, therefore making it difficult to interpret [134].
- Black boundary artifact, in which a black line is artificially created at water-fat interfaces such as muscle and fat. This effect can also be called 2$^{\text{nd}}$ order chemical shift artifact [134].
- Diastolic pseudogating, which happens when sequence timing is, by chance, in-sync with heart rate, therefore producing different blood signal intensity in large vessels across image slices [26].
- Magic angle effect, seen mostly with ligaments oriented at a 55-degree angle with the magnetic field; as a result, in T2 images this can produce a sudden increase in localised signal [49].

Many techniques and statistical procedures have been developed to extract the underlying signal from the raw images that come from these scanners. There is no standard method for preprocessing, although a number of common necessary steps are shared between preprocessing pipelines. The exact preprocessing steps depend on the type of imaging, the scanner strength and manufacturer, and the amount of inherent noise and motion artefact within the data. This variability is well illustrated in a recent study, where Botvinik-Nezer et al. [41] assessed the effect of the flexibility in analytical approaches available when analysing fMRI-based hypotheses by asking 70 independent teams to analyse the same dataset. Strikingly, no two teams chose identical workflows to analyse the data. This flexibility resulted in sizeable variation in the results of hypothesis tests, even for teams whose statistical maps were highly correlated at intermediate stages of the analysis pipeline. This study highlights the need for careful and principled selection of preprocessing steps for neuroimaging to improve transparency, reproducibility, and impact of research [126]. It is beyond the this thesis' scope to detail workflows; instead, I will provide the overall preprocessing steps used in this thesis in Section 2.3.3 and provide a more technical description of how they were processed in each chapter.

An important concept in neuroimaging is that of a *brain atlas*. In a map, regions are geographically delineated; likewise, in a brain atlas there are non-overlapping regions of interest (ROIs) dividing the brain. A brain atlas can be defined according to anatomical features, on regions known to functionally work together in some particular tasks (e.g. reasoning, perception), or even on known molecular properties [70, 83, 93].

As it is computationally expensive to analyse all the brain voxels, an atlas is useful to parcellate (i.e. divide) a brain captured in a scanner for downstream analysis. By applying a parcellation, one can analyse an average value calculated from all the voxels in a ROI. To use the analogy of a map, this would be similar to averaging the ages of all people living within one city rather than using each individual age. As these atlases are sometimes represented in a common brain in 3D space, one typical preprocessing step in brain imaging parcellation is to wrap the brain from the raw image into this common space by applying linear and non-linear calculations. Sometimes a brain parcellation could be data-driven instead of atlas-driven, in which ROIs are defined by clustering of information in the data.

An example of such an atlas is the Desikan-Killiany cortical atlas [70], which was used in chapters 4 and 5 and is depicted in Figure 2.6 for the cortex. This atlas was originally created based on 40 MRI scans and is a gyral-based atlas, that is, a gyrus[10] is defined as running between the bottoms of two adjacent sulci[11].

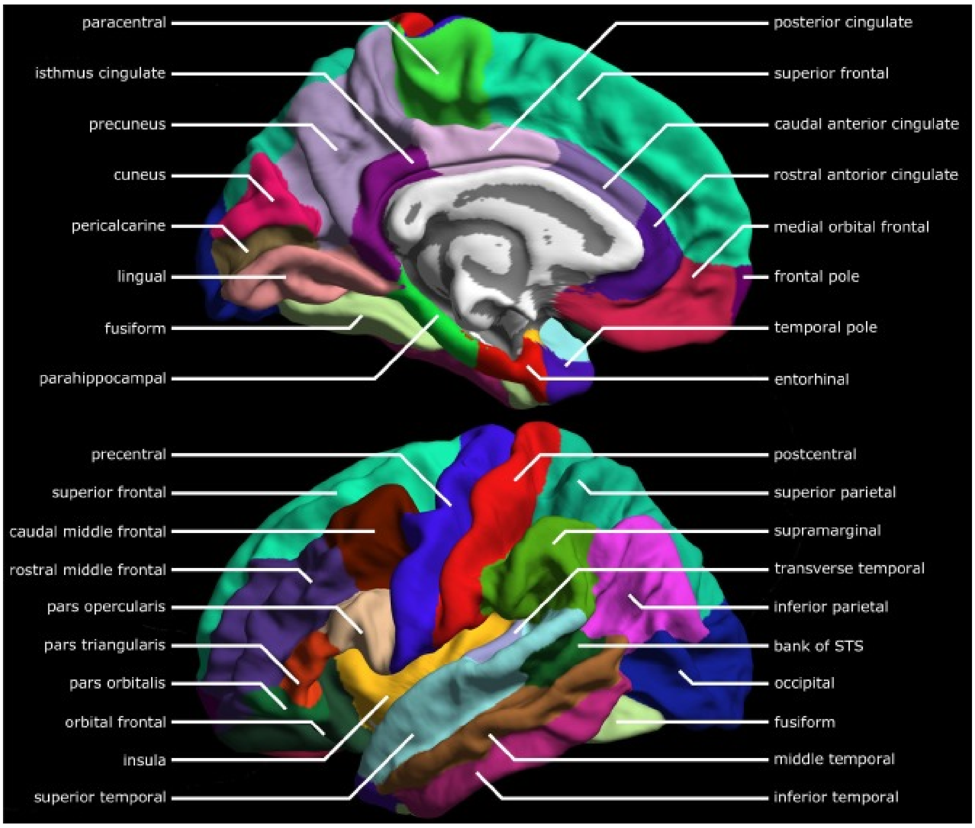


Figure 2.6: Cortical regions from the Desikan-Killiany atlas represented in one hemisphere, with subcortical regions in grey. Figure taken from Nagtegaal et al. [199] with permission from Elsevier.

2.3.3 Datasets

In this thesis I used two main neuroimaging datasets: the Human Connectome Project (in chapters 4 and 5) and the UK Biobank (in Chapter 4). The Human Connectome Project (HCP) is one of the most homogeneous and well-characterised open datasets for young healthy subjects, and the UK Biobank (UKB) is known to be subject to potential selection bias as its population tend to have a low risk for disease [108, 253] (for instance,

[10] A gyrus is the name given to the bumps ridges on the cerebral cortex.

[11] A sulcus is a shallow groove that surrounds a gyrus.

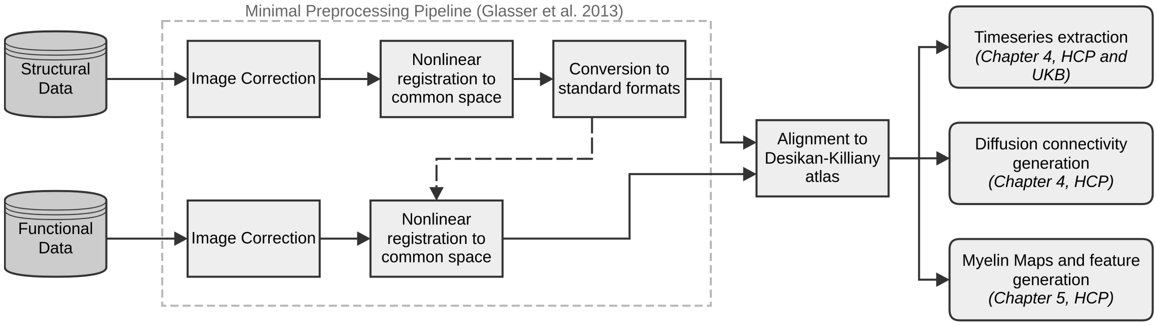

Minimal Preprocessing Pipeline (Glasser et al. 2013)
Structural Data
Image Correction
Nonlinear registration to common space
Conversion to standard formats
Functional Data
Image Correction
Nonlinear registration to common space
Alignment to Desikan-Killiany atlas
Timeseries extraction (Chapter 4, HCP and UKB)
Diffusion connectivity generation (Chapter 4, HCP)
Myelin Maps and feature generation (Chapter 5, HCP)

Figure 2.7: Main preprocessing pipeline followed by collaborators (see Section 1.4) on the Human Connectome Project (HCP) and UK Biobank (UKB) datasets. More specific details are provided in each chapter where the data is used. "Minimal preprocessing pipeline" was originally defined by Glasser et al. [122] and directly provided to download by each consortium, while remaining steps are run by collaborators.

the proportion of people currently smoking in the UKB is 10.7% compared to 14.7% in the general population[12]). In this thesis I used an HCP release with 1,200 subjects, and the UKB is a dataset of over 500,000 people aged around 40-80 years old who have undergone repeated cognitive testing, with approximately 35,000 having undergone neuroimaging with MRI.

Both datasets contain well-characterised population cohorts and have undergone consistent, standardised neuroimaging and clinical assessments [89, 191]. Having aligned and standardised acquisition protocols is crucial to ensure consistency across acquisition sites as well as to improve data quality across sites and scanner [73]. Furthermore, to support the need of large datasets for reproducible findings with minimal statistical errors [186], no selection criteria was applied on my side when acquiring the data. Despite this limitation, the size of the datasets, age of participants, and high quality neuroimaging data make the HCP and UKB ideal to assess the different modelling approaches applied in this thesis.

Both datasets followed the same preprocessing steps for the same type of data but were preprocessed by different collaborators before being shared with me (see Section 1.4 for collaborator details). Figure 2.7 depicts the main neuroimaging preprocessing steps followed in this thesis. Details and differences about the specific steps followed in each dataset and modality are provided in each chapter where the data is used.

As I mentioned in Section 2.3.2, there is no single preprocessing pipeline; however, the steps followed in this thesis are widely used in the field as the data preprocessed using the "minimal preprocessing pipeline" [122] are directly provided by the different consortia. The Desikan-Killiany atlas introduced in the previous section was chosen as a common atlas with enough granularity to balance ease of interpretation of results with consistency when compared, for instance, to parcellations with hundreds of regions or to parcellations without divisions based on known neuroanatomical knowledge. Furthermore, this is a generalised atlas that can be applied across populations and allowed the preprocessing of all the data modalities used in this thesis. Crucially, it is a widely used atlas in literature which enables the embedding and comparison of results to the wider scholarship.

[12] Data from the Office for National Statistics: `https://www.ons.gov.uk`

2.4 Summary

This chapter outlined and explained fundamental concepts and key literature that are important to the correct understanding of this thesis and its significance. The first section on machine learning fundamentals concerns all the main chapters of this thesis and therefore presented topics on model training steps, tasks, important characteristics, and others. This was followed by two sections clarifying concepts on the applied fields of molecular biology (needed for Chapter 3) and neuroimaging (needed for chapters 4 and 5).

Chapter 3

Multilayer Modelling and Analysis of the Human Transcriptome

The modern science of networks has contributed to notable advances in a range of disciplines, facilitating complex representations of biological, social, and technological systems; a key aspect of such systems is the existence of community structures, wherein groups of nodes are organised into dense internal connections with sparser connections between groups [103]. Community structure detection in genome-wide gene expression data may enable detection of regulatory relationships between regulators (e.g. transcription factors) and their targets, and capture novel tissue biology otherwise difficult to reach. Furthermore, it offers opportunities for data-driven discovery and functional annotation of biological pathways.

Together with collaborators (see Section 1.4), I hypothesised that community structure is an important organising principle of the human transcriptome, with critical implications for biological discovery and clinical applications. Co-expression networks, in fact, encode functionally relevant relationships between genes. These include gene interactions and coordinated transcriptional regulation [233], and provide an approach to elucidating the molecular basis of disease traits [117]. Therefore, reconstructing communities of genes in the transcriptome may uncover novel relationships between genes, facilitate insights into regulatory processes, and improve the mapping of the human diseasome.

Despite previous knowledge on the importance of transcriptomic network structures in distinct tissues of the human body [11, 27], to the best of my knowledge this is the first work exploring the multi-tissue modelling hypothesis we stated based on community structure from a Graph-Dimensional representation. Therefore, this represents a new perspective in the field of transcriptomics which I will argue to have further implications for biological discovery. This interesting perspective was probably not explored before due to only recently the last release of the GTEx dataset (see Section 3.1.1) was made available to researchers as the most comprehensive human transcriptome dataset [4].

As briefly introduced in Section 2.2.2, the hypothesis explored in this chapter follows other distinct lines of work that I bring together in this chapter and are therefore important to highlight. Multilayer networks were thoroughly used to integrate networks from many "-omics" in the past [130, 218, 292], whereas here I will only use transcriptomic networks. A paper developed after the work presented in this chapter used multiplex networks to integrate several networks of many "-omics", this time including networks generated from GTEx in a similar fashion like those in this chapter [47]. Previous works using gene co-expression networks from transcriptomic datasets were previously limited by the number of networks they could actually integrate together using multilayer networks or other approaches [183, 301]. In all these works, however, no communities were analysed in the same way as in this chapter, which I find a crucial difference given the importance of community structure in the hypothesis formulation.

This chapter presents a model of the human transcriptome as a multilayer network, and performs a comprehensive analysis of the communities obtained to further our understanding of its wiring diagram as well as facilitate research into improved disease diagnosis and profiling. I conduct a systematic analysis of the tissue-type specificity of the communities in the transcriptome to gain insights into gene function in the genome and enhance our ability to identify disease-associated genes. This study represents an effort to fill an important gap in our understanding of the role of gene expression in complex traits, i.e. how a gene's phenotypic consequence on disease or trait [111] is mediated by its membership in tissue-specific biological modules as molecular substrates. Finally, the inter-tissue analysis of the transcriptome holds promise for identifying novel regulatory mechanisms, enhancing our understanding of trait variation and pleiotropy[1], while opening up new possibilities for translational applications.

As I mentioned in Section 1.4, I aimed to publicly release the source code of all my work. This is of particular importance in this chapter as I describe a resource to catalyse further research by the scientific community, which I can only briefly summarise in some sections. Besides code and documentation to all experiments described in this chapter, I want to highlight that in a publicly available GitHub repository[2] it is possible to find: (1) more detailed information on the communities described in Section 3.2.2, (2) all the documentation on how to correctly include an example of an external dataset into the models described in Section 3.2.3, and (3) a full set of relationships between the communities and enriched pathways from Section 3.2.4.

All in all, the three main contributions of this chapter are: (1) identification of community structure as an important organising principal of the human transcriptome, thus with applications for biological discovery, (2) suggestion of a presence of a hierarchy of clusters in the transcriptome at increasingly finer scales, and (3) distribution of a publicly available rich resource of co-expression networks, communities, multiplex architectures and enriched biological pathways that can help catalyse hypothesis-driven research. These contributions therefore help tackling two gaps in literature as tissue-to-tissue regulatory studies are understudied when compared to intra-tissue ones, and I present a new approach to quantify UMAP global structure which is not based on a single run.

3.1 Methods

3.1.1 Dataset - GTEx

The GTEx V8 dataset [4, 60] is a genomic resource consisting of 948 donors and 17,382 RNA-Seq samples from 52 tissues and two cell lines. The resource provides a catalogue of genetic effects on the transcriptome and a broad survey of individual and tissue-specific gene expression. Of the 54 tissues and cell lines, 49 include samples with at least 70 subjects, forming the basis of the analysis of genetic regulatory effects [4]. Therefore, only those 49 tissues were used in this chapter.

[1] Pleiotropy occurs when one gene influences two or more seemingly unrelated phenotypic traits.

[2] `https://github.com/tjiagoM/gtex-transcriptome-modelling`

The analysis was restricted to protein-encoding genes based on the GENCODE Release 26 (GRCh38) annotation. Although the GTEx dataset had annotated genes with Ensembl IDs, those were converted to GENE IDs. During that process, duplicated and unmapped genes were removed from downstream analyses. After this preprocessing step, the resulting dataset is characterised by the following count statistics:

- Genes present in at least one tissue: 18,364
- Genes present in only one tissue: 412
- Genes present in all 49 tissues: 12,557

3.1.2 Accounting for Unmodelled Factors

Disambiguating true co-expression from artefacts is an important concern in the presence of hidden variables. In order to correct for batch effects and other unwanted variation in the gene expression data, I used the *sva* R package (v3.34.0). This package is specifically targeted for identifying surrogate variables in high-dimensional datasets [211] and investigating unmodelled and unmeasured sources of expression heterogeneity [171]. For each tissue gene expression matrix, the number of components (latent factors) was estimated using a permutation procedure, as described by Buja and Eyuboglu [45].

Subsequently, using the function *sva_network*, residuals were generated after regressing out the surrogate variables. The residual values, rather than the original gene expression values, were used in the downstream analyses. For convenience, I refer to the residual values as the ‘gene expression data’, since they represent the expression levels that have been corrected for (unwanted) confounders.

3.1.3 Community Detection on Co-Expression Networks

For each tissue, a correlation matrix $C = [z_{ij}]$ was created by calculating the Pearson correlation coefficient r_{ij} for every pair (i, j) of genes. Fisher z-transformation was then applied:

$$z_{ij} = 0.5 \times \ln\left(\frac{1 + r_{ij}}{1 - r_{ij}}\right), \tag{3.1}$$

where ln is the natural logarithm function.

For each correlation matrix, only the strongest correlations were retained (i.e. transformed z_{ij} less than -0.8 and greater than 0.8) to generate a co-expression network. An adjacency matrix $A = [A_{ij}]$ was defined, for each tissue, such that A_{ij} is equal to z_{ij} if gene i and gene j are co-expressed (retained), and zero otherwise. These networks are undirected and without self-loops, which implies $A_{ij} = A_{ji}$ and $A_{ii} = 0$.

I sought to detect groups of genes in each tissue to find communities whose internal connections are denser than the connections with the rest of the co-expression network. To that end, I applied the Louvain community detection method [37] in each tissue to generate a comprehensive atlas of communities. An asymmetric treatment for the negative correlations was used, thus inducing negatively correlated genes to belong to different communities [230]. The algorithm identifies communities by maximising the modularity index [201], Q^*, as the algorithm progresses:

$$Q^* = \frac{1}{v^+} \sum\nolimits_{ij} \left({w_{ij}}^+ - {e_{ij}}^+\right) \delta_{M_i M_j} - \frac{1}{v^+ + v^-} \sum\nolimits_{ij} \left({w_{ij}}^- - {e_{ij}}^-\right) \delta_{M_i M_j}. \tag{3.2}$$

Here, a positive connection between nodes i and j is denoted as ${w_{ij}}^+$ and has a value between 0 and 1; likewise, a negative connection is represented ${w_{ij}}^-$ and can also have a value between 0 and 1. ${e_{ij}}^{\pm}$ is the chance-expected within-module connection weight and calculated, for each positive/negative correspondent, as $\frac{s_i^{\pm} s_j^{\pm}}{v^{\pm}}$, where $s_i^{\pm}$ is the sum of positive or negative connection weights of node i. $v^{\pm}$ is the sum of all positive or negative edges, and $\delta_{M_i M_j} = 1$ when nodes i and j are in the same module or zero otherwise. The Louvain method initially assigns each node to its own community and iteratively evaluates the gain in modularity if one node is moved from one formed community to another of its neighbourhood. I have used the Brain Connectivity Toolbox Python package v0.5.0[3], where the resolution parameter γ was set to its default value, 1.

3.1.4 UMAP Embeddings

To produce a lower-dimensional representation of the original dataset, I applied Uniform Manifold Approximation and Projection (UMAP) [190] to check the embedded structure of all samples. The goal is to generate a map that reveals embedded structures and test whether biologically relevant clusters can be recovered from the gene expression data. UMAP was chosen because of its theoretical grounding in manifold theory [190], and the substantial improvement in running time on the data compared to t-SNE, with its known computational and quadratic memory complexity in sample size [180]. UMAP can also capture non-linear effects in gene expression, which is preferable over more traditional dimensionality reduction techniques such as Principal Component Analysis.

Towards this end, I analysed both the full master matrix $\mathbf{M}$ of scaled gene expression in the range $[0, 1]$ consisting of all genes (i.e. $18,364$), and a submatrix consisting of only those genes that belong to a community in at least one tissue (i.e. $3,259$). Similarly to all of the results in the rest of this chapter, I considered only Louvain communities with at least 4 genes.

Drawing conclusions about relationships between clusters (tissues) from UMAP and similar approaches must be done with caution due to some known caveats [71, "Caveats" Section], especially when trying to interpret Euclidean distances between points [281]. UMAP emphasises local distances over global distances, which means that disconnected clusters may change their relative positions even when running the algorithm with the same hyperparameters but different random seeds. With this in mind, I sought to quantify the conservation and variability of UMAP clusters (i.e. global structure), including the relation among biologically-meaningful clusters (tissues). Such structure was characterised using the matrix $[\mathrm{d}(i, j)]$ of pairwise distances for clusters i and j, representing in practice an estimate of the sampling distribution of the global structure.

This quantification problem was approached through a non-parametric bootstrapping procedure. From the master matrix $\mathbf{M}$ of gene expression, I generated a total of B bootstrapped manifolds, each of equal size. Here, each such sample was randomly drawn from 80% of the data points, i.e. rows in $\mathbf{M}$. For the k-th sample, I constructed the matrix $\mathbf{V_{(k)}} = [\widehat{\mathrm{d}(i, j)_{(k)}}]$ of pairwise distances derived from the UMAP embeddings for tissues i and j. Note that $\mathbf{V_{(k)}}$ is a symmetric matrix with zeros along the diagonal. The set $\{\widehat{\mathrm{d}(i, j)_{(k)}}\}_{k=1}^{B}$ allows us to calculate the mean and variance of the UMAP-derived

[3] https://github.com/aestrivex/bctpy

estimator for $\mathrm{d}(i, j)$:

$$\overline{\mathrm{d}(i,j)} = \frac{\sum_{k=1}^{B} \widehat{\mathrm{d}(i,j)_{(k)}}}{B}, \tag{3.3}$$

$$\widehat{\sigma^2_{d(i,j)}} = \frac{\sum_{k=1}^{B} \widehat{\mathrm{d}(i,j)_{(k)}}^2}{B-1} - \left(\frac{\sum_{k=1}^{B} \widehat{\mathrm{d}(i,j)_{(k)}}}{B-1} \right)^2 . \tag{3.4}$$

A heatmap can be used to visualise $\overline{\mathrm{d}(i,j)}$ for each tissue pair (i, j).

For two tissues i_0 and i_1, I define a "clustering conservation coefficient" to quantify the preservation of the clustering of tissues i_0 and i_1 relative to all tissues $\{j\}$:

$$C_{(i_0,i_1)} = \mathrm{corr}(\overline{\mathrm{d}(i_0,j)}, \overline{\mathrm{d}(i_1,j)}), \tag{3.5}$$

where *corr* is the correlation operator. The correlation is calculated for a pair of UMAP-derived distance estimates across all tissues $\{j\}$. In particular, this statistic allows us to formally test the null hypothesis of no conservation of global structure for a given pair of tissues. This coefficient can be extended to a larger set of tissues, $i_0, \ldots, i_l$ (e.g. the 13 brain regions), using the first order statistic:

$$C_{i_0,\ldots,i_l} = \min_{s,t \in 1,\ldots,l} C_{(i_s,i_t)}. \tag{3.6}$$

Collectively, this approach provides a way to perform statistical inference on the UMAP embedded structures.

To evaluate the relevance of the trained UMAP model generated from the GTEx communities, I passed previously unseen data to the model for embedding into the learned latent map. To that end, I used The Cancer Genome Atlas (TCGA) [282] gene expression data. The TCGA is a landmark cancer genomics program, molecularly characterised over 20,000 primary cancer samples spanning 33 cancer types.

3.1.5 Prediction Power on Tissues Gene Expression

I investigated the extent to which each community's gene expression profile was predictive of each of the tissues; as before, Louvain communities with less than 4 genes were filtered out from this analysis. The master matrix $\mathbf{M}$, representing the entire dataset under analysis, has $15,201$ rows representing each RNA-Seq sample from each tissue collected from all subjects, and $18,364$ columns representing the total number of genes available. If a value was non-existent (which may be due to the gene's expression being tissue-specific), a zero value is used, conveying no expression in that tissue.

For each community, the expression values of the member genes were selected from $\mathbf{M}$. With this sliced table, 49 binary classifications were performed using Support Vector Machine (SVM), wherein for each classification, each tissue was predicted. Essentially, the sliced table, which comprises the training data, for a k-member community can be viewed as a collection of vectors $\{(\vec{x_1}, y_1), \ldots, (\vec{x_n}, y_n)\}$, where $\vec{x_i} \in \mathbb{R}^k$ is the gene expression profile of the k genes for the i-th sample and $y_i \in \{1, 0\}$ indicates membership in the tissue to be predicted. The goal of the classification is to separate the tissue to be predicted from the other tissues via the largest margin hyperplane, which can be generically written as $\vec{w} \cdot \vec{x} + b = 0$, where $\vec{w}$ is normal to the hyperplane. SVM was used with a linear kernel and weights were adjusted to be inversely proportional to class frequencies in the input data (this corresponds to setting the *class_weight* parameter in *scikit-learn* to “balanced”). To avoid overfitting, each classification was performed using a stratified 3-fold cross-validation procedure, in which the F_1 score metric

$$F_1 = \frac{2}{(precision)^{-1} + (recall)^{-1}} \tag{3.7}$$

was used to report the prediction power across the three folds. The choice of the F_1 score instead of other metrics was because each binary classification was highly unbalanced, i.e., a given tissue is the positive outcome, and all the other 48 tissues are the negative outcome.

For comparison with the communities, it was also investigated how biologically meaningful sets of genes encoding current biological knowledge are predictive of tissues. For each Reactome pathway[4] the expression of member genes were selected from the master matrix $\mathbf{M}$. If a gene from a Reactome pathway was not present, that gene was ignored. The same stratified 3-fold cross-validation procedure was used here to perform 49 binary classifications.

3.1.6 Enrichment Analysis

To evaluate the degree to which a community corresponds to well-known biological pathways, enrichment analyses were performed using the Reactome 2016 database as a reference. Performing enrichment analysis is a useful statistical tool used to identify groups of over-represented genes in a large set of genes, in which such over-representation may be associated with known biological pathways or disease phenotypes. The *gseapy* python package[5] was used to send calls to the *Enrichr* web API [167]. As per the *Enrichr* official documentation, the p-value is computed using Fisher’s exact test (i.e. hypergeometric test). Those pathways with a Benjamini-Hochberg-adjusted p-value below 0.05 were considered significant. Louvain communities with less than 4 genes were considered not enriched.

3.1.7 Multilayer Analysis

In order to investigate the tissue-shared profiles of gene communities, as well as the relationships between gene expression traits across tissues, I proceeded to model the data as a multilayer network [164]. Formally, a multilayer network is defined as a pair $\mathbf{\Lambda} = (\mathbf{G}; \mathbf{D})$, where $\mathbf{G} = \{G_1, \dots, G_L\}$ is a set of L graphs and $\mathbf{D}$ consists of a set of interlayer connections existing between the graphs and connecting the different layers. Each graph $G_l \in \mathbf{G}$ is a “network layer” with its own associated adjacency matrix A_l. Thus, $\mathbf{G}$ can be specified by the vector of adjacency matrices of the L layers: $\mathbf{A} = (A_1, \dots, A_L)$. Multilayer networks allow us to represent complex relationships which would otherwise be impossible to describe using single-layer graphs separately.

A special case of multilayer networks is a multiplex network used in this chapter to model the GTEx transcriptome data. In this case, all layers are composed of the same set of nodes but may exhibit highly different topologies; in other words, the degree of node i is the vector $d^{[i]} = (d_1^{[i]}, \dots, d_L^{[i]})$, and $d_l^{[i]}$ may vary across the layers. Interlayer connections are established between corresponding nodes across different layers. Layers represent different

[4] The Reactome database is a free, open-source, curated and peer-reviewed pathway database [90].

[5] `https://github.com/zqfang/GSEApy`

tissues, nodes represent genes, and edges between two nodes are weighted according to the correlation weights. In the GTEx data, the correlation matrices define an adjacency matrix A_l for each layer l of the multiplex network.

Using the communities of co-expressed genes for each tissue defined in Section 3.1.3, the so-called *global multiplexity index* [141] was calculated to investigate the relationships of communities across different layers. This index quantifies how many times two nodes (i.e. genes) are clustered in the same communities across different layers. If, for example, gene i and gene j are clustered together in the layer of tissue T_1 and of tissue T_2, then the global multiplexity index is two. In the matrix $[\mathrm{gmi}(i,j)]$ of global multiplexity indices for a multiplex architecture, each element represents the number of times that two given genes, i and j, are clustered in the same community. More formally, if L is the number of layers, N the number of nodes for each layer, and c_i^g the community membership of gene i at graph g, then the global multiplexity index $\mathrm{gmi}(i,j)$ for gene i and gene j, with both i and $j \in \{0,\ldots,N\}$ is defined as follows:

$$\mathrm{gmi}(i,j) = \sum_{g=1}^{L} \delta(c_i^g, c_j^g), \tag{3.8}$$

where $\delta(c_i^g, c_j^g)$ represents the Kroenecker delta function[6]. The value of $\mathrm{gmi}(i,j)$ therefore increases by 1 if the two nodes are found to be part of the same community in a layer. If two genes share a high value of global multiplexity index, this may indicate a greater level of connectivity and suggest a greater functional similarity as they appear multiple times in the same community across different layers.

I tested whether the UMAP embeddings of the communities' transcriptome profiles in a multiplex architecture – a subset of all communities previously interrogated – could also recover biologically-meaningful clusters. This analysis allows an estimation of the high-dimensional transcriptome data's topology and tests whether additional clusters could be uncovered at increasingly finer scales.

3.2 Results

3.2.1 Spurious Co-expression and Confounding due to Unmodelled Factors

The number of factors or components identified by the *sva* package was significantly correlated ($r \approx 0.95$, $p \approx 5.4 \times 10^{-26}$) with the number of samples across tissues (see Figure 3.1a). Notably, the greater number of such surrogate variables regressed out for tissues with larger sample sizes recapitulates the approach used by the GTEx Consortium [4]. Specifically, the GTEx Consortium uses PEER, a related adjustment method, with 15 factors for tissue sample size $N < 150$ and up to 60 factors for $N \geq 350$.

The impact of confound correction in co-expression analysis can be seen in Figure 3.1b. The distribution of Pearson correlation values has more mass closer to zero with less variance after correction, suggesting that unmodelled factors may induce spurious (or artificially inflate) correlations in gene expression. The effect of unmodelled factors is further illustrated in Figure 3.1, Panels (B – D), where the distribution of correlation

[6] The Kroenecker delta function is 1 if the variables are equal, and 0 otherwise.

values for the covariate *Age* is shown for whole blood. Before correction, those values are spread between around -0.4 and 0.4, whereas after correction the corresponding values move towards the centre (zero) and become less dispersed. The enrichment for significant p-values for this covariate is greatly attenuated after the correction, suggesting again that unmeasured variables can induce spuriously significant correlations.

Notably, the variable *Cohort* seems to have undergone the largest change in the correction process (this variable's possible values are *Postmortem* and *Organ Donor* in available tissues, except for some which can also have the *Surgical* value). This change suggests that the estimation of cohort effect on gene expression can be substantially improved by accounting for unmodelled factors.

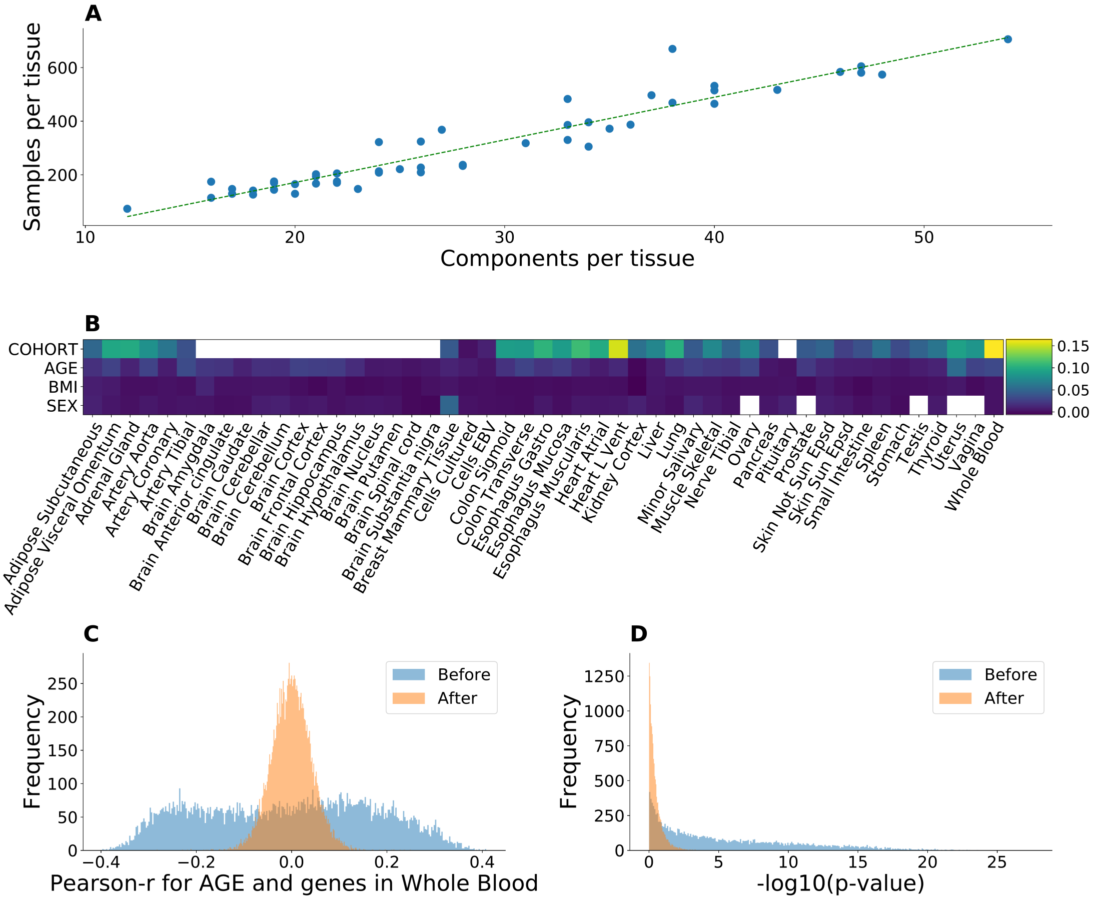


Figure 3.1: **Confounding due to unmodelled factors.** **(A)** Relationship between the number of inferred factors and tissue sample size. Fitted line ($r \approx 0.95$, $p \approx 5.4 \times 10^{-26}$) corresponds to a linear least-squares regression. The two-sided p-value is based on the null hypothesis that the slope is zero, using the Wald Test with t-distribution for the test statistic. **(B)** The difference in the variance of the distribution of Pearson correlation values for each tissue over all genes, before and after correction. Empty cells correspond to tissues in which only one value of the confounder is available. **(C)** Distribution of Pearson correlation between the expression of a gene in whole blood and age, before and after correction. After the correction, the correlation values move towards zero and show considerably less dispersion. **(D)** The p-value distribution from Panel (C)'s values, in logarithmic space.

3.2.2 Atlas of Communities across Human Tissues

The Louvain algorithm identified communities in the co-expression networks for each tissue to develop an atlas across human tissues. Summary statistics on these identified communities can be seen in Figure 3.2. On average, a tissue was found to have 108 communities (standard deviation [SD] = 31). The highest number of communities ($N = 251$) occurred in “Kidney Cortex”, and the lowest number ($N = 73$) in “Muscle Skeletal”. The non-solid tissues, consisting of “Cells EBV” and “Whole Blood”, have the highest number of genes that belong to a community (i.e. at least 4,300 for each). The size of a community varies considerably within each tissue and its distribution differs across tissues. Indeed, even though the median community size was equal to 2 for all tissues, some tissues had communities with size greater than 40, but always below 120. The brain tissues show significantly higher variability (median SD = 9.9, Mann-Whitney U test $p = 1.55 \times 10^{-4}$) than non-brain tissues (median SD = 5.18). Thus, tissues and tissue classes may differ in the overall topology of the communities in co-expression networks, which likely contains a lot of tissue information.

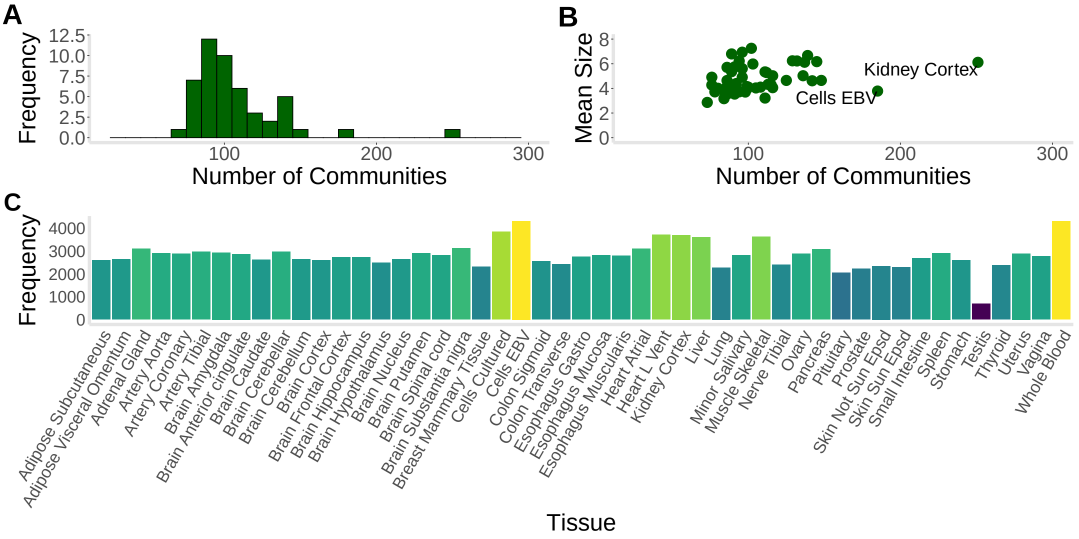


Figure 3.2: **Summary statistics on identified communities. (A)** The histogram shows the distribution of community count in the various tissues (mean = 108, SD = 31). **(B)** The scatter plot displays the community count and mean community size for each tissue, showing a significant correlation (Spearman $\rho = 0.39$, $p = 0.006$). The highest number of communities was observed in “Kidney Cortex” ($N = 251$). **(C)** The plot provides the number of genes that belong to a community in each tissue, with colour gradient used to highlight higher and lower values. The nonsolid tissues, “Cells EBV” and “Whole Blood”, show the highest number of genes with membership in a community.

After removing the weaker correlations ($-0.80 < z_{ij} < 0.80$), most of the subnetworks were already highly segregated from the rest of the network, indicating that just this removal process could almost completely form the Louvain communities. The number of connections coming out of communities of each size was calculated to evaluate the segregation of those communities: for every tissue the mode was zero, and the maximum number was never over 17. Given the thousands of genes in each tissue’s co-expression network, the observed maximum number of connections between different communities (i.e. at most 17) illustrates how strong the segregation is before applying the Louvain community analysis.

3.2.3 UMAP of Community-defined Gene Expression Manifold and Its Persistence

Around 17.7% of the genes belong to a community in at least one tissue. Notably, gene expression from this subset was able to recover the tissue clusters using UMAP, as illustrated in Figure 3.3a, which means that this subset contains sufficient information to recover the tissue clusters. The clustering of related tissues based on organ membership (such as the 13 brains regions), or the clustering of other related tissues based on shared function (such as the hypothalamus-pituitary complex), can be observed in the UMAP projection. When using the complete set of genes, a similar pattern was achieved.

Taken together, these results show that gene expression from the identified communities encodes sufficient information to distinguish the various tissues in a biologically-meaningful low-dimensional representation.

In theory, additional clusters may be present at different scales, such as within a tissue. Therefore, I performed UMAP analysis on the single-tissue "Whole Blood" to test for the presence of additional clusters. Notably, no well-defined clustering was observed concerning Cohort, body mass index (BMI), and other covariates, indicating that the *sva* analysis was successful in removing potential confounders (see Figure 3.1b).

External transcriptome data can be embedded into the trained model generated from the GTEx communities. Indeed, embedding TCGA data from 33 cancer types into the learned UMAP models showed clustering within the testis tissue. This outcome recapitulates two known results: (1) the GTEx finding that the testis is an outlier relative to the other GTEx tissues in transcriptome profile [59], and (2) the role of the so-called cancer-testis (CT) genes [55] that function as driver genes in cancer [244, 275]. Besides, UMAP representations of the genes that belong to a GTEx-derived community recovered the cancer types when using TCGA data to train a UMAP model. The resulting embeddings of one run are depicted in Figure 3.4, highlighting the cross-study relevance of this model.

Using 500 bootstrapped manifolds (i.e. variable B in Section 3.1.4), I found that, on average, related tissues tended to cluster closely together, as illustrated in Figure 3.3b. Examples of such clusters are the 13 brain regions, the colonic and oesophageal tissues, and various artery tissues. I also found a relationship between the average distance between tissue clusters and the variance in the distance, showing a significant positive correlation (Spearman $\rho \approx 0.38$, $p < 2.2 \times 10^{-16}$). Reassuringly, the tissue pairs ("Brain Cerebellum", "Brain Cerebellar") and ("Skin Not Sun Epsd", "Skin Sun Epsd") had the lowest average distance between clusters among all tissue pairs; the first pair consists of known duplicates of a brain region in the GTEx data [60] and is thus expected to cluster together. Among the tissue pairs with the highest average distance, "Adipose Subcutaneous" had an average distance greater than 17 with each of the colonic tissues ("Colon Sigmoid" and "Colon Transverse"), and a low variance comparable to tissue pairs with some of the smallest average distance. Additional global patterns can be easily observed. For example, as reflected in the heatmap, the relationship of the two skin tissues (i.e. "Skin Sun Epsd" and "Skin Not Sun Epsd") to all the other tissues is strongly preserved: the clustering conservation coefficient $C_{(i_0,i_1)}$ is approximately 0.62, with $p = 3.4 \times 10^{-5}$.

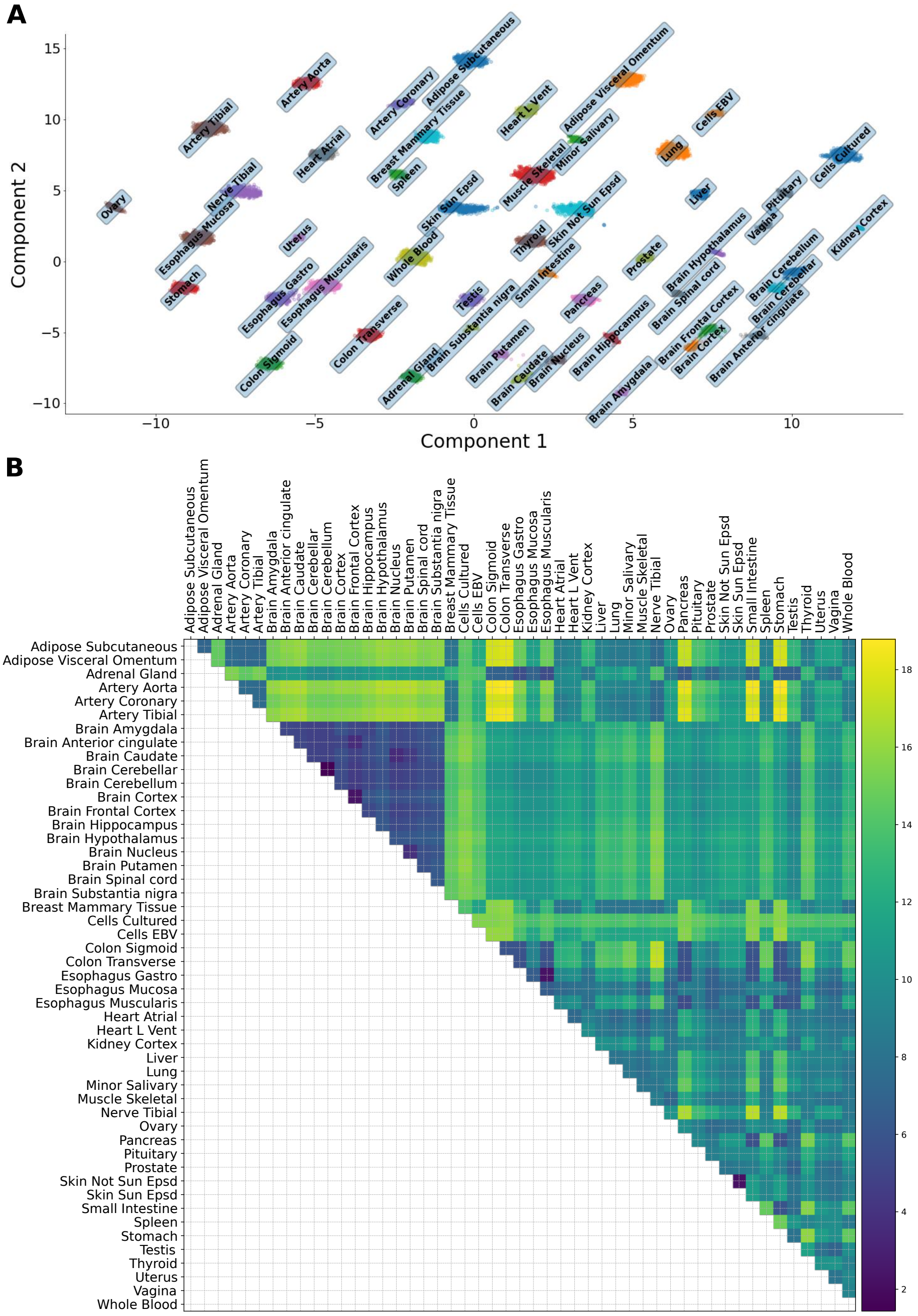


Figure 3.3: **Lower-dimensional UMAP representation of the transcriptome data restricted to the communities and conservation of global structure**. **(A)** UMAP generates embedded structures through a low-dimensional projection of the submatrix consisting of only the genes that belong to a community in at least one tissue ($N = 3,259$). Different colours are used to highlight samples belonging to distinct tissues. **(B)** Using bootstrapped manifolds, the persistence of the global structure and pairwise relationships across tissue clusters is estimated. Here, the upper-triangular matrix of the average pairwise distances across the bootstrapped manifolds is shown, with corresponding colours characterised by the coloured scale bar on the right.

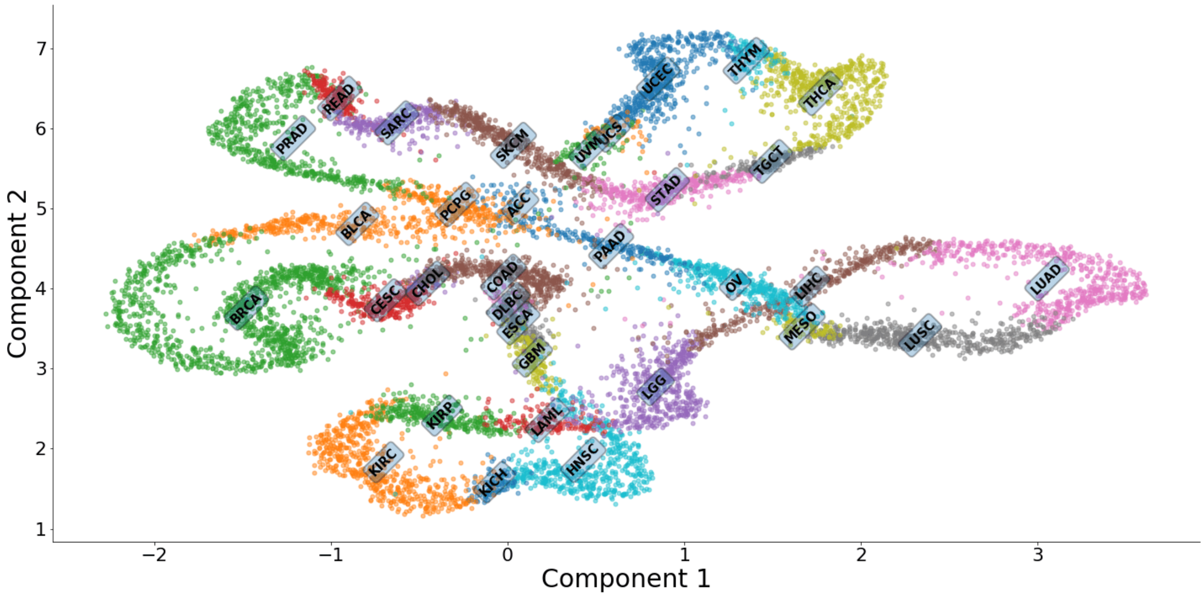


Figure 3.4: **Lower-dimensional representation of the GTEx-derived communities in TCGA transcriptome data.** The plot shows the UMAP embedded components through a low-dimensional projection of the submatrix consisting of only the genes that belong to a community in at least one GTEx tissue ($n = 3,259$). Different colours are used to highlight samples belonging to distinct TCGA cancer types.

The conservation and variability of the UMAP global structure using the TCGA data depicted in Figure 3.4 was also quantified and generated biologically consistent clusterings for each cancer types.

3.2.4 Relationship between Tissues, Communities, and Reactome Pathways

As described in Section 3.1.5, I tested individual communities for their ability to predict a tissue. This section considers that a set of genes can predict a tissue when the average F_1 score is above 0.80. Some broad patterns are noteworthy. Most of the communities from "Whole Blood" do not have predictive power for the other tissues (see Figure 3.5a) partly due to the stringency of the F_1 threshold, which is likely to produce false negatives. This observation indicates that the member genes in each such community from the source tissue ("Whole Blood") cannot "separate" the test tissue from the remaining tissues possibly due to lack of tissue specificity of the community's gene expression profile. However, a community of only five genes can predict the brain region nucleus accumbens (basal ganglia); for this community, the member genes, collectively, are "differentially expressed" between the test brain region and the remaining tissues. Thus, although the genes are present in "Whole Blood" (as a community), the expression profile in the test brain region is substantially different or tissue-specific.

Prediction of tissues by Reactome pathways varies substantially. For instance, consistent with observations for the communities, 197 Reactome pathways are not sufficient to predict any tissue, while 164 are tissue-specific (i.e. can predict only one tissue). However, some Reactome pathways can predict more than half of the tissues: *GPCR LIGAND BINDING*,

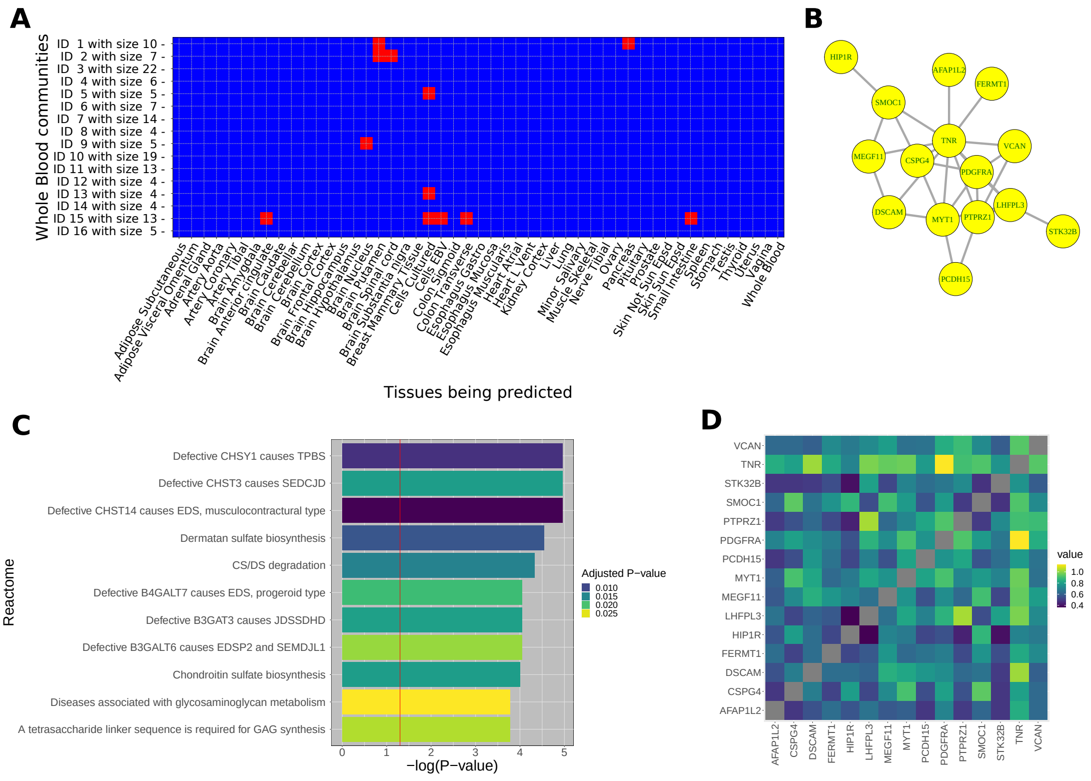


Figure 3.5: **Communities and their properties.** **(A)** Prediction power of "Whole Blood" communities, in F1 scores thresholded over 0.8. **(B)** A 15-member community in the hippocampus is shown here as an example. An edge indicates $A_{ij} > 0.80$ for genes i and j. **(C)** Enrichment analysis was performed on all communities to identify known biological processes. For example, the hippocampal community in panel (B) was found to be significantly enriched for Reactome pathways. P-value refers to raw p-value. Red line corresponds to the raw $p < 0.05$ threshold. Colour gradient reflects the adjusted p-value. All Reactome pathways shown meet adjusted $p < 0.05$. **(D)** Heatmap displays the correlation values for the member genes of the community in panel (B).

GPCR DOWNSTREAM SIGNALING, and *SIGNALING BY GPCR* predict 34, 33, and 32 tissues, respectively. This observation is perhaps expected: G-protein-coupled receptors (GPCRs) comprise a large family of cell surface receptors that form the essential sites of communication between the internal and external environments of cells, with a central and widespread role in human physiology [229]. Their gene expression profile in each of the predicted tissues differs from the remaining tissues, potentially reflecting their broad but tissue-specific function.

Some other patterns can also be seen. The brain tissues "Brain Caudate" (basal ganglia), "Brain Frontal Cortex", "Brain Hippocampus", and "Brain Nucleus" are not predicted by any Reactome pathway, likely reflecting the fact that our current understanding (as encoded in these pathways) have been hampered by the relative inaccessibility of these tissues. In contrast, the two tissues cells cultured fibroblasts and whole blood are the tissues most highly predicted by 269 and 330 Reactomes, respectively. Some tissues are predicted by less than 5 Reactome pathways, including the tissues "Brain Amygdala" (two), "Brain Anterior Cingulate" (one), "Brain Cortex" (two), and "Brain Hypothalamus" (two).

3.2.5 Enrichment of Communities for Known Biological Processes

There were 114 communities (8.28% of all the communities with more than three member genes) enriched for some Reactome pathway (i.e. at an adjusted $p < 0.05$ for level of enrichment), thus contributing in complex ways to multiple biomolecular processes. “Whole Blood” was the only tissue without any community enriched for known pathways, and the “Esophagus Mucosa” was the tissue with the most communities enriched for known pathways, with a total of 5 communities. Since the entire set of communities could fully recover all tissues as clusters in the UMAP embeddings, these results suggest that the remainder of the communities are likely to capture previously inaccessible and novel tissue biology.

Notably, this analysis may uncover the role of these communities in human diseases. For example, as depicted in Figure 3.5, a community of 15 genes in the “Brain Hippocampus” showed a significant enrichment for diseases associated with glycosaminoglycan metabolism. Glycosaminoglycans, which are major extracellular matrix components whose interactions with tissue effectors can alter tissue integrity, have been shown to play a role in brain development [187], modulating neurite outgrowth, and participating in synaptogenesis. Alterations of glycosaminoglycan structures from Alzheimer’s disease hippocampus have been implicated in impaired tissue homeostasis in the Alzheimer’s disease brain [146].

3.2.6 Multiplex Analysis of the Transcriptome

Five multiplex networks were used to model the various tissue interactions in the GTEx dataset. For each multiplex architecture, only the specific component tissues were used to construct the multiplex network, and consequently, the global community index was calculated separately for each multiplex architecture. The five architectures analysed were:

- **All Tissues**: Each layer represents one of the 49 tissues analysed. This architecture allows the investigation of gene communities shared across all tissues, with potentially universal function.

- **Brain Tissues**: The 13 layers correspond to the various brain regions. This architecture facilitates the identification of communities that may play a functional role throughout the central nervous system (CNS).

- **Brain Tissues and Whole Blood**: This multiplex model consists of the 14 layers corresponding to these tissues. This architecture allows the study of brain-derived communities for which the easily accessible whole blood can serve as a proxy tissue.

- **Brain and Gastrointestinal Tissues**: The 16 layers correspond to the brain tissues and three gastrointestinal tissues. This architecture may provide insights into the gut-brain axis, which has attracted recent attention in the literature, such as studies of neuropsychiatric processes and the interaction between the CNS and the enteric nervous system (ENS)[7] in neurological disorders [112, 228].

[7] The ENS is a large part of the autonomic nervous system that can control gastrointestinal behaviour [223].

- **Non-Brain Tissues**: The 36 layers consist of all tissues outside the brain. This architecture may stimulate investigations into developmental and pathophysiological processes outside the CNS.

The groups of genes with the maximal global multiplexity index were extracted in the five architectures, i.e., the groups of genes that share a value of 49, 13, 14, 16, and 36, respectively. These are the number of layers (tissues) in the respective architectures. Revealing the shared community structure across the layers improves the understanding of the functional and disease consequences of genes' clusters. I investigated the biological pathways in which such subgroups were involved for each architecture. The goal was to test the communities for enrichment for known biological pathways and therefore quantify the degree to which the communities capture the current understanding of biological processes as encoded in the knowledge base.

Among other communities, there was a 17-member community in the "Brain Tissues" multiplex that is significantly enriched for the nonsense-mediated decay (NMD) pathway (adjusted $p = 1.01 \times 10^{-37}$), which is known to be a critical modulator of neural development and function [148]. The pathway accelerates mRNAs' degradation with premature termination codons, limiting the expression of the truncated proteins with potentially deleterious effects. The community's presence in all brain regions underscores its crucial protective function throughout the CNS.

The "Brain and Gastrointestinal Tissues" multiplex can be used to illustrate the capacity of this approach to investigate the relationship between two distinct systems. Indeed, a 14-member community present in all 16 layers suggests a strong interaction and shared function across the CNS and the ENS. Consistent with this hypothesis, the community was found to be significantly enriched for the "metabolism of vitamins and cofactors" (adjusted $p = 6.5 \times 10^{-7}$), which has been shown to be responsible for altered functioning of the CNS and ENS [185]. Although the involvement of the individual member genes in this pathway is known, it is a novel finding that the genes are organised as a community structure within co-expression networks persisting across the entire 16 layers of the various brain regions and the gastrointestinal tissues.

The empirical distribution of the global multiplexity index is presented in Figure 3.6 for each of the five architectures. The maximal global multiplexity index in the five architectures represents the groups of genes that share a value of 49, 13, 14, 16, and 36 respectively, equal to the number of layers (tissues) in the respective architectures. These genes appear in the same community across *all* layers of the respective architectures. The proportion at each value k of the index is an estimate of the probability that two genes are clustered in the same communities across k layers.

For comparison with the UMAP embeddings generated from the set of all genes in communities, I performed a similar analysis in the various multiplex networks. For example, I tested whether the complete tissue clustering could be observed using just the subset of communities across all layers of the central nervous system multiplex. I discovered a different clustering pattern, with cultured fibroblasts clustering separately from the rest of the tissues, which in turn no longer show well-defined clustering. This finding suggests the presence of a hierarchy of clusters in the transcriptome at increasingly finer scales.

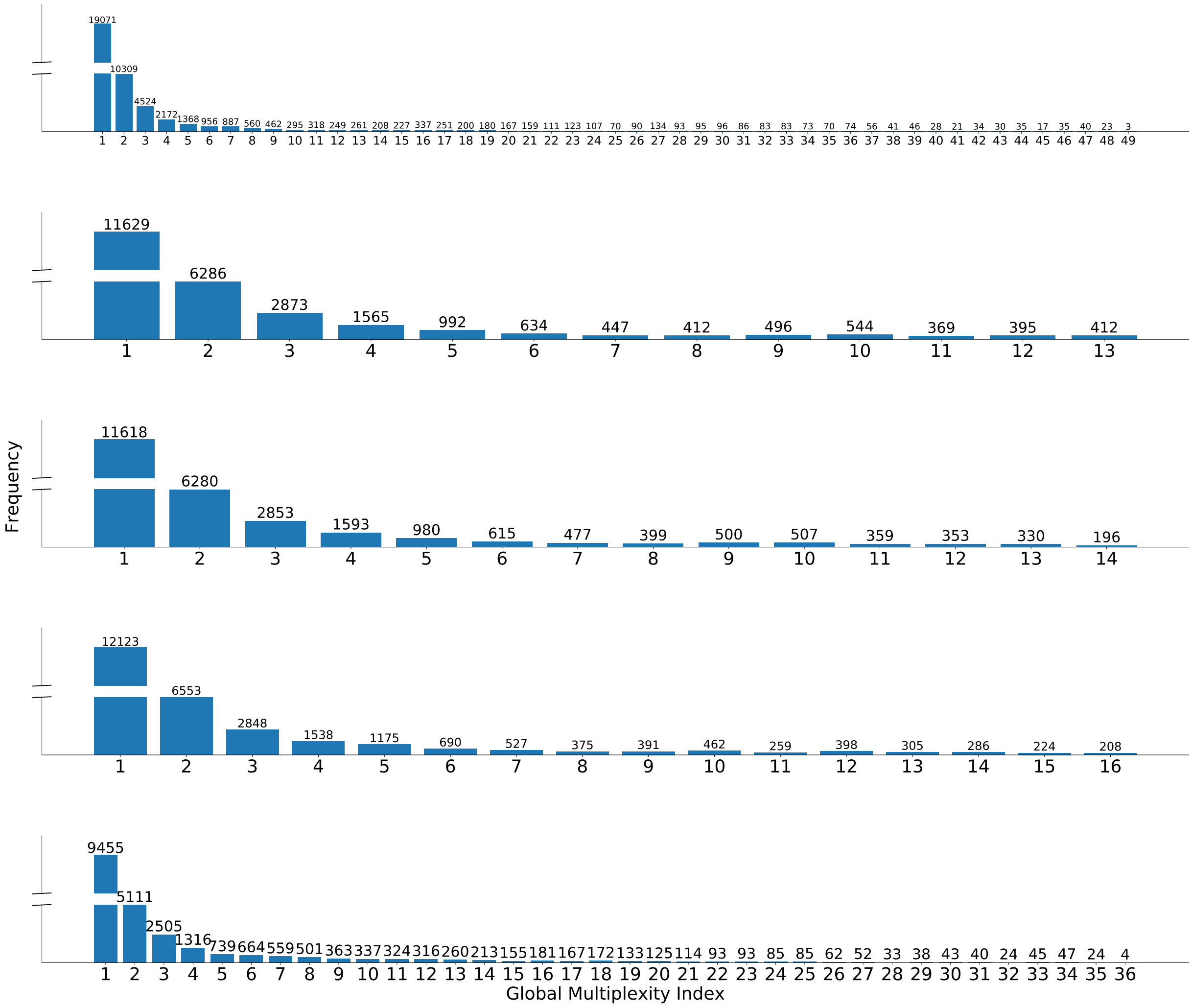


Figure 3.6: **Multiplex analysis.** Histograms show the empirical distribution of the global multiplexity index for each multiplex architecture. Histograms, from top to bottom, correspond to: “All tissues”, “Brain tissues”, “Brain tissues and whole blood”, “Brain and gastrointestinal tissues”, and “Non-brain tissues”.

3.3 Summary

This chapter developed an inter-tissue multiplex framework for the analysis of the human transcriptome. Given the complexity of pathophysiological processes underlying complex diseases, intra-tissue and inter-tissue transcriptome analysis should enable a more complete mechanistic understanding. For these phenotypes, studying the interaction among tissues may provide more significant insights into disease biology than an intra-tissue approach. Communities in co-expression networks were shown to be enriched for some known pathways, encoding current understandings of biological processes; however, it was also possible to identify other communities that are likely to contain novel or previously inaccessible functional information.

UMAP embeddings of the entire set of communities (representing only 17.7% of all genes) fully revealed the tissue clusters. The low-dimensional representation of the subset of communities that are in the multiplex networks did not recover the tissue clusters, but uncovered other clustering patterns, suggesting a hierarchy of clusters at increasingly finer scales. Instead of relying on a single UMAP run [75], an approach to quantify the conservation of, and uncertainty in, the UMAP global structure was developed. New gene expression data can be embedded into these models, facilitating integrative analyses of the large volume of increasingly available transcriptome data. Notably, in external TCGA data, UMAP representations of the genes that belong to a GTEx-derived community induced clustering by cancer type, demonstrating the cross-study relevance of this approach.

This chapter presented a reference atlas of communities in co-expression networks in each of 49 tissues and analysed them through various perspectives. Using the global multiplexity index, I investigated the tissue-sharedness of identified communities. In fact, communities that are shared across multiple tissues may suggest the presence of a tissue-to-tissue mechanism that controls the activity of member genes across the layers in the network. Such regulatory mechanisms have been relatively understudied in comparison with intra-tissue controls. Indeed, some of the communities are shared across multiple tissues; their dysregulation may thus lead to pleiotropic effects and contribute to known and novel comorbidities.

In summary, I performed a network analysis on the most comprehensive human transcriptome dataset available to gain insights into how structures in co-expression networks may contribute to biological pathways and mediate disease processes. I demonstrated the generalisability of this approach and its cross-study relevance through systematic testing in an external dataset. Thus, this chapter provides a publicly available rich resource of co-expression networks, communities, multiplex architectures, and enriched pathways in a broad collection of tissues. I hope that it can catalyse research into inter-tissue regulatory mechanisms and enable insights into downstream phenotypic and disease consequences.

Chapter 4

A Deep Graph Neural Network Architecture for rs-fMRI Data

Resting-state functional magnetic resonance imaging (rs-fMRI) is one of the most commonly used, noninvasive imaging techniques employed to gain insight into human brain function. The use of rs-fMRI data has proven extremely useful as an investigative tool in neuroscience and, to some extent, as a biomarker of brain disease diagnosis and progression [101]. As mentioned in Section 2.3.1, typical use of rs-fMRI data often involves using graph-theoretical measures (such as centrality measures and community structures) to summarise high-dimensional, whole-brain data for use in downstream tasks. As part of this process, it is common practice to reduce the dimensionality of the data in one of three main ways: (1) by collapsing the temporal dimension (e.g., into connectivity matrices between brain regions), (2) by reducing the spatial dimension (e.g., by aggregating voxel-wise signals into predefined brain regions) [277], and (3) by employing approaches that collapse both the temporal and spatial dimensions (e.g., in independent component analyses [30]). These feature engineering steps are performed mostly due to the considerable volume of data in a typical rs-fMRI dataset and its relatively low signal-to-noise ratio [246].

Although computationally beneficial, such dimensionality reduction steps inevitably involve disregarding large amounts of information which can be useful depending on the analysis task. For instance, collapsing the temporal dimension of rs-fMRI data reduces the brain to a static volume where the interactions between different brain regions are fixed over time. This stands in contrast to a growing body of research which shows that functional connectivity in the brain is dynamic and constantly changes over time [17, 174]. As another example, association measures most commonly used are still based on linear models, while it is well known that neuromonitoring data and brain signals, in particular, interact nonlinearly [78, 123].

To overcome such limitations, a different approach to the analysis of rs-fMRI data would be to devise a model that is able to combine both feature engineering and the learning of a low-dimensional representation of the brain's functional activity. Such a model would need to be able to accommodate both the spatial and temporal complexities of rs-fMRI data. To date, deep learning architectures have had great success at leveraging specific inductive biases from complex high-dimensional data. Convolutional neural networks (CNNs), for instance, are extremely effective at extracting shared spatial features such as corners and edges from grid-like data (e.g., 2D and 3D images). These features can then be combined into more complex concepts deeper within the network architecture [249]. Recurrent neural networks (RNNs), on the other hand, are able to learn features from data that are temporally organised as a sequence of steps [79, 80]. In contrast to both CNNs and RNNs, graph neural networks (GNNs) can learn from data that do not have a rigid structure like a grid or a sequence, and can be depicted in the form of unordered entities and relations which constitute graphs (see Section 2.1.8). The formulation of GNN models that deal with complex data structures has recently seen fast developments [288, 302] - such models are therefore strong candidates for the analysis of rs-fMRI data.

In this chapter, I propose a model that uses GNNs to account for spatial interrelationships between brain regions, and temporal convolutional networks (TCNs) to capture the intra-temporal dynamics of blood-oxygenated-level-dependent (BOLD) time series. By incorporating GNNs and CNNs in the same end-to-end architecture, I essentially combine intra- and inter-feature learning. In particular, GNNs can lift the limitation of assuming linearity in the interactions between brain region-specific time series by capturing higher-order interactions between regions of interest (ROIs). The architecture was further engineered to specifically retain edge weights (hence circumventing the common and arbitrary practice of thresholding and binarising adjacency matrices) and to contain elements of explainability [14, 235]. This was done specifically to provide advantages when a neuroscientific explanation of the inner model workings is desirable. To test the architecture, I use the publicly available UK Biobank dataset (see Section 2.3.3), which provides rs-fMRI scans from more than 30,000 distinct people. This dataset offers a unique opportunity to formulate novel architectures, while supporting the need of large datasets for reproducible findings with minimal statistical errors [186]. An ablation analysis was also conducted on a proof-of-concept binary sex prediction task to better evaluate the different contributions of each component of the model. Finally, to assess the effectiveness and flexibility of the architecture, I retrain it using the multimodal Human Connectome Project (HCP) dataset in two distinct experiments, one of which contains multimodal neuroimaging data (i.e., rs-fMRI and structural adjacency matrices derived from diffusion-weighted imaging).

The contribution of this chapter involves the fact that, to the best of my knowledge, this work is the first to leverage both the spatial and temporal information in rs-fMRI data in a single, end-to-end framework that: (1) includes temporal convolutions and graph neural networks, (2) provides the flexibility to extract human-readable, explainability-related patterns which are directly related to the neurobiology and neuroanatomy of the respective brains, and (3) is able to analyse the clusters created by the graph hierarchical pooling mechanism which turned out to carry sensible neurobiological insights. Importantly, the model includes edge features (i.e., weights) when leveraging the graph structure in the network; this information is often ignored in some papers which currently apply GNNs to the study of fMRI data [163].

4.1 Related Work

Previous work using deep learning for analysing rs-fMRI can be broadly grouped by how the spatial and temporal dimensions are processed. For the vast majority of methods, rs-fMRI is treated as euclidean data arranged on a image grid. A commonly used image representation within this domain is the functional connectivity matrix (FCM): a 2D matrix constructed by using a statistical measure of similarity between ROI-derived time series [276]. Both multilayer perceptrons (MLPs) and CNNs have been used extensively on FCMs to learn features in order to classify autism spectrum disorder [87, 135] and attention deficit hyperactivity disorder [225]. A major drawback of using FCMs is that they require an *a priori* choice of similarity measure, possibly introducing unrealistic bias into the data. For example, the often employed Pearson correlation coefficient can only measure linear associations between BOLD signals. More recently, in line with growing interest in dynamic functional connectivity [10, 219], CNNs have been combined with RNNs to learn from time windowed FCMs for tasks such as fluid intelligence prediction [92] as well as identifying major depression [290]. However, in addition to the choice of similarly measure, the construction of classical, dynamic FCMs requires the selection of a window length, which again is arbitrary and not trivial [143]. An alternative to the FCM representation is to use the entire 4D brain volume timeseries as input to convolutional RNNs [3, 32, 210]. Processing voxel-wise fMRI data, however, ignores the empirical evidence that functional brain activity may be localised depending on the task and exhibit very strong spatial correlations [252]. This would result in learning computationally expensive features which likely contain largely redundant information.

In line with the view of the human brain as a dynamical functional connectome, more recent deep learning approaches treat rs-fMRI data as a graph. Within this approach, ROIs are commonly employed to represent graph nodes, and edges between nodes are determined by a choice of similarity measure as per FCMs [2, 252]. In this framework, GNNs can be used to learn features between neighbouring ROIs by propagating information through the edges which connect them. Due to their scalability and interpretability, GNNs for rs-fMRI analysis have been widely used to model tasks such as gender classification [15, 109, 163], age prediction [109], as well as to find imaging biomarkers for brain disorders such as cognitive impairment [284] and autism spectrum disorder [172]. To date, the most common type of graph convolution used for rs-fMRI analysis has been spatial convolutions [109, 173] although spectral [166, 209] and edge convolutions [278] have also proven successful for classification tasks. A major limitation of existing works is that graph topology is estimated by taking a group average of FCMs [163, 278]. As a result, connectivity between subjects is assumed to be invariant. Furthermore, the initial choice of features used to represent ROIs is not trivial, ranging from graph theoretic measures to connectivity differences. These limitations are addressed through the novel combined GNN and CNN model architecture which is capable of learning from individual graph topologies as well as learning its own nodes features.

4.2 Methods

4.2.1 Problem Definition

To represent rs-fMRI data as an undirected weighted graph, the brain is spatially parcellated into N regions of interest (ROIs) representing graph nodes indexed by the set $\mathcal{V} = \{1, \ldots, N\}$. Let $\boldsymbol{x}_i \in \mathbb{R}^T$ represent the features of node i corresponding to the BOLD time series of length T. The connections between ROIs are represented by an edge set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ composed of $|\mathcal{E}|$ unordered pairs (i, j), where for every edge k connecting two nodes $(i, j) \in \mathcal{E}$ the connection strength is defined as $\boldsymbol{e}_k \in \mathbb{R}$. Let the tuple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote the resulting graph. Given the graph structure $\mathcal{G}$, let $\boldsymbol{X} \in \mathbb{R}^{N \times T}$, $\boldsymbol{E} \in \mathbb{R}^{|\mathcal{E}| \times 1}$, and $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ denote the nodes features, edge features and adjacency matrix, respectively.

4.2.2 Temporal Convolutional Networks

There has been evidence that a convolutional operator could perform equally well (or even better) as compared to RNNs for sequential data. Some advantages of the convolutional operator are, for instance: (1) lower requirements for long input sequences, especially compared to LSTMs and GRUs, which commonly consume big chunks of memory to store partial results for the multiple gates (convolutional kernels, in contrast, are shared across a layer), (2) better parallelisation because a TCN/CNN layer is processed as a whole instead of sequentially as in RNNs, and (3) easier to train (e.g., it is known that LSTM training can commonly encounter issues with vanishing gradients). Other teams in industry and academia have found similar results when using convolutional operations for sequential data, for instance, in sequence-to-sequence prediction/learning [85, 115], machine translation [156, 157], and others [58]. In summary, there is evidence that although LSTMs have historically been used for sequential data, CNNs can achieve similar or better performance at a significantly lower cost (I will empirically revisit this point in Section 4.4.2).

In order to learn a representation of the temporal dynamics contained in rs-fMRI time series, I use temporal convolutional networks (TCNs) [24]. These are a simplification over the original *WaveNet* architecture used for audio synthesis [266], which has been seen to provide significantly better results for sequence modelling in comparison to more traditional RNN architectures (e.g., LSTMs) across a range of tasks and datasets. In particular, Bai et al. [24] posit that convolutional networks should be seen as the natural starting point for sequence modelling tasks, which makes them ideal for extracting information from rs-fMRI time series.

TCNs are based on dilated causal convolutions [296], which are special 1D filters where the size of the receptive field exponentially increases over the temporal dimension of the data as the depth of the network increases. The padding of the convolution is 'causal' in the sense that an output at a specific time step is convolved only with elements from earlier time steps from the previous layers, thus preserving temporal order. More formally, given a single ROI time series $\boldsymbol{x}_i \in \mathbb{R}^T$ and a filter $\boldsymbol{f} \in \mathbb{R}^K$, the dilated causal convolution operation of $\boldsymbol{x}$ with $\boldsymbol{f}$ at time t is represented as

$$\boldsymbol{x}_i * \boldsymbol{f}(t) = \sum_{s=0}^{K-1} \boldsymbol{f}(s)\boldsymbol{x}_i(t - d \times s), \tag{4.1}$$

where $d = 2^{l-1}$ is the dilation factor which, depending on the layer l controls the number of time steps successively skipped. This relation between the dilation factor and the layer l is the one defined in the original paper [24], which I follow in this chapter. When compared with the original TCN architecture, batch normalisation was used instead of weight normalisation because it empirically provided a more stable training procedure in terms of loss evolution.

4.2.3 Graph Network Block

In this section I will present the formalisation of graph neural networks followed in this chapter, which can be seen as a specific case of the more general one introduced in Section 2.1.8.

Battaglia et al. [29] formalise a graph network (GN) framework through the definition of functions that work on graph-structured representations. The main unit of computation in the GN framework is called the *GN block* and contains two update functions and one aggregation function working on the edge and node levels.

The first operation of this GN block, which can be broadly defined as the *edge model*, concerns the update function ϕ^e, which computes updated edge attributes for each edge k based on the original edge's attributes $\boldsymbol{e}_k$ and the features of the connected nodes i and j:

$$\boldsymbol{e}'_k = \phi^e\left(\boldsymbol{e}_k, \boldsymbol{x}_i, \boldsymbol{x}_j\right). \tag{4.2}$$

Note that for rs-fMRI graph representations, each edge originally contains a single value (i.e., $\boldsymbol{e}_k \in \mathbb{R}$), but after this operation ϕ^e, the resulting dimensionality can be different: $\boldsymbol{e}'_k \in \mathbb{R}^M$, where $M >= 1$. Then, in what can be broadly defined as the *node model*, the block computes updated node features. Firstly, for each node i, it aggregates the edge features per node:

$$\overline{\boldsymbol{e}}'_i = \rho^{e\rightarrow v}\left(\mathcal{E}'_i\right), \tag{4.3}$$

where $\mathcal{E}'_i = \{(\boldsymbol{e}'_k, i, j)\}_{k=1}^{E}$ is the set of edges starting in node i, with node j connected with node i through edge k. Importantly, $\rho^{e\rightarrow v}$ needs to be invariant to edge permutations to account for the unordered structure of the data. Averaging and summation are examples of such operations invariant to edge permutations.

Finally, the updated node features are computed using another update function at the node level, for each node i:

$$\boldsymbol{x}'_i = \phi^v\left(\overline{\boldsymbol{e}}'_i, \boldsymbol{x}_i\right). \tag{4.4}$$

The aggregation function $\rho^{e\rightarrow v}$ needs to be invariant to edge permutations, but the update functions (i.e., ϕ^e and ϕ^v) are more flexible. For example, if the features are vectors in 1D space, the update functions could be implemented as multi-layer perceptrons (MLPs); however, a TCN or RNN could be more suitable if the features represent images or sequences, respectively. Section 4.3.2 details how these functions were implemented.

Although the rs-fMRI graph representation contains undirected edges, the GN block requires directed edges. To overcome this issue, every time there is a connection between any two nodes i and j, it is assumed the existence of two edges $(\boldsymbol{e}_k, i, j)$ and $(\boldsymbol{e}_k, j, i)$, one for each direction. The original GN block [29] further contains one update function and two aggregation functions for global (i.e., graph-level) features; however, this formalisation it is not applicable in the fMRI data representation of this chapter.

4.2.4 Graph Pooling

After the neural network processes the input as described in the previous sections, each node in the graph will contain a node-wise representation (i.e., a feature vector) as a result. For the prediction task described in this chapter, where a graph-level (as opposed to node-level) prediction is required, these representations need to be pooled (i.e., collated) to be used for a final downstream prediction task.

To this end, it is common practice to employ a global pooling mechanism, in which node features are pooled across the graph (e.g., by averaging or concatenating all node features), thus creating a final, low-dimensional embedding representation of the graph itself. Given the graphs used in this chapter all have the same number of nodes, a concatenation pooling mechanism is indeed possible.

However, assuming that distinct nodes (i.e., brain regions) have different levels of importance for the downstream prediction task [137, 161], a hierarchical (as opposed to flat) pooling mechanism could create richer embeddings. To this end, I employ the differentiable pooling operator introduced by Ying et al. [294], commonly called DiffPool, which learns how to sequentially collapse nodes into smaller clusters until only a single node with the final embedding exists.

When describing a Graph Network (GN) block, a sparse representation of nodes and edges is used to describe the operations that a GN block can have; however, DiffPool works on dense representations of a graph. In other words, a graph $\mathcal{G}$ is represented by a dense adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ and a feature matrix $\boldsymbol{X} \in \mathbb{R}^{N \times F}$, where N is the number of nodes and F the number of features in each node.

The DiffPool operator, at layer l, thus receives both an adjacency matrix and a node embedding matrix, and computes updated versions of both:

$$\boldsymbol{A}^{(l+1)}, \boldsymbol{X}^{(l+1)} = \text{DiffPool}\left(\boldsymbol{A}^{(l)}, \boldsymbol{X}^{(l)}\right). \tag{4.5}$$

To achieve this, the DiffPool operator uses a graph neural network (GNN) architecture. Specifically, the same GNN architecture is duplicated to compute two distinct representations: a new embedding $\boldsymbol{Z} \in \mathbb{R}^{N_{(l)} \times F'}$ and an assignment matrix $\boldsymbol{S} \in \mathbb{R}^{N_{(l)} \times N_{(l+1)}}$:

$$\boldsymbol{Z}^{(l)} = \text{GNN}_{l,\text{embed}}\left(\boldsymbol{A}^{(l)}, \boldsymbol{X}^{(l)}\right) \tag{4.6}$$

$$\boldsymbol{S}^{(l)} = \text{softmax}\left(\text{GNN}_{l,\text{pool}}\left(\boldsymbol{A}^{(l)}, \boldsymbol{X}^{(l)}\right)\right), \tag{4.7}$$

where $N_{(l)}$ is the number of nodes in layer l, $N_{(l+1)}$ the new number of nodes, each corresponding to a cluster ($N_{(l+1)} < N_{(l)}$), and F' the number of features per node, which can be different from the original size F from the matrix $\boldsymbol{X}$.

The operator ends with the creation of the new node embedding matrix and adjacency matrix, to be inputted to the next layer:

$$\boldsymbol{X}^{(l+1)} = \boldsymbol{S}^{(l)^T} \boldsymbol{Z}^{(l)} \tag{4.8}$$

$$\boldsymbol{A}^{(l+1)} = \boldsymbol{S}^{(l)^T} \boldsymbol{A}^{(l)} \boldsymbol{S}^{(l)}, \tag{4.9}$$

where $\boldsymbol{X}^{(l+1)} \in \mathbb{R}^{N_{(l+1)} \times F'}$ and $\boldsymbol{A}^{(l+1)} \in \mathbb{R}^{N_{(l+1)} \times N_{(l+1)}}$. Overall, equations 4.6-4.9 are the ones responsible to implement Equation 4.5.

4.3 Experiments Overview

4.3.1 Main Dataset - UK Biobank

Subject-level structural T1 and T2-FLAIR data as well as ICA-FIX [234] denoised rs-fMRI data were obtained from UK BioBank (application 20904) [48][1]. All data were acquired on

[1] https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf

a standard Siemens Skyra 3T scanner running VD13A SP4, with a standard Siemens 32-channel RF receive head coil. The structural data were further preprocessed with Freesurfer (v6.0)[2] using the T2-FLAIR weighted image to improve pial surface reconstruction, similarly to Glasser et al. [122]'s pipeline, as briefly mentioned in Section 2.3.3. Reconstruction included bias field correction, registration to stereotaxic space, intensity normalisation, skull stripping, and white matter segmentation. When no T2-FLAIR data were available, Freesurfer reconstruction was done using the T1-weighted image only. Following surface reconstruction, the Desikan-Killiany atlas [70] was aligned to each individual structural image, and ROIs were mapped into each individual's space for subsequent time series extraction. To this end, the same atlas was aligned to the functional denoised rs-fMRI data (490 volumes TR/TE = 735/39.00 ms, multiband factor 8, voxel size: 2.4×2.4×2.4, FA=52 deg, FOV 210x210 mm) using the warping parameters computed during the structural-to-functional alignment obtained using FSL's linear registration (FLIRT), and mean BOLD time series (490 timepoints per scan) were extracted for each ROI. The time series were then scaled subject-wise using the median and interquartile range according to the *RobustScaler* implementation in the *scikit-learn* [215] python package. Edge weights were defined as full correlations calculated with the Ledoit Wolf covariate estimator using the *nilearn* python package[3]. Figure 4.1 shows an example scaled time series and the resulting example graph from a single subject. The total number of subjects used from the UK Biobank was 35,159, in which 18,649 were females and 16,510 were males $(18,649/16,510 \approx 1.13)$. The median age was 64, with a minimum age of 44 and a maximum of 81.

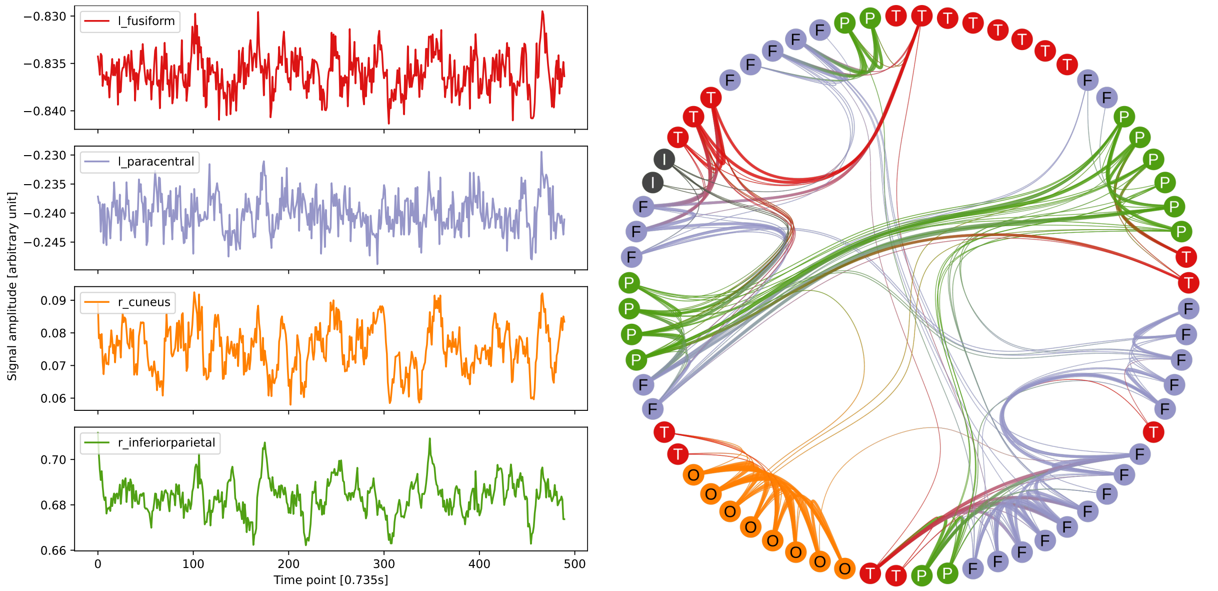


Figure 4.1: Left: Mean BOLD time series extracted from four brain regions (see legend) from one subject's data, after scaling. Each time point takes 0.735 seconds. Right: Graph representation of the same subject's data, at 10% threshold as described in Section 4.3.2. Thicker edges represent a stronger correlation between nodes, in this case with values between approximately 0.54 and 0.87. Each node is labelled and coloured according to the brain region it represents (i.e., T/F/O/P/I correspond to Temporal, Frontal, Occipital, Parietal, and Insula).

[2] http://surfer.nmr.mgh.harvard.edu/

[3] https://nilearn.github.io/

4.3.2 Model Implementation

A conceptual summary of the neural network architecture used in this chapter is shown in Figure 4.2, which was implemented using Pytorch [212], and Pytorch Geometric [96] for the specific graph neural network components. The edge feature matrix $\boldsymbol{E} \in \mathbb{R}^{E\times 1}$ defined in Section 4.2.1 was implemented as two sparse matrices: a sparse representation of the adjacency matrix $\mathbf{E}_i \in \mathbb{R}^{2\times E}$, and a sparse representation of the edge features $\mathbf{E}_a \in \mathbb{R}^{E\times 1}$ (i.e., there was only one feature per edge corresponding to the correlation value). The number of nodes N was 68 (corresponding to each brain region from the Desikan-Killiany atlas), the number of node features F was the number of timepoints (i.e., 490), and E is the number of edges in the graph. The number of edges depends on the threshold percentage used to retain only the strongest correlations. Given the non-conclusive evidence on the optimal threshold percentage in the vast majority of functional connectivity literature [114], in this work this threshold was included in the hyperparameters to be optimised.

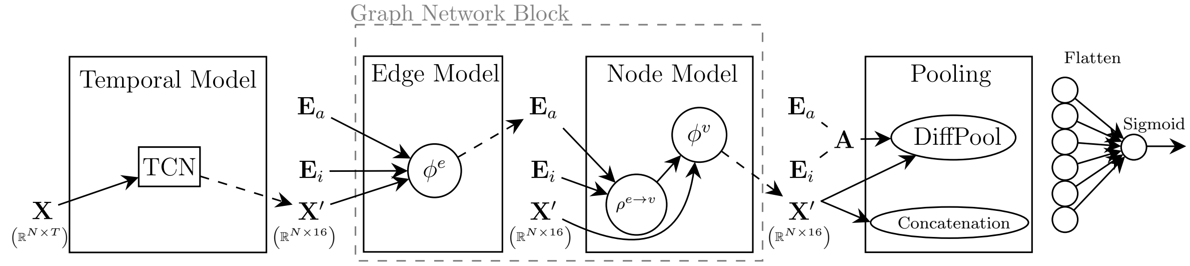


Figure 4.2: Main working blocks of the spatio-temporal model. The temporal model creates an initial representation from the original node features $\boldsymbol{X}$ (i.e., temporal dynamics). This is followed by transformations operated by the Graph Network Block which leverages the structure of data represented in edge features $\boldsymbol{E}_a$ and its sparse connectivity $\boldsymbol{E}_i$. Finally, a Pooling mechanism (either DiffPool or concatenation) creates a graph representation which is flattened and employed for a final prediction task.

The full list of hyperparameters and respective value ranges were as follows:

- `dropout`: $[0, 0.5]$ (uniform distribution)
- `threshold`: $\{5, 10, 20, 30, 40\}$ (categorical)
- `learning rate`: $[1\mathrm{e}{-5}, 1\mathrm{e}{-1}]$ (log uniform distribution)
- `weight decay`: $[1\mathrm{e}{-12}, 1\mathrm{e}{-1}]$ (log uniform distribution)

The model starts by employing a temporal convolutional network (TCN) architecture [24] to extract a lower-dimensional embedding representation from the rs-fMRI time series in each node. This was implemented by using three blocks, each of which containing two layers of *1D convolutions*, *1D batch normalisation*, *ReLU activation*, and *dropout*. Each block uses a kernel with size 7 (i.e., $K = 7$ in Equation 4.1), containing a skip connection, and increases the number of output channels at each block, specifically 8, 16, and 32. Dilation factor d was set to $d = 2^{l-1}$, where l is the block (i.e. $l \in \{1, 2, 3\}$). After these three blocks (i.e., six layers), node features from all channels are flattened to form the input to a linear transformation which reduces each node representation to a fixed embedding of size 16. These transformations thus reduce the original node feature matrix from size $N \times T$ to size $N \times 32 \times T$ after the three blocks, and finally to size $N \times 16$, corresponding to the final embedding.

The Graph Network (GN) block is then applied, in which the update functions ϕ^e and ϕ^v in equations 4.2 and 4.4 are multi-layer perceptrons (MLPs), and the function $\rho^{e \to v}$ in Equation 4.3 is a set of aggregators following Corso et al. [62]'s work (i.e., edge-wise mean, min, max, standard deviation, and sum). I stack 3 GN blocks, after each of which an *1D batch normalisation* over the node's features and a *ReLU activation* are applied. The original dimensions of $\boldsymbol{X}$, $\boldsymbol{E}_i$, and $\boldsymbol{E}_a$ before the GN block are kept after these transformations.

Two types of pooling mechanisms were analysed, both of which reduce the node feature matrix from a size of $N \times 16$ to a size of 1×16: a concatenation over all node's features followed by a single-layered MLP, and the hierarchical pooling mechanism (i.e., DiffPool). For DiffPool, which expects a dense graph representation, data are first transformed into a symmetric adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$, which is a weighted matrix when considering edge features, and binary otherwise. Similarly to the original DiffPool paper [294], DiffPool employs three layers of the graph neural network operator from Morris et al. [195] (to make use of weighted adjacency matrices) followed by a *1D batch normalisation*, with a final skip connection.

4.3.3 Training Procedure

In order to assess the validity of the model, I performed proof-of-concept experiments through the well-known binary sex prediction task [151, 283]. I used a 5-fold stratified cross-validation procedure: the UK Biobank dataset was divided into training and test sets five times, in which each test set corresponds to 20% of the original size, and a sample would only belong to a test set once (i.e., all test sets are mutually exclusive). This division was done in a stratified fashion considering the sex label, bucketised age, and bucketised BMI measures (for each variable, eight equal-sized buckets were created based on sample quantiles). For each test set, the training set is further divided once to generate single inner training and validation sets, using the same stratification strategy as for the training/test case.

The neural network was trained over 150 epochs with the RMSprop optimiser [262] and Binary Cross-Entropy loss function. The training procedure was set to stop early if the validation loss did not reduce further after 33 consecutive epochs. Learning rate is reduced by a factor of 0.1 after 30 epochs of validation loss not improving (i.e. when the learning plateaus for 30 consecutive epochs). A hyperparameter search was included in the inner training/validation sets, in which 25 random runs were launched exploring random values of dropout, edge threshold, learning rate, and weight decay (see Section 4.3.2 for ranges explored). In each random run, the model with the smallest validation loss was saved, and the model with the smallest validation loss across the 25 runs was selected to be evaluated in the test set. This procedure is done separately for each test set, and metrics are then averaged across the five test sets.

Weights & Biases [35] was used to log the training procedure and generate the random hyperparameters for all the 25 models in each inner sweep. These inner sweeps were run across two different servers, and each model took between 20 minutes and 11 hours to train depending on GPU type and early stopping. All these details are stored using *Weights & Biases*, and can be accessed through my public repository (see Section 1.4). Figure 4.3

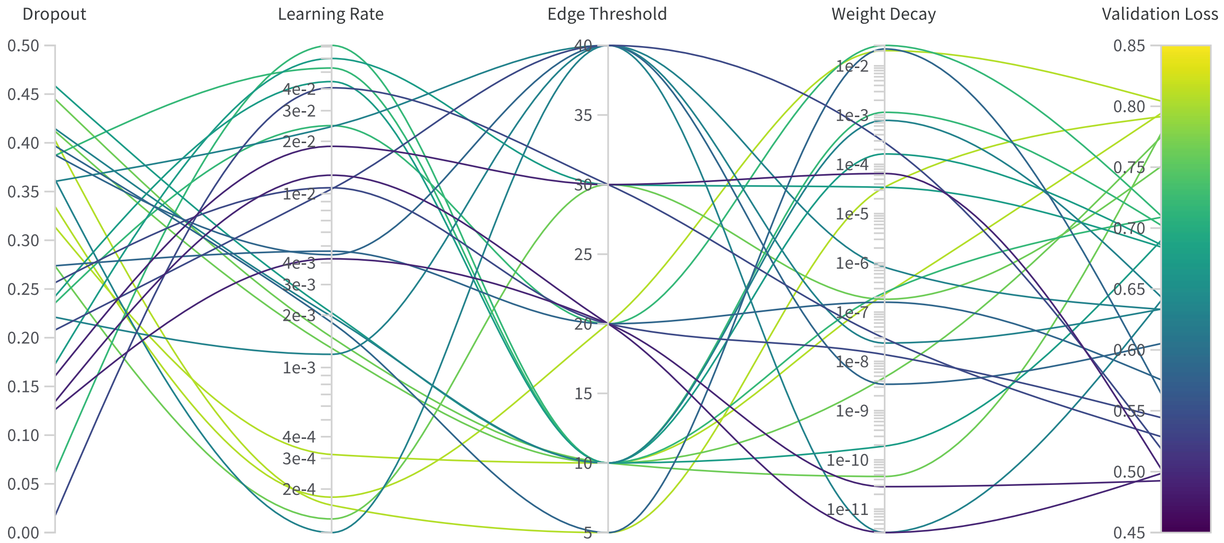


Figure 4.3: Values of hyperparameters corresponding to each validation loss achieved for one illustrative inner sweep of one fold. For each one of the 25 training runs (each represented by a curved line), a set of random values is chosen for dropout, learning rate, edge threshold and weight decay, which ultimately results in the model's validation loss.

shows the results for the inner sweep of one of the folds for illustrative purposes. While a certain amount of variability is visible, some trends are evident in this particular split: the best models (i.e., with lower validation loss) tend to be achieved with higher edge thresholds, higher learning rates, and lower dropout rates. It should be highlighted that different sweeps could result in different trends.

4.3.4 Evaluation

As shown in Figure 4.2, the model consists of (1) a TCN block that learns intra-temporal features from the mean BOLD time series of each ROI, followed by (2) a GN block which leverages the spatial inter-relationships between ROIs, and finally (3) a pooling mechanism which leverages all the information in the input, from the temporal rs-fMRI dynamics to the graph structure and the edge features of that graph.

To understand the inner workings of this combination, I conducted an ablation analysis to quantify the contributions of each component of the model for the specific prediction task. Firstly, the two cases where the GN block is not used are considered, hence essentially evaluating the importance of edge weights for this prediction task. In one case the graph structure is completely ignored (i.e., no GN block and concatenation pooling), and in another case a binary graph is used only for the final hierarchical pooling part (i.e., no GN block and DiffPool applied to a binary graph).

In order to investigate the influence of the different GN components, I consider not only the case where both *node model* and *edge model* are used in the GN Block, but also a case where only the *node model* is applied. For each of these two cases, both a concatenation pooling and DiffPool with weighted adjacency matrices are considered.

The model is compared with two deep learning models. The first one, by Gadgil et al. [109] (named CNSLAB) is based on a voting scheme across timesteps, and the second, by Wang et al. [278] (named cGCN), uses averaged FCMs. For both, I used the best hyperparameters selected from each paper/repository and trained those models on my preprocessed data.

The model is also compared with baseline models where data structure is not leveraged; here, the entire data representation is flattened and fed into two non-deep learning models, namely: (1) a support vector machine (SVM) classifier with a linear kernel and hyperparameter search over the regularisation parameter, and (2) a XGBoost [57] classifier with hyperparameter search over several parameters.

4.3.5 External Multimodal Dataset - Human Connectome Project

To further evaluate the effectiveness and flexibility of the end-to-end architecture, its behaviour is analysed in a multimodal setting, i.e. when adjacency matrices and time-series are derived from distinct imaging procedures (rs-fMRI and diffusion-weighted MRI, respectively). I employed the preprocessed Human Connectome Project (HCP) rs-fMRI data. This dataset consists of four 15-minute-long fMRI sessions (TR = 0.72s) per subject, acquired on a 3T scanner with isotropic spatial resolution of 2mm in 1,003 healthy subjects, and preprocessed according to Glasser et al. [122]. For each subject, this results in 4 distinct sessions/samples per subject with 1,200 timesteps for each sample and component. In order to ensure comparability to the UK Biobank experiments, every timeseries was truncated to 490 timepoints. Similarly to the steps described in Section 4.3.1, the Desikan-Killiany atlas [70] was aligned to each individual structural image, warped into single subject space, and employed to extract ROI- and subject-wise timeseries which were scaled subject-wise. Diffusion data was processed locally using multi-tissue, multishell constrained spherical deconvolution [150] to obtain orientation distribution function estimates, which were then passed to probabilistic fiber tracking (10^8 tracks, subsampled to 10^7 tracks through Spherical-deconvolution Informed Filtering of Tractograms [245]). Structural connectivity matrices were obtained by length-normalised streamline counts between the same ROIs described above. A total of 3,668 graphs were used in which 1,976 were females and 1,692 were males ($1,976/1,692 \approx 1.17$). Nodes correspond to Desikan-Killiany ROIs, node features correspond to 490 time points, and the adjacency matrix corresponds to the structural connectivity extracted from tractography.

All training and evaluation steps were kept identical across all datasets.

4.4 Results

4.4.1 General Results

Table 4.1 shows the results of the ablation analysis across three different backbones - no graph block, only *node model*, and full graph network block - each with two different aggregators (i.e., concatenation and DiffPool). For notation purposes, each one of these cases is described in the form "Backbone → Aggregator", in which *Aggregator* can be "Concat" (i.e. Concatenation) or "DiffPool", and *Backbone* can be "N" for only *node model*, "N + E" for both *node model* and *edge model* (i.e., full GN Block), and empty otherwise. The table also includes results from the baselines experiments.

The models developed in this chapter perform significantly better as compared to all baselines but in which the model without a GNN block (i.e., "→ Concat') is similarly good. The SVM baseline performs worse overall and involves an increase in the standard deviation of performance parameters, possibly indicating that the deep learning model

Table 4.1: Ablation analysis, with metrics averaged across the five test sets, with standard deviation in parenthesis. Aggregator on the right-hand side of the arrow, "N" corresponds to only *node model*, and "N + E" corresponds to full Graph Network block. **Params** stands for number of parameters.

Model	**AUC**	**Accuracy**	**Sensitivity**	**Specificity**	**Params**
N + E → Concat	**0.92** (0.004)	**0.85** (0.006)	**0.85** (0.006)	0.84 (0.012)	291,898
N + E → DiffPool	0.82 (0.020)	0.75 (0.016)	0.72 (0.030)	0.77 (0.025)	287,420
N → Concat	**0.92** (0.003)	0.84 (0.004)	0.84 (0.028)	**0.85** (0.029)	291,337
N → DiffPool	0.84 (0.020)	0.76 (0.020)	0.75 (0.013)	0.77 (0.038)	286,859
→ DiffPool	0.84 (0.010)	0.76 (0.008)	0.75 (0.019)	0.77 (0.023)	278,843
→ Concat	**0.92** (0.012)	0.84 (0.013)	0.84 (0.024)	0.84 (0.023)	283,321
CNSLAB [109]	0.86 (0.003)	0.78 (0.005)	0.76 (0.024)	0.79 (0.018)	198,937
cGCN [278]	0.77 (0.021)	0.70 (0.018)	0.66 (0.028)	0.74 (0.040)	45,065
XGBoost	0.89 (0.003)	0.81 (0.005)	0.80 (0.008)	0.82 (0.006)	-
SVM	0.79 (0.015)	0.79 (0.017)	0.82 (0.098)	0.76 (0.101)	-

is more robust to different dataset divisions (i.e., folds), while retaining the flexibility and representation ability described above. Using DiffPool as a final aggregator appears to result in worse overall performance when compared to the concatenation counterpart and, in some metrics, to some baselines. Using the *edge model* did not bring significantly better results when compared to using the *node model* only, possibly indicating that the information contained in the edge attributes is successfully leveraged by the *node model* alone for this particular prediction task.

The results presented so far consider an adjacency matrix threshold below 50% as a hyperparameter at training time, a common data reduction practice in the connectivity analysis field. These results were further analysed when using no threshold at all, and the type of activation function was explored as a hyperparameter instead (i.e., *ReLU* or *tanh* activations). This choice was made explicitly since retaining 100% of the adjacency matrix elements results in a share of negative correlation elements, whose physiological significance is likely to be important in brain connectivity [299]. The results of this analysis are presented in Table 4.2.

The performance was slightly lower for most cases which did not include a threshold, especially for the N + E → DiffPool model. A possible explanation would be the excessive "noise" (i.e., low, possibly spurious correlations) not allowing the graph's dominating spatial structure to be successfully leveraged in a practical timeframe, in turn possibly resulting in some degree of overfitting. However, performance metrics remain comparable or better to what is illustrated in Table 4.1, suggesting that these models are still able to extract spatial information from the data after training despite of the significant increase in memory usage.

Table 4.2: Results with no thresholded graphs, with metrics averaged across the five test sets, with standard deviation in parenthesis. Aggregator on the right-hand side of the arrow, "N" corresponds to only *node model*, and "N + E" corresponds to full Graph Network block.

Model	**AUC**	**Accuracy**	**Sensitivity**	**Specificity**
N + E → Concat	0.92 (0.002)	0.84 (0.004)	0.85 (0.014)	0.83 (0.017)
N + E → DiffPool	0.77 (0.012)	0.70 (0.011)	0.68 (0.080)	0.72 (0.067)
N → Concat	0.93 (0.003)	0.85 (0.003)	0.83 (0.017)	0.86 (0.019)
N → DiffPool	0.85 (0.007)	0.77 (0.008)	0.77 (0.026)	0.77 (0.017)

4.4.2 Evaluating Architectural Choices

To better understand the utility of TCNs when compared to the more traditional LSTMs, I reran six ablations using the UK Biobank dataset, in which the TCN block was substituted with a LSTM block. Striving for a fair comparison between LSTM and TCN, I used the same number of layers in both (i.e., three layers), and chose the feature dimension in the hidden state such that the total number of learnable parameters would be similar. The 25 runs per fold have the same hyperparameter ranges in both the TCN and LSTM cases. Table 4.3 shows that the LSTM models achieve similar performance to the TCN models; however, this comes at a significantly higher computational cost. Due to computational constraints, I am not able to fairly compare the runtimes among all models because of the use of different servers with different GPU cards. However, there are two folds in the "N → DiffPool" model (i.e., folds 4 and 5) which were run in the same GPU for both the TCN and LSTM cases; in this case, the average runtime per model training went from 1 hour and 35 minutes (fold 4) and 1 hour and 38 minutes (fold 5) in the TCN case, to an average of 3 hours and 14 minutes (fold 4) and 2 hours and 47 minutes (fold 5) in the case of the LSTM. Given that these specific four folds were run on the most recent NVIDIA A100 GPUs, which are able to speedup runtimes by a large factor when compared to older GPUs, I expect these differences to be more striking when running the models on more commonly used hardware.

In summary, these experiments confirm findings that RNNs and TCNs can provide similar performance, but the former come with a significantly higher computational cost.

The impact of including the TCN block in the model was further evaluated. In this "no TCN" experiment, I omitted the TCN block and therefore only the GNN components are present, with a much larger temporal feature representation (i.e., 490 raw timepoints instead of the 16 features created by the TCN block). Table 4.4 shows that performance metrics were similar between the TCN and "no TCN" versions, but the latter resulted in an almost 100-fold increase in the number of parameters. This means that removing the TCN block came at a very significant cost of an unnecessary explosion in the number of learnable parameters, making the model unnecessarily complex both at training and test time. The important task of finding a good representation in machine learning goes therefore beyond the simple performance analysis (i.e., metrics), and by using a TCN block it is possible to find a lower embedding in a realistic time/complexity frame.

Table 4.3: Results when using an LSTM instead of a TCN in the temporal block (UK Biobank rs-fMRI dataset).

Model	**AUC**	**Accuracy**	**Sensitivity**	**Specificity**
N + E → Concat	0.93 (0.002)	0.85 (0.004)	0.85 (0.006)	0.85 (0.005)
N + E → DiffPool	0.84 (0.014)	0.76 (0.015)	0.74 (0.051)	0.77 (0.031)
N → Concat	0.93 (0.004)	0.85 (0.007)	0.85 (0.020)	0.86 (0.019)
N → DiffPool	0.84 (0.020)	0.76 (0.017)	0.78 (0.025)	0.74 (0.035)
→ DiffPool	0.82 (0.035)	0.73 (0.034)	0.73 (0.083)	0.74 (0.090)
→ Concat	0.91 (0.003)	0.83 (0.003)	0.82 (0.020)	0.83 (0.012)

Table 4.4: Results when no TCN block is used to train and evaluate on the UK Biobank dataset.

Model	**AUC**	**Accuracy**	**Sensitivity**	**Specificity**	**Params** (Before)
N → Concat	0.93 (0.004)	0.85 (0.004)	0.85 (0.014)	0.86 (0.013)	23,541,071 (291,337)
N + E → Concat	0.93 (0.002)	0.85 (0.003)	0.84 (0.017)	0.86 (0.015)	24,022,742 (291,898)

4.4.3 Visualisation of TCN Kernels

The weights of the TCN layers can be visually inspected. I visualised the first two layers of one of the trained N + E → Concat models. Figure 4.4 shows the weights learned from the first TCN layer (each row corresponding to one of the 8 output channels of that layer), while Figure 4.5 depicts the same for the second TCN layer (each row corresponding to one of the 8 output channels and the columns corresponding to the 8 kernels of size 7 coming from the previous 8 channels).

In both figures, and with little exceptions, it can be seen that the output channels in the first two TCN convolutional layers will be a non-trivial weighted multiplication of input channels, as illustrated by the patterns in the kernel weights. Given the qualitative variability observed in these weights (which are learned at training time), I argue that these kernels might be filtering and selecting different, non-redundant patterns present in the original time series; however, further work is needed to actually prove that these kernels are not redundant patterns. One possible counterexample is the kernel for the 7th output channel in the first TCN convolutional layer illustrated in Figure 4.4, which is essentially applying a simple low pass filter by smoothing the original time series from the input channel. It could be possible in potential future work that quantitative analysis and comparison of the kernel weights could yield interpretable information on which type of brain dynamics may contribute most to the final prediction. Given that these weights are also influenced by additional factors such as normalisation strategy and subsequent non-linear operations, further research is needed in order to establish a framework to fully exploit this information.

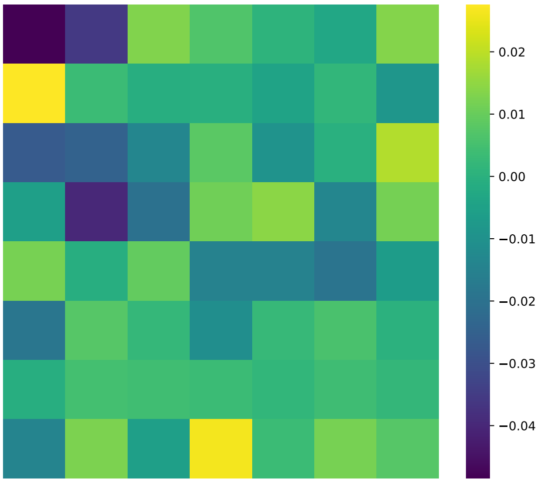


Figure 4.4: Weights of the kernels in the first TCN convolutional layer in a N + E → Concat model. Rows correspond to the 8 output channels of this layer, and each column is a position in the kernel array of size 7.

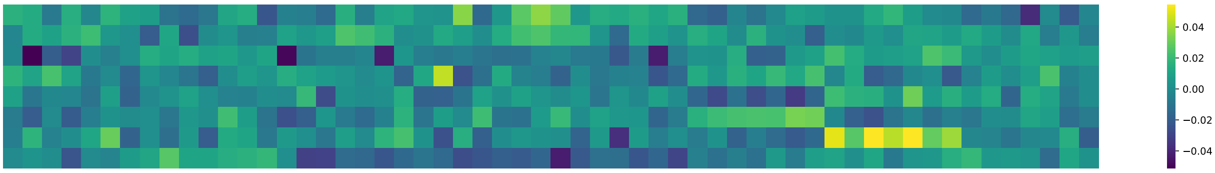


Figure 4.5: Weights of the kernels in the second TCN convolutional layer in a N + E → Concat model. Rows correspond to the 8 output channels of this layer, and each column is a position in the 8 kernels of size 7 that come from the 8 input channels (56 columns in total).

4.4.4 Explainability of DiffPool Clusters

Although deep neural networks are usually regarded as "black boxes", I strived to inject explainability elements by inspecting which mechanisms were learned during training. To this end, I designed a strategy to inspect the hierarchical spatial pooling mechanism provided by the DiffPool architecture. The assignment matrices from the first DiffPool layer $\boldsymbol{S}^{(1)}$ (see Equation 4.7) are analysed over all participants across all test sets. This is of particular interest because it corresponds to an aggregation of subsets of brain regions which the architecture has considered optimal while learning a particular prediction task. These aggregations can therefore be considered "optimal" for that task within this architecture, and provide insight into the neurophysiology which may drive the formation of such patterns. An assignment matrix corresponds to how the original nodes in the graph will be mapped into new nodes. In this respect, a simple and useful way of summarising this behaviour across individuals is to count how many times two ROIs have ended up in the same cluster, regardless of cluster size and number. More formally, an association matrix $\boldsymbol{S'} \in \mathbb{R}^{68\times68}$ is created, where each element $S'_{i,j}$ is the number of times brain regions i and j have been assigned together in the first DiffPool layer. This means that the higher the value of $S'_{i,j}$, the more often information from brain regions i and j is pooled when learning to predict binary sex. It is important to note that matrix thresholding (see `threshold` hyperparameter in Section 4.3.2) can - and often will - introduce disconnected nodes in the graph. Since the number of disconnected nodes would vary across individuals, this would introduce unrealistic imbalances/biases in the association matrix $\boldsymbol{S'}$; therefore, in this section, I only employed unthresholded matrices. In specific, I used the best performing DiffPool model (i.e., N → DiffPool) described in Section 4.3.2.

Figure 4.6 depicts the association matrix $\boldsymbol{S'}$ for the best performing DiffPool model (i.e., N → DiffPool) trained on unthresholded matrices, with dendrograms resulting from hierarchical clustering of this latter matrix (performed for visualisation purposes). The hierarchical clustering algorithm and the corresponding dendrograms are calculated using the *seaborn* [280] python package. In addition, a more traditional brain connectivity visualisation is generated by selecting the four main clusters defined by the dendrograms for the N → DiffPool model and overlaying their anatomical correspondence on a sample brain surface in Figure 4.7.

An advantage of this explainability strategy (i.e., the use of the association matrix $\boldsymbol{S'}$) is the flexibility inherent in the multiple granularities provided by hierarchical clustering. When choosing large clusters (e.g., four like in Figure 4.7) one can illustrate the general aggregation patterns across the brain's anatomy, while by selecting smaller clusters (e.g. twelve clusters) one can reveal more local patterns in the data. I consider that these different levels of granularity are an advantage of using DiffPool to help explain the model, in practice revealing different scales of explainability.

When looking at how the GNNs clustered the brain regions to optimise and achieve best sex prediction, it is possible to find that clustering into four sets of brain regions showed interesting properties in terms of neurobiological explainability. More specifically, the brain regions were grouped in a manner that mirrors to a certain degree the well-known cytoarchitectural and functional properties of the cerebral cortex. For example, in Figure 4.7, cluster 1 (dark blue) included the bilateral frontal cortex as well as occipito-parietal regions that have a well-known role in working-memory, executive functions, and visuo-spatial processing, amongst many other cognitive functions. The left temporal cortex grouped with the paracentral lobule, while the right temporal cortex clustered with the pre-cuneus (light green and light blue, respectively). Cluster 3 (dark green) included several midline cortical areas that form the classic limbic-emotional system.

I do not wish to overinterpret these results or make "reverse neuroscience" inferences in the sense of interpreting *post hoc* the behavioural meaning of a set of regions without having directly analysed their behavioural relevance. However, I speculatively note that the clusters emerged may have some neurobiological relevance in terms of explaining some of the behavioural differences described between males and females in terms of cognitive, motor and emotional skills [202, 216]. Future work, particularly directed at investigating the links between brain and behavioural measures, is warranted to confirm whether the clustering of regions that this model has generated to achieve optimal sex classification is relevant at phenotypical level.

To evaluate the robustness of the DiffPool clusters, I compared the association matrices $\boldsymbol{S'}$ for all the five folds for the best performing DiffPool model (i.e., "N → DiffPool") trained on the unthresholded matrices (see Figure 4.8). Despite some differences, it is possible to see a similar overall qualitative structure across the folds. To quantify this difference, the normalised difference between every pair of association matrices i and j is defined as:

$$\text{Normalised difference} = \frac{\boldsymbol{S'_{\text{i}}} - \boldsymbol{S'_{\text{j}}}}{\boldsymbol{S'_{\text{i}}} + \boldsymbol{S'_{\text{j}}}}. \tag{4.10}$$

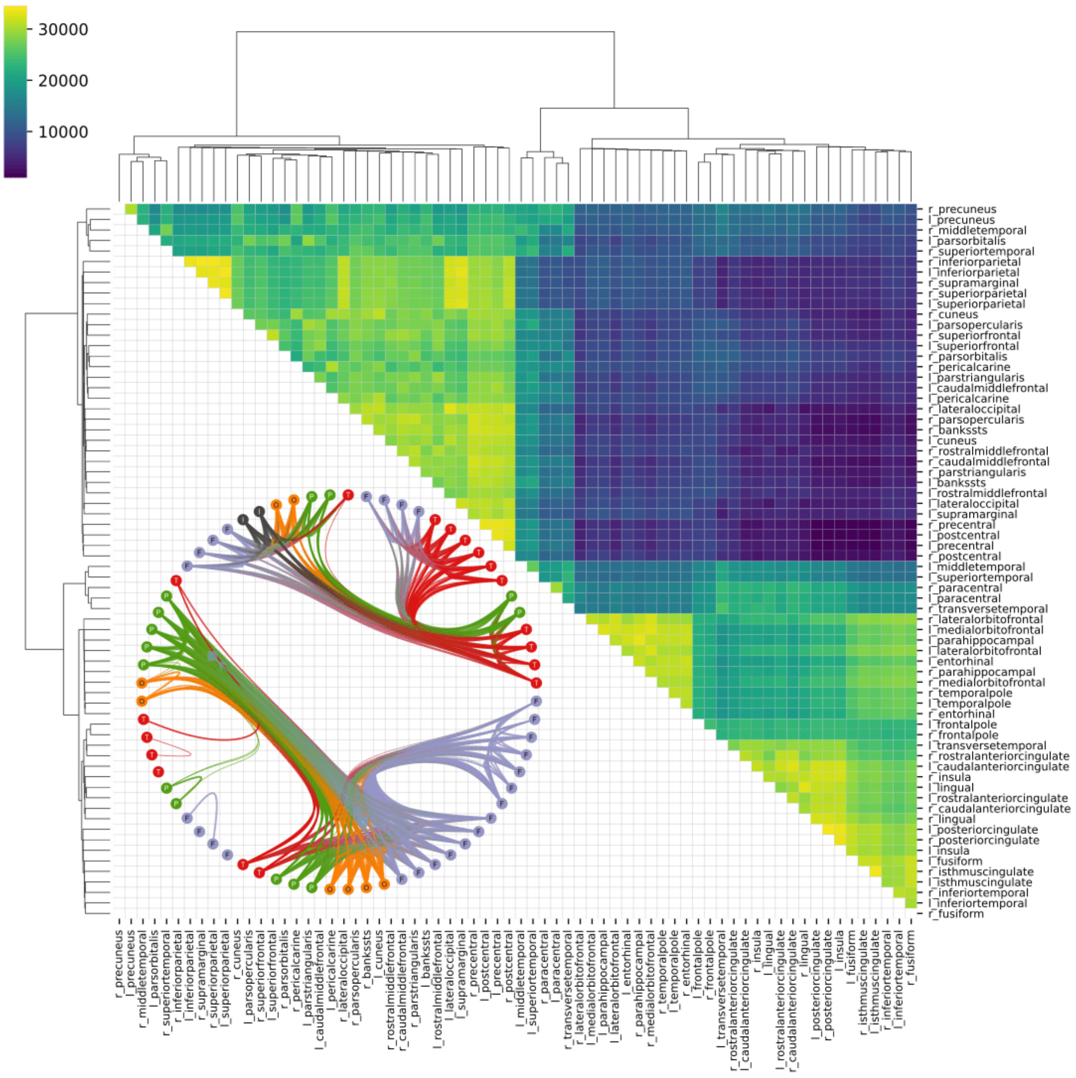


Figure 4.6: Upper-triangle of the association matrix $\mathbf{S}'$ for N → DiffPool model generated when predicting binary sex on unthresholded matrices, with dendrograms from hierarchical clustering. Each element $S'_{i,j}$ indicates how many times brain regions i and j are pooled together. On the lower left corner, a graph representation of the same association matrix $\mathbf{S}'$, thresholded at 25% with nodes identified and coloured according to their general brain region (i.e., T/F/O/P/I correspond to Temporal, Frontal, Occipital, Parietal, and Insula); thicker edges represent a higher $S'_{i,j}$ value, in this graph representation ranging from $23,911$ to $34,503$.

The various normalised differences can be seen in Figure 4.9, with every pair showing an average normalised different below 30%, therefore demonstrating an acceptable stability and robustness of the clusters learned by DiffPool across folds.

4.4.5 Evaluation on an External Multimodal Dataset

Table 4.5 shows the results when training and evaluating the architecture on the Human Connectome Project (HCP) dataset, both for the multimodal (rs-fMRI and diffusion data) and unimodal (only rs-fMRI data) cases. Performance metrics of the developed model

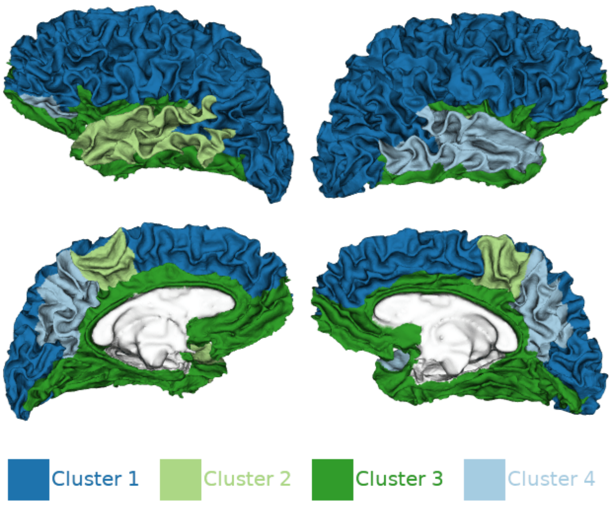


Figure 4.7: Four main brain clusters on association matrix $\boldsymbol{S'}$ generated from N → DiffPool model predicting binary sex on unthresholded matrices. Each colour corresponds to one cluster.

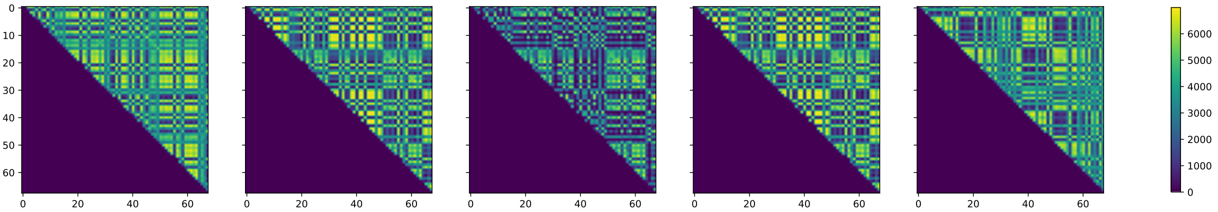


Figure 4.8: Association matrices $\boldsymbol{S'}$ for all the five folds for the model "N → DiffPool" trained on the unthresholded matrices.

are lower, as compared to the UK Biobank analyses, when considering only rs-fMRI data (i.e., 3-5% difference for concatenation and around 20% difference when using DiffPool). This may illustrate known concerns about the behaviours of deep learning models in general, and graph learning models in particular, when data is scarce; indeed, in the case of rs-fMRI data only, the non-DL baselines reach similar, or slightly better performances when compared to all DL models (both my model and the DL baselines), confirming that DL models can struggle with smaller datasets. However, in the multimodal case, when complementary information from both rs-fMRI (i.e. functional data) and diffusion-weighted MRI (i.e. structural data) are used, my model performs notably better than all baselines. This highlights how the model can flexibly leverage multiple data sources, achieving performances that in some cases are higher than the unimodal results obtained with the much larger UK Biobank dataset. This also emphasises the anticipated outcome that even DL models perform better in the presence of richer and varied data rather than when merely increasing dataset size, provided the model is able to leverage data richness. This does not happen with the non-DL baselines, which perform almost equally when comparing unimodal and multimodal data.

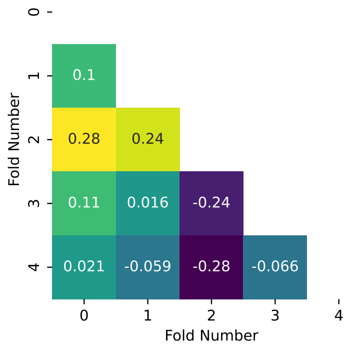


Figure 4.9: Averaged normalised differences between association matrices across the five folds of the model "N → DiffPool" depicted in Figure 4.8.

4.5 Extensions of Graph Neural Networks to Multimodal Brain Graphs

At the end of this chapter I would like to revisit how I specifically tackled *Research Question 3 (Graph)* (see Section 1.2) in this chapter. I employed GNNs and TCNs to integrate the spatial and temporal components of fMRI data in a single architecture. While initially focused on a single data modality (using the UK Biobank), I later showed the advantages of the model using multiple modalities of data and therefore showing different ways to tackle the research question. However, there are other ways to integrate multiple modalities of data to tackle *Research Question 3 (Graph)*; in this sense, I would like to highlight two Part II works which I have supervised at the Department of Computer Science and Technology at the University of Cambridge (see Section 1.4).

The first work, by Alexandru-Catalin Filip [97], used brain graph representations in a similar fashion as the ones I have used in this chapter; however, he specifically extended a very successful GNN architecture called Graph Attention Network (GAT) [270]. He handled a set of graphs provided with node features and non-binary edge weights, and we demonstrated his architecture's effectiveness by training it on seemingly integrated multimodal data. This adaptation provided a powerful and flexible deep learning tool to integrate multimodal neuroimaging connectomics data in a predictive context. We matched state-of-the-art results in a binary sex classification task while confirming the previously reported difficulties in predicting personality scores using brain data.

The work by Kamilė Stankevičiūtė [254] goes one step further and employs even more data, by including both non-imaging and brain imaging data directly in the same architecture. She was able to do so by using another data representation in the form of a population graph, as described in Figure 4.10. As it is possible to see, she combined several imaging and non-imaging modalities into a population graph to predict the apparent brain age for a large and diverse dataset of subjects. The population graph representation allows to control for confounding effects through pairwise similarities (i.e. the population graph edges) rather than fitting of separate models, and train the entire dataset at once without extensive filtering of the data (i.e. conditions that are closer to real clinical settings). Consistent and unified processing of the different data modalities is important, regardless of the downstream task or analysis method, as there is a widespread community effort to

Table 4.5: Results when training and evaluating on the HCP dataset, both for the multimodal (rs-fMRI and diffusion data) and unimodal (only rs-fMRI data) cases. Metrics averaged across the five test sets, with standard deviation in parenthesis. Aggregator on the right-hand side of the arrow, "N" corresponds to only *node model*, and "N + E" corresponds to full Graph Network block.

Model	**AUC**	**Accuracy**	**Sensitivity**	**Specificity**
		no GNN		
→ Concat	0.89 (0.034)	0.81 (0.038)	0.79 (0.050)	0.83 (0.031)
		Using both rs-fMRI and diffusion data		
N + E → Concat	0.94 (0.010)	0.85 (0.016)	0.83 (0.047)	0.87 (0.058)
N + E → DiffPool	0.89 (0.019)	0.81 (0.019)	0.78 (0.047)	0.84 (0.053)
N → Concat	**0.95** (0.012)	**0.88** (0.018)	**0.86** (0.045)	**0.90** (0.026)
N → DiffPool	0.93 (0.018)	0.85 (0.024)	0.79 (0.044)	**0.90** (0.035)
CNSLAB [109]	0.81 (0.029)	0.74 (0.022)	0.69 (0.051)	0.79 (0.033)
cGCN [278]	0.62 (0.019)	0.57 (0.027)	0.51 (0.205)	0.61 (0.220)
XGBoost	0.88 (0.018)	0.81 (0.021)	0.77 (0.036)	0.84 (0.017)
SVM	0.82 (0.020)	0.82 (0.022)	0.79 (0.044)	0.85 (0.058)
		Using only rs-fMRI data		
N + E → Concat	0.88 (0.025)	0.81 (0.030)	**0.80** (0.056)	0.82 (0.037)
N + E → DiffPool	0.63 (0.027)	0.59 (0.012)	0.47 (0.080)	0.70 (0.059)
N → Concat	**0.89** (0.019)	0.82 (0.019)	**0.80** (0.046)	0.83 (0.054)
N → DiffPool	0.68 (0.018)	0.64 (0.014)	0.59 (0.041)	0.68 (0.057)
CNSLAB [109]	0.82 (0.031)	0.75 (0.023)	0.70 (0.053)	0.79 (0.040)
cGCN [278]	0.65 (0.039)	0.59 (0.024)	0.41 (0.175)	0.75 (0.167)
XGBoost	**0.89** (0.014)	0.82 (0.019)	0.78 (0.025)	0.85 (0.030)
SVM	0.83 (0.022)	**0.83** (0.024)	0.78 (0.044)	**0.87** (0.064)

combat the reproducibility crisis in both neuroimaging [126] and machine learning[4]. This crisis is caused by, among other factors, the lack of transparency in preprocessing methods and software errors due to bad software engineering practices [217]. While the efforts to improve reproducibility in neuroimaging are currently targeted at consistent processing of functional and structural MRI with open-source libraries, this work is, to the best of our knowledge, one of the first to additionally incorporate non-imaging data modalities. Although this pipeline was designed to prepare the data specifically for population graphs,

[4] `https://reproducibility-challenge.github.io/neurips2019/`

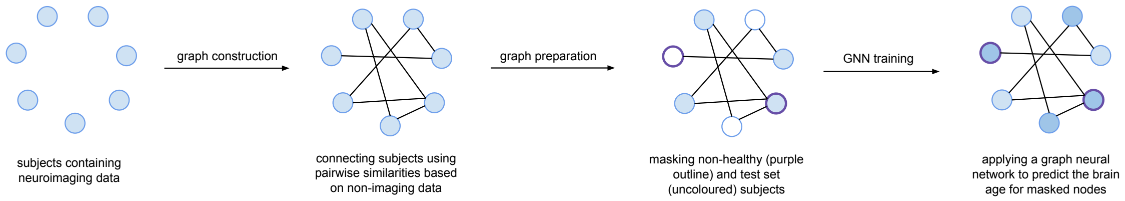

subjects containing neuroimaging data
graph construction
connecting subjects using pairwise similarities based on non-imaging data
graph preparation
masking non-healthy (purple outline) and test set (uncoloured) subjects
GNN training
applying a graph neural network to predict the brain age for masked nodes

Figure 4.10: Overview of the population graph preparation and graph neural network training procedure for brain age prediction. Figure taken from Stankevičiūtė et al. [254].

it has sufficient flexibility to be extended to more preprocessing options and adapted to work independently from the downstream analysis method.

These two works highlight the practical interest of *Research Question 3 (Graph)* while showing a different perspective on how to tackle it. They also highlight how, despite all the discoveries and improvements in research, there are still many avenues to explore in leveraging graph-like data with interests in scientific and applied developments.

4.6 Summary

This chapter provided a novel deep learning architecture which can successfully use the high-dimensional and noisy rs-fMRI data, by leveraging their temporal dynamics and spatial associations represented by what is commonly called the connectivity between brain locations. In contrast with previous literature, I use TCNs to model intra-subject temporal dynamics and combine them with GNNs to model inter-regional associations. I illustrated and analysed the effectiveness of the model in a proof-of-concept binary sex prediction task, which also included an ablation analysis with variations of the spatial pooling mechanism. The ablation study showed how the graph network block successfully leveraged the weights of the spatial dynamics, indicating the importance of designing an architecture specifically targeted for spatio-temporal rs-fMRI data. These results point to an advantage of using subject-specific FCMs because the baseline obtained using group-averaged FCMs (i.e. cGCN) consistently performed worse against all other models, including non-DL baselines. Contrary to my initial hypothesis, using a hierarchical pooling mechanism (i.e., DiffPool) did not provide an improvement in overall performance when compared to concatenation pooling and, in some cases, to baselines. The most notable exception is the multimodal setting with the HCP dataset, in which the hierarchical pooling mechanism occasionally provides similar results to my best model. Still, I posit that the compelling explainability potential of DiffPool is advantageous in settings such as neuroscience investigation. In this context, additional explorations of hierarchical pooling mechanisms could represent an exciting future research direction.

One of the aims of this work was to provide additional contributions beyond the goal of end-to-end modelling of functional brain activity, hoping to provide a tool that can be tailored to the analysis of medical images. For instance, the set of experiments using unique multimodal data from the HCP dataset illustrate how this approach can be of interest in the emerging multimodal brain connectivity community. Also, I am not aware of any other work in the neuroimaging field which uses a hierarchical pooling mechanism for the purpose of generating explainable patterns from fMRI data - a crucial aspect when interacting with the neuroscience community. While temporal convolutional networks (TCNs) and graph neural networks (GNNs) have been successfully introduced in previous literature, the contribution of this chapter also lies in the combined use of these building blocks for the specific case of modelling rs-fMRI data. Importantly, I have also motivated the choice of TCN kernels with respect to LSTM models through a head-to-head comparison in Section 4.4.2.

I hope this chapter can lay the groundwork for future exploration of flexible architectures which are able to leverage the entirety of neuromonitoring data arising from the extremely complex spatio-temporal interplay of groups of firing neurons. By demonstrating improved performance in a task which is commonly employed in benchmarking models for functional brain data, comparing to both non-DL and DL baselines, and sharing all code and implementation details, I hope that this work will have an impact on future research which will further improve spatio-temporal modelling specific to fMRI data. As demonstrated with the multimodal Human Connectome Project (HCP) dataset, the architecture can very easily include other types of data (e.g., multimodal structural data) and be further extended to include possible confounds (e.g. age, IQ, cognitive status) that could drive the prediction task in other brain disorders. Furthermore, additional analyses can be conducted to study the robustness of the architecture to finer parcellations beside the Desikan-Killiany atlas, possibly leading to additional neurobiological insights depending on which regions are represented in the parcellations. Another exciting recent trend that can be included in this architecture is to allow the network to learn the underlying connectivity from scratch [152, 160] instead of computing associations or other handcrafted features such as the ones used in this and other works [172, 173].

Chapter 5

Interpreting Differences in Cognition Using Brain Features

Predicting variability in cognitive functioning can have important consequences in delineating a person's life trajectory. Individual differences in cognition have been related to important outcome measures like education, occupational achievement, general health, longevity, or risk to develop dementia. One way of predicting cognitive performances at the single-subject level is to use different neuroimaging data sources that assess distinct aspects of brain function and structure (i.e. anatomy). Resting-state functional connectivity [77, 203] or neuroanatomically extracted features [226] are examples of such sources of neuroimaging data. However, almost no work in this field has tried to capitalise on multimodal neuroimaging, i.e. on the possibility to predict individual differences in cognition by simultaneously using different types of information regarding the brain structure or function.

This chapter combines different structural neuroimaging modalities: T1-weighted and T1/T2- derived intracortical myelin estimates. They were used to predict subject-specific scores in a series of cognitive and demographic measures derived from a data-reduction analysis of a large set of behavioural measures, drawn from the Human Connectome Project (HCP).

As shown in Chapter 2, machine learning has recently gained attention in several applied biomedical disciplines due to the increased prediction power it can provide. Specifically, deep learning methods are the most well-known ones, but they usually bring some complexity when one wants to interpret the results or avoid overfitting [67, 243, 271]. Instead of using neural networks, this chapter focuses on a type of gradient boosting decision tree algorithm, XGBoost [57], which has recently achieved top results in applied machine learning competitions using tabular data (see Section 2.1.5.2). As it is built on top of decision trees, it also has the advantage that there is space to understand how the model makes its decisions.

To interpret the neuroanatomical basis of cognitive measures, I used recently developed algorithms that improve machine learning models' interpretability capacity without neglecting their prediction power. Specifically, I used an adaptation of SHAP (SHapley Additive exPlanations) [179] which is a unified framework for interpreting predictions without falling in common mistakes that do not bring consistent and trustworthy interpretations [178].

Interpretation of 3D neuroimages can be complex and usually relies on visual plots and specific neuroanatomical knowledge. When interpreting machine learning models that use 3D images, it is possible to visually identify areas of the brain responsible for the prediction;

however, it is not evident whether those regions are related, for instance, to actual values of thickness or surface area. Therefore, by reducing multimodal neuroimaging measures to a well-characterised set of features, the interpretation of such models is direct as each feature is self-explanatory *per se*. In this chapter, the different structural neuroimaging modalities were used as subject-wise features in XGBoost.

This chapter's main contribution is thus to show how one can interpret the neuroanatomical basis of cognition by applying state-of-the-art machine learning methods that might not yet be fully and correctly explored in the field. Regarding data representation, this is the typical flatten, direct one, which I try to capitalise for interpretability purposes.

5.1 Methods

5.1.1 Dataset - Human Connectome Project

Preprocessed structural magnetic resonance images as well as demographic and cognitive data from 1,200 subjects were obtained from the HCP public repository[1].

After accounting for missing information, the total number of subjects included in this analysis was 905, where around 53% were females, and 47% were males. All participants were young and healthy adults, with a median age of 29, with no hypertension, alcohol misuse, panic disorder, depression, or other psychiatric and neurologic disorder, or history of childhood conduct problems. The majority of people were right-handed white Americans with a non-Hispanic or Latino background.

HCP structural T1w images were collected from a 3-Tesla Siemens Skyra unit (housed at Washington University in St. Louis) using an axial T1-weighted sequence (TR = 2400 ms, TE = 2.14 ms, flip angle=8°, voxel-size 0.7 × 0.7 × 0.7 mm3). The T1w data were passed through the full *Freesurfer* (v. 5.3)[2] reconstruction stream to calculate cortical thickness, surface area, grey volume, integrated rectified mean curvature, integrated rectified Gaussian curvature, folding index, and intrinsic curvature index. The optimal pipeline used to obtain these variables was the one developed by Glasser et al. [122], as briefly mentioned in Section 2.3.3. To map all subjects' brains to a common space, namely the Desikan-Killiany atlas [70], reconstructed surfaces were registered to an average cortical surface atlas released by HCP using a non-linear procedure that optimally aligned sulcal and gyral features across subjects [100].

The HCP consortium generated myelin[3] maps according to Glasser and Essen [121]. Transformation of myelin map data from the individual subject's native mesh to the right *fsaverage* template (LR) standard mesh involves two deformation maps, one representing registration from the native mesh to the *fsaverage* left mesh (L) and *fsaverage* right mesh (R) and another representing registration between L and R and LR. The two deformation maps were concatenated into a single deformation map using Caret software applied to the individual subject's myelin map data, cortical thickness data, and surface curvature data. The individual myelin map data were normalised to a group global mean and then averaged at each surface node to achieve anatomical correspondence to other structural features.

[1] https://www.humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release

[2] http://surfer.nmr.mgh.harvard.edu/

[3] Myelin is an insulating substance that is present around the nerves. It allows electric impulses to transmit more efficiently along nerve cells, and therefore its damage could potentially cause diseases.

5.1.2 SHapley Additive exPlanations (SHAP)

SHAP algorithm is based on Shapley values, a concept from cooperative game theory. Shapley values have been introduced in 1953 by Lloyd Shapley [242], and used in the past to compute explanations on model predictions [176, 208, 258]. In cooperative game theory, a coalition game consists of a set of N players and a *value function* v that maps subset of players $S \subseteq \{1, 2, \ldots, N\}$ to a real value. This real value represents the collective pay-off of a subset of players gained by "cooperating" as a set. Therefore, the outcome of each possible combination (i.e. coalition) of players should be considered to determine the importance of a single player.

We can find the marginal contribution of player i with respect to a coalition S, by calculating the additional value generated by including player i in the coalition:

$$\Delta_v(i, S) = v(S \cup \{i\}) - v(S). \tag{5.1}$$

A Shapley value is then the weighted average of all the player's marginal contributions, which can be written as (for a player i):

$$\phi_i = \sum_{S \subseteq \{1,2,\ldots,N\} \setminus \{i\}} \frac{|S|!(N - |S| - 1)!}{N!} \Delta_v(i, S). \tag{5.2}$$

Intuitively, over all possible ways to go from the empty set $\emptyset$ to the entire set of players, the Shapley value generates the average player i's contributions.

I will exemplify this mechanism with an adaptation from the glove game. Let A and B be two players that sell gloves for £1 each, with a restriction that gloves must be sold in pairs. Player A has 3 gloves, player B has 11 gloves, and the value of the game is the gloves revenue. If player A plays alone, it can sell 2 gloves alone and therefore its value is £2. Player A's marginal contribution is different depending on whether it is joining the empty coalition or joining the coalition of B. The following representation illustrates this process, in which each arrow represents the inclusion of a player not present in the previous coalition:

$$\begin{tikzcd}
v(\emptyset) = 0 \arrow[r, "+2"] \arrow[d, "+10"] & v(A) = 2 \arrow[d, "+12"] \\
v(B) = 10 \arrow[r, "+4"] & v(\{A, B\}) = 14
\end{tikzcd}$$

$$\phi_A = \frac{\Delta_v(A,\emptyset) + \Delta_v(A,\{B\})}{2} = 3$$

$$\phi_B = \frac{\Delta_v(B,\emptyset) + \Delta_v(B,\{A\})}{2} = 11$$

By looking on all possible ways to go from $\emptyset$ to $\{A, B\}$, the averaged marginal contribution for player A is £3, and for player B is £11, which represent the number of gloves that each player brought.

For a machine learning model, this means that the game is reproducing the outcome of a model $f(x_1, x_2, \ldots, x_N)$, and the players are the features x_i inputted to the model. We are then trying to quantity the contribution that each player brings to the game, that is, the contribution that each single feature brings to the model's outcome.

Lundberg and Lee [178] proposes SHAP values as a unified measure of feature importance by defining a class of additive feature attribution methods. Their definition of SHAP values are derived by making the value function dependent on a specific input instance $\boldsymbol{x}$, making this a "local" method. These SHAP values are therefore derived from defining the value function as the expected model output conditioned for a specific data point when only the features in S are known:

$$v_{\boldsymbol{x}}(S) = E\left[f(\boldsymbol{X}) | \boldsymbol{X}_S = \boldsymbol{x}_S\right]. \tag{5.3}$$

Tree-SHAP [179] is an efficient adaptation of SHAP values on additive tree-based models such as XGBoost used in this chapter. It brings some advantages when compared to other previously popular feature importance methods such as Gini coefficients. Two of such advantages are individualised explanations (i.e. per input) and solving issues that popular feature attribution methods are inconsistent and incapable of reporting the real impact of features in tree ensemble methods. Figure 5.1 illustrate these cases with two tree-based models.

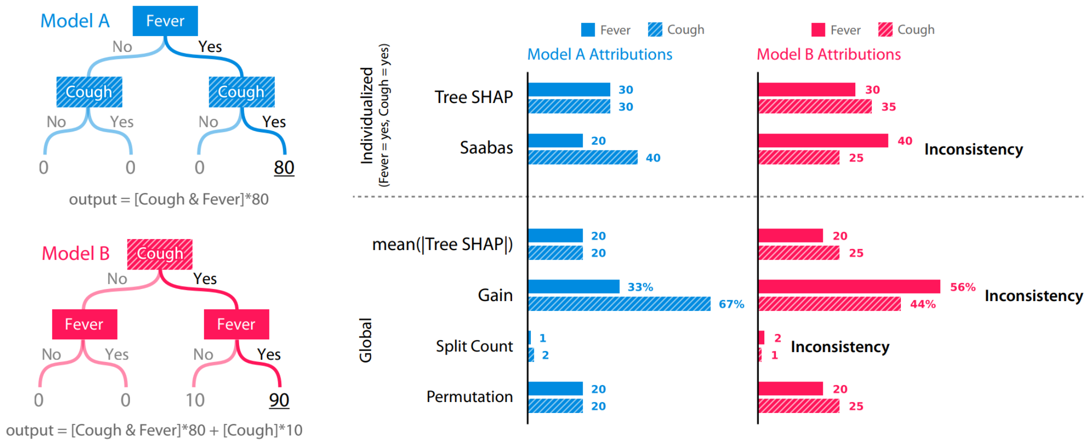


Figure 5.1: Inconsistencies in previously widely-used feature attribution methods (i.e. Saabas [232], Gain [236], Split Count [57], and Permutation-based methods [16]). The Cough feature has a larger impact in Model B than Model A, but is attributed less importance in Model B for most of other methods. Similarly, the Cough feature has a larger impact than Fever in Model B, yet is attributed less importance for most of other methods. Gini coefficients (a Gain method) were previously widely used by default to explain XGBoost models, but are also the methods showing greater inconsistencies for tree ensemble models. Figure taken from Lundberg et al. [179].

As it is possible to see, the feature importance values from the Gain, split count, and Saabas methods are all inconsistent for this example. As without consistency it is impossible to reliably compare feature attribution values, the guaranteed consistency of SHAP values help explain why it gained so much popularity in recent years. Indeed, these and other advantages have been highlighted in recent literature reviews [14, 38, 136, 272]. SHAP is presented by Linardatos et al. [175] as "the most complete method, providing explanations for any model and any type of data, doing so at both a global and local scope", and "[together with LIME], by far, the most comprehensive and dominant across the literature methods for visualising feature interactions and feature importance".

Despite all the recent successes with SHAP (a lot due to the solid theoretical foundation in game theory), there are still improvements being developed in a very dynamic research field. Some of its current drawbacks driving research developments include: (1) computation time, as the possible feature combinations exponentially increases with the number of features, though this is not a critical issue in this thesis as I use the faster implementation for tree-based models, (2) not being able to simulate scenarios of how changes in particular features will impact the output, given that, unlike LIME, SHAP does not return a model, and (3) order of feature selection might impact the SHAP values as in real-world datasets there are non-linear correlations among different features, even if they are independent.

5.1.3 Factor Analysis

Factor analysis (FA) is a broad term in multivariate statistics which has the objective of finding latent variables (i.e. not directly measured variables) in a dataset. It starts from the idea that there are a certain number of factors in a dataset and that each measured sample captures a part of one or more of those factors[4]. The goal of FA is to model the interrelationships among elements, and is accomplished by following two steps: (1) factor extraction, and (2) factor rotation.

One possible factor extraction method is Principal Component Analysis (PCA), a well-known unsupervised and efficient way to reduce the number of dimensions in a dataset into certain number of "components". These components are created in such a way that the first component contains the maximum variation of the original dataset, the second component contains the second-largest amount of variation, and so forth. Thanks to this, we are able to retain only a few components for downstream analysis while retaining a good amount of the original variation. Specifically, PCA tries to find a linear combination with maximum variance of the original variables. For a specific dataset $\boldsymbol{X} \in \mathbb{R}^{n \times k}$ with n samples/observations, and k columns/features, it solves a new linear combination $\boldsymbol{X}c_i$, where c_i is a column vector. To find the first principal component, it calculates:

$$c_1 = \underset{c_1}{\operatorname{argmax}}[Var(\boldsymbol{X}c_1)], \quad \text{s.t.} \quad c_1^T c_1 = 1, \tag{5.4}$$

in which we add the constrain $c_1^T c_1 = 1$ to keep every element in c_1 different from infinity. For the remaining components, there is an orthogonality constrain (i.e. $c_i^T c_j = 0$) which is added for all the corresponding preceding components. For the second component, the calculations would then be:

$$c_2 = \underset{c_2}{\operatorname{argmax}}[Var(\boldsymbol{X}c_2)], \quad \text{s.t.} \quad c_2^T c_2 = 1 \quad \text{and} \quad c_1^T c_2 = 0. \tag{5.5}$$

Instead of iteratively solving this maximisation problem, it is known that the eigenvectors of the covariance matrix of the data are the principal components [107]. Therefore, to get the ordered principal components, one just needs to order the eigenvectors of the sample covariance matrix in descendent way by their corresponding eigenvalues. The eigenvalues represent the total amount of variance that can be explained by the corresponding principal component[5]. An important concept which will be needed for FA is the notion of a component *loading*, which is equal to the eigenvector times the square root of the eigenvalue, and can be interpreted as the correlation of each element to the principal component.

[4] There are, in fact, two types of factor analysis: exploratory and confirmatory. As I never consider a fixed number of *a priori* factors, I am actually using exploratory factor analysis. However, to avoid verbosity, I refer to it as simply factor analysis (FA).

[5] In theory, eigenvalues can be negative, but as in the case of PCA they explain variance, they are always positive in practice.

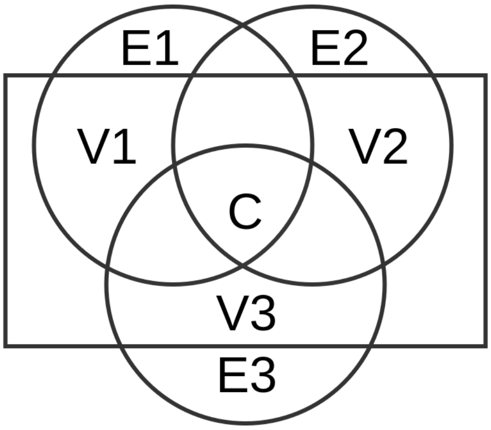


Figure 5.2: Example of variance over three elements, in which each circle represents one element's variance, and the box corresponds to a possible factor. Each element has its specific variances Vi; C is the common variance of all the elements, and Ei are the error variances (e.g. due to measurement errors).

When conducting FA, PCA's formulation is changed to what is called "common factor analysis". In practice, this means that instead of maximising the total variance, it assumes the factors are linear combinations maximising the shared portion of variance, therefore underlying *latent constructs*. For intuition, Figure 5.2 illustrates the different types of variances that play a role in these calculations.

With common factor analysis, calculations are now iterative and dependent on specific optimisation routines. For a more detailed explanation of the mathematical concepts and implications of this fundamental differences in variance, I point the reader to Brown [44].

After this first step (factor extraction), the next one is factor rotation, in which one wants to give the most useful interpretations in the newly defined factor loadings, rather than only maximising measures of variance. In other words, without rotation, the first factor will be the one explaining the largest amount of variance and therefore onto which most elements will load, hindering interpretability by an expert.

There are two general types of rotations: (1) orthogonal, which assumes that factors are independent among each other, and (2) oblique, in which factors are not independent and are therefore correlated. A popular orthogonal rotation scheme that I use in this thesis is the Varimax, introduced in 1958 by Kaiser [154] in which all factors remain uncorrelated with one another. Its name come from the fact that it tries to *maximise* the *variance* of the squared loadings in each factor. As a result, the orthogonal basis is rotated to disperse the loading scores across factors, thus simplifying its analysis by an expert.

A common application of FA exists in psychology, for instance when one tries to identify personality dimensions represented in distinct questionnaires/tests. Given both PCA and FA *reduce* the original dataset dimension into a smaller set of components/factors, it is common to wonder when to use one or another. Despite that sometimes it is possible to see that PCA and FA loadings can be similar, it is important to make the distinction that their fundamental aim is different (besides the mathematics explored in this section). With PCA, one tries to find variables that are composite of the observed variables. For instance, one could define socioeconomic status (z) as a linear combination of level of education (x_e) and wealth (x_w): $x = x_e + x_w$. With FA, instead, we assume the existence of latent factors underlying the data that are not directly measured. For example, given a set of questionnaires, one could assume that there is a latent construct called “prejudice” (Ψ) *influencing* how people answer the questions. Possible questions could be ‘I feel negatively towards people of other colour” (x_1) or “I cannot perceive me having a friend who is Black” (x_2). Assuming residuals ϵ_i, we would have $x_1 = c_1\Psi + \epsilon_1$ and $x_2 = c_2\Psi + \epsilon_2$.

5.1.4 Adequacy of Factor Analysis

There are usually two tests to measure data adequacy to proceed with factor analysis: Barlett’s test of sphericity [28] and the Kaiser-Meyer-Olkin (KMO) test [155].

In the Barlett’s test of sphericity, the null hypothesis is that a correlation matrix is equal to an identity matrix. Intuitively, if the correlation matrix produced from a dataset is the same as the identity matrix, there are no correlations/redundancies among variables and thus no factors to find. The test statistic consists in calculating:

$$\chi^2 = -\ln(\det(R)) * \left(N - 1 - \frac{2p+5}{6}\right), \tag{5.6}$$

where N is the sample size, p the number of variables (e.g. the 23 behavioural variables from Section 5.2.1), R the correlation matrix, ln() the natural logarithm, and det() the determinant of a matrix.

Kaiser-Meyer-Olkin (KMO) test is a test specifically targeted to measure data adequacy to factor analysis. It measures the proportion of variance among variables which could be common variance by using the following formula:

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} u_{ij}^2}, \tag{5.7}$$

where r_{ij} is the correlation between variables i and j, and u_{ij} is the value in the partial covariance matrix between variables i and j (i.e. partial correlation). The value below which the test indicates factor analysis is not adequate is either 0.5 or 0.6 depending on different authors.

5.2 Experimental Setup

5.2.1 Factor Analysis on Behavioural Data

Besides neuroimaging data, the HCP contains a rich resource of questionnaires and phenotypes related to the problem I want to explore in this chapter (i.e. cognitive functioning). Therefore, factor analysis is a strong candidate to simplify these variables into more concise and easier to interpret set of variables. Indeed, factor analysis “is (...) appropriate if the stated objective is to reproduce the intercorrelations of a set of indicators with a smaller number of latent dimensions, recognising the existence of measurement error in the observed measures” [44], which nicely connects with the other phenotypes available in the HCP that I want to explore here.

To explore cognitive functioning, an expert neuroscientist (see Section 1.4) identified twenty-three measures of affect, cognition, and health, including self-report questionnaires, neuropsychological tasks, and other behavioural indices from the HCP. Using factor analysis (FA), these twenty-three variables were grouped into nine independent components which explained 70% of cumulative variance. Upon inspection by the expert neuroscientist of the corresponding factor loadings, he was able to assign coherent interpretations to each factor, namely, fluid intelligence, visual memory, sex, executive functions, sustained attention, waiting impulsivity, linguistic skills, verbal episodic memory, and visuo-spatial skills. Subject-specific loading values were employed as dependent variables. It is important to acknowledge that a further analysis on the influence of the number of factors (or the threshold used for the cumulative variance) defined by the expert neuroscientist would be necessary to understand the robustness of this approach, but this is beyond the scope of this thesis.

When applying the Barlett's test of sphericity to the 23 behavioural variables, the resulting statistic test is approximately 9384.21, corresponding to a p-value < 0.00001, therefore rejecting the null hypothesis at alpha=0.05 significance level while at the same time providing evidence that the correlation matrix is not an identity matrix and we can proceed with factor analysis with these variables. When applying the Kaiser-Meyer-Olkin test to the 23 behavioural variables, the resulting value is approximately 0.69, thus indicating that it is adequate to run factor analysis on this data.

The loadings of each phenotypic variable are specified in Table 5.1. A detailed explanation of these phenotypic variables and how they were collected are available online[6]. As it is possible to see, a single test from the HCP cannot capture a whole *trait* (e.g. fluid intelligence), underscoring the usefulness of FA as not only to simplify analysis, but also to have more complex, informative traits represented in a single variable.

Practically speaking, one could use other phenotypes available in the HCP to run factor analysis; however, in this chapter I will focus the analysis only on the phenotypes selected by the expert neuroscientist to not go beyond the scope of this chapter's objectives, and therefore focus on the cognitive/emotional dimensions. If one wants to include other phenotypes for other hypothesis-driven questions, the methodology described in this chapter could be followed in the same way.

5.2.2 Training Procedure

XGBoost [57] is used as the supervised machine learning method to predict each factor (see Section 2.1.5.2).

I will use the mean squared error (MSE) and Pearson-r correlation to report the performance of the models averaged across the five outer folds for each factor. For a sample i (from a total of N samples), corresponding ground truth value $\hat{y}_i$, and predicted prediction y_i, MSE is the average squared difference between ground truth and predicted values, calculated as:

$$MSE = \frac{1}{N} \sum_{i}^{N} \left(y_i - \hat{y}_i \right). \tag{5.8}$$

The Pearson-r correlation is a measure of linear correlation between two sets of data and is calculated as follows for samples x with ground-truth values $\hat{y}$:

$$r = \frac{\sum_i^N \left(x_i - \overline{x}\right)\left(y_i - \overline{\hat{y}}\right)}{\sqrt{\sum_i^N \left(x_i - \overline{x}\right)^2 \sum_i^N \left(y_i - \overline{\hat{y}}\right)^2}}. \tag{5.9}$$

[6] https://wiki.humanconnectome.org/display/PublicData/HCP-YA+Data+Dictionary-+Updated+for+the+1200+Subject+Release

Table 5.1: Loading values for 23 behavioural variables on the estimated factors from the factor analysis. The colour scale represents the item loading scores on the independent factors: blue for the most negative, and red for the most positive ones.

	Fluid Intelligence	Visual Memory	Sex-related Factor	Executive Functions	Sustained Attention	Waiting Impulsivity	Linguistic Skills	Verbal Episodic Memory	Visuo-spatial Skills
Total Intracranial Volume	0.121	0.138	−0.772	0.107	0.042	0.071	−0.033	−0.005	0.093
Gender	−0.048	0.047	0.841	−0.041	−0.058	−0.010	0.005	−0.082	−0.001
Age	−0.084	0.182	0.455	0.077	0.137	0.067	−0.019	0.471	0.058
Memory - Picture Sequence Memory Test Score	0.124	0.560	0.220	0.130	0.036	−0.034	−0.035	−0.269	0.215
Executive Function/Cognitive Flexibility - Card Sort Test Score	0.102	0.187	0.001	0.790	0.058	−0.025	−0.029	−0.008	−0.006
Executive Function/Inhibition - Flanker Inhibitory Control and Attention Test Score	−0.016	0.041	−0.156	0.800	0.005	0.025	−0.109	−0.011	0.011
Fluid Intelligence - Penn Progressive Matrices: Number of Correct Responses	0.898	0.272	−0.117	0.110	0.030	0.083	−0.105	−0.127	−0.010
Penn Progressive Matrices: Total Skipped Items	−0.901	−0.260	0.111	−0.106	−0.021	−0.065	0.088	0.109	0.007
Penn Progressive Matrices: Median Reaction Time for Correct Responses	0.875	0.071	−0.050	−0.096	0.028	0.118	−0.017	0.103	−0.003
Oral Reading Recognition Test Score	−0.070	−0.083	−0.028	−0.013	0.018	0.002	0.865	0.016	0.035
Picture Vocabulary Test Score	−0.076	−0.003	0.061	−0.018	−0.023	−0.027	0.871	0.016	−0.010
Processing Speed Test Score	−0.005	0.101	0.021	0.732	0.060	0.018	0.103	−0.182	−0.039
Delay Discounting: Area Under the Curve for Discounting of $200	0.101	0.047	−0.035	−0.005	0.015	0.900	−0.038	−0.059	0.036
Delay Discounting: Area Under the Curve for Discounting of $40,000	0.111	0.102	−0.037	0.021	0.006	0.898	0.011	−0.023	−0.010
Spatial Orientation - Variable Short Penn Line Orientation: Total Number Correct	0.215	0.665	−0.413	0.094	0.062	0.102	−0.042	−0.013	−0.182
Variable Short Penn Line Orientation: Median Reaction Time Divided by Expected Number of Clicks for Correct	−0.017	0.013	−0.050	−0.034	−0.013	0.029	0.027	0.042	0.940
Variable Short Penn Line Orientation: Total Positions Off for All Trials	−0.242	−0.668	0.420	−0.092	−0.061	−0.125	0.029	0.008	0.189
Sustained Attention - Short Penn Continuous Performance Test: Sensitivity	0.065	0.152	−0.021	0.095	0.905	0.027	0.005	−0.027	−0.017
Short Penn Continuous Performance Test: Specificity	0.113	0.566	0.248	0.066	−0.064	0.048	0.067	0.155	−0.039
Short Penn Continuous Performance Test: Longest Run of Non-Responses	0.000	0.101	0.053	−0.025	−0.923	0.006	0.009	0.063	−0.003
Verbal Episodic Memory - Penn Word Memory Test: Total Number of Correct Responses	0.085	0.252	0.148	−0.007	0.158	0.073	−0.029	−0.643	−0.019
Penn Word Memory Test: Median Reaction Time for Correct Responses	0.033	0.006	0.036	−0.228	−0.007	−0.049	0.015	0.761	0.008
List Sorting Working Memory Test Score	0.115	0.586	−0.066	0.124	0.028	0.043	−0.104	−0.132	0.112

These two metrics are chosen because all the factors are real numbers and from the central limit theorem we can reasonably assume that the mean values approximately follow a normal distribution given the large sample size (over 190 per outer fold). A model that fits the data well will have a lower MSE value and a higher Pearson-r correlation. There are other metrics that could have been used such as mean absolute error and root mean squared error, but given the normality assumption, and for the purposes of this chapter, these two metrics are good enough summary metrics to convey the analysis required on model performance. Furthermore, instead of Pearson-r correlation one could have used the Kendall and Spearman correlation coefficients, but they are non-parametric and do not assume normality.

To predict each factor using the features generated from HCP, I employed a nested cross-validation procedure as depicted in Figure 5.3 to avoid overly-optimistic scores, with five outer folds, and three inner folds. Essentially, the data are divided into five folds, and selecting each fold once, the other four folds are selected to run an inner loop. For each inner loop, the data are divided in three folds, where two of these folds are used for hyperparameter search, selecting the hyperparameters that yield the best MSE in the remaining fold. After running this process for each of the three folds, the model that yields the lowest mean MSE is selected and used in the fold selected in that outer loop. In other words, this process selects five different models for each outer fold.

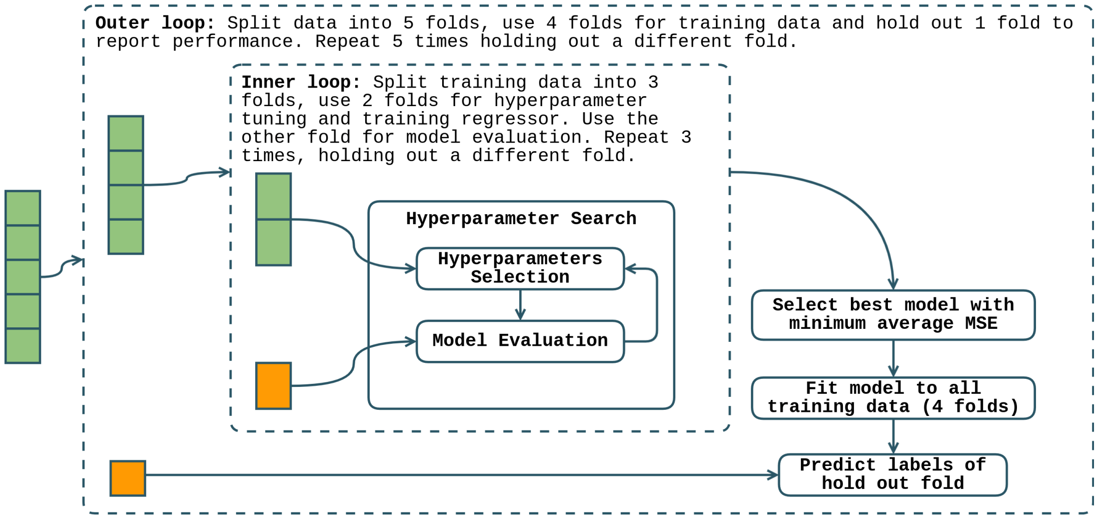

Outer loop: Split data into 5 folds, use 4 folds for training data and hold out 1 fold to report performance. Repeat 5 times holding out a different fold.
Inner loop: Split training data into 3 folds, use 2 folds for hyperparameter tuning and training regressor. Use the other fold for model evaluation. Repeat 3 times, holding out a different fold.
Hyperparameter Search
Hyperparameters Selection
Model Evaluation
Select best model with minimum average MSE
Fit model to all training data (4 folds)
Predict labels of hold out fold

Figure 5.3: Illustration of the nested cross-validation procedure used in this chapter, where the coloured boxes represent the data. The outer loop is responsible to report final performance, while the inner loop is responsible for hyperparameter search and model selection. Inside each loop, the green boxes are the ones in which the model is fit, and the orange box is where evaluation takes place.

To better evaluate the significance of these two metrics, I calculated a permutation-based p-value [206] for each metric at each factor. For this permutation test, the null hypothesis is that the regressor fails to leverage any dependency between the features and corresponding labels. Specifically, for each kth permutation ($k = 100$ in this chapter), the labels are stochastically permuted and the entire training/evaluation procedure is run again. The empirical p-value is then the fraction of randomised permutations in which the final averaged metric was better than in the original data. Intuitively, this process is measuring how likely the observed metric can be obtained by chance, therefore this test can be seen as a "control" to show with more confidence that the method is working.

To examine each factor in separate, I will analyse the most important features from each model interpretability. To that end, I will be looking to the rank of those features across the different folds. Given these are (ranked) ordinal values with a possibility of existing some outliers, reporting the median is then more adequate than reporting the mean across the folds. As a consequence, the median absolute deviation (MAD) is also reported instead of standard deviation.

5.3 Results

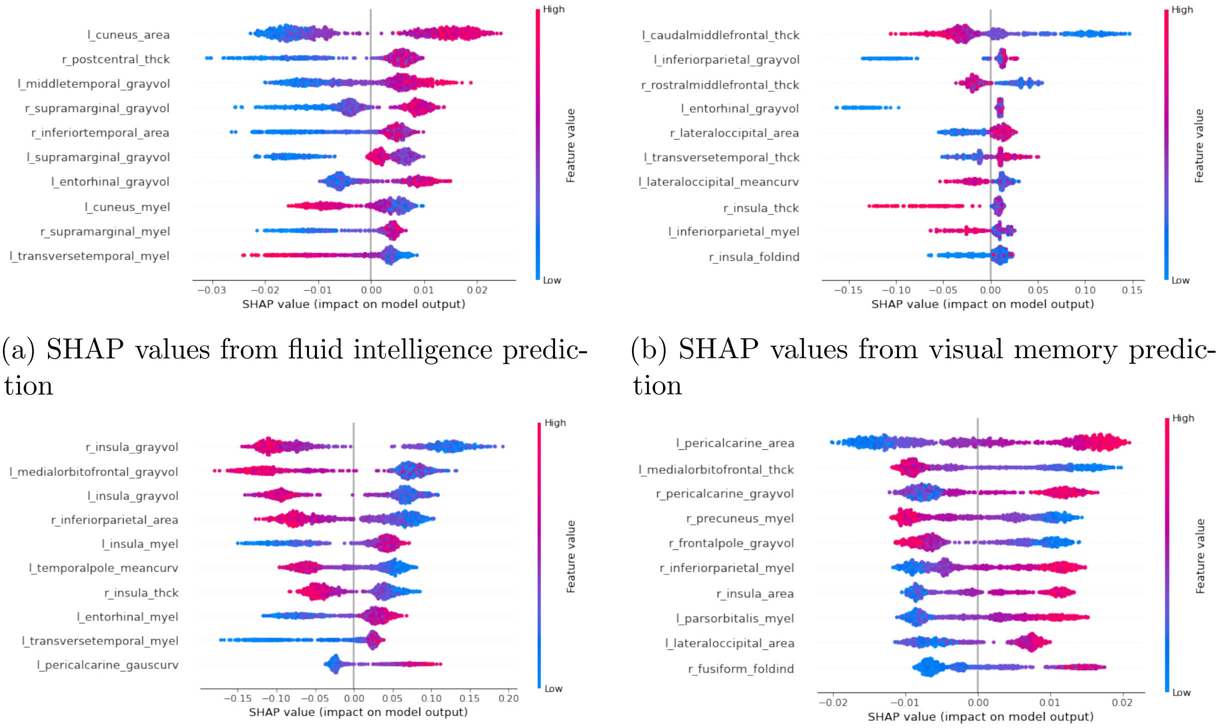
Table 5.2 summarises the results of the nested cross-validation procedure over the nine factors by reporting the averaged mean squared errors (MSE) and Pearson-r correlations between actual and predicted values, with the respective standard deviations. This section will only analyse the results on factors 1, 2, 3, and 4 (in bold in the table) as the remaining factors had a Pearson-r correlation below 0.15. To simplify explanations, this section will use a specific naming convention on the brain features extracted from HCP, consisting of three parts separated by underscores. The first part is the hemisphere (**l** for left and **r** for right). The second part is a short form of a brain region as defined by the Desikan-Killiany atlas [70]. The third part is the specific feature extracted from that region of the brain, namely **thck** (cortical thickness), **area** (surface area), **grayvol** (grey volume), **meancurv** (integrated rectified mean curvature), **gauscurv** (integrated rectified Gaussian curvature), **foldind** (folding index), **curvind** (intrinsic curvature index), and **myel** (myelin estimate). For instance, **l_precuneus_foldind** corresponds to the folding index value in the precuneus region on the left hemisphere.

Table 5.2: Prediction power regarding mean squared error (MSE) and Pearson-r correlation. In bold the only factors that are further analysed in this chapter. One asterisk means that the p-value of the corresponding test (see Section 5.2) is below 0.05 significance level and therefore we can reject the null hypothesis and confidently say that the model was able to significantly leverage a dependency between the features and labels. Two asterisks are for p-value < 0.01 significance level.

Factor	**Mean MSE (std deviation)**	**Mean Pearson-r (std deviation)**
Factor 1 - Fluid Intelligence	0.95 (0.069)**	0.15 (0.049)**
Factor 2 - Visual Memory	0.91 (0.080)**	0.20 (0.019)**
Factor 3 - Sex-related Factor	0.40 (0.060)**	0.76 (0.040)**
Factor 4 - Executive Functions	0.98 (0.045)**	0.15 (0.042)**
Factor 5 - Sustained Attention	1.04 (0.839)	−0.01 (0.109)
Factor 6 - Waiting Impulsivity	1.02 (0.054)	0.12 (0.052)**
Factor 7 - Linguistic Skills	0.97 (0.131)**	0.14 (0.060)**
Factor 8 - Verbal Episodic Memory	0.99 (0.052)	0.11 (0.094)**
Factor 9 - Visuo-spatial Skills	0.98 (0.284)	0.05 (0.041)*

Figure 5.4 shows the aggregated feature impact on the model output on each of the four factors. For each subfigure, a dot represents a sample from the dataset and its colour is the value of that feature rather than the importance on the model output. The y-axis contains the 10 most important input features, ranked by the aggregated magnitude of impact on the model output across all the samples. Each feature is assigned a SHAP value (in the x-axis) which represents the marginal impact (i.e., importance) on model output or, in other words, both the magnitude and direction of the feature's contribution. For instance, a higher SHAP value means that that feature contributed towards a higher predicted value on the model's output.

It is possible to learn different patterns from the data. For instance, from Figure 5.4a, the surface area of the left hemisphere's cuneus area has almost a symmetrical effect in the model output: a higher value of this feature drives up fluid intelligence prediction with a similar magnitude as when a lower value of this feature drives the prediction down. In contrast, the surface area of the right hemisphere's inferior temporal region contains a more asymmetrical pattern: it primarily drives the prediction to a lower fluid intelligence for lower values of this feature but drives the prediction up to a much lesser extent.



(a) SHAP values from fluid intelligence prediction

(b) SHAP values from visual memory prediction

(c) SHAP values from the sex-related factor prediction

(d) SHAP values from the executive functions prediction

Figure 5.4: Contribution of the most important features in four factors, from one of the respective outer folds. For each feature represented in each row, vertical dispersion stands for the data points which share the same SHAP value for that feature. Each feature value is colour-coded from the highest (i.e. red) to the lowest value (i.e. blue). Higher SHAP values, which are distinct from the actual feature values, mean they contribute in a positive direction to the final predicted variable.

The median rank of brain features when predicting fluid intelligence is not very consistent across the outer folds. This evidence can be seen in Table 5.3 where most of the features identified were ranked, in general, above 10, and the median absolute deviation is relatively high as well. The most important regions are situated towards the posterior part of the brain (parietal and cuneus regions), but some other brain regions situated around the temporal lobes (inferior, entorhinal, transverse) are also selected. The left hemisphere is selected more times than the right hemisphere. As one can see from the most important features in one outer fold when predicting fluid intelligence (see Figure 5.4a), a higher value of that feature generally contributes to a higher value of fluid intelligence, with a clear exception of the myelin of the cuneus and transverse temporal regions on the left hemisphere.

The most important features selected across the five outer folds for the prediction of visual memory are much more consistent across the outer folds, as the median ranks of the ten most important features are below 15 and the median absolute deviation is generally below 10 (see Table 5.4). For this factor, the most important feature is the cortical thickness of the caudal middle frontal region of the left hemisphere, and the entorhinal cortex's grey volume of both hemispheres. In contrast with fluid intelligence,

Table 5.3: Median rank of the ten most important features when predicting fluid intelligence. In parenthesis, MAD stands for median absolute deviation. Feature nomenclature explained in the text.

Brain Feature	**Median Rank (MAD)**
r_inferiorparietal_myel	9 (8)
l_superiorparietal_meancurv	13 (6)
l_cuneus_area	14 (12)
l_cuneus_grayvol	18 (12)
l_inferiortemporal_area	18 (8)
r_inferiorparietal_foldind	21 (20)
l_entorhinal_grayvol	21 (14)
r_postcentral_thck	22 (20)
r_supramarginal_grayvol	23 (19)
l_transversetemporal_myel	32 (11)

the right hemisphere is selected more times to predict visual memory. Although very speculative, this finding is consistent with previous literature linking the right hemisphere to visual memory [158]; still, the grey volumes of the entorhinal region were selected in both hemispheres. As one can see from Figure 5.4b, some variables will contribute with different directions to the predicted visual memory, therefore not being able to extract a general pattern like with fluid intelligence.

Table 5.4: Median rank of the ten most important features when predicting visual memory. In parenthesis, MAD stands for median absolute deviation. Feature nomenclature explained in the text.

Brain Feature	**Median Rank (MAD)**
l_caudalmiddlefrontal_thck	3 (2)
r_entorhinal_grayvol	5 (3)
l_entorhinal_grayvol	7 (3)
r_lateraloccipital_area	8 (3)
l_inferiorparietal_myel	9 (4)
l_insula_myel	13 (10)
r_middletemporal_area	14 (2)
r_insula_thck	14 (6)
r_bankssts_myel	14 (8)
r_insula_foldind	15 (5)

Prediction of the sex-related factor yielded not only the best Pearson-r correlation values but the most consistent median ranks across the outer folds, with the ten most important features always being in the top 15 (see Table 5.5). The insula plays a significant role in distinguishing a male- or female-like brain, specifically regarding its grey volume in both hemispheres and the myelin estimate in the left hemisphere. There is a dominance

Table 5.5: Median rank of the ten most important features when predicting the sex-related factor. In parenthesis, MAD stands for median absolute deviation. Feature nomenclature explained in the text.

Brain Feature	**Median Rank (MAD)**
r_insula_grayvol	1 (0)
l_medialorbitofrontal_grayvol	2 (0)
r_inferiorparietal_area	3 (0)
l_insula_grayvol	3 (1)
l_superiorfrontal_grayvol	5 (2)
l_insula_myel	6 (1)
l_temporalpole_meancurv	8 (2)
r_inferiorparietal_grayvol	11 (3)
r_middletemporal_area	12 (2)
l_entorhinal_myel	12 (4)

of the left hemisphere in predicting this factor, as well as grey volumes. In Figure 5.4c it is possible to find a clear trend in which a lower feature value corresponds to a higher sex-related factor or, in other words, to a brain looking more like a female one.

Finally, the median ranks of the ten most important features when predicting executive functions are not very consistent across the outer folds, but not as inconsistent as to when predicting fluid intelligence (see Table 5.6). The presence of variables from the pericalcarine cortex is significant, though mostly from the right hemisphere. However, some of the most significant regions occur towards the anterior region of the brain with both the medial and lateral orbitofrontal regions and the frontal pole. There is a dominance of the left hemisphere in predicting this factor, and, in general, lower feature values tend to contribute to a lower value of executive functions (see Figure 5.4d).

5.4 Further Exploring SHAP Capabilities

The results analysed in the previous sections illustrate how one could use XGBoost and SHAP to leverage multimodal brain data to predict individual differences in cognitive functioning. Those results allowed to highlight specific and well-defined features that can help with the identification of brain regions for downstream analysis by an expert neuroscientist. To facilitate the communication of these results to a neuroscientist, more analysis can be explored using other SHAP interpretability capabilities. This is even more important due to the nature of XGBoost which can contain many weak learners (see Section 2.1.5.2); therefore, even though each decision tree can be very easy to interpret (see for instance Figure 5.5), it is not humanely possible to analyse a hundred decision trees from a hundred weak learners.

To that end, in this section I want to further illustrate two interpretability capabilities allowed by SHAP which can be used in interaction with an expert neuroscientist interested in further validating and understanding these results.

The first part concerns the analysis of overall patterns among different brain regions, illustrated by the two dependence plots in Figure 5.6. In these plots, the x-axis contains

Table 5.6: Median rank of the ten most important features when predicting executive functions. In parenthesis, MAD stands for median absolute deviation. Feature nomenclature explained in the text.

Brain Feature	Median Rank (MAD)
l_medialorbitofrontal_thck	5 (4)
l_pericalcarine_area	5 (4)
l_lateralorbitofrontal_myel	7 (6)
r_pericalcarine_grayvol	9 (5)
r_pericalcarine_area	10 (7)
r_pericalcarine_foldind	13 (7)
l_frontalpole_thck	15 (6)
l_postcentral_myel	15 (9)
l_temporalpole_area	17 (10)
l_cuneus_grayvol	19 (15)

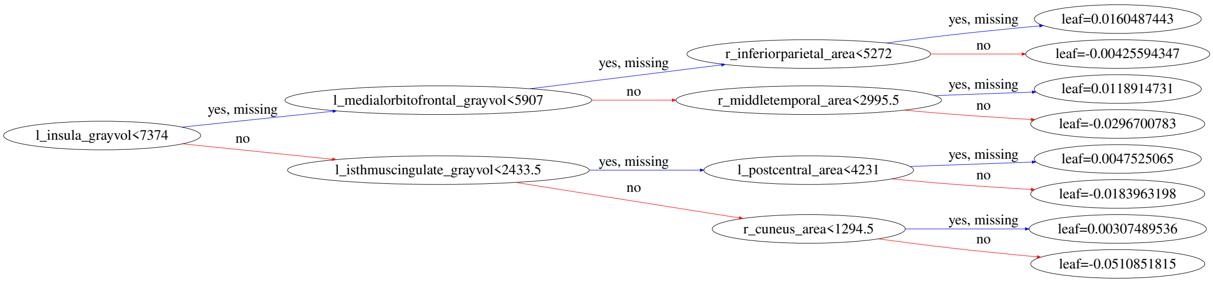


Figure 5.5: One of the XGBoost decision tree for the prediction of the sex-related factor.

the range of values for that feature, the y-axis shows the corresponding SHAP value for each sample/dot marginalised for that feature, and the colour represents the corresponding value of another feature (on the right) for that same sample. The plot on the left shows an expected correlation between the grey volumes of the two insula brain regions from both hemispheres; indeed, higher grey volumes in one hemisphere typically correspond to higher values on the other hemisphere as well, with the exception of a very few cases. Besides making it clearer that there is an approximate value of the *r_insula_grayvol* feature which separates the direction of the contribution towards the model output (around 7300 in this case), it allows the discovery of potential correlations in the data. Moving to the plot on the right (Figure 5.6b), the pattern is a bit more complex and shows a different type of interactions between two other variables: for instance, for lower values of *l_isthmuscingulate_grayvol* approximately below 2600, the SHAP value (i.e. contribution towards model output) is higher when *l_insula_grayvol* is higher. All these insights can not only allow for better discussions with an expert, but also highlight possible sources of improvements in the data and model[7].

Finally, besides allowing the interpretation and analysis of output drivers on an aggregated (i.e. global) level, SHAP also enables the analysis of individual samples to have a more granular understanding of prediction drivers for different people. As a reminder,

[7] https://neurips.cc/Conferences/2021/ScheduleMultitrack?event=21860

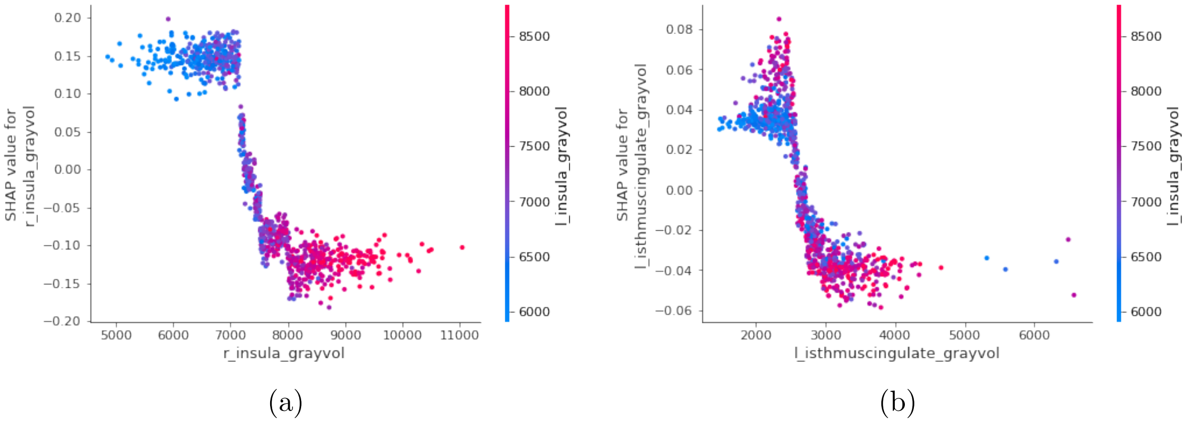


Figure 5.6: Dependence plots between the insula's grey volume in the left hemisphere with two other input features, highlighting different types of dependences and interactions. (a) dependence with the insula's grey volume in the right hemisphere. (b) dependence with the isthmus-cingulate cortex's grey volume in the left hemisphere.

SHAP values represent the change in the expected model prediction conditioned on each feature, therefore explaining the contribution of that feature towards the difference between the average model prediction and the actual final prediction.

In Figure 5.7 it is possible to see an example of a similar final output (around -0.6) for the same model applied on two different people. These plots decompose the drivers of predictions for one single sample each. The y-axis contains the most important features driving the prediction and the corresponding raw value in lighter grey, and the x-axis contains the SHAP value corresponding to the impact on final prediction from the baseline prediction across the population (represented by $E[f(X)]$). The SHAP value of each individual feature is detailed in the arrows that move the prediction from the $E[f(X)]$ baseline. A striking difference between the two plots is that for the bottom one, the most important features drive most of the output value, but in the sample on the top, the remaining 534 other features (in total) are driving most of the changes. This could point an expert to a more wider analysis on the whole brain (in the top case), while the analysis on the bottom case can possibly be more focused on a handful of brain regions.

5.5 Evaluating Model Choice

The purpose of this chapter was to show how one could interpret the neuroanatomical basis of cognition by applying state-of-the-art machine learning methods; therefore, I focused the analysis on SHAP and its specific implementation for the state-of-the-art XGBoost method. A natural question arising from this approach is how useful XGBoost actually is for this dataset in specific.

In this section I employed support vector regression models (adapted from SVM, see Section 2.1.5.1) to approach this chapter's problem in the same way as what was done using XGBoost. I used the same nested cross-validation procedure using the *scikit-learn* [215] python package with a random search over 100 hyperparameters including kernel (polynomial or radial basis), epsilon (a range between 0.1 and 0.5), gamma (scale, auto, and a range between 1e-7 and 1e-4), C (range between 0.1 and 10), and shrinking (boolean variable).

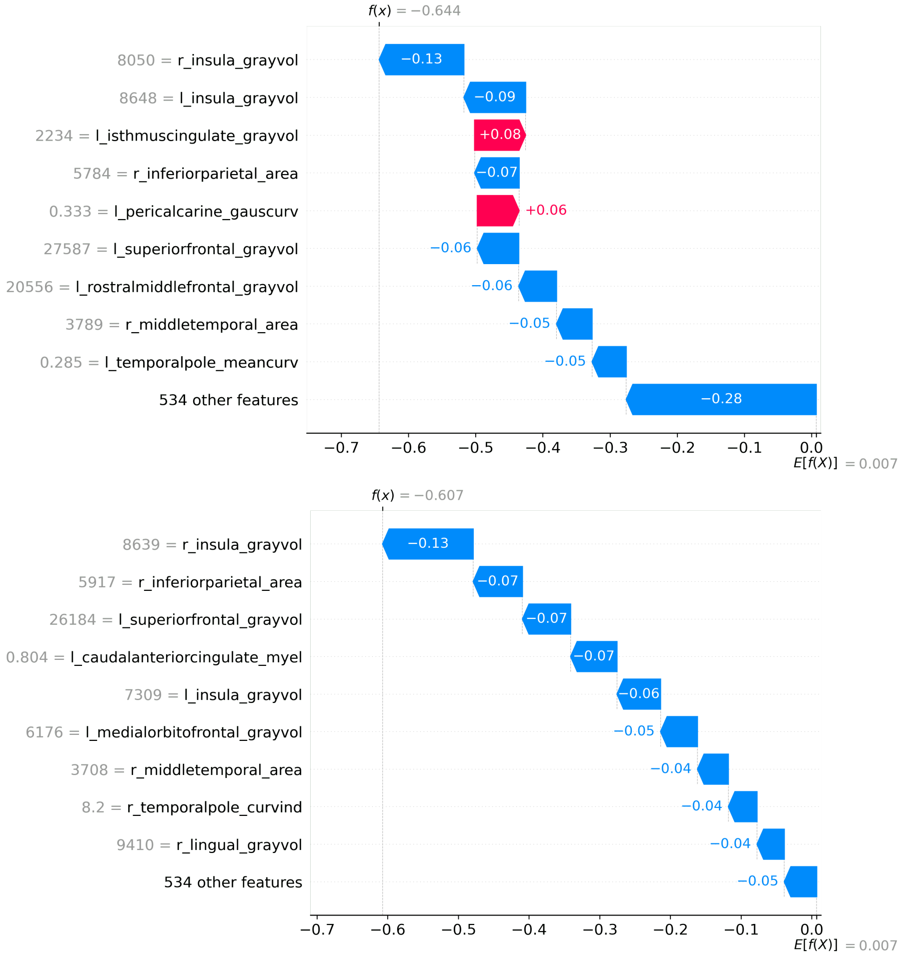


Figure 5.7: Most important features driving a similar final output (around -0.6) on two different people. The marginal contribution of each feature is not the same in the two cases.

The results from this implementation are detailed in Table 5.7 which for readability also includes the results from XGBoost. NaNs occurred when the model learned to output a single value for all samples and therefore it is not possible to calculate a Pearson-r correlation.

Although the results are slightly worse when using SVR (on average), the difference is not significant, and shows once again the challenge of the prediction tasks tackled in this chapter. The lack of robustness presented by SVR when predicting a single value for all test samples in certain factors (illustrated by the NaNs reported for the Pearson-r correlation) shows one strength of XGBoost for these challenging tasks. One possible explanation for a SVR model to output a single value for all test samples could be that the factor loadings are approximately normally distributed, therefore indicating that SVR models can underfit more easily than XGBoost for very complex tasks.

All in all, the main benefit of using a XGBoost model in this case concerns the specific interpretability advantages that SHAP can bring (see Section 5.1.2) instead of using model-agnostic interpretability tools on SVR models. As I tried to show in the previous Section 5.4, the interpretability analysis can be both on an aggregated and granular levels and potentially highlight further developments in cognitive functioning prediction tasks.

Table 5.7: Prediction power regarding mean squared error (MSE) and Pearson-r correlation between XGBoost (XGB) and Support Vector Regression (SVR).

Factor	**Mean MSE (std deviation)**		**Mean Pearson-r (std deviation)**	
	XGB	SVR	XGB	SVR
Factor 1 - Fluid Intelligence	0.95 (0.069)	0.95 (0.047)	0.15 (0.049)	0.13 (0.044)
Factor 2 - Visual Memory	0.91 (0.080)	0.92 (0.086)	0.20 (0.019)	0.19 (0.056)
Factor 3 - Sex-related Factor	0.40 (0.060)	0.46 (0.060)	0.76 (0.040)	0.73 (0.045)
Factor 4 - Executive Functions	0.98 (0.045)	0.99 (0.041)	0.15 (0.042)	0.12 (0.028)
Factor 5 - Sustained Attention	1.04 (0.839)	1.02 (0.854)	-0.01 (0.109)	NaN
Factor 6 - Waiting Impulsivity	1.02 (0.054)	1.01 (0.061)	0.12 (0.052)	0.07 (0.033)
Factor 7 - Linguistic Skills	0.97 (0.131)	0.99 (0.136)	0.14 (0.060)	NaN
Factor 8 - Verbal Episodic Memory	0.99 (0.052)	0.99 (0.084)	0.11 (0.094)	0.09 (0.023)
Factor 9 - Visuo-spatial Skills	0.98 (0.284)	0.97 (0.293)	0.05 (0.041)	NaN

Indeed, this connects quite well with the recent trend in the machine learning community regarding “data-centric” AI approaches to move from a focus on modelling to a focus on the underlying data to improve performance. Given that both XGBoost and SVR seem to provide similar performance in practice, a data-centric approach based on the advantages that SHAP can bring together with XGBoost modelling looks like a potentially interesting future direction in the field.

5.6 Summary

This chapter aimed to leverage multimodal data in the form of a 1-D data representation from brain surface-based morphometry and cortical myelin estimates. I tackled the specific problem of predicting individual differences in cognitive functioning using these two data modalities. The regression model yielded good performance (Pearson-r correlation of almost 0.8) in predicting a sex-related factor; furthermore, the model was able to significantly leverage a dependency between the features and labels for fluid intelligence, visual memory, executive functions, and linguistic skills, but with a corresponding weak performance of a Pearson-r correlation below 0.20 (see Table 5.2). It is important to highlight that I could not achieve an acceptable performance or stability in predicting sustained attention, waiting impulsivity, verbal episodic memory and visuo-spatial skills, hinting that structural neuroimages alone cannot predict these cognitive factors and other types of data are probably needed.

I argue that the approach described in this chapter shows that the use of specific, well-defined and self-explanatory features can help with the quick identification of meaningful regions of interest that are important to interpret machine learning models. Ultimately, this avoids the visual interpretation of neuroimages, which do not convey as many semantics as the directly extracted features.

It was possible to see that while some regions of interest appear more often, others do not appear so much, thus suggesting they might not be as important for understanding general cognition. Some specific features do not seem important as they never appeared in the lists of most important features (eg. integrated rectified Gaussian curvature and intrinsic curvature index). It was also interesting that some factors showed a difference in which hemisphere contributes the most to the prediction task.

Although most factors did not achieve a good performance, these results highlight the potential of these type of features being used to predict cognitive performances. The significance of these results are supported by the fact that the dataset used was the HCP. This dataset is one of the most homogeneous and well-characterised open datasets for healthy subjects, and has a considerable number of people to be analysed (i.e. around a thousand), bringing more credibility to the results presented here. I hope this work could inspire researchers in the field to better study differences of cognition using well-implemented machine learning models that can be correctly interpreted.

Chapter 6

Conclusion

I conclude this thesis by highlighting its key contributions resulting from a combination of different knowledge fields, including machine learning, molecular biology, and neuroimaging. I will also offer some thoughts on future research directions based on the main limitations.

6.1 Contributions

This thesis has explored data-driven representations and methods targeted for brain data in the gene expression and neuroimaging domains. This was conducted having three main research questions - outlined in Section 1.2 - driving the developments. In a first part composed of two chapters, I tackled cases of Graph-Dimensional representations, while in the second part I explored a 1-Dimensional (i.e. flatten) data representation.

Chapter 3 modelled gene expression data on the most comprehensive human transcriptome dataset by using a Graph-Dimensional representation in the form of a multiplex framework. I used unsupervised approaches to find intra- and inter-tissue profiles of gene expression, with a focus on inter-tissue profiles stemming from the brain. I provided a rich resource of co-expression networks, communities, multiplex architectures, and enriched biological pathways, while illustrating the cross-study relevance through analysis on external datasets. The identification of community structure as an important organising principle of the human transcriptome, and the suggestion of a presence of a hierarchy of clusters at increasingly finer scales might help catalyse research into inter-tissue regulatory insights with disease consequences.

In Chapter 4, I proposed a novel deep spatio-temporal model for the analysis of resting-state fMRI data. In this supervised model, temporal convolutional networks capture intra-temporal dynamics, and graph neural networks account for spatial inter-relationships of brain connectivity. I investigated the model's explainability capacities by capitalising on a dataset with over 35,000 individual brain scans, and the robustness and applicability of this model on an external multimodal dataset with almost a thousand brain scans. This model is, to the best of my knowledge, the first end-to-end deep learning model that is able to capitalise on both spatial and temporal information from rs-fMRI data by using temporal convolutions and graph neural networks while providing the flexibility to extract human-readable explanations from a hierarchical pooling mechanism.

In Chapter 5, I presented an approach to correctly identify meaningful regions of interest in the brain for a supervised task by leveraging semantic, well-characterised 1-Dimensional features. The interpretation of this approach was important to understanding the driving forces in the notoriously hard problem of predicting cognitive functioning, ultimately avoiding the visual interpretation of neuroimages.

These three main chapters helped highlight the contributions of gene expression and neuroimaging data when modelling the brain and its intra- and inter-dynamics across the human body. As a consequence, this thesis makes three main contributions while tackling the research questions defined in Section 1.2:

- **Scientific contribution**, as the work conducted in the three main chapters helped explore the three research questions posed at the beginning of this thesis, across the scientific fields of machine learning, molecular biology, and neuroscience. This contribution is further illustrated in the publications outlined in Section 1.4.

- **Technological contribution**, by integrating tools from well-established and open-source technologies, while publicly releasing the code to the community. The tools that were integrated included mostly those which used the Python programming language and various specific packages.

- **Applied contribution**, which expanded from the technological contribution. By publicly releasing all my code and resources whenever possible, I provided documented platforms that enable applied scientific research and development in the various fields explored.

I conclude this section by humbly highlighting that this thesis tries to tackle two Sustainable Development Goals (SDGs). Given the practical implications of molecular biology and neuroscience, this thesis can indirectly catalyse research needed to improve healthcare. This implication directly connects with **SDG 3 - Good Health and Well-being**, the goal of which is to “ensure healthy lives and promote well-being for all at all ages”. Indeed, a better knowledge of how the brain works and interacts with other parts of the body can lead to better counselling for patients with both general and neurological diseases [120, 198]. For example, the likely novel functional information present in the communities of Chapter 3 may have critical implications in mapping the human diseasome.

As outlined in Section 1.4, this thesis contains a considerable amount of multidisciplinary research and interaction, which was able to create multi-stakeholder partnerships across different countries and knowledge fields. It is likely that these partnerships will not end with the conclusion of this thesis and, given the importance to SDG 3, I consider that this thesis also helps to tackle **SDG 17 - Partnerships for the Goals**. These partnerships will surely continue to mobilise and share expert knowledge on technological and applied resources to directly or indirectly achieve the other SDGs.

6.2 Limitations and Future Research Directions

It is quite an exciting time to be a researcher. We are witnessing astounding developments in all areas of science while still leaving many questions unanswered. This section will offer some concluding thoughts on what I believe are open avenues of research from this thesis, and which are already finding their place in the community. I will focus on research topics that correspond to what I consider to be the specific limitations of this thesis, rather than trying to give broad ideas. By no means do I intend this to be an exhaustive section, but I also hope it can highlight some interesting paths on which I am sure the field of artificial intelligence will play a critical role. When short-term topics are presented, they are meant to be set on a time frame of approximately between 6 to 12 months.

The first topic I should highlight is the **multimodal alignment of genetic and neuroimaging data**, the two scales I have explored in this thesis. There have been some approaches in recent years tackling this connection. For instance, a previous study linked brain connectivity with genetic levels to find evidence for different hypotheses on tau distribution [61], but the lack of samples hindered the statistical power of such connections. We are seeing other articles [34, 220, 255] coming into play, and this is supported by the increased number of very big datasets providing both neuroimaging brain scans, as well as genetic information. Take, for instance, the UK Biobank which I have used in Chapter 4, already with more than 500,000 people involved with some genotype data. Likewise, the Human Connectome Project is a traditional neuroimaging open dataset that I have exploited in chapters 4 and 5; this dataset now provides SNP genotypes since 2018[1]. These numbers were unimaginable a few years ago, and I am sure the field will use these and other datasets to produce a broader framework linking all brain connectivity, genetic profiles, and phenotypes. The field of machine learning and, more particularly, deep learning, has matured several multimodal approaches just waiting to be fully explored in the next stage of big data coming from large consortia. To be more specific, a direct, short-term objective would be to bring these other modalities to the architecture I presented in Chapter 4 given the promising results leveraging the multimodal brain data in Section 4.4.5; this could be achieved by defining the global level functions from the Graph Network block [29] which were not used in this thesis and could include subject-specific genetic information. From here, a further short-term extension would be to go beyond a binary output and instead include more complex labels to be predicted by the architecture [113].

Another topic, more closely related to machine learning itself, is **model complexity**, which needs to be considered as the fields of machine learning and neuroimaging keep growing. With the advent of more potent MRI scans [205] at unprecedented resolution [82], models will need to accompany this increase. Both data and models will likely become increasingly complex; however, there is the need to focus on the opposite direction, by reducing model complexity to improve inference time and ultimately to allow a full democratisation of artificial intelligence in resource-constrained devices. While tied to environmental advantages [239], this is already drawing interest in neuroscience [269]. In a short-term perspective, there is an important follow-up work directly stemming from this thesis which concerns exploring hardware-software co-design techniques such as quantisation and pruning. Despite being widely popular with other types of neural networks, only recently these techniques started to be adapted to graph neural networks [56, 259]. I believe it is then paramount to explore these techniques for model complexity reduction with the GNN architecture explored in Chapter 4 and understand how they can affect performance. This approach can then be extended to another short-term follow-up work: to understand whether the unprecedented resolution allowed with 7 Tesla MRI machines actually translate in performance improvements. Recent papers showed promising results that indeed moving from 3 Tesla to 7 Tesla brings advantages [50, 196], but more experiments would be needed to validate these results, for example using the models from chapters 4 and 5.

A third, short-term topic regards the use of probabilistic **Bayesian models to mea-**

[1] `https://www.humanconnectome.org/study/hcp-young-adult/article/hcp-releases-snp-genotypes-collected-hcp-young-adult-subjects-dbgap`

sure uncertainty instead of using single deterministic outputs[2]. Despite the existence of many such models and varied research directions [6, 265], most of these applications on real-world data are not focused on 1-D Dimensional data; therefore, given the practical utility of this type of data (see Chapter 5), exploring how recent developments on probabilistic machine learning can be applied is important. Indeed, I was already personally involved in another work using Bayesian modelling for 1-D Dimensional neuroimaging data to identify healthy individuals with Alzheimer neuroimaging phenotypes, with promising results [22]. A natural, short-term extension would be to use recent extensions to model probabilistic outputs with XGBoost [188] on the learning task from Chapter 5.

I would like to finish with a long-term future research direction related to the issue of **dealing with confounders**. This issue is of paramount importance in any applied research field and reasonably left untackled in the machine learning literature. The existence of confounders is a problem in any serious statistical analysis, as it could drive the prediction task without the researcher detecting its existence. I tried to indirectly avoid issues with confounders in this thesis. I used a highly homogeneous dataset in chapters 4 and 5 (i.e., HCP) and exploited a complete unsupervised tool to deconfound unwanted variation in Chapter 3. In Section 4.5, possible confounders were directly included in the architectures, and in Chapter 4 I used a dataset with only healthy people (i.e., UK Biobank) while being careful with stratification. Nevertheless, to further validate my findings, one would need to analyse confound effects in more detail to test causality rather than correlation. The neuroimaging field does not seem to have a widely accepted approach. For example, it is common to linearly regress out variables of interest and develop models on the residuals, which was widely analysed in a recent study on the UK Biobank [9]. Some more recent developments are being directly implemented in deep learning models, for cases such as unlearning the scanner bias in MRI scans [73]. In molecular biology the use of unsupervised tools such as *sva* (see Chapter 3) seems to be widely adopted; nevertheless, we might be removing real signal from the data by only correcting values instead of bringing those variables into the model itself. I should highlight that this can have direct health and economic effects on people's lives. For instance, at the beginning of the COVID-19 pandemic there were studies indicating the potential protector effect of vitamin D, which seemed to be corroborated in people on the UK Biobank; however, after making adjustments for confounders, Vitamin D insufficiency was not independently associated with either COVID-19 infection or linked mortality [133].

Tackling all these and other issues will ultimately help establish artificial intelligence as a necessary tool to help achieve the sustainable development goals [273]. I will personally try to frame my future career following these goals, and I urge all researchers to do likewise. This will certainly unleash the true potential of applied artificial intelligence, whatever the researcher or research group using it.

[2] Model uncertainty can be framed in different ways, though in this case I am referring to the simple generic case of understanding how certain/confident a model is of its output. In practice, this means that a distribution over predictions should be returned instead of a single prediction, and this confidence range can indicate when one can trust a model's prediction.

References

[1] Due credit. *Nature*, 496(7445):270–270, April 2013. Cited on page 36.

[2] Graph theory methods: applications in brain networks. *Dialogues in Clinical Neuroscience*, 20(2):111–121, June 2018. Cited on page 67.

[3] A Hybrid 3DCNN and 3DC-LSTM based model for 4D Spatio-temporal fMRI data: An ABIDE Autism Classification study. *arXiv preprint arXiv:2002.05981*, 2020. Cited on page 67.

[4] The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, September 2020. Cited on pages 15, 47, 48, and 53.

[5] Sergi Abadal, Akshay Jain, Robert Guirado, Jorge López-Alonso, and Eduard Alarcón. Computing graph neural networks: A survey from algorithms to accelerators. *ACM Computing Surveys*, 54(9):1–38, December 2022. Cited on page 35.

[6] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, December 2021. Cited on page 110.

[7] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018. Cited on page 33.

[8] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. Cited on page 14.

[9] Fidel Alfaro-Almagro, Paul McCarthy, Soroosh Afyouni, Jesper L.R. Andersson, Matteo Bastiani, Karla L. Miller, Thomas E. Nichols, and Stephen M. Smith. Confound modelling in UK Biobank brain imaging. *NeuroImage*, 224:117002, January 2021. Cited on page 110.

[10] Elena A. Allen, Eswar Damaraju, Sergey M. Plis, Erik B. Erhardt, Tom Eichele, and Vince D. Calhoun. Tracking Whole-Brain Connectivity Dynamics in the Resting State. *Cerebral Cortex*, 24(3):663–676, March 2014. Cited on page 67.

[11] Dominic J Allocco, Isaac S Kohane, and Atul J Butte. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, 5(1): 18, 2004. Cited on page 47.

[12] Christopher J. Anders, Leander Weber, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. Finding and removing clever hans: Using explanation methods to debug and improve deep models. *Information Fusion*, 77: 261–295, January 2022. Cited on page 34.

[13] Augusto Anguita-Ruiz, Alberto Segura-Delgado, Rafael Alcalá, Concepción M. Aguilera, and Jesús Alcalá-Fdez. eXplainable artificial intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. *PLOS Computational Biology*, 16(4):e1007792, April 2020. Cited on page 34.

[14] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, June 2020. Cited on pages 33, 66, and 90.

[15] Salim Arslan, Sofia Ira Ktena, Ben Glocker, and Daniel Rueckert. Graph saliency maps through spectral convolutional networks: Application to sex classification with brain connectivity. In *Lecture Notes in Computer Science*, pages 3–13. Springer International Publishing, 2018. Cited on page 67.

[16] Lidia Auret and Chris Aldrich. Empirical comparison of tree ensemble variable importance measures. *Chemometrics and Intelligent Laboratory Systems*, 105(2): 157–170, February 2011. Cited on page 90.

[17] Andrea Avena-Koenigsberger, Bratislav Misic, and Olaf Sporns. Communication dynamics in complex brain networks. *Nature Reviews Neuroscience*, 19(1):17–33, December 2017. Cited on page 65.

[18] Tiago Azevedo, Luca Passamonti, Pietro Lió, and Nicola Toschi. A machine learning tool for interpreting differences in cognition using brain features. In *IFIP Advances in Information and Communication Technology*, pages 475–486. Springer International Publishing, 2019. Cited on page 17.

[19] Tiago Azevedo, Luca Passamonti, Pietro Liò, and Nicola Toschi. Towards a predictive spatio-temporal representation of brain data. In *ICLR Workshop on AI for Affordable Healthcare (AI4AH)*, February 2020. Cited on page 17.

[20] Tiago Azevedo, Luca Passamonti, Pietro Lio, and Nicola Toschi. A deep spatiotemporal graph learning architecture for brain connectivity analysis. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, July 2020. Cited on page 17.

[21] Tiago Azevedo, Giovanna Maria Dimitri, Pietro Lió, and Eric R. Gamazon. Multilayer modelling of the human transcriptome and biological mechanisms of complex diseases and traits. *npj Systems Biology and Applications*, 7(1), May 2021. Cited on page 17.

[22] Tiago Azevedo, Richard A.I. Bethlehem, David J. Whiteside, Nol Swaddiwudhipong, James B. Rowe, Pietro Lio, and Timothy Rittman. Identifying healthy individuals with alzheimer neuroimaging phenotypes in the UK Biobank. *medRxiv preprint*, January 2022. Cited on page 110.

[23] Tiago Azevedo, Alexander Campbell, Rafael Romero-Garcia, Luca Passamonti, Richard A.I. Bethlehem, Pietro Liò, and Nicola Toschi. A deep graph neural network architecture for modelling spatio-temporal dynamics in resting-state functional MRI data. *Medical Image Analysis*, 79:102471, July 2022. Cited on page 17.

[24] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. Cited on pages 68 and 72.

[25] Muhammet Balcilar, Guillaume Renton, Pierre Héroux, Benoit Gaüzère, Sébastien Adam, and Paul Honeine. Analyzing the expressive power of graph neural networks in a spectral perspective. In *International Conference on Learning Representations*, 2021. Cited on page 35.

[26] J. Ballinger. Diastolic pseudogating, April 2013. Cited on page 42.

[27] S. Bandres-Ciga, S. Saez-Atienzar, J. J. Kim, M. B. Makarious, F. Faghri, M. Diez-Fairen, H. Iwaki, H. Leonard, J. Botia, M. Ryten, D. Hernandez, J. R. Gibbs, J. Ding, Z. Gan-Or, A. Noyce, L. Pihlstrom, A. Torkamani, A. R. Soltis, C. L. Dalgard, S. W. Scholz, B. J. Traynor, D. Ehrlich, C. R. Scherzer, M. Bookman, M. Cookson, C. Blauwendraat, M. A. Nalls, and A. B. Singleton. Large-scale pathway specific polygenic risk and transcriptomic community network analysis identifies novel functional pathways in parkinson disease. *Acta Neuropathologica*, 140(3): 341–358, June 2020. Cited on page 47.

[28] M. S. BARTLETT. The effect of standardization on a $\chi 2$ approximation in factor analysis. *Biometrika*, 38(3-4):337–344, 1951. Cited on page 93.

[29] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. Cited on pages 68, 69, and 109.

[30] Christian F Beckmann, Marilena DeLuca, Joseph T Devlin, and Stephen M Smith. Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457): 1001–1013, May 2005. Cited on page 65.

[31] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8):1798–1828, August 2013. Cited on pages 13 and 32.

[32] Marcel Bengs, Nils Gessert, and Alexander Schlaefer. 4D Spatio-Temporal Deep Learning with 4D fMRI Data for Autism Spectrum Disorder Classification. *arXiv preprint arXiv:2004.10165*, 2020. Cited on page 67.

[33] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305, 2012. Cited on page 27.

[34] Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R. Vanderburg, Asa Segerstolpe, Meng Zhang, Inbal Avraham-Davidi, Sanja Vickovic, Mor Nitzan, Sai Ma, Jason Buenrostro, Nik Bear Brown, Duccio Fanelli, Xiaowei Zhuang, Evan Z. Macosko, and Aviv Regev. Deep learning and alignment of spatially-resolved whole transcriptomes of single cells in the mouse brain with tangram. August 2020. Cited on page 109.

[35] Lukas Biewald. Experiment tracking with weights and biases, 2020. URL `https://www.wandb.com/`. Software available from wandb.com. Cited on page 73.

[36] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 2006. Cited on page 29.

[37] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. Cited on page 49.

[38] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *arXiv preprint arXiv:2102.13076*, 2021. Cited on page 90.

[39] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*. ACM Press, 1992. Cited on page 28.

[40] Nick Bostrom and Eliezer Yudkowsky. The ethics of artificial intelligence. *The Cambridge Handbook of Artificial Intelligence*, 1:316–334, 2014. Cited on page 34.

[41] Rotem Botvinik-Nezer, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A Mumford, R Alison Adcock, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, pages 1–7, 2020. Cited on page 42.

[42] G.E.P. Box. Robustness in the strategy of scientific model building. In *Robustness in Statistics*, pages 201–236. Elsevier, 1979. Cited on page 14.

[43] Leo Breiman, Jerome Friedman, Charles J Stone, and R A Olshen. *Classification and Regression Trees*. Chapman & Hall/CRC, January 1984. Cited on page 30.

[44] Timothy A Brown. *Confirmatory factor analysis for applied research*. Guilford publications, 2015. Cited on pages 92 and 93.

[45] Andreas Buja and Nermin Eyuboglu. Remarks on parallel analysis. *Multivariate Behavioral Research*, 27(4):509–540, October 1992. Cited on page 49.

[46] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018. Cited on page 34.

[47] Pisanu Buphamalai, Tomislav Kokotovic, Vanja Nagy, and Jörg Menche. Network analysis reveals rare disease signatures across multiple levels of biological organization. *Nature Communications*, 12(1), November 2021. Cited on page 47.

[48] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, October 2018. Cited on page 70.

[49] Mark Bydder, Andres Rahal, Gary D. Fullerton, and Graeme M. Bydder. The magic angle effect: A source of artifact, determinant of image contrast, and technique for imaging. *Journal of Magnetic Resonance Imaging*, 25(2):290–300, 2007. Cited on page 42.

[50] Yuxuan Cai, Shir Hofstetter, Wietske van der Zwaag, Wietske Zuiderbaan, and Serge O. Dumoulin. Individualized cognitive neuroscience needs 7t: Comparing numerosity maps at 3t and 7t MRI. *NeuroImage*, 237:118184, August 2021. Cited on page 109.

[51] Michael Caldera, Pisanu Buphamalai, Felix Müller, and Jörg Menche. Interactome-based approaches to human disease. *Current Opinion in Systems Biology*, 3:88–94, June 2017. Cited on page 38.

[52] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, July 2019. Cited on page 33.

[53] Davide Castelvecchi. Can we open the black box of AI? *Nature*, 538(7623):20–23, October 2016. Cited on page 14.

[54] Giacomo Cavalli and Edith Heard. Advances in epigenetics link genetics to the environment and disease. *Nature*, 571(7766):489–499, July 2019. Cited on page 37.

[55] Yuting Chang, Xuewei Wang, Yide Xu, Liu Yang, Qufei Qian, Sihan Ju, Yao Chen, Shuaizhou Chen, Na Qin, Zijian Ma, et al. Comprehensive characterization of cancer-testis genes in testicular germ cell tumor. *Cancer Medicine*, 8(7):3511–3519, 2019. Cited on page 56.

[56] Tianlong Chen, Yongduo Sui, Xuxi Chen, Aston Zhang, and Zhangyang Wang. A unified lottery ticket hypothesis for graph neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1695–1706. PMLR, 18–24 Jul 2021. Cited on page 109.

[57] Tianqi Chen and Carlos Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. ACM Press, 2016. Cited on pages 29, 30, 31, 75, 87, 90, and 94.

[58] Yitian Chen, Yanfei Kang, Yixiong Chen, and Zizhuo Wang. Probabilistic forecasting with temporal convolutional neural network. *Neurocomputing*, 399:491–501, July 2020. Cited on page 68.

[59] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015. Cited on page 56.

[60] GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017. Cited on pages 48 and 56.

[61] Thomas E Cope, Timothy Rittman, Robin J Borchert, P Simon Jones, Deniz Vatansever, Kieren Allinson, Luca Passamonti, Patricia Vazquez Rodriguez, W Richard Bevan-Jones, John T O'Brien, and James B Rowe. Tau burden and the functional connectome in Alzheimer's disease and progressive supranuclear palsy. *Brain*, 141 (2):550–567, January 2018. Cited on page 109.

[62] Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. In *Advances in Neural Information Processing Systems*, 2020. Cited on page 73.

[63] Francis HC Crick. On protein synthesis. In *Symposia of the Society for Experimental Biology*, volume 12, page 8, 1958. Cited on page 36.

[64] Brian M Dale, Mark A Brown, and Richard C Semelka. *MRI: basic principles and applications*. John Wiley & Sons, 2015. Cited on page 39.

[65] Christos Davatzikos. Machine learning in neuroimaging: Progress and challenges. *NeuroImage*, 197:652–656, August 2019. Cited on page 15.

[66] Stéphanie Debette, Sabrina Schilling, Marie-Gabrielle Duperron, Susanna C. Larsson, and Hugh S. Markus. Clinical significance of magnetic resonance imaging markers of vascular brain injury. *JAMA Neurology*, 76(1):81, January 2019. Cited on page 39.

[67] Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, May 2021. Cited on page 87.

[68] Mary F Dempsey, Barrie Condon, and Donald M Hadley. MRI safety review. *Seminars in Ultrasound, CT and MRI*, 23(5):392–401, October 2002. Cited on page 39.

[69] Austin Derrow-Pinion, Jennifer She, David Wong, Oliver Lange, Todd Hester, Luis Perez, Marc Nunkesser, Seongjae Lee, Xueying Guo, Brett Wiltshire, Peter W. Battaglia, Vishal Gupta, Ang Li, Zhongwen Xu, Alvaro Sanchez-Gonzalez, Yujia Li, and Petar Velickovic. ETA prediction with graph neural networks in google maps. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. ACM, October 2021. Cited on page 35.

[70] Rahul S. Desikan, Florent Ségonne, Bruce Fischl, Brian T. Quinn, Bradford C. Dickerson, Deborah Blacker, Randy L. Buckner, Anders M. Dale, R. Paul Maguire, Bradley T. Hyman, Marilyn S. Albert, and Ronald J. Killiany. An automated

labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3):968–980, July 2006. Cited on pages 42, 43, 71, 75, 88, and 97.

[71] Alex Diaz-Papkovich, Luke Anderson-Trocmé, and Simon Gravel. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genetics*, 15(11), 2019. Cited on page 50.

[72] Botty Dimanov. *Interpretable Deep Learning: Beyond Feature-Importance with Concept-based Explanations*. PhD thesis, University of Cambridge, 2020. Cited on pages 32 and 34.

[73] Nicola K. Dinsdale, Mark Jenkinson, and Ana I.L. Namburete. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. *NeuroImage*, 228:117689, March 2021. Cited on pages 44 and 110.

[74] Xiaowen Dong, Dorina Thanou, Laura Toni, Michael Bronstein, and Pascal Frossard. Graph signal processing for machine learning: A review and new perspectives. *IEEE Signal Processing Magazine*, 37(6):117–127, November 2020. Cited on page 15.

[75] Michael W. Dorrity, Lauren M. Saunders, Christine Queitsch, Stanley Fields, and Cole Trapnell. Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nature Communications*, 11(1), March 2020. Cited on page 63.

[76] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. Cited on page 33.

[77] Julien Dubois, Paola Galdi, Yanting Han, Lynn K. Paul, and Ralph Adolphs. Resting-state functional brain connectivity best predicts the personality dimension of openness to experience. *Personality Neuroscience*, 1, July 2018. Cited on page 87.

[78] Andrea Duggento, Luca Passamonti, Gaetano Valenza, Riccardo Barbieri, Maria Guerrisi, and Nicola Toschi. Multivariate granger causality unveils directed parietal to prefrontal cortex connectivity during task-free MRI. *Scientific Reports*, 8(1), April 2018. Cited on page 65.

[79] Andrea Duggento, Maria Guerrisi, and Nicola Toschi. Recurrent neural networks for reconstructing complex directed brain connectivity. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, July 2019. Cited on page 66.

[80] Nicha C. Dvornek, Pamela Ventola, Kevin A. Pelphrey, and James S. Duncan. Identifying autism from resting-state fMRI using long short-term memory networks. In *Machine Learning in Medical Imaging*, pages 362–370. Springer International Publishing, 2017. Cited on page 66.

[81] Mr Amir Ebrahimighahnavieh, Suhuai Luo, and Raymond Chiong. Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review. *Computer Methods and Programs in Biomedicine*, 187:105242, April 2020. Cited on page 15.

[82] Brian L. Edlow, Azma Mareyam, Andreas Horn, Jonathan R. Polimeni, Thomas Witzel, M. Dylan Tisdall, Jean C. Augustinack, Jason P. Stockmann, Bram R. Diamond, Allison Stevens, Lee S. Tirrell, Rebecca D. Folkerth, Lawrence L. Wald, Bruce Fischl, and Andre van der Kouwe. 7 tesla MRI of the ex vivo human brain at 100 micron resolution. *Scientific Data*, 6(1):1–10, October 2019. Cited on page 109.

[83] Simon B. Eickhoff, Klaas E. Stephan, Hartmut Mohlberg, Christian Grefkes, Gereon R. Fink, Katrin Amunts, and Karl Zilles. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, 25 (4):1325–1335, May 2005. Cited on page 42.

[84] Shaker El-Sappagh, Jose M. Alonso, S. M. Riazul Islam, Ahmad M. Sultan, and Kyung Sup Kwak. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for alzheimer's disease. *Scientific Reports*, 11(1), January 2021. Cited on page 34.

[85] Maha Elbayad, Laurent Besacier, and Jakob Verbeek. Pervasive attention: 2D convolutional neural networks for sequence-to-sequence prediction. *arXiv preprint arXiv:1808.03867*, 2018. Cited on page 68.

[86] Frank Emmert-Streib, Zhen Yang, Han Feng, Shailesh Tripathi, and Matthias Dehmer. An introductory review of deep learning for prediction models with big data. *Frontiers in Artificial Intelligence*, 3, February 2020. Cited on page 14.

[87] Taban Eslami, Vahid Mirjalili, Alvis Fong, Angela R. Laird, and Fahad Saeed. ASD-DiagNet: A hybrid learning approach for detection of autism spectrum disorder using fMRI data. *Frontiers in Neuroinformatics*, 13:70, 2019. Cited on page 66.

[88] Achraf Essemlali, Etienne St-Onge, Maxime Descoteaux, and Pierre-Marc Jodoin. Understanding alzheimer disease's structural connectivity through explainable ai. In Tal Arbel, Ismail Ben Ayed, Marleen de Bruijne, Maxime Descoteaux, Herve Lombaert, and Christopher Pal, editors, *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, pages 217–229. PMLR, July 2020. Cited on page 34.

[89] D.C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T.E.J. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S.W. Curtiss, S. Della Penna, D. Feinberg, M.F. Glasser, N. Harel, A.C. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S.E. Petersen, F. Prior, B.L. Schlaggar, S.M. Smith, A.Z. Snyder, J. Xu, and E. Yacoub. The human connectome project: A data acquisition perspective. *NeuroImage*, 62(4):2222–2231, October 2012. Cited on page 44.

[90] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, Marija Milacic, Corina Duenas Roca, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Guilherme Viteri, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, 46(D1):D649–D655, November 2017. Cited on page 52.

[91] Maria I. Falcon, Viktor Jirsa, and Ana Solodkin. A new neuroinformatics approach to personalized medicine in neurology: The virtual brain. *Current Opinion in Neurology*, 29(4):429–436, August 2016. Cited on page 41.

[92] Liangwei Fan, Jianpo Su, Jian Qin, Dewen Hu, and Hui Shen. A deep network model on dynamic functional connectivity with applications to gender classification and intelligence prediction. *Frontiers in Neuroscience*, 14:881, 2020. Cited on page 67.

[93] Lingzhong Fan, Hai Li, Junjie Zhuo, Yu Zhang, Jiaojian Wang, Liangfu Chen, Zhengyi Yang, Congying Chu, Sangma Xie, Angela R. Laird, Peter T. Fox, Simon B. Eickhoff, Chunshui Yu, and Tianzi Jiang. The human brainnetome atlas: A new brain atlas based on connectional architecture. *Cerebral Cortex*, 26(8):3508–3526, May 2016. Cited on page 42.

[94] Bryn Farnsworth. Eeg vs. mri vs. fmri - what are the differences?, July 2019. URL `https://imotions.com/blog/eeg-vs-mri-vs-fmri-differences/`. Cited on page 39.

[95] Scott H Faro and Feroze B Mohamed. *Functional MRI: basic principles and clinical applications*. Springer Science & Business Media, 2006. Cited on page 39.

[96] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. Cited on page 72.

[97] Alexandru-Catalin Filip, Tiago Azevedo, Luca Passamonti, Nicola Toschi, and Pietro Lio. A novel graph attention network architecture for modeling multimodal brain connectivity. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, July 2020. Cited on pages 17 and 83.

[98] Massimo Filippi and Filippi. *fMRI techniques and protocols*, volume 830. Springer, 2016. Cited on pages 40 and 41.

[99] Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam, and Isaac S. Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, March 2019. Cited on page 14.

[100] Bruce Fischl, Martin I. Sereno, Roger B.H. Tootell, and Anders M. Dale. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, 8(4):272–284, 1999. Cited on page 88.

[101] Alex Fornito, Andrew Zalesky, and Michael Breakspear. The connectomics of brain disorders. *Nature Reviews Neuroscience*, 16(3):159–172, February 2015. Cited on page 65.

[102] Alex Fornito, Andrew Zalesky, and Edward Bullmore. *Fundamentals of brain network analysis*. Academic Press, 2016. Cited on page 41.

[103] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, February 2010. Cited on page 47.

[104] Rosalind E Franklin and Raymond G Gosling. Molecular configuration in sodium thymonucleate. *Nature*, 171(4356):740–741, April 1953. Cited on page 36.

[105] P. Frasconi, M. Gori, and A. Sperduti. A general framework for adaptive processing of data structures. *IEEE Transactions on Neural Networks*, 9(5):768–786, 1998. Cited on page 35.

[106] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2), April 2000. Cited on page 30.

[107] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001. Cited on page 91.

[108] Anna Fry, Thomas J Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American Journal of Epidemiology*, 186(9):1026–1034, June 2017. Cited on page 43.

[109] Soham Gadgil, Qingyu Zhao, Adolf Pfefferbaum, Edith V. Sullivan, Ehsan Adeli, and Kilian M. Pohl. Spatio-temporal graph convolution for resting-state fMRI analysis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 528–538. Springer International Publishing, 2020. Cited on pages 67, 74, 76, and 84.

[110] Ilaria Boscolo Galazzo, Federica Cruciani, Lorenza Brusini, Ahmed Salih, Petia Radeva, Silvia Francesca Storti, and Gloria Menegaz. Explainable artificial intelligence for magnetic resonance imaging aging brainprints: Grounds and challenges. *IEEE Signal Processing Magazine*, 39(2):99–116, March 2022. Cited on page 34.

[111] Eric R Gamazon, Ayellet V Segrè, Martijn van de Bunt, Xiaoquan Wen, Hualin S Xi, Farhad Hormozdiari, Halit Ongen, Anuar Konkashbaev, Eske M Derks, François Aguet, et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease-and trait-associated variation. *Nature Genetics*, 50(7):956–967, 2018. Cited on page 48.

[112] Eric R Gamazon, Aeilko H Zwinderman, Nancy J Cox, Damiaan Denys, and Eske M Derks. Multi-tissue transcriptome analyses identify genetic mechanisms underlying neuropsychiatric traits. *Nature Genetics*, 51(6):933–940, 2019. Cited on page 60.

[113] Jianliang Gao, Tengfei Lyu, Fan Xiong, Jianxin Wang, Weimao Ke, and Zhao Li. MGNN: A multimodal graph neural network for predicting the survival of cancer patients. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, July 2020. Cited on page 109.

[114] Kathleen A. Garrison, Dustin Scheinost, Emily S. Finn, Xilin Shen, and R. Todd Constable. The (in)stability of functional brain network measures across thresholds. *NeuroImage*, 118:651–661, September 2015. Cited on page 72.

[115] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR, August 2017. Cited on page 68.

[116] Michael A. Gelbart, Jasper Snoek, and Ryan P. Adams. Bayesian optimization with unknown constraints. *arXiv preprint arXiv:1403.5607*, 2014. Cited on page 27.

[117] Zachary F. Gerring, Eric R. Gamazon, and Eske M. Derks. A gene co-expression network-based analysis of multiple brain tissues reveals novel genes and molecular pathways underlying major depression. *PLOS Genetics*, 15(7):e1008245, July 2019. Cited on page 47.

[118] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, November 2021. Cited on page 34.

[119] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1263–1272. JMLR.org, 2017. Cited on page 35.

[120] Matthew J. Girgenti, Jiawei Wang, Dingjue Ji, Dianne A. Cruz, Murray B. Stein, Joel Gelernter, Keith A. Young, Bertrand R. Huber, Douglas E. Williamson, Matthew J. Friedman, John H. Krystal, Hongyu Zhao, and Ronald S. Duman. Transcriptomic organization of the human brain in post-traumatic stress disorder. *Nature Neuroscience*, 24(1):24–33, December 2020. Cited on page 108.

[121] M. F. Glasser and D. C. Van Essen. Mapping human cortical areas in vivo based on myelin content as revealed by t1- and t2-weighted MRI. *Journal of Neuroscience*, 31 (32):11597–11616, August 2011. Cited on page 88.

[122] Matthew F. Glasser, Stamatios N. Sotiropoulos, J. Anthony Wilson, Timothy S. Coalson, Bruce Fischl, Jesper L. Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R. Polimeni, David C. Van Essen, and Mark Jenkinson. The minimal preprocessing pipelines for the human connectome project. *NeuroImage*, 80: 105–124, October 2013. Cited on pages 44, 71, 75, and 88.

[123] Gadi Goelman, Rotem Dan, and Tarek Keadan. Characterizing directed functional pathways in the visual system by multivariate nonlinear coherence of fMRI data. *Scientific Reports*, 8(1), November 2018. Cited on page 65.

[124] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016. Cited on pages 21, 32, and 33.

[125] Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17 (6):333–351, May 2016. Cited on page 37.

[126] Krzysztof J. Gorgolewski and Russell A. Poldrack. A practical guide for improving transparency and reproducibility in neuroimaging research. *PLOS Biology*, 14(7): e1002506, July 2016. Cited on pages 42 and 84.

[127] M. Gori, G. Monfardini, and F. Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* IEEE, 2005. Cited on page 35.

[128] Assaf Gottlieb, Roxana Daneshjou, Marianne DeGorter, Stephane Bourgeois, Peter J Svensson, Mia Wadelius, Panos Deloukas, Stephen B Montgomery, and Russ B Altman. Cohort-specific imputation of gene expression improves prediction of warfarin dose for african americans. *Genome Medicine*, 9(1):98, 2017. Cited on page 37.

[129] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):1–42, September 2019. Cited on page 33.

[130] Arda Halu, Manlio De Domenico, Alex Arenas, and Amitabh Sharma. The multiplex network of human diseases. *npj Systems Biology and Applications*, 5(1), April 2019. Cited on pages 38 and 47.

[131] Jing-Dong J. Han, Nicolas Bertin, Tong Hao, Debra S. Goldberg, Gabriel F. Berriz, Lan V. Zhang, Denis Dupuy, Albertha J. M. Walhout, Michael E. Cusick, Frederick P. Roth, and Marc Vidal. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430(6995):88–93, June 2004. Cited on page 38.

[132] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome Biology*, 18(1), May 2017. Cited on pages 36 and 37.

[133] Claire E. Hastie, Jill P. Pell, and Naveed Sattar. Vitamin d and COVID-19 infection and mortality in UK Biobank. *European Journal of Nutrition*, August 2020. Cited on page 110.

[134] Sabine Heiland. From a as in aliasing to z as in zipper: Artifacts in MRI. *Clinical Neuroradiology*, 18(1):25–36, March 2008. Cited on page 42.

[135] Anibal Sólon Heinsfeld, Alexandre Rosa Franco, R. Cameron Craddock, Augusto Buchweitz, and Felipe Meneguzzi. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clinical*, 17:16–23, 2018. Cited on page 66.

[136] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214:106685, February 2021. Cited on page 90.

[137] Claus C. Hilgetag and Alexandros Goulas. ‘Hierarchy’ in the organization of brain networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1796):20190319, February 2020. Cited on page 70.

[138] Barbara Hollunder, Nanditha Rajamani, Shan H. Siddiqi, Carsten Finke, Andrea A. Kühn, Helen S. Mayberg, Michael D. Fox, Clemens Neudorfer, and Andreas Horn. Toward personalized medicine in connectomic deep brain stimulation. *arXiv preprint arXiv:2109.12327*, 2021. Cited on page 41.

[139] Katherine Hollywood, Daniel R. Brison, and Royston Goodacre. Metabolomics: Current technologies and future trends. *PROTEOMICS*, 6(17):4716–4723, September 2006. Cited on page 38.

[140] Elizabeth A. Holm. In defense of the black box. *Science*, 364(6435):26–27, April 2019. Cited on page 14.

[141] Desislava Hristova, Alex Rutherford, Jose Anson, Miguel Luengo-Oroz, and Cecilia Mascolo. The international postal network and other global flows as proxies for national wellbeing. *PloS One*, 11(6), 2016. Cited on page 53.

[142] Yongli Hu, Takeshi Hase, Hui Peng Li, Shyam Prabhakar, Hiroaki Kitano, See Kiong Ng, Samik Ghosh, and Lawrence Jin Kiat Wee. A machine learning approach for the identification of key markers involved in brain development from single-cell transcriptomic data. *BMC Genomics*, 17(13):1025, 2016. Cited on page 37.

[143] R. Matthew Hutchison, Thilo Womelsdorf, Elena A. Allen, Peter A. Bandettini, Vince D. Calhoun, Maurizio Corbetta, Stefania Della Penna, Jeff H. Duyn, Gary H. Glover, Javier Gonzalez-Castillo, Daniel A. Handwerker, Shella Keilholz, Vesa Kiviniemi, David A. Leopold, Francesco de Pasquale, Olaf Sporns, Martin Walter, and Catie Chang. Dynamic functional connectivity: Promise, issues, and interpretations. *NeuroImage*, 80:360–378, October 2013. Cited on page 67.

[144] Matthew Hutson. Eye-catching advances in some AI fields are not real. *Science*, May 2020. Cited on page 19.

[145] Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *International Conference on Learning and Intelligent Optimization*, pages 507–523. Springer, 2011. Cited on page 27.

[146] Minh Bao Huynh, Mohand Ouidir Ouidja, Sandrine Chantepie, Gilles Carpentier, Auriane Maïza, Ganlin Zhang, Joao Vilares, Rita Raisman-Vozari, and Dulce Papy-Garcia. Glycosaminoglycans from Alzheimer's disease hippocampus have altered capacities to bind and regulate growth factors activities and to bind tau. *PloS One*, 14(1), 2019. Cited on page 60.

[147] Johannes Jaeger and Nick Monk. Dynamical modules in metabolism, cell and developmental biology. *Interface Focus*, 11(3), April 2021. Cited on page 38.

[148] Samie R Jaffrey and Miles F Wilkinson. Nonsense-mediated RNA decay in the brain: emerging modulator of neural development and disease. *Nature Reviews Neuroscience*, 19(12):715–728, 2018. Cited on page 61.

[149] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, October 2000. Cited on page 38.

[150] Ben Jeurissen, Jacques-Donald Tournier, Thijs Dhollander, Alan Connelly, and Jan Sijbers. Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion MRI data. *NeuroImage*, 103:411–426, December 2014. Cited on page 75.

[151] Rongtao Jiang, Vince D Calhoun, Lingzhong Fan, Nianming Zuo, Rex Jung, Shile Qi, Dongdong Lin, Jin Li, Chuanjun Zhuo, Ming Song, Zening Fu, Tianzi Jiang, and Jing Sui. Gender differences in connectome-based predictions of individualized intelligence quotient and sub-domain scores. *Cerebral Cortex*, July 2019. Cited on page 73.

[152] Biao Jie, Mingxia Liu, Chunfeng Lian, Feng Shi, and Dinggang Shen. Designing weighted correlation kernels in convolutional neural networks for functional connectivity based brain disease diagnosis. *Medical Image Analysis*, 63:101709, July 2020. Cited on page 86.

[153] M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, July 2015. Cited on page 14.

[154] Henry F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, September 1958. Cited on page 92.

[155] Henry F. Kaiser and John Rice. Little jiffy, mark iv. *Educational and Psychological Measurement*, 34(1):111–117, April 1974. Cited on page 93.

[156] Lukasz Kaiser, Aidan N. Gomez, and Francois Chollet. Depthwise separable convolutions for neural machine translation. In *International Conference on Learning Representations*, 2018. Cited on page 68.

[157] Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aäron van den Oord, Alexander Graves, and Koray Kavukcuoglu. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*, 2016. Cited on page 68.

[158] R. F. Kaplan, M.-E. Meadows, M. Verfaellie, E. Kwan, B. L. Ehrenberg, E. B. Bromfield, and R. A. Cohen. Lateralization of memory for the visual attributes of objects: Evidence from the posterior cerebral artery amobarbital test. *Neurology*, 44 (6):1069–1069, June 1994. Cited on page 99.

[159] Philipp Kapranov, Aarron T. Willingham, and Thomas R. Gingeras. Genome-wide transcription and the implications for genomic organization. *Nature Reviews Genetics*, 8(6):413–423, May 2007. Cited on page 37.

[160] Anees Kazi, Luca Cosmo, Nassir Navab, and Michael Bronstein. Differentiable graph module (DGM) for graph convolutional networks. *arXiv preprint arXiv:2002.04999*, 2020. Cited on page 86.

[161] Stefan J. Kiebel, Jean Daunizeau, and Karl J. Friston. A hierarchy of time-scales and the brain. *PLoS Computational Biology*, 4(11):e1000209, November 2008. Cited on page 70.

[162] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. Cited on page 33.

[163] Byung-Hoon Kim and Jong Chul Ye. Understanding graph isomorphism network for rs-fMRI functional connectivity analysis. *Frontiers in Neuroscience*, 14, June 2020. Cited on pages 66 and 67.

[164] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 2014. Cited on pages 38 and 52.

[165] Cassie Kozyrkov. Explainable AI won't deliver. Here's why. *Hackernoon, November*, 16, 2018. Cited on page 34.

[166] Sofia Ira Ktena, Sarah Parisot, Enzo Ferrante, Martin Rajchl, Matthew Lee, Ben Glocker, and Daniel Rueckert. Metric learning with spectral graph convolutions on brain connectivity networks. 169:431–442, April 2018. Cited on page 67.

[167] Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, Michael G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Ma'ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97, May 2016. Cited on page 52.

[168] Vadim Kuperman. *Magnetic resonance imaging: physical principles and applications*. Elsevier, 2000. Cited on page 39.

[169] Sebastian Lapuschkin, Alexander Binder, Gregoire Montavon, Klaus-Robert Muller, and Wojciech Samek. Analyzing classifiers: Fisher vectors and deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016. Cited on page 34.

[170] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1), March 2019. Cited on page 34.

[171] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007. Cited on page 49.

[172] Xiaoxiao Li, Yuan Zhou, Nicha C. Dvornek, Muhan Zhang, Juntang Zhuang, Pamela Ventola, and James S Duncan. Pooling regularized graph neural network for fMRI biomarker analysis. *arXiv preprint arXiv:2007.14589*, 2020. Cited on pages 67 and 86.

[173] Xiaoxiao Li, Yuan Zhou, Siyuan Gao, Nicha Dvornek, Muhan Zhang, Juntang Zhuang, Shi Gu, Dustin Scheinost, Lawrence Staib, Pamela Ventola, and James Duncan. BrainGNN: Interpretable brain graph neural network for fMRI analysis. May 2020. Cited on pages 67 and 86.

[174] Xuhong Liao, Miao Cao, Mingrui Xia, and Yong He. Individual differences and time-varying features of modular brain architecture. *NeuroImage*, 152:94–107, May 2017. Cited on page 65.

[175] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1):18, December 2020. Cited on page 90.

[176] Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001. Cited on page 89.

[177] Zachary C. Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016. Cited on page 33.

[178] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. Cited on pages 87 and 89.

[179] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018. Cited on pages 87 and 90.

[180] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. Cited on page 50.

[181] Lloyd Mabonga and Abidemi Paul Kappo. Protein-protein interaction modulators: advances, successes and remaining challenges. *Biophysical Reviews*, 11(4):559–581, July 2019. Cited on page 38.

[182] Brenda Maddox. The double helix and the 'wronged heroine'. *Nature*, 421(6921): 407–408, January 2003. Cited on page 36.

[183] S. Mahapatra, R. Bhuyan, J. Das, and T. Swarnkar. Integrated multiplex network based approach for hub gene identification in oral cancer. *Heliyon*, 7(7):e07418, July 2021. Cited on page 48.

[184] Carlo Maj, Tiago Azevedo, Valentina Giansanti, Oleg Borisov, Giovanna Maria Dimitri, Simeon Spasov, Pietro Lió, and Ivan Merelli. Integration of machine learning methods to dissect genetically imputed transcriptomic profiles in Alzheimer's disease. *Frontiers in Genetics*, 10:726, September 2019. Cited on page 20.

[185] M Majewski, A Kozlowska, M Thoene, E Lepiarczyk, and WJ Grzegorzewski. Overview of the role of vitamins and minerals on the kynurenine pathway in health and disease. *Journal of Physiology and Pharmacology*, 67(1):3–19, 2016. Cited on page 61.

[186] Scott Marek, Brenden Tervo-Clemmens, Finnegan J. Calabro, David F. Montez, Benjamin P. Kay, Alexander S. Hatoum, Meghan Rose Donohue, William Foran, Ryland L. Miller, Eric Feczko, Oscar Miranda-Dominguez, Alice M. Graham, Eric A.

Earl, Anders J. Perrone, Michaela Cordova, Olivia Doyle, Lucille A. Moore, Greg Conan, Johnny Uriarte, Kathy Snider, Angela Tam, Jianzhong Chen, Dillan J. Newbold, Annie Zheng, Nicole A. Seider, Andrew N. Van, Timothy O. Laumann, Wesley K. Thompson, Deanna J. Greene, Steven E. Petersen, Thomas E. Nichols, B.T. Thomas Yeo, Deanna M. Barch, Hugh Garavan, Beatriz Luna, Damien A. Fair, and Nico U.F. Dosenbach. Towards reproducible brain-wide association studies. August 2020. Cited on pages 44 and 66.

[187] RU Margolis, RK Margolis, LB Chang, and C Preti. Glycosaminoglycans of brain during development. *Biochemistry*, 14(1):85–88, 1975. Cited on page 60.

[188] Alexander März. Xgboostlss – an extension of xgboost to probabilistic forecasting. *arXiv preprint arXiv:1907.03178*, 2019. Cited on page 110.

[189] Liam G. McCoy, Connor T.A. Brenna, Stacy S. Chen, Karina Vold, and Sunit Das. Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. *Journal of Clinical Epidemiology*, November 2021. Cited on page 34.

[190] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. UMAP: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018. Cited on page 50.

[191] Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper L R Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Thomas W Okell, Peter Weale, Iulius Dragonu, Steve Garratt, Sarah Hudson, Rory Collins, Mark Jenkinson, Paul M Matthews, and Stephen M Smith. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, 19(11):1523–1536, September 2016. Cited on page 44.

[192] TM Mitchell. Machine learning, mcgraw-hill higher education. *New York*, 1997. Cited on page 21.

[193] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018. Cited on page 32.

[194] Christoph Molnar. Interpretable machine learning: A guide for making black box models explainable, 2021. URL `https://christophm.github.io/interpretable-ml-book/`. Cited on page 33.

[195] Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4602–4609, 2019. Cited on page 73.

[196] Laurel S. Morris, Prantik Kundu, Sara Costi, Abigail Collins, Molly Schneider, Gaurav Verma, Priti Balchandani, and James W. Murrough. Ultra-high field MRI reveals mood-related circuit disturbances in depression: a comparison between 3-tesla and 7-tesla. *Translational Psychiatry*, 9(1), February 2019. Cited on page 109.

[197] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, October 2019. Cited on page 33.

[198] Monika A. Myszczynska, Poojitha N. Ojamies, Alix M. B. Lacoste, Daniel Neil, Amir Saffari, Richard Mead, Guillaume M. Hautbergue, Joanna D. Holbrook, and Laura Ferraiuolo. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature Reviews Neurology*, 16(8):440–456, July 2020. Cited on page 108.

[199] Steven H.J. Nagtegaal, Szabolcs David, Arthur T.J. van der Boog, Alexander Leemans, and Joost J.C. Verhoeff. Changes in cortical thickness and volume after cranial radiation treatment: A systematic review. *Radiotherapy and Oncology*, 135: 33–42, June 2019. Cited on page 43.

[200] Eric J. Nestler. Transgenerational epigenetic contributions to stress responses: Fact or fiction? *PLOS Biology*, 14(3):e1002426, March 2016. Cited on page 37.

[201] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004. Cited on page 49.

[202] Rudolf Nieuwenhuys, Jan Voogd, and Christiaan van Huijzen. *The Human Central Nervous System*. Springer Berlin Heidelberg, 2008. Cited on page 80.

[203] Alessandra D. Nostro, Veronika I. Müller, Deepthi P. Varikuti, Rachel N. Pläschke, Felix Hoffstaedter, Robert Langner, Kaustubh R. Patil, and Simon B. Eickhoff. Predicting personality from network-based resting-state functional connectivity. *Brain Structure and Function*, 223(6):2699–2719, March 2018. Cited on page 87.

[204] KL Novik, I Nimmrich, B Genc, S Maier, C Piepenbrock, A Olek, and S Beck. Epigenomics: genome-wide study of methylation phenomena. *Current Issues in Molecular Biology*, 4(4):111–128, 2002. Cited on page 37.

[205] Anna Nowogrodzki. The world's strongest MRI machines are pushing human imaging to new limits. *Nature*, 563(7729):24–26, October 2018. Cited on page 109.

[206] Markus Ojala and Gemma C Garriga. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11(6), 2010. Cited on page 96.

[207] Shane O'Sullivan, Fleur Jeanquartier, Claire Jean-Quartier, Andreas Holzinger, Dan Shiebler, Pradip Moon, and Claudio Angione. Developments in AI and machine learning for neuroimaging. In *Artificial Intelligence and Machine Learning for Digital Pathology*, pages 307–320. Springer International Publishing, 2020. Cited on page 34.

[208] Art B. Owen and Clémentine Prieur. On shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):986–1002, January 2017. Cited on page 89.

[209] Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrerro Moreno, Ben Glocker, and Daniel Rueckert. Spectral Graph Convolutions for

Population-Based Disease Prediction. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, Lecture Notes in Computer Science, pages 177–185, Cham, 2017. Springer International Publishing. ISBN 978-3-319-66179-7. Cited on page 67.

[210] Harshit Parmar, Brian Nutter, Rodney Long, Sameer Antani, and Sunanda Mitra. Spatiotemporal feature extraction and classification of Alzheimer's disease using deep learning 3D-CNN for fMRI data. *Journal of Medical Imaging*, 7(5):056001, October 2020. Cited on page 67.

[211] Princy Parsana, Claire Ruberman, Andrew E. Jaffe, Michael C. Schatz, Alexis Battle, and Jeffrey T. Leek. Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome Biology*, 20(1), May 2019. Cited on page 49.

[212] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. Cited on page 72.

[213] THDJ Patil and T Davenport. Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, 90(10):70–76, 2012. Cited on page 19.

[214] Scott D. Patterson and Ruedi H. Aebersold. Proteomics: the first decade and beyond. *Nature Genetics*, 33(S3):311–323, March 2003. Cited on page 37.

[215] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. Cited on pages 71 and 102.

[216] David Poeppel, G. R. Mangun, and Michael S. Gazzaniga, editors. *The cognitive neurosciences*. The MIT Press, Cambridge, MA, sixth edition edition, 2020. ISBN 9780262043250. Cited on page 80.

[217] Russell A Poldrack, Chris I Baker, Joke Durnez, Krzysztof J Gorgolewski, Paul M Matthews, Marcus R Munafò, Thomas E Nichols, Jean-Baptiste Poline, Edward Vul, and Tal Yarkoni. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2):115, 2017. Cited on page 84.

[218] Ehsan Pournoor, Zaynab Mousavian, Abbas Nowzari Dalini, and Ali Masoudi-Nejad. Identification of key components in colon adenocarcinoma using transcriptome to interactome multilayer framework. *Scientific Reports*, 10(1), March 2020. Cited on pages 38 and 47.

[219] Maria Giulia Preti, Thomas AW Bolton, and Dimitri Van De Ville. The dynamic functional connectome: State-of-the-art and perspectives. *NeuroImage*, 160:41–54, October 2017. Cited on page 67.

[220] Anqi Qiu, Han Zhang, Brian K. Kennedy, and Annie Lee. Spatio-temporal correlates of gene expression and cortical morphology across lifespan and aging. *NeuroImage*, 224:117426, January 2021. Cited on page 109.

[221] Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2847–2854. PMLR, 06–11 Aug 2017. Cited on page 32.

[222] J. Rajendhran and P. Gunasekaran. Human microbiomics. *Indian Journal of Microbiology*, 50(1):109–112, March 2010. Cited on page 38.

[223] Meenakshi Rao and Michael D Gershon. The bowel and beyond: the enteric nervous system in neurological disorders. *Nature Reviews Gastroenterology & Hepatology*, 13 (9):517, 2016. Cited on page 60.

[224] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018. Cited on pages 22, 23, 24, and 25.

[225] Atif Riaz, Muhammad Asad, Eduardo Alonso, and Greg Slabaugh. DeepFMRI: End-to-end deep learning for functional connectivity and classification of ADHD using fMRI. *Journal of Neuroscience Methods*, 335:108506, April 2020. Cited on page 67.

[226] Roberta Riccelli, Nicola Toschi, Salvatore Nigro, Antonio Terracciano, and Luca Passamonti. Surface-based morphometry reveals the neuroanatomical basis of the five-factor model of personality. *Social Cognitive and Affective Neuroscience*, page nsw175, January 2017. Cited on page 87.

[227] David Robinson. The incredible growth of python. *Stack Overflow*, 6, September 2017. Cited on page 19.

[228] GB Rogers, DJ Keating, RL Young, ML Wong, Julio Licinio, and S Wesselingh. From gut dysbiosis to altered brain function and mental illness: mechanisms and pathways. *Molecular Psychiatry*, 21(6):738–748, 2016. Cited on page 60.

[229] Daniel M Rosenbaum, Søren GF Rasmussen, and Brian K Kobilka. The structure and function of g-protein-coupled receptors. *Nature*, 459(7245):356–363, 2009. Cited on page 59.

[230] Mikail Rubinov and Olaf Sporns. Weight-conserving characterization of complex functional brain networks. *NeuroImage*, 56(4):2068–2079, June 2011. Cited on page 49.

[231] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215, May 2019. Cited on page 34.

[232] Ando Saabas. Interpreting random forests, 2014. URL `http://blog.datadive.net/interpreting-random-forests/`. Accessed: 2021/04/19. Cited on page 90.

[233] Ashis Saha, Yungil Kim, Ariel D.H. Gewirtz, Brian Jo, Chuan Gao, Ian C. McDowell, Barbara E. Engelhardt, and Alexis Battle. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Research*, 27(11):1843–1858, October 2017. Cited on page 47.

[234] Gholamreza Salimi-Khorshidi, Gwenaëlle Douaud, Christian F. Beckmann, Matthew F. Glasser, Ludovica Griffanti, and Stephen M. Smith. Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage*, 90:449–468, April 2014. Cited on page 70.

[235] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019. Cited on page 66.

[236] Marco Sandri and Paola Zuccolotto. A bias correction algorithm for the gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, 17(3):611–628, September 2008. Cited on page 90.

[237] Emine U. Saritas, Samantha J. Holdsworth, and Roland Bammer. Susceptibility artifacts. In *Quantitative MRI of the Spinal Cord*, pages 91–105. Elsevier, 2014. Cited on page 42.

[238] Michael Schirner, Simon Rothmeier, Viktor K. Jirsa, Anthony Randal McIntosh, and Petra Ritter. An automated pipeline for constructing personalized virtual brains from multimodal neuroimaging data. *NeuroImage*, 117:343–357, August 2015. Cited on page 41.

[239] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green AI. *arXiv preprint arXiv:1907.10597*, 2019. Cited on page 109.

[240] Younjoo Seo, Andreas Loukas, and Nathanaël Perraudin. Discriminative structural graph classification. *arXiv preprint arXiv:1905.13422*, 2019. Cited on page 35.

[241] James A. Shapiro. Revisiting the central dogma in the 21st century. *Annals of the New York Academy of Sciences*, 1178(1):6–28, October 2009. Cited on page 36.

[242] L. S. Shapley. 17. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28), Volume II*, pages 307–318. Princeton University Press, December 1953. Cited on page 89.

[243] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. In *8th ICML Workshop on Automated Machine Learning (AutoML)*, 2021. Cited on page 87.

[244] Andrew JG Simpson, Otavia L Caballero, Achim Jungbluth, Yao-Tseng Chen, and Lloyd J Old. Cancer/testis antigens, gametogenesis and cancer. *Nature Reviews Cancer*, 5(8):615–625, 2005. Cited on page 56.

[245] Robert E. Smith, Jacques-Donald Tournier, Fernando Calamante, and Alan Connelly. SIFT: Spherical-deconvolution informed filtering of tractograms. *NeuroImage*, 67: 298–312, February 2013. Cited on page 75.

[246] Stephen M. Smith and Thomas E. Nichols. Statistical challenges in "big data" human neuroimaging. *Neuron*, 97(2):263–268, January 2018. Cited on page 65.

[247] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. Cited on page 27.

[248] Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Mr Prabhat, and Ryan Adams. Scalable bayesian optimization using deep neural networks. In *International Conference on Machine Learning*, pages 2171–2180. PMLR, 2015. Cited on page 27.

[249] Simeon Spasov, Luca Passamonti, Andrea Duggento, Pietro Liò, and Nicola Toschi. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer's disease. *NeuroImage*, 189:276–287, April 2019. Cited on page 65.

[250] A. Sperduti and A. Starita. Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks*, 8(3):714–735, May 1997. Cited on page 35.

[251] Alessandro Sperduti. Encoding labeled graphs by labeling raam. In J. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann, 1994. Cited on page 35.

[252] Olaf Sporns. The human connectome: a complex network. *Annals of the New York Academy of Sciences*, 1224(1):109–125, January 2011. Cited on page 67.

[253] Emmanuel Stamatakis, Katherine B. Owen, Leah Shepherd, Bradley Drayton, Mark Hamer, and Adrian E. Bauman. Is cohort representativeness passé? poststratified associations of lifestyle risk factors with mortality in the UK Biobank. *Epidemiology*, 32(2):179–188, January 2021. Cited on page 43.

[254] Kamilė Stankevičiūtė, Tiago Azevedo, Alexander Campbell, Richard Bethlehem, and Pietro Liò. Population graph GNNs for brain age prediction. In *ICML Workshop on Graph Representation Learning and Beyond (GRL+)*, June 2020. Cited on pages 17, 83, and 85.

[255] Eva-Maria Stauffer, Richard A. I. Bethlehem, Varun Warrier, Graham K. Murray, Rafael Romero-Garcia, Jakob Seidlitz, and Edward T. Bullmore. Grey and white matter microstructure is associated with polygenic risk for schizophrenia. *Molecular Psychiatry*, August 2021. Cited on page 109.

[256] Leon Stefanovski, Amna Ghani, Anthony Randal McIntosh, and Petra Ritter. Linking connectomics and dynamics in the human brain. *e-Neuroforum*, 7(3):64–70, September 2016. Cited on page 41.

[257] Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.e13, February 2020. Cited on page 35.

[258] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3): 647–665, August 2013. Cited on page 89.

[259] Shyam Anil Tailor, Javier Fernandez-Marques, and Nicholas Donald Lane. Degree-quant: Quantization-aware training for graph neural networks. In *International Conference on Learning Representations*, 2021. Cited on page 109.

[260] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484, May 2019. Cited on page 37.

[261] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33, 2020. Cited on page 14.

[262] Tijmen Tieleman, Geoffrey Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012. Cited on page 73.

[263] Christian Tinauer, Stefan Heber, Lukas Pirpamer, Anna Damulina, Reinhold Schmidt, Rudolf Stollberger, Stefan Ropele, and Christian Langkammer. Interpretable brain disease classification and relevance-guided deep learning. *medRxiv preprint*, September 2021. Cited on page 34.

[264] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813, November 2021. Cited on page 34.

[265] Rens van de Schoot, Sarah Depaoli, Ruth King, Bianca Kramer, Kaspar Märtens, Mahlet G. Tadesse, Marina Vannucci, Andrew Gelman, Duco Veen, Joukje Willemsen, and Christopher Yau. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1), January 2021. Cited on page 110.

[266] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016. Cited on page 68.

[267] Bas H.M. van der Velden, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, and Max A. Viergever. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, July 2022. Cited on page 34.

[268] Kush R. Varshney and Homa Alemzadeh. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *arXiv preprint arXiv:1610.01256*, 2016. Cited on page 34.

[269] Sagar Vaze, Weidi Xie, and Ana I. L. Namburete. Low-memory CNNs enabling real-time ultrasound segmentation towards mobile deployment. *IEEE Journal of Biomedical and Health Informatics*, 24(4):1059–1069, April 2020. Cited on page 109.

[270] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *International Conference on Learning Representations*, 2018. Cited on page 83.

[271] Sandra Vieira, Walter H.L. Pinaya, and Andrea Mechelli. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, 74:58–75, March 2017. Cited on page 87.

[272] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, December 2021. Cited on page 90.

[273] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, 11(1), January 2020. Cited on page 110.

[274] Marinus T Vlaardingerbroek and Jacques A Boer. *Magnetic resonance imaging: theory and practice*. Springer Science & Business Media, 2013. Cited on page 39.

[275] Cheng Wang, Yayun Gu, Kai Zhang, Kaipeng Xie, Meng Zhu, Ningbin Dai, Yue Jiang, Xuejiang Guo, Mingxi Liu, Juncheng Dai, et al. Systematic identification of genes with a cancer-testis expression pattern in 19 cancer types. *Nature Communications*, 7(1):1–12, 2016. Cited on page 56.

[276] Huifang E. Wang, Christian G. Bénar, Pascale P. Quilichini, Karl J. Friston, Viktor K. Jirsa, and Christophe Bernard. A systematic framework for functional connectivity measures. *Frontiers in Neuroscience*, 8, 2014. Cited on page 66.

[277] Lebo Wang, Kaiming Li, Xu Chen, and Xiaoping P. Hu. Application of Convolutional Recurrent Neural Network for Individual Recognition Based on Resting State fMRI Data. *Frontiers in Neuroscience*, 13, 2019. Cited on page 65.

[278] Lebo Wang, Kaiming Li, and Xiaoping P. Hu. Graph convolutional network for fMRI analysis based on connectivity neighborhood. *Network Neuroscience*, 5(1): 83–95, February 2021. Cited on pages 67, 74, 76, and 84.

[279] Likai Wang, Yanpeng Xi, Sibum Sung, and Hong Qiao. Rna-seq assistant: machine learning based methods to identify more transcriptional regulated genes. *BMC Genomics*, 19(1):546, 2018. Cited on page 37.

[280] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. Cited on page 80.

[281] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-SNE effectively. *Distill*, 2016. Cited on page 50.

[282] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, September 2013. Cited on page 51.

[283] Susanne Weis, Kaustubh R Patil, Felix Hoffstaedter, Alessandra Nostro, B T Thomas Yeo, and Simon B Eickhoff. Sex classification by resting state brain connectivity. *Cerebral Cortex*, June 2019. Cited on page 73.

[284] Dong Wen, Zhenhao Wei, Yanhong Zhou, Guolin Li, Xu Zhang, and Wei Han. Deep learning methods to process fMRI data and their application in the diagnosis of cognitive impairment: A brief overview and our opinion. *Frontiers in Neuroinformatics*, 12, April 2018. Cited on page 67.

[285] Markus R. Wenk. The emerging field of lipidomics. *Nature Reviews Drug Discovery*, 4(7):594–610, July 2005. Cited on page 38.

[286] Matthias Wilms, Pauline Mouches, Jordan J. Bannister, Deepthi Rajashekar, Sönke Langner, and Nils D. Forkert. Towards self-explainable classifiers and regressors in neuroimaging with normalizing flows. In *Lecture Notes in Computer Science*, pages 23–33. Springer International Publishing, 2021. Cited on page 34.

[287] Jie Wu and Yiqiang Zhao. Machine learning technology in the application of genome analysis: A systematic review. *Gene*, 705:149–156, July 2019. Cited on page 15.

[288] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596*, 2019. Cited on pages 35 and 66.

[289] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. Cited on page 35.

[290] Baoyu Yan, Xiaopan Xu, Mengwan Liu, Kaizhong Zheng, Jian Liu, Jianming Li, Lei Wei, Binjie Zhang, Hongbing Lu, and Baojuan Li. Quantitative Identification of Major Depression Based on Resting-State Dynamic Functional Connectivity: A Machine Learning Approach. *Frontiers in Neuroscience*, 14, 2020. Cited on page 67.

[291] Bencheng Yan, Chaokun Wang, Gaoyang Guo, and Yunkai Lou. TinyGNN: Learning efficient graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, July 2020. Cited on page 35.

[292] Dengcheng Yang, Yi Jin, Xiaoqing He, Ang Dong, Jing Wang, and Rongling Wu. Inferring multilayer interactome networks shaping phenotypic plasticity and evolution. *Nature Communications*, 12(1), September 2021. Cited on pages 38 and 47.

[293] Melvyn Yap, Rebecca L. Johnston, Helena Foley, Samual MacDonald, Olga Kondrashova, Khoa A. Tran, Katia Nones, Lambros T. Koufariotis, Cameron Bean, John V. Pearson, Maciej Trzaskowski, and Nicola Waddell. Verifying explainability of a deep learning tissue classifier trained on RNA-seq data. *Scientific Reports*, 11 (1), January 2021. Cited on page 34.

[294] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *Advances in neural information processing systems*, pages 4800–4810, 2018. Cited on pages 70 and 73.

[295] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating explanations for graph neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. Cited on page 34.

[296] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. Cited on page 68.

[297] Yang Yu, Pathum Kossinna, Qing Li, Wenyuan Liao, and Qingrun Zhang. Explainable autoencoder-based representation learning for gene expression data. *bioRxiv preprint*, December 2021. Cited on page 34.

[298] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey. *arXiv preprint arXiv:2012.15445*, 2020. Cited on page 34.

[299] Liang Zhan, Lisanne M. Jenkins, Ouri E. Wolfson, Johnson Jonaris GadElkarim, Kevin Nocito, Paul M. Thompson, Olusola A. Ajilore, Moo K. Chung, and Alex D. Leow. The significance of negative correlations in brain connectivity. *Journal of Comparative Neurology*, 525(15):3251–3265, July 2017. Cited on page 76.

[300] Chengxin Zhang, Wei Zheng, Micah Cheng, Gilbert S. Omenn, Peter L. Freddolino, and Yang Zhang. Functions of essential genes and a scale-free protein interaction network revealed by structure-based function and interaction prediction for a minimal genome. *Journal of Proteome Research*, 20(2):1178–1189, January 2021. Cited on page 38.

[301] Ji Zhang, Meige Guan, Qianliang Wang, Jiajun Zhang, Tianshou Zhou, and Xiaoqiang Sun. Single-cell transcriptome-based multilayer network biomarker for predicting prognosis and therapeutic response of gliomas. *Briefings in Bioinformatics*, 21(3):1080–1097, April 2019. Cited on page 48.

[302] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018. Cited on pages 35 and 66.

[303] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):457–466, 2018. Cited on page 38.