Latent navigation for building better predictive models for neurodevelopment research

Citation: Carvalho de Paula Ferreira da Costa, Pedro Henrique (2022) Latent navigation for building better predictive models for neurodevelopment research. [Thesis] (Unpublished)

Birkbeck, University of London

Doctoral Thesis

Latent navigation for building better predictive models for neurodevelopment research

Author:
Pedro Henrique Carvalho de Paula Ferreira da Costa

Supervisors:
Prof. Emily J. H. Jones
Prof. Robert Leech

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy at

Birkbeck College, University of London
King's College London, University of London

September, 2022

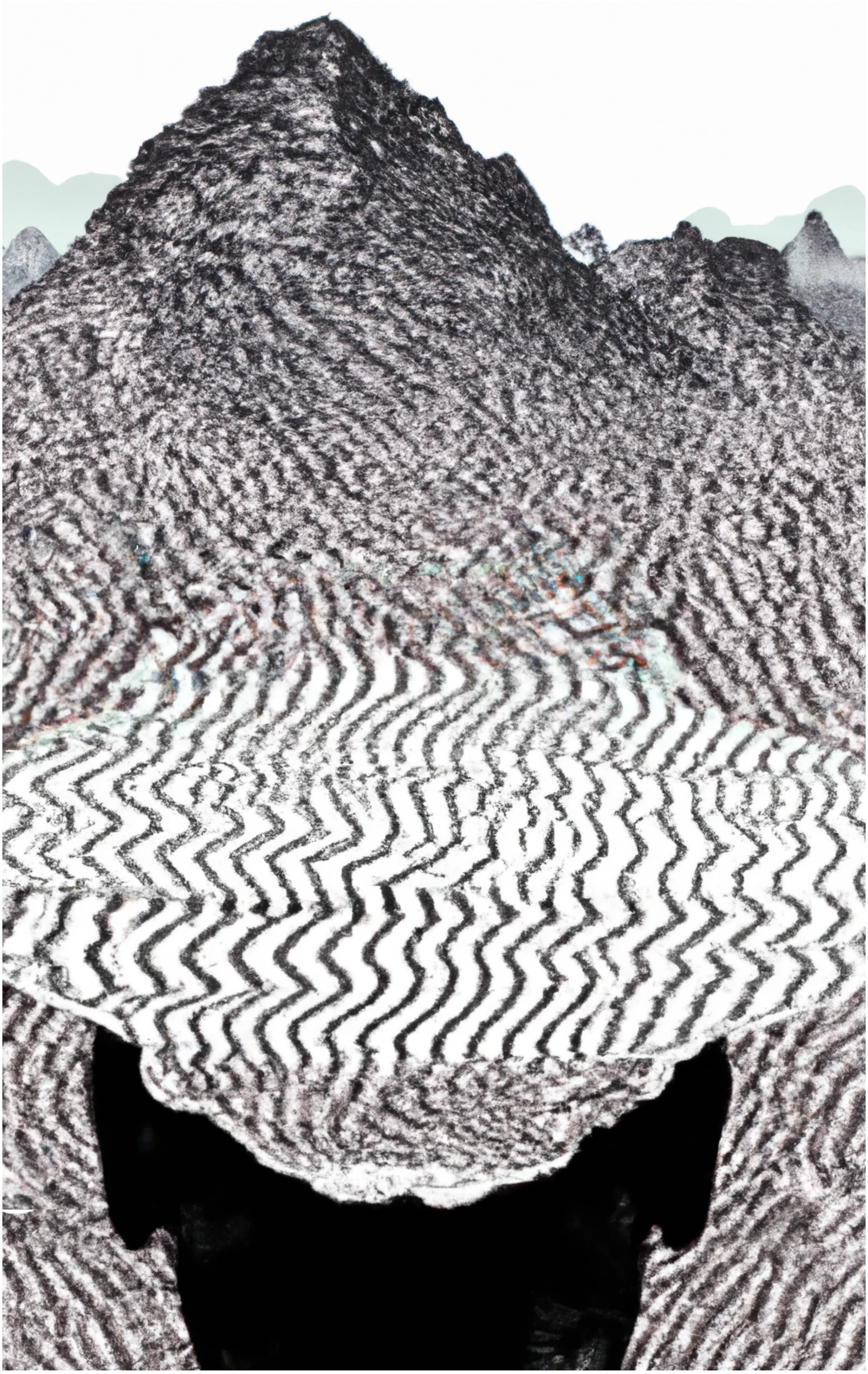

Figure 1: Peaks of individuality. *Artificially generated image using DALL-E2(Ramesh et al., 2022) and a human-in-the-loop approach.*

Para o meu Pai

Originality Statement

I, Pedro Ferreira da Costa, hereby declare that, except where explicit attribution is made, the work presented in this thesis is my own.

The following contributions to the work presented in this thesis are acknowledged:

In Chapter 4:
Elena Throm collected infant data.
Rianne Haartsen contributed with EEG processing scripts.

In Chapter 5:
Prof. Robert Leech and Jessica Dafflon contributed to the concept, scripting of the code and write-up.

This work was funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Actions: Innovative Training Network [SAPIENS, grant number 814302].

Portions of this Thesis have been adapted from published work and preprints being presently reviewed for publication. Chapter 3 was peer-reviewed and presented in the autoML workshop at ICML2020 (da Costa et al., 2020). Chapter 4 is currently available as a pre-print (da Costa et al., 2021). Chapter 5 is adapted from the journal publication (Dafflon et al., 2022). Chapter 7 has been peer-reviewed and presented in the PAI4MH workshop at NeurIPS 2022. (da Costa et al., 2022).

- da Costa, P. F., Lorenz, R., Monti, R. P., Jones, E., and Leech, R. (2020). Bayesian Optimization for real-time, automatic design of face stimuli in human-centred research

- da Costa, P. F., Haartsen, R., Throm, E., Mason, L., Gui, A., Leech, R., and Jones, E. J. H. (2021). Neuroadaptive electroencephalography: a proof-of-principle study in infants

- Dafflon, J., F. Da Costa, P., Váša, F., Monti, R. P., Bzdok, D., Hellyer, P. J., Turkheimer, F., Smallwood, J., Jones, E., and Leech, R. (2022). A guided multiverse study of neuroimaging analyses. *Nature Communications*, 13(1):3758

- da Costa, P. F., Dafflon, J., Mendes, S. L., Sato, J. R., Cardoso, M. J., Leech, R., Jones, E., and Pinaya, W. H. L. (2022). Transformer-based normative modelling for anomaly detection of early schizophrenia. In *Empowering Communities: A Participatory Approach to AI for Mental Health*

Abstract

In recent decades, replication efforts in research have found that many findings are not reproducible. Many of these studies serve as the basis for others that might be relying on false assumptions. This replication crisis stands out in neurodevelopment research where heterogeneity in the typical human brain cannot, in most cases, be probed directly and relies on proxy measures of brain activity. This thesis develops three methodological frameworks for more robust research paradigms. I employ machine learning algorithms to navigate and optimise spaces of hidden variables, such as outcome variation between individual participants or data processing pipelines.

The first framework builds a closed-loop experiment where an experimental space is explored automatically to maximise an individual's brain response. Generative modelling is used to create spaces of face stimuli to be explored in visual self-recognition. The framework is extended to EEG experiments with a mum-stranger paradigm run with infant participants. This allows the researcher to learn each individual's responses across many stimuli.

The second framework builds a searchable space of different analysis. These spaces are used to model how robust each approach is within the multiverse of different analysis options. First, the multiverse of preprocessing pipelines is explored for functional connectivity data with the task of predicting brain age from adolescent developmental data. Second, a multiverse of predictive models is explored for an EEG face processing task predicting autism.

The third framework is a normative modelling approach that uses state-of-the-art machine learning algorithms to model normal variability in brain structure. This approach generalises to different cohorts characterised by deviations from typical brain structure, detecting them as outliers. We illustrate its use by successfully predicting a neurodevelopmental psychiatric condition.

This work intends to explore different avenues to build new gold standards in methodology that can improve the robustness of neurodevelopment and neuropsychiatry research.

Acknowledgements

I am grateful beyond words of having people around me that inspire and shape me towards where I am today and where I will be tomorrow. These past few years have been the most challenging I've lived through, and without you I certainly wouldn't be where I am today.

In that respect, no person has been more responsible for me being here professionally than my supervisor Rob Leech. Thank you, Rob, for taking the chance on a random kid from Portugal 5 years ago and for continuously supporting me throughout this journey. One of my main motivations has always been to make you not regret your choice. Thank you for everything you have taught me, thank you for showing me how to look broadly for novel ideas while the whole field chases the same ball. Thank you for giving me space to find my own research directions but always being available to suggest the best course of action. Thank you for pushing me up when my motivation was down and for clearing my head whenever I got confused. Above all else, thank you for always being in my corner. Your loyalty and kindness to your students make you a great leader, and I will continue to follow you wherever you may go. N.O. lab is better!

I also want to express my unfettered gratitude to my supervisor Emily Jones. Thank you for your continuous support throughout my PhD, both in research and outside of it, even in the most difficult circumstances. Thank you for all you have taught me in our discussions and for always having good suggestions to overcome the research roadblocks we found along the way. Your extensive knowledge, helpfulness and availability were instrumental in my understanding of the field of neurodevelopment. I am grateful to have had the chance to work for both of you. I could not have asked for better supervisors in my PhD.

I am also eternally grateful for my non-official third supervisor, mentor, and friend Walter Pinaya. I am the luckiest person in the world to have been matched with you on the Neuroimaging KCL competition. In these past years you have *been my Deep Learning guru, with an infinite patience and goodwill with no obligation to do so and without ever asking anything in return. Thank you for welcoming me into your projects and for taking me under your wing. I will always be indebted to you for all you have helped me and all I have learned from you. I hope our partnership continues for many years to come.*

I would also like to thank Jessica Dafflon for her influence and friendship these past few years. It has been a pleasure to work with you in all our projects. Thank you for introducing me to what is the life of a PhD researcher and how to manage expectations in our long conversations at the CNS. The MindThePinapple™ best years are yet to come!

I also want to thank all the people I have had the luck and pleasure of collaborating with. I have learned from all of you, and you have made me a better researcher. Thank you, Rianne, my EEG guru, for all your patience in our long days at the CBCD collecting data in between the many bugs we had to solve live. Thank you, Anna, for welcoming me to the CBCD with such kindness and open-arms, and for always making it such an enjoyable endeavour to work with you. Thank you, Elena, for all your patience throughout these years and all the hard work you put into every project we did. I would also like to express my gratitude to the European Union for their very generous MSCA-ITN programme that gave me all the necessary conditions and more to be able to focus solely on my research.

Thank you to all my lab mates at the CNS and especially to Caitlin, František and Erik. Sadly, our time together got shortened by the pandemic, but I cherished the lab environment we had and all the conversations we had that helped me as a researcher and as a person. I also want to thank Romy and Ricardo with whom I had the pleasure to collaborate with at the beginning of my PhD. Your academic work has built the basis for what I have developed here, and your academic and professional paths serve me as inspiration for my own.

Doing a PhD is never an easy endeavour, but to do it during a global pandemic that made people shelter at home for a significant part of two years makes the challenge even harder. Fortunately, I had a very supportive environment, at a professional level but also at a personal one. Thank you to all my friends in London that have accompanied me through this journey, Filipe, Patricia, Cartucho, Maria, Pedro, Ricardo, Ji, Crispim, Gui, Cristina, Alex, Laura, Tom. I also want to thank the support from my friends at home, thank you Zé Pedro, Fred, Gui, Ricardo Lopes, Ricardo Lopes, Bruno and the família Imurtal. I am sorry for my absence at times when I was trying to push my work forward.

Finally, and most importantly, I want to recognise the role my family had in me being here through their love, support and, in many times, personal sacrifice. Thank you, mãe, for being my greatest defender, for teaching me to be ambitious, to never limit my horizons and for giving me every opportunity to succeed in life. Thank you, pai, for teaching me the importance of family, for being my biggest supporter and my hero. I wish you were here to share this moment with me. Thank you, Lena, for being my spiritual guide throughout the PhD. You are one of my biggest influences and your academic path has inspired a lot of my decisions. Thank you for always being ready to give me advice or support at a phone call away. I am grateful to Nuno, Bárbara, Gustavo, Luis, Mila and Inês for being there for me.

Words fail to express how thankful I am to you, Margarida, for being part of my life and for being the rock that keeps me whole. You bring me peace, stability, and love and I am the most grateful man to be at your side. You are my better half and make me a better person in the process. Thank you for your eternal patience and understanding for my late working hours. This work is finished because of you.

Pedro Ferreira da Costa - September, 2022

Contents

List of Figures

List of Tables

Symbols and Abbreviations

ANTs	Advanced Normalisation Tools
API	Application Programming Interface
ASD	Autism Spectrum Disorder
autoML	Automated Machine Learning
AUROC	Area Under the Receiver Operating Characteristic curve
BCI	Brain-Computer Interface
BLR	Bayesian Linear Regression
BO	Bayesian Optimisation
CASH	Combined Algorithm Selection and Hyperparameter optimisation problem
CMS	Common Mode Sense
CNN	Convolutional Neural Network
DRL	Driven Right Leg
DSM	Diagnostics and Statistical Manual of Mental Disorders
dVAE	discrete Variational Autoencoder
EEG	Electroencephalogram
EI	Expected Improvement
ERP	Event-Related Potential

FC	Functional Connectivity
fMRI	functional Magnetic Resonance Imaging
GAN	Generative Adversarial Network
GPR	Gaussian Process Regression
HCP-EP	Human Connectome Project - Early Psychosis
HCP-D	Human Connectome Project - Development
HCP-YA	Human Connectome Project - Young Adults
KNN	K-Nearest Neighbours
LLE	Local Linear Embedding
MDS	Multidimensional Scaling
MAE	Mean Absolute Error
MEG	Magnetoencephalography
ML	Machine Learning
MRI	Magnetic Resonance Imaging
Nc	Negative central
NIRS	Near Infrared Spectroscopy
PCA	Principal Component Analysis
PCN	Predictive Clinical Neuroscience
RBF	Radial Basis Function
ROC	Receiver Operator Characteristic
SD	Standard Deviation
SFCN	Simple Fully Convolutional Network
sMRI	structural Magnetic Resonance Imaging
SVM	Support Vector Machine
t-SNE	t-distributed Stochastic Neighbor Embedding

UCB	Upper Confidence Bound
UMAP	Uniform Manifold Approximation and Projection
VAE	Variational Autoencoder
VQVAE	Vector-Quantised Variational Autoencoder

1 Introduction

1.1 Challenges with Hypothesis testing in neurocognitive paradigms

In 1710, Dr John Arbuthnot used 82 years of birth records in London, where male births consistently exceeded the number of female births, to show that the probability of equal likelihood of birth between both genders was exceedingly small ($\frac{1}{2^{82}}$) and that this hypothesis could not be attributed to chance (Arbuthnott, 1710). It was credited as the first example of hypothesis testing. Centuries later, much due to Fisher's work in inventing and establishing modern statistical science (Box, 1978, Fisher, 1915, 1992), hypothesis testing is today the tried-and-true method in accepting or rejecting a hypothesis in all fields of scientific research, as is the case of cognitive neuroscience. It is responsible for moving science from a more Bayesian perspective, involving subjective prior probabilities, to an objective, entirely data-driven direction, where statistics determine the likelihood of a given hypothesis being true in a determined set of conditions.

1.1.1 The Replication & Generalisability crisis

However, in the past couple of decades, the reliability of cognitive and medical science findings has been questioned (Pashler and Wagenmakers, 2012, Simmons et al., 2011). Statistical studies have shown that the percentage of false positive findings is vastly above the threshold for significance, generally set at 5% (Ioannidis, 2005). A study trying to replicate high-impact research in psychology has shown that from a set of 100 studies, less than half of significant findings held when replicated and that the average effect size across various studies was only half of the initially reported values (Collaboration, 2015). Similar results have been found in medicine research (Prinz et al., 2011), behavioural economics (Camerer et al., 2016), genetic research (Munafó, 2009) and neuroscience research (Button et al., 2013). The failure to replicate major research findings that serve as the cornerstone for many other research studies has presented a methodological reckoning to the field. It is widely recognised that the fault lies with questionable research practices that lead to overinflated significance results (John et al., 2012, Simmons et al., 2011). Despite the statistical soundness of hypothesis testing, the flexibility in data collection, analysis and reporting of results allows the researcher's biases to creep in, which leads to questionable findings (Simmons et al., 2011). These include but are not limited to selectively discarding data through questionable outlier removal techniques; repeating statistical analysis in slightly different conditions until a significant result is obtained (Head et al., 2015); defining or changing the hypothesis after the analysis of the results (Kerr, 1998); selecting regions-of-interest and hypothesised areas of activity after the analysis of results, so as not to have to correct for multiple significance measures (Poldrack et al., 2017). These concerns have led to a movement for better standards and practices in research by promoting full transparency in the data and code used in published works as a standard practice, making them fully available online (Westfall et al., 2017) and publishing in open-access venues to promote information accessibility; and, introducing pre-registration of studies, dividing the publication process into two steps, publishing the hypothesis, methodology and proposed analysis, and following peer-review publishing the results using the pre-defined methodology (Foster and Deardorff, 2017). These steps limit data collection and processing flexibility to avoid the researcher's biases or poor practices.

In developmental neurocognitive research (particularly with infants and challenging neurodevelopmental populations), these problems are exacerbated by other challenges that are hard to overcome: recruitment is challenging, leading to low sample sizes, data quality is often reduced because the participants are often active or have short attention spans or are not compliant. Therefore, the signal-to-noise ratio in any given experiment tends to be low, promoting false-positive results and low-powered studies. These types of problems have led to an increasing use of pre-registration in developmental cognitive neuroscience studies and a movement to many labs collaborating to increase sample sizes and robustness of results (Frank et al., 2017).

While the hacking of results may suggest ill-intent from many researchers, the lack of reproducibility cannot be assigned to data misuse in its entirety. Traditional developmental cognitive neuroscience methodologies map many signal metrics to limited hypotheses (e.g., in ERP research, we can consider the latency, peak-to-peak amplitude, post-stimulus average signal, and power analysis, all for the same collected study (Woodman, 2010)). Cognitive paradigms also tend to be overly narrow, considering only one experimental condition at a time while addressing broad hypotheses (e.g., what face elicits a stronger signal from an infant? can we use a given event-related potential (ERP) as a biomarker for a given developmental population?). They are reliant on classical paradigms and stimuli sets that are limited in scope, which has the potential to lead to overinflated test statistics (Westfall et al., 2017). Finally, most paradigms and stimuli sets were built based on a restricted range of typically developing participants to test a given response. These not only fail to capture the variance across the whole population but can be suboptimal when studying responses from participants with psychiatric conditions, whose variable of interest is not considered while designing the paradigm. Throughout this thesis, the term response is used to refer to the variable of interest in a given experiment (e.g., maximum amplitude in an EEG experiment; phenotypic data in behavioural studies; structural brain data in neuroimaging studies) on both predictive and statistical analysis.

Another important direction in discussing methodological best practices in psychology research comes from Yarkoni (2022), who introduces the generalisability crisis. In it, he addresses the concern with the overuse of "random effects" to justify all variability present in data. Yarkoni relates how linear models are inadequate to model brain responses as they do not capture the variance present in data but are used pervasively across the field. This inability to account for variability dismisses fundamental intra-group differences and overestimates confidence intervals, leading to incorrect significant group-wise differences. It also leads to failures in replicating the results obtained, as they fail to generalise to a different setting or just different participants. The solution consists of modelling responses with more complex algorithms that account for uncertainty on measured data and on trying to capture the variability in data by accounting for the possible covariates (e.g., recording systems, neurodiverse population, state-of-mind). The issue of generalisability has been acknowledged as a limitation in neurodevelopmental research (Visser et al., 2022).

In sum, the fields of cognitive neuroscience and psychology are going through transformative years, where new gold standards of research methodology are being devised and embraced by the community; this is equally true for neurodevelopmental research where many of the challenges underlying the reproductibility crisis are more acute. There is a particular focus on limiting the researcher's degrees of freedom when testing their hypothesis. I argue that this leaves a

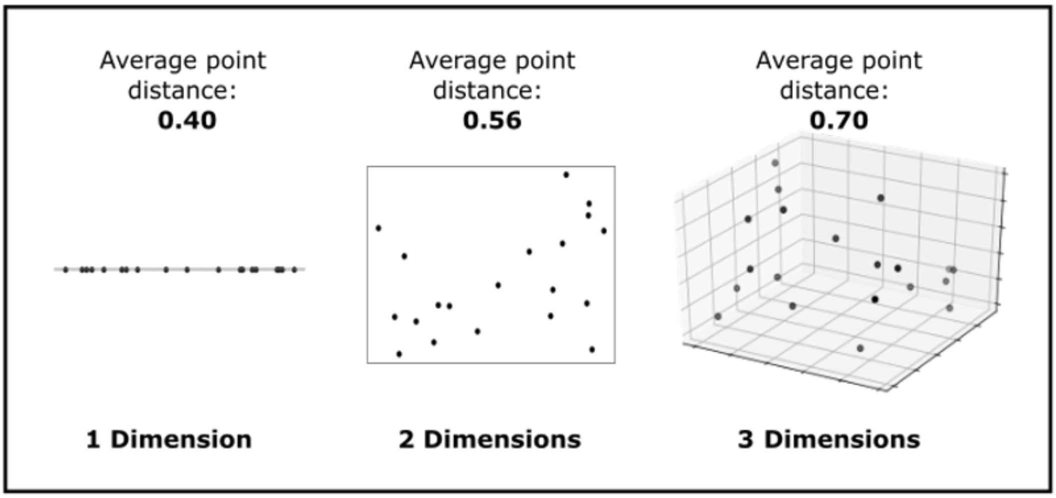

Figure 1.1: The sparseness of high-dimensional spaces. *This effect can be seen in as few as three dimensions. From a random sampling of 20 data points in unit dimensions ($x \in [0, 1[$), the distance between data points increases with the number of dimensions considered in the space. For 1000 unitary dimensions, the average data point distance is 12.89. For this random sampling. This sparseness of the data will directly affect the assertions that it is possible to do with the data and the capacity of predictive models to infer information from it.*

gap to be filled for best practices in more explorative studies, where too narrow hypotheses and paradigms hinder the capacity to address broader research questions and fail to consider the substantial variation in brain responses across a neurally diverse population. The challenge is to build new methods that can account for broader hypotheses and paradigms that cover a more extensive range of experimental conditions while maintaining the best methodological practices and without falling into the curse of dimensionality.

1.1.2 Curse of dimensionality

If I am to sample a unit interval at evenly spaced points of 0.01 distance, I will sample it 100 times. If I instead sample a 10-dimensional hypercube at the same distance between points, I will need to sample the hypercube 10^{20} times. The consideration that the number of data points to sample from increases exponentially with the number of dimensions in Euclidean space is referred to as the curse of dimensionality (Bellman, 1957). The curse of dimensionality is a known problem in data modelling, as increasing the number of features (i.e., dimensions) while maintaining the same number of data points leads to a sparser space of sampled points (see Figure 1.1). This sparsity leads to a bad generalisation of the data model as emptier regions of the space are predicted to behave according to data points that are far apart in the data space. Interpolation fails due to the distance between interpolated points.

This phenomenon brings challenges to numerical analysis, sampling and data modelling that are relevant when addressing the generalisability and replication limitations of current cognitive science methods:

a) Complex experimental paradigms that consider more than one condition will require a factorial higher number of data points to extensively sample the paradigm space (Lorenz et al., 2017)

If three factors are considered for a paradigm, e.g., in face recognition tasks, directed-averted gaze; emotion; and familiarity, I need to consider the number of possible combinations between these factors. One way to circumvent this limitation is to maximise information retrieved when sampling instead of aiming to sample all combinations extensively (i.e., perform active sampling). This can be done with classical optimisation methods that leverage uncertainty to guide where to sample next (Baptista and Poloczek, 2018, Lorenz et al., 2017).

b) The large assortment of methods available for data preprocessing, each with different tweakable hyperparameters, creates a highly dimensional space of preprocessing outputs resulting in the same data leading to different conclusions (Bzdok and Yeo, 2017).

A large amount of methods impacts the replication of studies and the reliability of findings in general. By not treating it as a high-dimensional problem, we ignore the variability associated with the preprocessing methods used (Bzdok and Yeo, 2017, Carp, 2012). It is unfeasible and undesirable that each analysis is run on all existing preprocessing pipelines as this compromises inferential and predictive power. However, preprocessing pipelines' outputs are correlated with each other, and dimensionality reduction can be used to find a lower-dimensional manifold that approximates the whole space of preprocessing outputs. The same problem arises when building data models of response or training predictive systems. The number of heterogeneous algorithms can lead to many different model representations (Bell et al., 2022). As before, it is relevant to find lower dimensional representations that can capture all the possible variability explained by choice of modelling algorithm.

c) Outlier detection studies that rely on proximity measures to build

models of normal response fail with a higher number of features as the sparsity of the data points increases exponentially with the number of dimensions (Beyret et al., 2019, Zimek et al., 2012).

Larger feature space dimensions increase data sparsity due to the exponential increase in space volume. In this setting, proximity measures become irrelevant as the distance between the nearest data point converges to the distance of the furthest data point (Aggarwal and Yu, 2001). In this scenario, every data point can be identified as an outlier. This problem is relevant to cognitive neuroscience when trying to identify clinical conditions as deviations from the norm because many features of relevance are highly dimensional (e.g., structural brain data; raw brain activity recordings). To address this limitation, some studies have relied instead on metrics of outlier detection that do not depend on Euclidean distance, such as likelihood estimation between samples (Pinaya et al., 2022). Other works have focused on building lower-dimensional projections of the high-dimensional spaces through classical dimensionality reduction techniques or extracting relevant high-level features from brain data, such as brain volume of longitudinal cortical thickness (Marquand et al., 2019).

Building lower dimensional representations of original data tackle several limitations with high dimensional feature spaces. The challenge is maintaining the relevant information that summarises the original high-dimensional data and discarding the non-informative dimensions.

1.1.3 Distilling information

Consider a dataset of hundreds of black-and-white images of faces where each image contains 1000 pixels. Each image can be interpreted as a data point with 1000 dimensions, where each pixel's brightness represents the value of the coordinate in a single dimension. They populate the space of possible data points in this hyper-dimensional space, and, due to the curse of dimensionality, they do so very sparsely. Additionally, as this space accounts for any combination of brightness from the 1000 pixels that make up each image, including all possible white-noise combinations, all images of faces will only cover a small and specific region of the hyper-dimensional space. All the data points in our dataset live in a lower-dimensional manifold (i.e., a topological closed surface that is locally Euclidean) of the much larger hyper-dimensional space (see Figure 1.2). It is the objective of dimensionality reduction to find lower-dimensional manifolds that capture the relevant components of a set while losing the least amount of

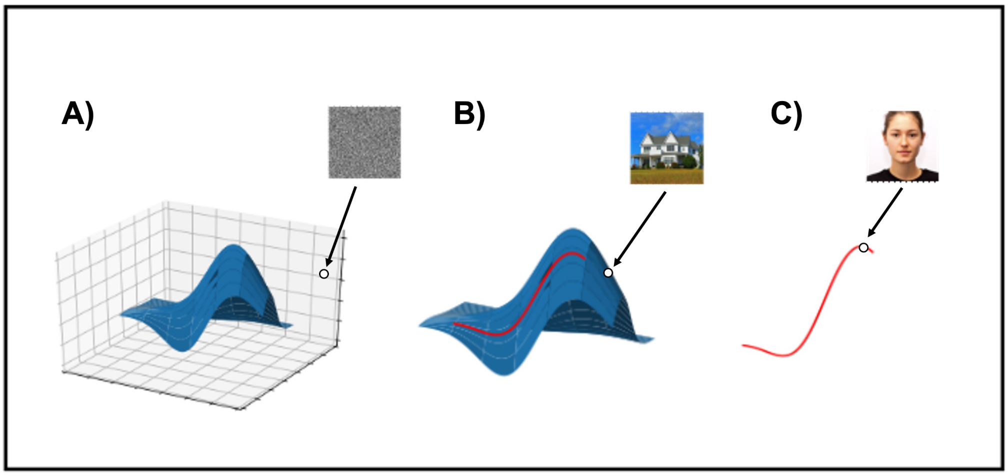

Figure 1.2: Illustration of natural images manifold in hyper-dimensional space of possible pixel combinations. *A) is a schematic representation of the hyper-dimensional space of all possible values an image can take (illustrated in three dimensions for representation purposes). Most images in the space are incomprehensible, but there is a small multidimensional surface that captures all possible natural images (pictured as a blue surface). That is the manifold of natural images that can be represented in a smaller number of dimensions. The same is true when considering only the manifold of images of faces, which is contained in the manifold of natural images (here represented by the red line). The manifold of faces has a lower-dimensional representation than the manifold of natural images that, in turn, has a lower-dimensional representation than the full space of possible values. It is then fundamental to find the lower-dimensional representation for the problem being addressed to combat the challenges created by the curse of dimensionality through statistical methods or density function estimators.*

information possible. The visual cortex does something similar when processing visual information - it is simpler to distinguish between the faces of two siblings than between two white-noise images, even though the latter example is at a much larger distance in the space of possible stimuli (Pang et al., 2016). Importantly, we can reduce the dimensionality of the original space by focusing on our variable of interest (i.e., faces). If our variable of interest were the estimated age of a given face, we could reduce the data points in a dataset to only one dimension to encode the estimated age.

Dimensionality reduction is then the transformation of data from a higher dimensional space to a lower-dimensional space so that the relevant properties of data are retained, $x \in \mathbb{R}^n$, $y = F(x)$, $y \in \mathbb{R}^k$, where $F(.)$ is the transformation function and $k < n$. We can subdivide dimensionality reduction techniques into statistical methods and density estimators. Statistical methods focus on the statistical properties of the data (e.g., variance across dimensions) and can be linear, such as Principal Component Analysis (PCA), or nonlinear, such as Multidimensional Scaling (MDS). Density estimators (e.g., Variational Auto-encoder (VAE); Generative Adversarial Networks (GAN)) are models that try to learn the unobservable probability density function of the data that generated a given dataset. In machine learning, they are known as generative models. Large advances in the field of machine learning have led to the development of several generative models that can capture lower-dimensional manifolds of data (most commonly images) while implicitly or explicitly learning its density function (Tomczak, 2022). By learning the probability density function, these models permit us to sample unseen points from the lower-dimensional manifold and generate realistic synthetic new data without a cost.

The use of dimensionality reduction algorithms is common practice when building data models of responses or as a preprocessing step to analyse results and draw conclusions (Ayesha et al., 2020) as they promote more generalisable and robust results. Nevertheless, these algorithms' use can be extended further in the research pipeline as a mechanism to build more informative paradigms shaped by data-driven insights into the relationship between experimental conditions (Lorenz et al., 2017) or as a mechanism to study the impact of the multiverse of options between data processing or data modelling algorithms.

1.2 Machine Learning as a tool for predictive research

Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Samuel, 1959

Despite the term being coined in 1959, it was only in the early 2000s that machine learning started gaining more visibility and popularity. The advent of big data and large computational resources (Alon Halevy et al., 2009, Bell et al., 2005) has proved crucial for machine learning to leap from an experimental field to building tools for the betterment of society. Today, its quantitative tools are ubiquitous in every sector of society and many research fields due to their capacity to map complex data patterns to make predictions in unseen data (Bishop, 2006). They are the backbone of search engines, recommender systems, face recognition algorithms and spam identifiers (Sarker, 2021). In research, they are a powerful tool for data modelling and finding complex relationships in data, allowing the researcher to obtain insights that were not accessible before. In neuroscience research specifically, machine learning algorithms have been primarily used as predictive models to predict individual outcomes from brain patterns (Livezey et al., 2019) (e.g., building a model to predict if a given functional MRI scan is from a participant with schizophrenia (Lai et al., 2021, Pereira et al., 2009)). In infant research, these models allow us to build individual development profiles, find biomarkers in heterogeneous conditions and objectively identify digressions to normal development – an especially crucial step to capture clinical conditions early. Other applications of machine learning tools in neuroscience research include improving brain-computer interfaces (Sussillo et al., 2016) and building computational models of the brain (Hassabis et al., 2017, Lindsay, 2020).

1.2.1 Individual predictions & profiling in neuroscience & psychiatry

Predictive models can generalise to out-of-sample distributions by interpolating between complex patterns in the data that it is given (i.e., the training data) (Efron and Tibshirani, 1991). Commonly, it does so by adjusting its parameters to minimise a given loss function that tends to represent the error between prediction and known labels. This setting is known as supervised learning. These models are useful in research because, by being able to predict unseen data points, they hold insights into what differentiates two categories in a fully data-driven manner (Vu et al., 2018). Furthermore, developing robust and generalisable predictive models that perform above the human level in a clinical and psychiatric setting holds the key for the shift to personalised healthcare and new diagnostic tools. For example, many diagnoses of psychiatric conditions still rely on phenotyping patients based on a question-answer approach (American Psychiatric Association, 2013) that depends on patient's reliability, an approach with low inter-rater reliability and ambiguous descriptions (Wakefield, 2013). Recent studies have focused on finding structural changes in the brain associated with psychiatric and neurologic disorders using predictive models to build robust biomarkers that can be used as objective measures of disorders (Vieira et al., 2017).

In research settings, it is common to pit this methodology against statistical inference's hypothesis testing methods (Anderson and Perona, 2014) as they aim to extract new knowledge from mathematical models. However, they address different questions and require different assumptions. Whereas hypothesis testing focuses on building theories of causal underpinnings of human behaviour, predictive models focus instead on predicting new examples of behaviour (Yarkoni and Westfall, 2017). Where hypothesis testing requires strong pre-defined assumptions, predictive models benefit from having as few assumptions as possible. Predictive models are primarily data-driven, whereas hypothesis testing is a fully hypothesis-driven method that requires mathematical rigour to describe interdependencies. Statistical models are often tractable and interpretable, whereas predictive models are less so and more parameterised, making them more expressive. It is possible to have a predictive model that correctly identifies two categories that do not have a significant p-value when submitted to the null hypothesis and *vice-versa* (Arbabshirani et al., 2017). So, it is fairer to see these approaches as complementary. Statistical modelling takes a more confirmatory and exploitative approach to the data, whereas predictive modelling is more explorative. This comes with its drawbacks. As mentioned in Yarkoni (2022), the generalisability crisis revealed how predictive models mostly fail to capture the natural variance in the data and do not account for hidden confounders, assuming a deterministic response for a given set of features. In practice, this is never the case (Vieira et al., 2020). Their exploratory nature makes them brittle as they learn to map the data distribution correctly but fail to generalise for slightly different settings (Kelly et al., 2019) (e.g., a different device manufacturer; a neurodivergent population). In practice, it is impossible to control for all variables, and the predictive model's performance falters.

One solution that can draw from both methodologies while capturing the inherent variability in data acquisition is to build individual predictive models of each participant (i.e., individual profiles of response) using algorithms that account for uncertainty. Instead of having a model predicting the response of a whole group to a pre-defined paradigm, the researcher can build individual predictive models for each participant based on their performance in the same paradigm (Cusack et al., 2012, Vu et al., 2018). This approach is especially sensitive to the inherent neurodiversity of in-group participants that gets watered-down in both predictive analysis and statistical inference studies. Furthermore, it sits at the intersection of both methodologies. It allows the researcher to perform statistical analysis on the group-level differences and do hypothesis testing with the different predicted responses. It also allows for more robust predictive models by building an ensemble of individual predictive models (i.e., weighing the many predicted responses into one consensus) (Zhang and Ma, 2012). One drawback of taking an individual approach to building predictive models is that they require a large number of data samples to perform efficiently, limiting individual measurements in an experimental setting. It is then crucial to sample the paradigm efficiently, maximising samples with higher measurement uncertainty.

1.2.2 Active Learning

Commonly, machine learning algorithms are trained with observational data, i.e., using datasets that are collected *a priori*, in what is termed passive learning. The dataset is shuffled and partitioned into training, validation, and testing sets to minimise the chance of overfitting the data it was trained on. The disadvantage of passive learning is that not all samples are equally relevant for training. Data points close to the decision border will be more impactful for a correct predictive system (Konyushkova et al., 2017), but they will be sampled at the same rate as any other data point. Furthermore, passive learning is not always possible. Some problems in machine learning, as in reinforcement learning, are naturally sequential and require the model to be trained as it collects data from its environment (Barto and Sutton, 1992). In these scenarios, as the data collection is online, the algorithm can inform what is the most informative data point to sample next based on the results of previous iterations. This is called active learning or active sampling (Cohn et al., 1994). The informed decision on where to sample next becomes crucial when exploring a high dimensional feature space, as an extensive exploration becomes intractable due to the curse of dimensionality (Bellman, 1957). The same is true when sampling a data point entails high financial costs, computational burdens, low participant attention span, or a diminishing significance threshold per measure. In these scenarios, efficiently sampling the feature space to extract as much information as possible becomes paramount.

What is defined as the most informative data point to sample depends logically on the optimisation problem that is being addressed. The diversity of options for where to sample next is called query strategies, and they tend to be controlled by a mathematical algorithm that works as a "model of the model", receiving as input the relevant state of the model for the optimisation task (Settles, 2009). This algorithm is termed the acquisition function as it guides the search through the feature space. If the model's goal is to predict a response in any region of the feature space with high reliability, the next point to sample should be the one with the highest expected error reduction or the one with the highest uncertainty measure. This is an explorative strategy. If the goal is to find the maximum or minimum of the underlying function of response in as few samples as possible, then the regions around the previously sampled extrema should be privileged. This is an exploitative strategy. In practice, strategies tend to be a trade-off between these two approaches, as overly explorative strategies are expensive and overly exploitative strategies tend only to find local extrema (i.e., an extreme value which is not the maximum or minimum of the space, a common problem in non-convex spaces) instead of the desired global one. The acquisition function can control a spectrum between exploration and exploitation depending on the problem it is optimising for.

Active learning approaches have been previously employed in neuroscientific research, going as far back as 1987, when Jones and Palmer (1987) searched for the set of stimuli that would maximise the neuron's firing rate in animal studies. This neuroadaptive approach has been more recently explored in humans by Lorenz et al., where brain activity of regions-of-interest was maximised in fMRI studies in a closed-loop setting using an organised space of experiments that were presented to the participant based on their brain responses in previous experiments (Lorenz et al., 2015, 2017, 2018). Active learning requires a feature space to sample from, and in neuroscientific and psychology research, our features of relevance are stimuli or experiments, depending on the paradigm. In these cases, we name the spaces, respectively, stimuli spaces and experiment spaces. These are the spaces the active learning approach optimises over to build individual predictors of response in a given paradigm efficiently. For an in-depth review of active learning, please refer to Settles (2009).

1.2.3 Finding outliers

The supervised learning setting is defined by the process of training a predictive model to learn the underlying function that maps the data points to their known labels by minimising a loss function that penalises wrong mappings. The expectation is that the learned function generalises to predicting labels of unseen data points, so the dataset is divided into three components (i.e., training, validation and testing set) so that the model learns on the first, it is regularly evaluated on the second and only evaluated once on the latter to test how well it generalises its performance. This general setting in machine learning has shown remarkable results in important clinical tasks such as tumour detection and segmentation, skin cancer detection (Lundervold and Lundervold, 2019), or even diagnosing psychiatric conditions (Vieira et al., 2017), overcoming expert performances in many cases. Despite these published results, most of these algorithms fail to move to a clinical setting because they fail to generalise their performance consistently (Pinaya et al., 2016, Vieira et al., 2020). The main limitation in medical machine learning is the lack of large datasets required for algorithms to capture the full data distribution. Supervised learning algorithms excel at interpolating results between data points they were trained on but are notoriously bad at extrapolating for out-of-distribution data points (McCartney et al., 2020). Medical datasets are limited due to information privacy restrictions, high costs of data collection, such as an MRI structural scan, and the limited number of patients that can be recorded for any given disorder. This low number of data points leads to unbalanced datasets and is prone to overfitting or failing to capture the variability of the population.

To address these limitations, there have been efforts to move to other machine learning settings, such as unsupervised learning. In an unsupervised learning setting, the labels are unknown, and the predictive model learns instead to build internal representations of the data through pattern recognition and grouping. One unsupervised approach is outlier detection by building normative models (i.e., a model of statistically expected responses) of healthy participants to try to capture the normal heterogeneity in participants' responses. This model can then detect any data point that is an outlier to the expected response and classify it as an anomaly. As such, any participants with conditions that impact the expected response will be flagged while not requiring the use of their data to train the model. This alleviates the limited dataset constraint, as collecting large swathes of data becomes more manageable if only healthy participants are required. This approach has been specifically explored in psychiatric research using neuroimaging data, where patients are characterised by heterogeneous patterns both in symptoms and in subtle brain changes (Shenton et al., 2001), making it challenging to build supervised predictive models. In schizophrenia, for example, inter-individual differences between schizophrenic patients mask group-level differences to healthy participants (Wolfers et al., 2018). Taking an outlier-detection approach to the problem makes it possible to identify distinct anomaly patterns without requiring the pattern to be replicated across all subjects with the condition. For a detailed review of normative modelling in psychiatry research, please refer to Marquand et al. (2019).

1.3 Outline and aims of this thesis

Research in development cognitive neuroscience, and neuroscience and psychology more generally, are at a transformative moment. There is a wide acceptance of the shortcomings in current methodologies that lead to the failure of replicating and generalising research findings. The ongoing debate has focused on pre-registration and open science initiatives, leaving a gap for new gold standards for more open-ended questions and personalised experiments. At the same time, computer science and machine learning developments have brought forward promising new tools that can be leveraged for better research practices. These tools allow for tackling the curse of dimensionality by learning relevant lower-dimensional representations; they can be used to navigate the multiverse of preprocessing and postprocessing methods by learning their interdependencies; they can be used to build more robust detection methods based on outlier detection. This thesis aims to present, validate and apply three new frameworks for improving cognitive neuroscience's replicability and generalisability studies by borrowing from recent advances in machine learning.

Chapter 2 introduces the general methodologies leveraged throughout the three presented frameworks. It is subdivided between the algorithms used throughout the thesis for building lower representations of the original data and the algorithms used for optimising these representations. For the algorithms that build lower-dimensional representations, classical dimensionality reduction techniques (i.e., PCA and MDS) and density estimator deep generative models (i.e., VAE and GAN) are described. For the response optimisation algorithms, we describe

Bayesian optimisation, an increasingly recognised technique for neuroadaptive experiments and closed-loop sequential optimisations, and auto-regressive models, with a special focus on Transformer artificial neural networks, that learn the probability distribution of representations and are shown to be strong outlier detectors.

Chapter 3 presents an extension of the neuroadaptive optimisation framework to qualitative studies using generative models to create expressive stimuli spaces. It explores the human response to face stimuli, optimising over feature attributes of a face space learned by the generative model. In the presented proof-of-concept, I use the neuroadaptive framework to measure on an individual basis how self-recognition varies for perceived variations of age and emotion in the space of face stimuli. This work was performed to design and optimise the adaptive algorithms, in a simpler setting, before applying it to the more challenging domain of infant EEG, below.

Chapter 4 extends the neuroadaptive framework to EEG studies. It further explores how individual response profiling can be used to build future biomarkers of neurodevelopment conditions. The introduced proof-of-concept uses the neuroadaptive design to optimise over EEG signals. More specifically, the negative-central (NC), an ERP associated with facial recognition in infants, in a mom-stranger paradigm with infant participants. I show what information can be extracted from these individual models of response.

Chapter 5 presents an active learning multiverse analysis framework to study robustness and variability across methodology choices in the data processing pipelines. To account for the variability in results introduced by choice of data processing, this chapter explores building informative spaces of an extensive range of methods and visualising how results change across the space. Specifically, I introduce a multiverse analysis of preprocessing methods in functional MRI data pipelines, where it is shown how the choice of methodology can impact age prediction in a developmental functional connectivity dataset of adolescents.

Chapter 6 presents a generalisation of the framework introduced in chapter 5. The presented study is a multiverse analysis of predictive models and hyperparameters, showing how navigating a space of organised predictive models can maximise the predictive power on a given dataset. I show how it outperforms results in the literature of predicting autism in infants from volumetric data.

Chapter 7 explores outlier detection methods for robust and generalisable predictions in settings with a small number of examples. Instead of taking the supervised approach to predictive modelling, we investigate unsupervised normative models to detect outliers without training on the data of non-typically developing participants. Specifically, auto-regressive models are explored as a solution for predicting early-stage schizophrenia detection while building models of neutotypical individuals.

Chapter 8 concludes the thesis by discussing how new frameworks are required to tackle the methodological challenges present in research and how the frameworks presented here can contribute to the goal of improving research's gold standards. It provides an overarching view of where these frameworks fit into the full research pipeline and suggests the next direction to build stronger methodologies.

2 | Methods

2.1 Building Lower Representations of information

As described in Section 1.1.2, the curse of dimensionality poses a concrete challenge to the replication and generalisability efforts in hypothesis testing research and data modelling. To that end, several possible methods can be used to extract the relevant variables from data, for example, dimensionality reduction techniques and generative models. Classical dimensionality reduction techniques find lower-dimensional spaces where the redundant data is eliminated, but the full data variability is kept as intact as possible. Generative models, instead, learn hidden spaces that well represent the whole lower-dimensional manifold that is represented on a given dataset. In this Section, I will present a short introduction to some of these techniques and models.

2.1.1 Classical dimensionality reduction

2.1.1.1 Principal Component Analysis

Principal Component Analysis (PCA) (Minka, 2000) is a method that aims to capture the most variance of a dataset in a lower number of dimensions, here named Principal Components. It does so by applying orthogonal linear transformations to the feature space to create linearly independent components that capture the most variance while reducing dimensionality. This process uses a specific type of matrix decomposition, the eigendecomposition. Eigendecomposition factorises the covariance matrix of the dataset into a product of two matrices. One of the matrices contains per column the eigenvectors, unit vectors which are a restructuring of the axis of the original space into independent basis. The other matrix is a diagonal matrix containing the eigenvalues, which determine the magnitude or scaling of the original covariance regarding the basis defined by the eigenvectors. The relative value of the eigenvalues determines the explained variance of each eigenvector and so can be sorted to maximise the explained variance in a smaller number of eigenvectors while dropping the ones with small eigenvalues. This is the process of PCA, and the eigenvectors ordered by explained variance are named principal components. This allows the user to apply feature extraction to a dataset into n linearly independent principal components while knowing how much explained variance is being lost in the dimensionality reduction.

2.1.1.2 Multidimensional Scaling (MDS)

MDS (Kruskal, 1964) is a non-linear dimensionality reduction technique that captures similarities between high-dimensional data points as distances in a lower-dimensional representation. It can be subdivided into metric or classical MDS, where similarities are measured as the distance in the coordinate space using a linear scale, and nonmetric MDS, where the similarity is measured with rank and only the order of similarity between data points matters. The first is optimal for numeric problems and the latter for ordinal data. Metric MDS is specifically computed by first calculating the Euclidean distance between pairs of samples in a dataset,

$$d(p,q) = \sqrt{\sum (q_i - p_i)^2}, \tag{2.1}$$

and then iteratively optimising the coordinates in a lower-dimensional space that minimises a Stress function:

$$Stress_D(x_1, \dots, x_n) = \sqrt{\sum_{i \neq j} (d_{ij} - ||x_i - x_j||^2)}, \tag{2.2}$$

where d_{ij} is the Euclidean distance between sample i and sample j in the higher-dimensional space and $||x_i - x_j||^2$ is the distance in the lower-dimensional space. Minimising the Stress function by adjusting the coordinates position x will force the lower-dimensional space to correctly represent the distances between data points in the higher-dimensional space. This method is optimal for preserving the global and local structures of the original dataset, which is paramount when the similarity between data points is the most important factor in a given problem.

2.1.2 Generative models for Density estimation

Generative models are a subclass of machine learning where the stated goal is to learn the probability density function that underlies the data it represents, $p_{data}(x)$. In doing so, it is possible to sample the learned function to make up new data examples that appear to be from the original training set. Importantly, the probability density function lies in a lower-dimensional manifold of all possible data points, as it only accounts for the ones representing a realistic example of the variables of interest.

The advent of deep neural networks has been paramount to the recent progress in generative modelling research. Its hierarchical and modular structure allied to scalable gradient learning methods has brought forth models with a higher information processing capacity that have successfully approximated many intractable probabilistic computations. The use of these models in generative modelling has led to the subclass of deep generative models. Two examples of deep generative model architectures are GANs and autoregressive models. In both cases, the model learns to map a lower-dimensional hidden representation, z, to higher-dimensional samples of the learned manifold, y (i.e., given a learned model $G,(z) = y\,; z \in R^n\,; y \in R^m\,; n < m$). This hidden representation is appropriately named latent representation and can be exploited to manipulate the original data using a lower-dimensional space.

2.1.2.1 Generative Adversarial Networks (GAN)

A GAN is is a deep neural network composed of two modules that are trained concurrently in an adversarial process (Goodfellow et al., 2014). One module, the generator G, transforms a noise vector input into a high-fidelity representation of original data. The second module, the discriminator D, receives as input the outputs of the generator interspersed with real data points from the training set. The discriminator is then trained to classify the real data points (i.e., from the training set) and the fake ones (i.e., from the generator). The components try to optimise their parameters following a two-player minimax game with the following value function:

$$\min_{G}\max_{D} V(D, G) = \mathop{\mathbb{E}}_{x}[log(D(x))] + \mathop{\mathbb{E}}_{z}[log(1 - D(G(z))], \tag{2.3}$$

where z is the noise vector passed as input to the generator and functions as a lower-dimensional latent representation of the learned manifold. The two modules learn and improve their performance in tandem through stochastic gradient descent. Progressively the generator learns the probability density function of the real data distribution and can generate high-quality examples of artificial data not present in the training set by sampling randomly from z (see Figure 2.1.A). These algorithms have been extensively applied to generate realistic faces (Karras et al., 2017), music (Engel et al., 2019), naturalistic images (Wang et al., 2018) and many other complex data distributions.

Because of the general setting of GANs, there have been several variations of the algorithm, with different focuses on the value loss and architecture of both the generator and the discriminator (Arjovsky et al., 2017). In this thesis, the implementation used was the one introduced by Karras et al. (2019), named StyleGAN.

2.1.2.2 StyleGAN

StyleGAN achieves state-of-the-art performance on the generation of realistic faces and is able to control disentangled attributes in the generated images. The StyleGAN training is done in a progressive growing GAN (Karras et al., 2018), where the model is trained for 4x4 images and, when stable, trained for images with double the size by adding another block of layers to both the generator and the discriminator. This is done sequentially until the model is trained to generate and discriminate images with size 1024x1024. On the generator side, StyleGAN also added a mapping network to process the latent representation before feeding its output, y, to the generator's different progressively growing blocks. This is done by using an adaptive instance normalisation (AdaIN) at the end of each generator's block where, for a given generator's instance x_i, and the respective scaling and bias parameters from the mapping network $y_{s,i}$ and $y_{b,i}$:

$$AdaIN(x_i, y) = y_{s,i}\frac{y_{s,i} - \mu(x_i))}{\sigma(x_i)} + y_{b,i}, \tag{2.4}$$

Where $\mu(.)$ and $\sigma(.)$ stand for the mean and standard deviation of the variable. The output of the mapping network is, in essence, parameterised affine transformations that specialise to styles or attributes in the data. This results in more flexibility in controlling the image attributes using the latent representation. Finally, noise is added to the output of each convolutional layer of the generator's network (Figure 2.1.B). This model fits well into the developed work as we can use the latent representations in the mapping network to control for variables of interest (e.g., emotion; age) for any given paradigm.

2.1.2.3 Controlling disentangled attributes in latent representations

It has been previously demonstrated that the organisation of the latent representation can be exploited to control different attributes of the image (Pumarola et al., 2018, Radford et al., 2016). It is possible to manipulate specific features while maintaining the image identity by generating thousands of images from a generative model and labelling each according to a binary categorical variable (e.g., happy vs neutral). For the case of face images, it is possible to automate the labelling process by using a trained classifier such as Microsoft's Face recognition API[*]. For each categorical variable, it is possible to fit a logistic regression using the image's lower-dimensional latent representation as input and the label as expected output. The coefficients (c) from the fitted regression can then be used in the latent space of representations, d_{latent}, to shift the generated image to change the original image according to the categorical variables while maintaining the face identity. The obtained coefficients c are obtained concept vectors that navigate axes in the hyperdimensional space that control specific variations (e.g., the happy/sad vector). The degree of change can be controlled by a scalar magnitude multiplied by the whole vector. In this thesis, a bounded magnitude, $x\{x \in \mathbb{R} : -2 < x < 2\}$ was explored for each attribute.

$$d_{latent_{new}} = d_{latent} + c * x \tag{2.5}$$

One benefit arising from using linear transformations to vary a categorical variable while controlling for the image identity (i.e., maintaining all other variables constant) is that it reduces the problem's dimensionality to as many dimensions as the number of linear transformations. When the goal is to reduce the space of stimuli in a paradigm, this mechanism can greatly address the curse of dimensionality.

2.1.2.4 Autoregressive models

GANs perform implicit probability density estimation. The output of the generator follows the manifold of the training data (i.e., it learns what are real representations of the data or not), but the model does not have knowledge of the probability of the observations, nor can it specify the conditional likelihood function inherent to the generated data. These limitations are not present in autoregressive models where the probability density estimation is modelled explicitly. This is relevant for problems where it is important to assert how likely a

[*] https://azure.microsoft.com/en-gb/services/cognitive-services/face/

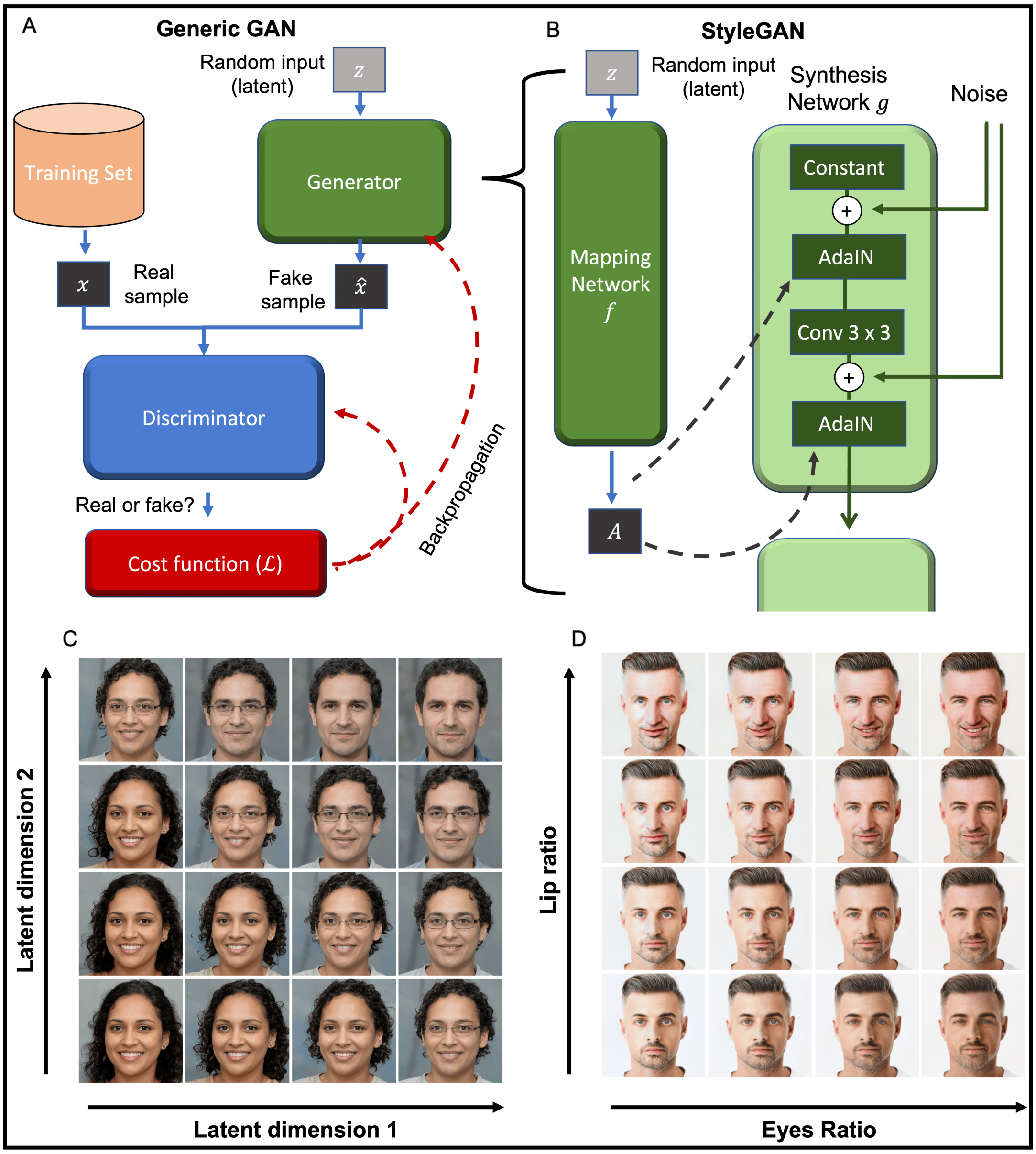

Figure 2.1: Schematic representation of a generic GAN. *The Generator and the Discriminator optimise concurrently in a minimax game by having the generator create samples that can be identified by the discriminator as real samples and by having the discriminator progressively improve its classification of real and fake samples. B. Schematic representation of the Generator component of the StyleGAN. The Generator in StyleGan is composed of a Mapping Network and a Synthesis Network, where the first controls an affine transformation on the latter. C. Example of images obtained by sampling a region of the StyleGAN latent. This represents a high-dimensional patch of the manifold of faces learned by the algorithm. These faces are not real, they are artificially generated by the algorithm. D. Variation of the semantic attributes 'lip ratio' and 'eye ratio' for a given face using StyleGAN. This GAN allows control of specific attributes of the face, which gives the researcher a more controlled environment when testing specific variables, as face identity is maintained.*

given data point is, such as the case of outlier detection. Autoregressive models take the problem of estimating the joint distribution of feature variables in a given data observation, $p(x)$, as the estimation of the product of conditional distributions over each individual feature. This is done because modelling the joint distribution by itself is not a computationally tractable problem. The autoregressive approach uses the chain rule to break down the likelihood estimation of x into the product of 1-dimensional distributions. It is then possible to consider the joint probability estimation as a sequence problem, where the probability estimation of data point, x, given feature i, is dependent on all previous features, $p(x_i|x_1, \dots, x_{i-1})$. Mathematically the estimation can be decomposed as such:

$$p(x) = p(x_1, x_2, \dots, x_n) \tag{2.6}$$

$$p(x) = p(x_1)p(x_2) \dots p(x_n) \tag{2.7}$$

$$p(x) = \prod_{i=1}^{n} p(x_i) \tag{2.8}$$

$$p(x) = \prod_{i=1}^{n} p(x_i|x_1, \dots, x_{i-1}) \tag{2.9}$$

This makes modelling a data point inherently sequential, which can be challenging for modelling data with a large number of features. Dimensionality reduction also plays an important role in this problem as it reduces the number of features drastically. With this goal, (Van Den Oord et al., 2017) created the VQVAE model.

2.1.2.5 Vector-Quantized Variational Autoencoder (VQVAE)

VQVAE (Van Den Oord et al., 2017) is a deep neural network model that is composed of two large modules: a discrete-representation variational auto-encoder model to create the lower-dimensional latent representation and an autoregressive model that explicitly learns the probability density function of the latent representation. Both modules are trained separately. The discrete latent variational autoencoder ($dVAE$) is trained to learn a transformation where the input is the same as the output (i.e., $dVAE(X) = X$), where there is a bottleneck of information in the middle of the model. This forces the model to learn a lower-dimensional representation from which we can still recover the original data, and this is the latent representation of the $dVAE$. The module comprises an Encoder that maps $x \in \mathbb{R}^D$ to a latent space $x \in \mathbb{R}^{d*n_z}$, with n_z relating to the dimensionality of the latent vector and d the number of latent codes. As the latent space is discrete, each element of z is quantised into its nearest vector $e_k \in \mathbb{R}^{n_z}, k \in 1, \ldots, K$ from a codebook with K elements. The codebook positioning in the latent space is learned jointly with the parameters of the $dVAE$. A Generator G can reconstruct the original observation x from the discrete latent codes. The module is trained through gradient descent by minimising the following composite loss function:

$$L_{VQVAE} = L_{recons} + L_{codebook} + \beta L_{commit} \tag{2.10}$$

$$L_{recons} = ||x - \widehat{x}||_2^2 \tag{2.11}$$

$$L_{codebook} = ||sg[z_e] - e_k||_2^2 \tag{2.12}$$

$$L_{commit} = ||sg[e_k] - z_e||_2^2 \tag{2.13}$$

Where $sg[.]$ denotes the stop-gradient operator. L_{recons} controls the quality of the reconstruction of the original data points and $L_{codebook}$ and L_{commit} try to minimise the distance between the codebook vectors and the latent output of the Encoder (Figure 2.2.A). The second module of VQVAE is an autoregressive model that benefits from the discreteness of the element-wise discreteness of the latent representation to learn the underlying probability density function of the data distribution sequentially. In the original VQVAE paper (Van Den Oord et al., 2017) a PixelCNN is employed (Tim Salimans, Andrej Karpathy, Xi Chen, Diederik P. Kingma, 2017), but novel autoregressive models such as the transformers (Vaswani et al., 2017) have been demonstrated to outperform the other autoregressive models in this setting (Esser et al., 2020).

2.1.2.6 Decoder-only Transformer

Decoder-only transformers, here named only as transformers for simplicity, are a subclass of autoregressive models that leverage blocks of stacked self-attention layers to model data dependencies without regard to their distance in the input sequence, resulting in an improved performance when compared to other autoregressive models (Vaswani et al., 2017). The attention layer consists of the dot-product of a Query matrix, Q, with the Key matrix, K, divided by a scaling factor, $\sqrt{d_k}$ and passed through a softmax function. This output is used as weighting for the Value matrix, V. The Query, Key and Value are all parameterised matrices obtained from the dot-product of the layer's input and the respective weight matrix, W_Q, W_K and W_V. The output of the attention mechanism is then obtained through the following mathematical function, where d_k is the number of rows in K:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{2.14}$$

The model has no preconception of the data ordering, as all previous inputs are preprocessed simultaneously, so a positional encoding is added to the input sequence. In an autoregressive setting, the transformer tries to predict the next value in a sequence x_i, while receiving as input all previous values in the sequence x_0, ... , x_{i-1}. The output of the transformer is then passed through a softmax function where the values of each possible output sum up to one and can be interpreted as the probability of being the correct output. The autoregressive transformer is trained to maximise the log-likelihood of the input sequence, s, by minimising the following loss function:

$$Loss_{Transformer} = \mathbb{E}_x[-log\, p(s)] \tag{2.15}$$

The trained autoregressive transformer outputs a probability distribution over the possible next values in a given sequence (see Figure 2.2.B). For a generative problem, the VQVAE uses the autoregressive algorithm (in this case, the transformer) to sample a sequence of discrete latent encodings that follows the probability distribution of the training data. These sequences of the latent codes are then passed through the generator to generate realistic representations of the data without copying any real example. Another example of the application of the VQVAE is for outlier detection. As the Transformer learns the explicit probability distribution of the latent code, it is possible to use the likelihood associated with the real data to assess if the Transformer is identifying the sequence as unlikely. If the transformer is only trained on a specific distribution of data, inferencing on a data point outside of the distribution (i.e., an anomaly or an outlier) will identify the specific latent sequence with low probability.

2.2 Bayesian Optimisation

Bayesian Optimisation (BO) is a powerful sampling algorithm that efficiently finds extrema of unknown functions, $f(x) = y$. It does so by employing a statistical model that is fitted to the sampled values and a function that guides where

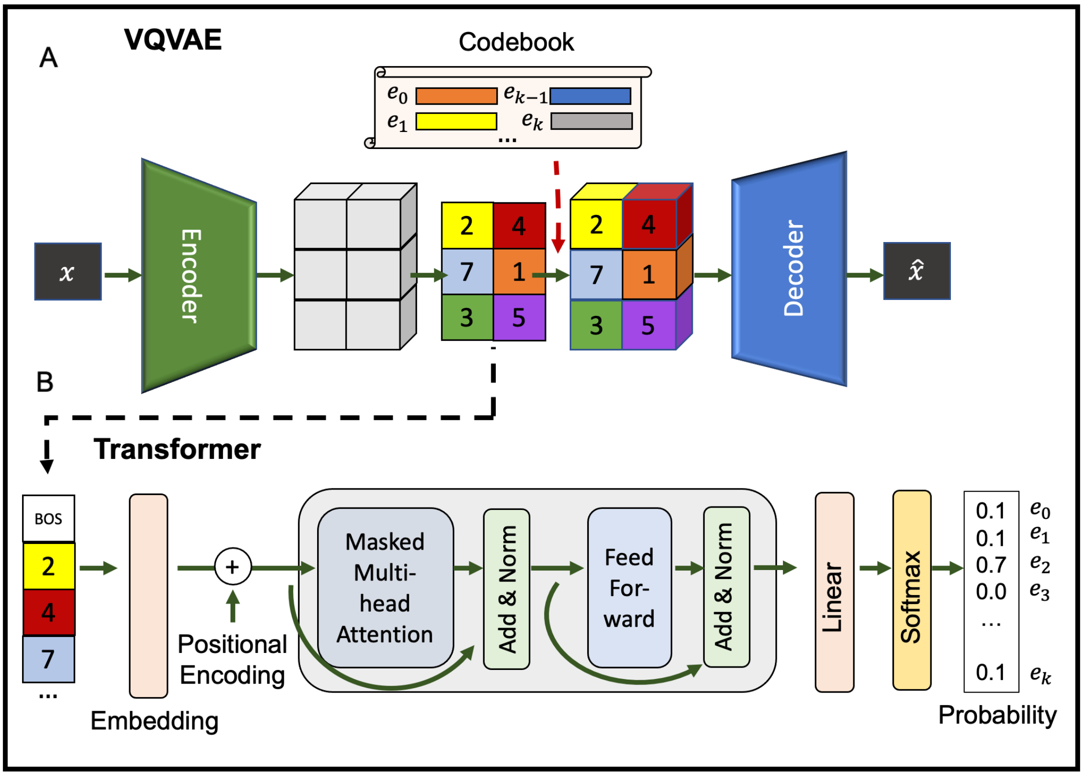

Figure 2.2: Schematic representation of the VQVAE and Transformer algorithms. *Schematic representation of the VQVAE encoder-decoder algorithm. The dVAE learns a discrete latent representation by optimising the reconstruction of the input (x) and the codebook to minimise the samples' distance to the latent codes. B. Schematic representation of the autoregressive Transformer. The Transformer receives sequentially as input the latent codes of the dVAE and learns the probability distribution of the next value, conditioned on the previous values of the sequence and starting with the BOS (begining-of-sentence) token. It does so using a multi-head attention mechanism.*

to sample next based on a balance between the values' uncertainty and the previously obtained values (Brochu et al., 2010). It predicts a posterior distribution across the space of functions using the available evidence (i.e., previously sampled points) and a prior. This method is particularly useful for optimising over costly functions where it is expensive to sample any given point. This is because its balance between exploration and exploitation allows the algorithm to find the function extrema in a small number of samples. The method performs global optimisation and does not require the unknown function to be convex, nor does it depend on gradient-based methods. These two characteristics benefit greatly the application of BO to real-world problems as the optimised unknown functions rarely are derivable and smoothly convex. Bayesian optimisation is composed of two main parts: 1) the surrogate model that fits the data in a supervised learning framework; 2) the acquisition function that determines which point to be sampled next based on maximising utility.

2.2.1 Surrogate model

The surrogate model is a statistical model of the unknown objective function, $f(x)$, which learns a mapping of the input x to the output y and the distribution of uncertainty around the output predictions (Figure 2.3.A). Typically, a Gaussian Process Regressor (GPR) is used as the surrogate model. This algorithm is used to build the statistical model based on previously sampled values, $GP(x) = p(x|y)$ (Rasmussen and Williams, 2018). The GPR fits a multivariate normal distribution to the feature variables $(x_0, x_1, \ldots, x_i)$, using the covariance matrix, $\Sigma = cov(x)$, to define how the features are correlated with each other. This allows the model to interpolate between data points following the probability density function of the multivariate normal distribution, with dimension D:

$$N(x|\mu, \Sigma) = \frac{1}{2\pi^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} exp[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)], \tag{2.16}$$

Where μ is the mean response vector, and the covariance matrix Σ contains the pairwise covariance of all jointly modelled random variables. In order to make the GPR extend to stochastic processes defined on a continuum, the covariance matrix is replaced by a covariance function or kernel $k(x_i, x_j)$ that predicts the covariance between any two data points. There are many different kernels that represent different relationships between data points. In our systems, I use the stationary Mátern kernel with a smoothness parameter ν=2.5 and an added white noise term. The Mátern kernel is a commonly used stationary kernel (Rasmussen and Williams, 2018), and the white noise term will simulate the global noise level as data acquisition is inherently non-deterministic and noisy. The kernel's hyperparameters are optimised during the fitting of the model by maximising the log-marginal-likelihood. For every new sample of the Bayesian optimisation, a GPR is fitted with all available data and the prediction of $f(x)$, and its standard deviation is retrieved. These are passed to the acquisition function to control where to sample next.

2.2.2 Acquisition Function

The acquisition function controls where BO should sample next by balancing both goals of minimising uncertainty (i.e., exploration) and finding the maximum of the predicted function (i.e., exploitation). The maximum of the acquisition function will determine the data point to sample next (Figure 2.3.B). We want to sample $f(x)$ at $\underset{x}{\operatorname{argmin}}\, u(x|\mu, \Sigma)$, where $u(.)$ represents any acquisition function. Two commonly used acquisition functions are Expected Improvement (EI) and Upper Confidence Bound (UCB).

EI is a function that leverages information about the expected best candidate and the uncertainty of the estimations in an exploration-exploitation trade-off (Kandasamy et al., 2016). It is given by the expectation of improvement function, I:

$$\mathbb{E}[I[x]] = \mathbb{E}[\,max(f(x*) - y, 0)] \tag{2.17}$$

As the surrogate function follows a normal distribution, we can compute EI in closed form:

$$\mathbb{E}[I[x]] = (\mu(x) - f(x*))\,\Phi(z) + \sigma(z)\phi(z), \tag{2.18}$$

where

$$z = \frac{\mu(x) - f(x*) + \kappa}{\sigma(x)} \tag{2.19}$$

Where Φ is the standard normal density, ϕ is the standard normal distribution function, and κ is a hyperparameter that controls how much the acquisition function privileges exploration over exploitation.

The UCB acquisition function was introduced by (Cox and John, 1992) and provides a cleaner balance between the aim for the global maxima and the minimisation of uncertainty:

$$UCB(x) = \mu(x) + \kappa\sigma(x), \tag{2.20}$$

Where κ is a hyperparameter that controls for the degree of exploration as it weights the importance of the standard deviation for the choice of the next point to sample.

The BO setting can be used for problems that focus only on minimising uncertainty and not on optimising a function, by controlling the κ hyperparameter to only account for uncertainty. This setting is known as active learning. Because the acquisition function requires a statistical model of the target metric across the space to guide where to sample next, we need to pre-define what the first samples of the algorithm will be (i.e., burn-ins). This approach is adapted from Automatic Machine Learning research (Misir and Sebag, 2013) and allows relevant heuristics to be added to a model prior to being optimised by the BO algorithm. The number of burn-ins and where they sample is user-defined and should be adapted depending on the paradigm being addressed.

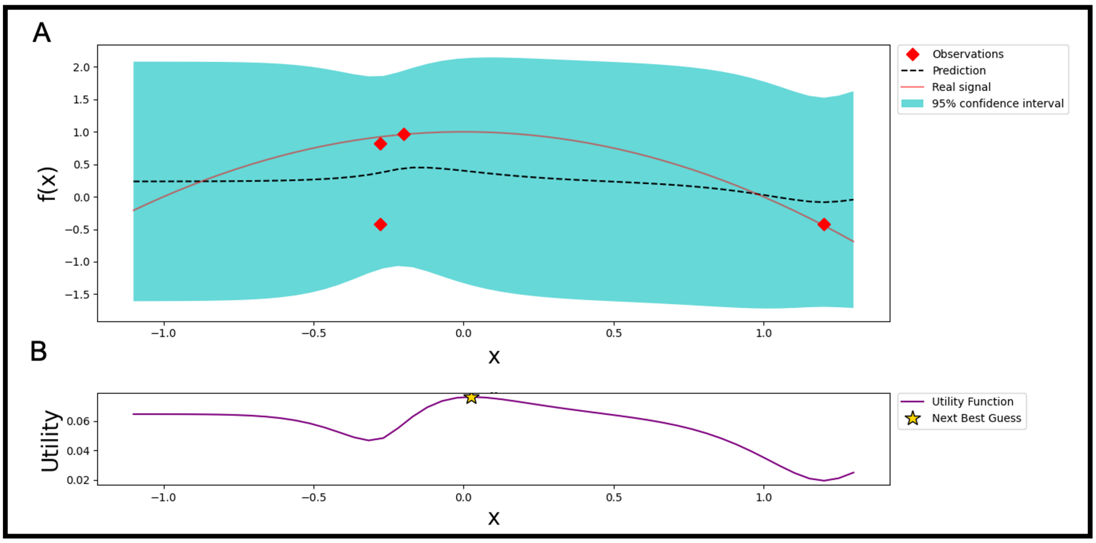

Figure 2.3: Fifth iteration of a Bayesian optimisation algorithm to find the maximum of a noisy parabolic function **$(-x^2 + 1)$*.*** *In A., the previous samples from the algorithm are marked with red diamonds, the Bayesian optimisation prediction of the real function is marked in a dashed black line, with a 95% certainty boundary marked in blue, and the real function that is being sampled is marked in red. The prediction of Bayesian optimisation consists on fitting a Gaussian Process to the already sampled points. Although the prediction of Bayesian optimisation is not a good match to the real signal, mostly due to the noisy samples, the algorithm correctly predicts the maxima to be around 0. The uncertainty around the areas already sampled is lower than those that were not yet sampled. B. presents the acquisition function, which controls where Bayesian optimisation should sample next. The golden star is the maxima of the utility function and the position that will next be sampled in* $f(x)$*. Because this is an exploitative run, it prioritises the region with the highest value and is sampling very close to the real function maxima (i.e., 0).*

Table 2.1: *Benefit and project applicability to each method described in this chapter.*

Method	**Benefit**	**Chapters**
PCA	Simple linear dimensionality reduction	5, 6
MDS	Non-linear dimensionality reduction	5, 6
StyleGAN	Learning manifolds with relevant features from hyperspace of available datapoints	3 4
VQVAE	Data compression and decompression	7
Transformer	Likelihood estimation for finding unlikely samples	7
Bayesian Optimisation	Active learning and optimisation on a Euclidean space	3 4, 5, 6

3 Neuroadaptive Optimisation of face stimuli in human-centred research

3.1 Introduction

The replication crisis and the generalisability crisis previously introduced in 1.1.1 are challenges faced by most natural science research. Some of the practices that lead to these causes include researchers having too many degrees of freedom when conducting their research, especially in the case of exploratory studies, and most studies not accounting for individual variability in their population of study when doing group-wise analysis. As introduced in 1.2.2 the neuroadaptive optimisation approach attempts to tackle these limitations by employing active learning methods to automate the process of data collection, data processing and result analysis in more open-ended hypotheses and retrieving individual profiles of response for each participant in the study. These topics have been explored in fMRI studies (Lorenz et al., 2016) and this chapter extends this approach for behavioural responses. It. explores responses over a space of face stimuli, as faces have a key role in our understanding of brain representations (Tsao and Livingstone, 2008) and have been considered a relevant biomarker for psychiatric conditions (Jones et al., 2019). Furthermore, this chapter introduces a novel mechanism to enrich the space of stimuli using generative models to create continuous variations between variables of interest.

Face perception and processing is fundamental for human survival. Within a fraction of a second, faces reveal to us information about the emotions, gender, age, trustworthiness or intention of another human. Therefore, faces are among the most important visual stimuli in the natural world and, consequently, a large portion of neuroscience and psychology research has been dedicated to studying face processing mechanisms (Eimer, 2012, Kanwisher et al., 1997, Kanwisher and Yovel, 2006, Tsao and Livingstone, 2008). As a result, we now know humans have a specialised neural mechanism to process faces that is in influenced by their individual experience (Pascalis et al., 2011) Furthermore, neuroimaging studies have shown that different face stimuli elicit different brain response patterns (Kriegeskorte et al., 2007). This heterogeneity in our neural response to faces presents a challenge to current methodology in the field, where the status-quo consists of using the same set of pre-selected face stimuli for every individual and then drawing conclusions from group-level results. The absence of personalised stimuli presents a serious limitation as it fails to account for how each individual face processing system is tuned to cultural embeddings or how it is disrupted in disease. By performing group-level analysis on a subset of the general population, results may fail systematically to generalise to different populations or even different acquisition devices. Besides not allowing to tailor face stimuli to specific research questions (e.g., what kind of face stimuli maximise response in each brain region), this approach overlooks inter-individual differences in face processing by averaging over the individual signal across a predefined group. If we want to better understand the mechanisms underlying face processing, how it develops and how it is disrupted (e.g., autism spectrum disorder or fronto-temporal dementia), we need an approach sensitive to individual responses.

To address this shortcoming, this project presents a framework that leverages the neuroadaptive framework and generative models to tailor face stimuli with the aim to maximise a particular response from an individual subject (e.g., neural, behavioural or subjective)*. By requiring a small number of iterations, this approach bypasses the inherent limitation of participants' attention and familiarity effects from repeated testing. This closed-loop and automated approach measures how the manipulation of face stimuli alters evoked measures. For this, a continuous space was created, where each dimension manipulates a facial semantic attribute orthogonally from the other facial attributes. The algorithm automatically searches through this ”face space” using Bayesian optimisation that queries only the most informative points in that space in order to find the maximum of a target function. The target function can be neural, such as the participant's brain signal while processing the face stimulus or a behavioural evaluation, such as similarity to a target face or aesthetic judgement. The face stimuli are generated by a generative adversarial network (GAN) previously introduced in Section 2.1.2.1. GANs are effective at image manipulation because

*Code access to the full framework is available at github.com/PedroFerreiradaCosta/FaceFitOpt

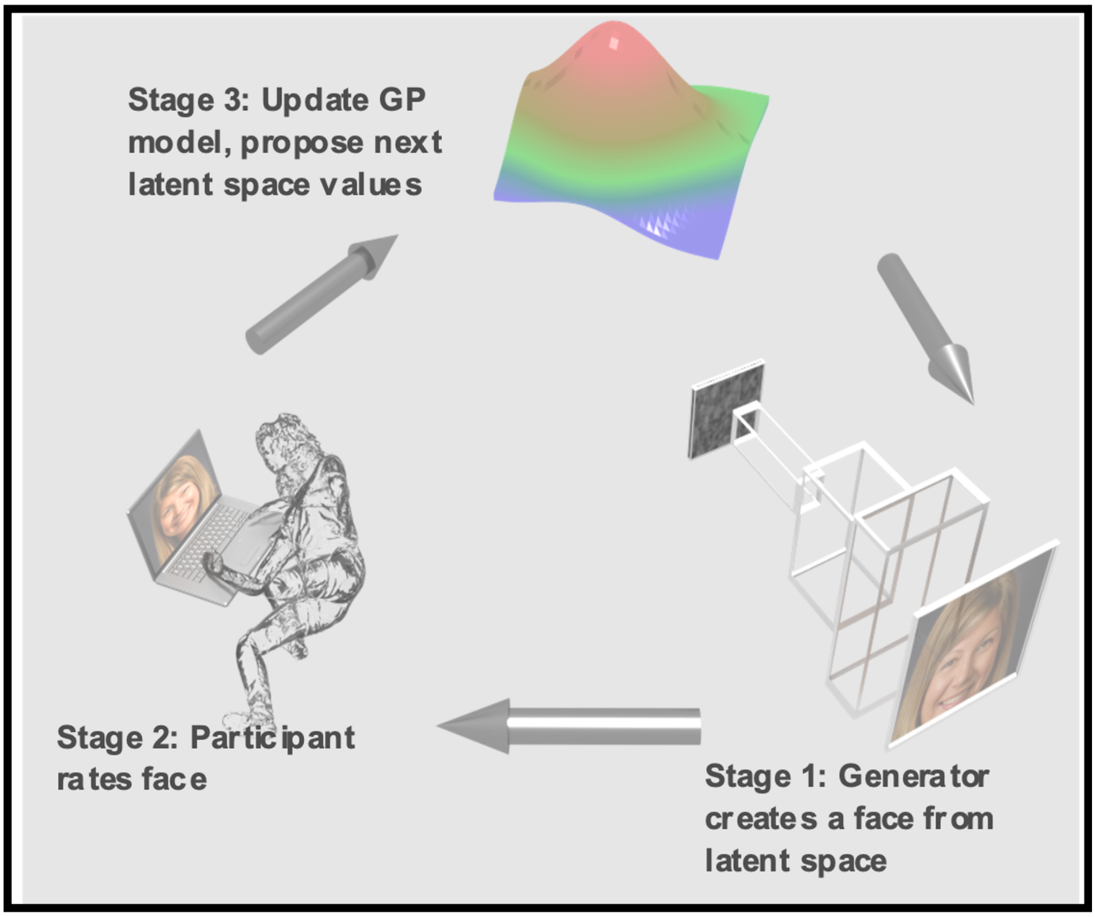
Stage 3: Update GP model, propose next latent space values
Stage 2: Participant rates face
Stage 1: Generator creates a face from latent space

Figure 3.1: Schematic Representation of the framework.

they create an unsupervised separation of semantic features (such as gender, age, etc) (Goodfellow et al., 2014). When the network is trained on images of faces, the latent space is transformed by the generator component of the GAN into a point in the low-dimensional manifold of realistic faces (Goodfellow et al., 2014, Karras et al., 2019). By moving the point along a vector in the low-dimensional manifold, we can manipulate the image along certain facial attributes while maintaining face identity (Pumarola et al., 2018), presenting a continuous mapping of these attributes, which would be impossible to obtain from any dataset.

This proof-of-concept tests whether the approach can identify an individual's own face by manipulating the age and emotion of an original photograph and considering the ground truth to be the non-manipulated image in the space. It is shown how the algorithm can efficiently locate an individual's optimal face while mapping out their response across different semantic transformations of a face. Finally, inter-individual analyses suggest how the approach can provide rich information about individual differences in face processing.

3.2 Methods

The introduced algorithm has four main components: 1) a pre-trained GAN to sample from the face manifold; 2) a face encoder that allows to obtain the latent representation for any real face in order to find its position in the manifold; 3) learned attribute directions from the latent space to manipulate images; 4) a Bayesian optimisation algorithm that efficiently samples the space.

3.2.1 GANs for object generation

GANs algorithms are introduced in Section 2.1.2.1 where it is described how this generative model implicitly learns the density function of a variable of interest (in this case faces), from the large hyper-dimensional space of possible images. In essence, the algorithm learns the lower-dimensional manifold using an adversarial process. By sampling from the learned density function, it is possible to generate realistic samples of the data it is trained on. Here, StyleGAN (Karras et al., 2018) was used, which was pre-trained on the Flickr-Faces-HQ dataset (FFHQ). The input is set to the intermediate latent space of the mapping network ($d_{latent} \in \mathbb{R}^{18.512}$) since it provides a more disentangled representation of the features. This is possible because StyleGAN is optimised for maximising the disentanglement of features in the latent space by using a network that controls the generator through affine transformations.

3.2.2 Latent Space Encoder

In order to manipulate images that the generator model was not trained on (e.g., a photo of the participants' faces taken by their webcam), we need to project the image from the manifold of face images, towards the learned latent space. GANs consist of multiple layers of non-linear transformations, which makes it challenging to invert the model (Abdal et al., 2019, Creswell and Bharath, 2019). Using *styleganencoder*[†], this is done by projecting both the image we want to transform and the generated images into a common feature space of a perceptual model - *conv3_2* of the VGG16 pre-trained on the ImageNet dataset (Simonyan and Zisserman, 2015). Then, the latent values are optimised through Gradient Descent on the perceptual loss for 500 iterations, where $F(I)$ is the output of the feature space, and I is the image input.

[†] github.com/Puzer/stylegan-encoder

$$L_{percept}(I_1, I_2) = ||F(I_1) - F(I_2)||^2 \tag{3.1}$$

This process results in a latent representation that, when fed into the generative network, outputs an image that is nearly identical to the original one (Bojanowski et al., 2018). As a demonstration of the reliability of the encoder, the results of an encoding from the face space to the latent space are presented in Figure 3.2. By learning the latent representation of a figure, the researcher can manipulate it by applying linear transformations in the latent space to transform semantic attributes of the original image. The constructed space is not limited to the semantic directions described in the chapter as Figure 3.2 demonstrates.

3.2.3 Attributes across latent space

In order to manipulate specific features while maintaining facial identity, I follow the methodology introduced in 2.1.2.3 where a large compendium of faces is generated from the GAN and automatically labelled to identify concept vectors of use. In this work the vectors of interest were age (i.e., older vs. younger) and emotion (i.e., sad vs. happy). This results in a 2-dimensional paradigm space where the faces are mapped to, covering emotion on the x axis and age on the y axis. By maintaining facial identity while varying these two features, the algorithm is setting all other facial variables that are uncorrelated to the variables of interest as fixed.

3.2.4 Bayesian Optimisation

Bayesian optimisation is ideally suited to perform optimal stimulus selection in the context of neuroscientific research (Lorenz et al., 2016, 2017, 2018) because a) evaluating all possible stimuli is not feasible with human participants, b) the target functions are unknown (e.g., structure, concavity, number of maxima or linearity), c) the sampled values are "derivative-free", limiting the use of any gradient descent approaches, and d) the neural or behavioural samples will inherently be affected by stochastic noise. The set A, the face space, is a hyper-rectangle $\{x \in \mathbb{R}^d : a_i < x_i < b_i\}$, where each dimension d manipulates one disentangled facial feature across its axis. The choice of features is flexible and should be adapted depending on the specific research question. This is the space that the Bayesian optimisation will navigate, where a point in the space represents an image generated by the generator network with the given latents $(d_{latent} + \sum c_i\, x_i)$. As for the acquisition function the upper confidence bound

(UCB) was chosen (Cox and John, 1992) with a more explorative hyperparameter.

3.2.5 The framework

The proposed framework is a combination of these four algorithms to automatically explore the face space and find the maximum of a target function that varies across the chosen feature manipulations. It can use as input any real face, which is automatically encoded into its latent representation by minimising the differences of the generated image and the real face in a perceptual space. The target function is first evaluated after 5 burn-in samples, uniformly chosen at random to fill the space. Each point is converted to an image through the generator network, displayed to the participant and the response is measured and fed back to the algorithm. After the initial five iterations, the loop is closed by the Bayesian optimisation algorithm automatically choosing the points to sample for the next 20 iterations, with each point sample following the same steps as before (see Algorithm 1). A combination of a Mátern $\frac{5}{2}$ kernel and a white noise kernel was used to allow for noisy inputs (Rasmussen, 2004). The algorithm was wrapped around a GUI that was run with Google Colab to take advantage of Google hardware to run the generative model. Its automated and flexible process allows any user with an internet connection to run the software from end-to-end.

Algorithm 1 Framework pseudo-code

```
procedure BURN-IN
    while $n < 5$ do
        randomly sample $x_n$ from $A$.
        $dlatent_{new} = dlatent + c * x_n$.
        Run $G(dlatent_{new}) = image\_stimulus$.
        Observe $y_n = f(x_n)$.
        Increment $n$.
procedure B.O.
    while $n < 25$ do
        Update posterior probability distribution on $f$ using sampled points.
        Let $x_n$ be the maximiser of the acquisition function.
        $dlatent_{new} = dlatent + c * x_n$.
        Run $G(dlatent_{new}) = image\_stimulus$.
        Observe $y_n = f(x_n)$.
        Increment $n$.
    return sampled point with largest posterior mean
```

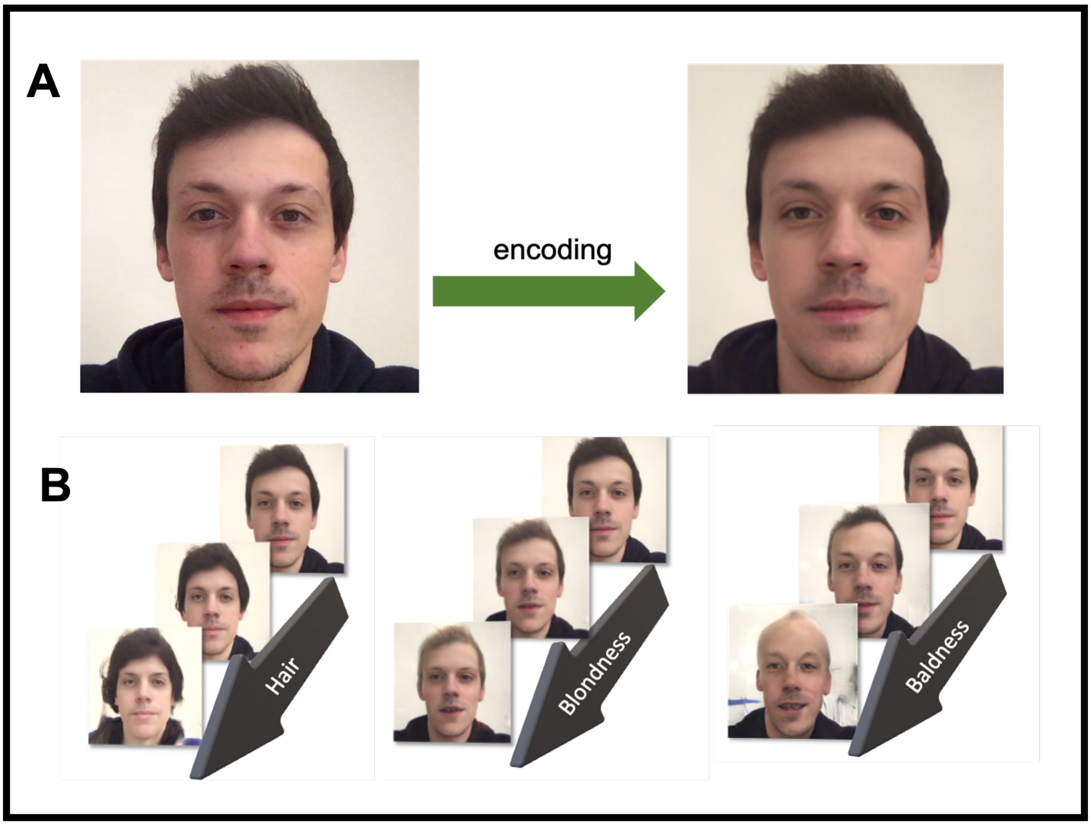

Figure 3.2: Encoding of the real photo into the learned representation. *A) The image on the left is a photograph. The image on the right is generated from a multidimensional datapoint in the latent space that was chosen by optimising the latent values through gradient descent on a perceptual loss. These results were obtained with 500 epochs, which took 7 minutes to run on Google Colab. B) Examples of axis in the latent space resulting in different semantic transformations of the generated image.*

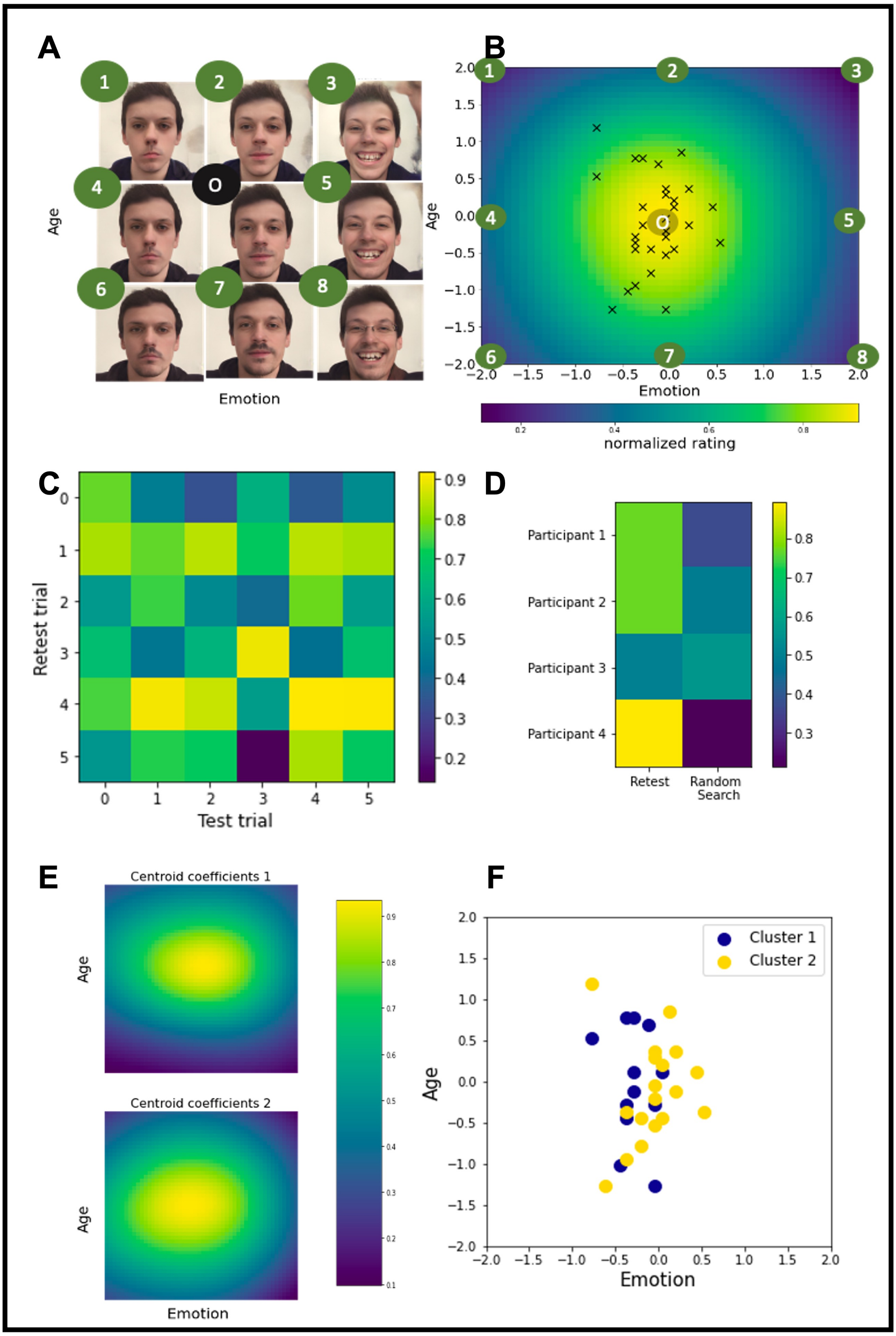

Figure 3.3: Obtained responses mapped by individual and global patterns.

Figure 3.3: *A. Images displaying the extremes of the space (1-8) and the origin point (O). B. Mean response of the 30 participants. Each individual maximum response is marked with a cross. C. Similarity matrix between the target space of different runs of the same participant (matrix diagonal) and between different participants. The mean Pearson correlation between trials of the same participant is 0.76, where the correlation between trials of different participants is 0.64. D. Correlation matrix between the predicted target space of the first run and the second run (first column) and between the first run and a run using random search across the space instead of Bayesian optimisation (second column). E. Cluster Analysis using K-means clustering on the full space of the participants' evaluations, the centroids for the two clusters are identified. The first centroid captures a higher dispersion towards positive values of age (younger images) and the second centroid captures a higher dispersion towards negative values of emotion (angrier images). F. Maximum of each participant labelled according to its cluster.*

3.3 Proof of concept study

To demonstrate the framework, a web-based behavioural study was conducted with 30 participants (14 female, mean ± sd age: 31.33 ± 13.94 years) in which they had to rate manipulated photos of themselves. The aim was to quickly identify the face stimuli that maximally resembled their original, non-manipulated photo as an exercise in self-perception. The hypothesis being tested was that the capacity for self-perception of a person's own face varies across individuals (Strauss and Kaplan, 1980). For this, a face space composed of two dimensions, age and emotion, was defined where each axis is a linear variation of these features across the latent vector. A negative value corresponded to an older version in the age axis and to an angrier version in the emotion axis. Each participant took a photo that was encoded into the latent space. At each of the 25 iterations, participants were shown a manipulated image of their original photo and were instructed to rate the similarity between them (0 being nothing alike and 10 being exactly like their original photo). Each manipulated image corresponded to the transformation associated with the point sampled in the space. This study design allowed to benchmark the algorithm's performance against a known ground-truth (the non-manipulated image in the space). In addition, for six participants, further runs were conducted to assess the algorithm's test-retest reliability (first and second run) and compare the algorithm's performance against random search (third run).

The results showed that the maximum is more dispersed on the age axis than on the emotion axis, although the median response tends to be near the origin of the space (median ± sd for emotion: -0.04 ± 0.15; age: -0.06 ± 0.30). The test-retest reliability analysis showed a high intra-subject spatial correlation (mean Pearson correlation coefficient across participants ± sd: 0.76 ± 0.14); higher than the mean inter-subject correlation between participants' response patterns (0.64 ± 0.19). This result seems to sustain the argument that the framework might be able to capture personalised responses on self-perception. To analyse this further, k-means clustering was performed for two clusters on the full space predictions of the participant's response. The silhouette score was 0.17. The results are displayed in Figure 3.3.E. An analysis of the correlation of the test runs with the re-test runs and a run not using the sampling algorithm (i.e., using a random search algorithm) shows that the correlation between the two formers (mean ± sd.: 0.74 ± 0.14) is higher between the test and the random-search patterns (mean ± sd.: 0.41 ± 0.13). The results are presented in Figure 3.3.D.

3.4 Discussion

This project proposes a new tool to automatically generate and manipulate face stimuli across several semantic directions in a well-controlled manner. It was shown that after only a few iterations it is possible to identify the optimal face stimulus to maximise a target response and can accurately predict the individual's response across the entire face space. Importantly, it was shown that response patterns are more stable within individuals than across participants. This suggests that there might indeed be inter-individual response patterns. This is relevant as high intra-subject reliability is a critical prerequisite for this method's validity as one of the objectives is to create a profile of a person's response. Despite the results agreeing with our hypothesis that inter-individual variation in face self-perception could be captured through this method, further studies and larger cohorts are required to better asses these individual variations.

This approach is relevant for a wide range of disciplines interested in an individual's response to faces (e.g., neuroscience, psychology, psychiatry, marketing). In a clinical setting, altered response patterns to faces could be used to guide diagnosis or patient stratification for neuropsychiatric conditions known to affect face processing (e.g., autism spectrum disorder (Golarai et al., 2006), fronto-temporal dementia, eating disorders (Phillipou et al., 2015) or schizophrenia (Bortolon et al., 2017)). In experimental neuroscience, it allows us to identify a set of face stimuli that evoke similar brain responses but bypass effects of habituation. For psychology, it could be used to investigate how different emotions or personality traits might result in different response patterns.

The space is not limited to be 2-dimensional and there is no limitation on the types of images that can be presented. GANs have been used to learn different manifolds (e.g., houses, animal faces), which could be used to create a navigable space following the same framework. Equally, sounds could also be optimised in the same way. Regarding limitations of the stimuli, the extremes of the space will sometimes display distorted images. One reason is that we are interpolating linearly between categories in the latent space, where a non-linear transformation would be able to better capture the transition across the axis. In conclusion, this framework offers a novel tool for human-centred research that can address limitations from group analysis by including the assessment of individual responses.

4 | Neuroadaptive electroencephalography

4.1 Introduction

The general methodology behind the neuroadaptive framework does not limit it to a single data modality or type of data space. It can be extended to any measured response given that it can be evaluated automatically and in real-time. In this chapter, I extend the neuroadaptive optimisation framework for EEG studies. Specifically, this chapter explores an infant's paradigm where an ERP associated with face stimuli is studied along the spectrum of images of mum and a stranger. Here it is studied the individual response across the space and how efficient sampling of the space contributes for avoiding problems of stimulus habituation. Understanding how the brain processes and represents the physical and social environment is one of the fundamental goals of functional neuroimaging. Decades of research have yielded a range of methodologies for studying the electrical activity (electroencephalography; EEG), magnetic activity (magnetoencephalography; MEG) and oxygenated haemoglobin changes (functional magnetic resonance imaging/MRI and near infrared spectroscopy/NIRS) that are associated with neuronal activity in the human brain. Further, cognitive neuroscience has generated a rich tapestry of neural metrics suitable for assessing response to environmental stimulation. These include measurement of task-induced changes in oxygenated haemoglobin concentrations in spatially defined regions or networks (fMRI/NIRS, and timing-defined event-related neural potentials or oscillations (EEG/MEG). These metrics were identified through experiments in which a researcher measured a broad spectrum of brain responses (e.g., whole brain EEG or fMRI) to a small pre-selected stimulus set. Often, the selected stimuli are important environmental cues like faces (Tsao et al., 2006, Tsao and Livingstone, 2008), or basic auditory or visual features designed to represent the building blocks of perception (like checkerboards or tones). Such research has yielded many critical insights into metrics relevant to brain function (Kropotov and Kropotov, 2016).

Despite this progress, there are increasing questions over the value of traditional stimulus-driven methods for understanding the 'meaning' of these brain metrics (i.e., what dimensions they represent); and for studying how brain function differs in populations who are not the modally studied groups of heteronormative White western young adults. First, understanding what each brain metric 'means' involves defining the range of stimuli by which each brain metric is modulated; doing this sequentially through separate experiments that each focus on one or two stimuli is slow and inefficient. Second, stimulus selection is often guided by theory or previous empirical study; but the replication crisis (Ioannidis, 2014, Open Science Collaboration, 2015) and the overwhelming focus on a narrow subset of the world's population in the neuroimaging literature (Westfall et al., 2016) means that trying to select the right stimuli to study individual differences or brain function in broader populations may be little better than guessing. The substantial analytic flexibility afforded by allowing post-experiment analysis of brain data creates an overwhelming risk of false positives and can only be partially addressed by written pre-registration (Nosek et al., 2015). Finally, the traditional approach embeds a deficit model in our approach to studying individuals with neurodevelopmental or psychiatric disorders. A typical research project in this area involves studying how people with neurodevelopmental or psychiatric disorders respond differently to a stimulus selected based on normative preferences. Reframing this work within a neurodiversity framework (Singer and Willett, 2009) prompts the researcher to ask not how people's brains respond differently to stimuli that 'typical' brains prefer, but to identify what stimuli are preferred by people whose brains work differently. Longer term, this may prove more fruitful for the design of individualised interventions that build on strengths, rather than aim to address weaknesses.

To address these problems, Leech and colleagues developed a complementary approach that inverts the traditional experimental paradigm (Lorenz et al., 2017). Instead of preselecting one or two stimuli and measuring a broad spectrum of brain responses, neuroadaptive Bayesian optimisation is a method through which the experimenter selects one or two brain responses, and measures how they are modulated by a broad spectrum of environmental stimuli in a real-time closed loop design. This method has been used successfully to study fMRI responses in the frontal cortex in neurotypical adults (Lorenz et al., 2017), and to identify cognitive difficulties in stroke patients (Lorenz et al., 2021). Here, this approach is extended to EEG and the study of infant brain function. EEG is a particularly fruitful method for real-time analysis because its high temporal resolution allows rapid feedback loops of response-guided stimulus selection. Further, the problems of traditional approaches are particularly acute when studying the developing brain, where the cognitive topography is likely to be substantially different than in adulthood and where practical challenges of working with infants make data collection slow and difficult. However, without studying the brain as it develops, we cannot move beyond studying the correlation between environmental features and brain response. To understand the mechanisms of causation through which the environment is represented and shaped we must study change over developmental time in how the brain processes information and controls behaviour. Insights from translational work in animal models and computational modelling approaches to studying learning are also most likely to be effectively mapped onto preverbal infants, where common mechanisms are most likely to be conserved. Thus, in the present study I present a proof of principle of the use of neuroadaptive techniques to study the developing brain.

As a test case, a stimulus space in which we could make a strong prediction about individual level brain 'preferences' was selected. Specifically, a face space within which the face of the infant's mother was positioned. If the algorithm can automatically converge to identify the infant's own mother based on real-time closed loop analysis of their brain activity, this provides a strong proof-of-principle that the technique can be used to study how a particular brain metric reacts to a broad environment. Notably, this is different to a typical BCI approach in which an algorithm would first be trained to distinguish (for example) two faces on the basis of multimodal brain data; here, the neural feature is selected based on the previous literature for greater interpretability. In this case, the negative central (Nc) event-related potential response was selected because of its demonstrated links to attention (Richards et al., 2010) and its modulation by face familiarity (including differentiation between mother and stranger (de Haan and Nelson, 1999, M. De Haan and C. A. Nelson, 1997, Webb et al., 2011)).The infant Nc (negative component) is a negative deflection occurring between 300 and 800ms at the frontal midline after the stimulus onset. Previous studies showed that the amplitude of the Nc was larger (more negative) in response to the mother's face (de Haan and Nelson, 1999, M. De Haan and C. A. Nelson, 1997) likely reflecting elevated attention (Conte et al., 2020). These studies use the traditional approach of preselecting two stimuli (mother and stranger) and analysing the resulting data with highly variable study-specific parameters (scalp location, window timing, peak or latency etc). To invert this paradigm, facial stimuli was generated based on the mothers and strangers' faces and created a configuration space. Then, it was presented one of the faces to the infants while measuring their EEG. Using real time EEG analyses, their Nc amplitude response to the stimuli was analysed. The Nc responses were then forwarded to the neuroadaptive algorithm that predicted which face would elicit the optimal Nc response (i.e., larger Nc amplitude) in the individual infant. When the algorithm converged on the stimulus eliciting the optimal Nc response, the experiment was automatically terminated. In the following sections, I describe the method and proof-of-principle results.

4.2 Material and Methods

The neuroadaptive method consists of multiple steps, outlined in the following sections. These are: generating stimuli and creating a configuration space (here a face space, 4.2.1); recording and performing real-time analysis of the neural (EEG) responses to the stimuli 4.2.2; using a sampling algorithm (Bayesian optimisation) across the space 4.2.3 and re-iterating steps 2 and 3 until a stopping criterion has been reached 4.2.4. This allows for a rapid prediction of the target EEG metric across the space despite sampling only a limited number of stimuli. Figure 4.1 presents an overview of these steps. In what follows the development and choice of parameters for each step are described in more detail. All scripts are available in an open-source repository*.

4.2.1 Generating stimuli and creating a configuration space

This method maps the peak and topography of the modulation of the pre-selected neural response across a large stimulus space. As such, a continuous experimental space needs to be generated that can be characterised along one or more stimulus dimensions (da Costa et al., 2020). Since the test case involved faces (a common focus of cognitive neuroscience because of their importance to social function), to allow for creating a smooth space interpolating between faces of strangers and participant's mothers, stimuli were artificially generated using StyleGAN2. StyleGAN2 is a state-of-the-art generative adversarial network (GAN), a deep learning algorithm for generative image modelling (Karras et al., 2019). One of the examples of successful trained GANs on a given data distribution is the

*github.com/PedroFerreiradaCosta/NeuroadaptiveEEG

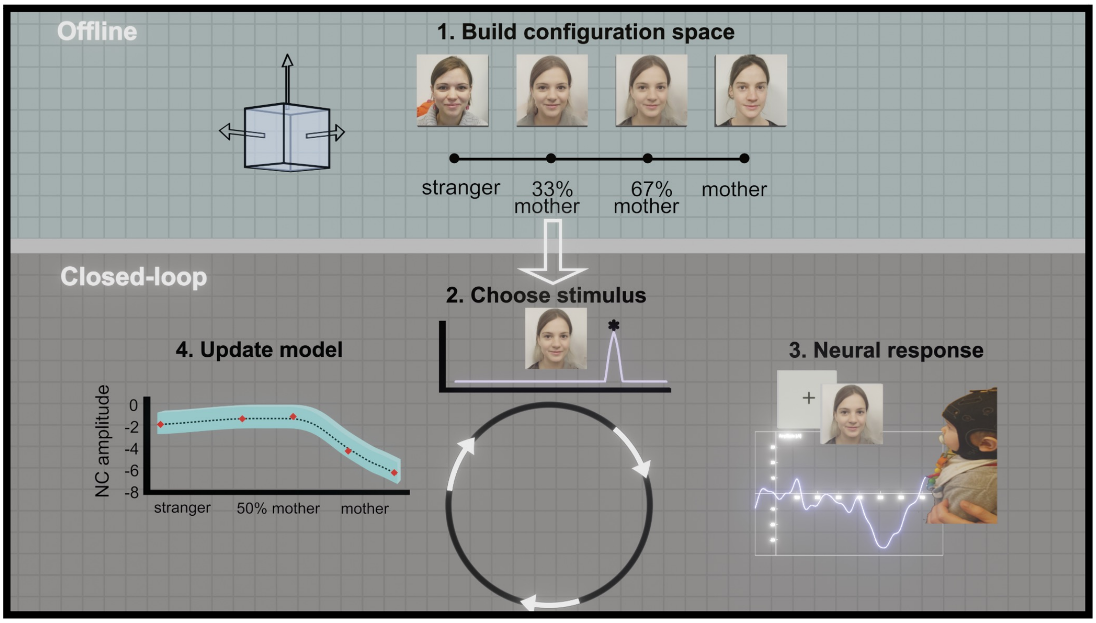

Figure 4.1: Schematic of the EEG neuroadaptive framework. *The first step is taken before the testing of a participant and consists of building the configuration space of stimuli (1.) – in this paradigm, mother-stranger interpolation. However, this could represent a stimulus space across a large number of dimensions. During testing, the framework is run in a closed-loop, where a stimulus in the configuration space is chosen to be sampled (2.). The stimulus is displayed to the participant and the EEG response is processed automatically to retrieve the target metric (3.) - in this paradigm, the Nc amplitude. A statistical model is built by fitting a Gaussian process to the sampled data (4.), guiding where to sample in the next iteration (2.).*

generation of realistic faces (Karras et al., 2018, 2020). The generative algorithm can create a large and diverse compendium of non-existing faces that are indistinguishable from real photos. Exploring the trained latent manifolds of generated faces allows to create smooth spaces defined by relevant dimensions along which faces continuously vary. By using artificially generated faces we are not limited by a bounded dataset and can maximise diversity of the generated faces, important to expanding research studies to diverse cultures and ethnicities. Furthermore, the generated images can be manipulated to create realistic representations of faces that progressively change across a given semantic dimension, which provides an alternative way to define a 'space' (e.g., changing a given face's perceived age) (Radford et al., 2016). Finally, as the generated faces do not depict real people, there is no risk of privacy infringement or limitations on proprietary datasets.

An additional use of GANs is to project real images of faces to the latent manifold. This facilitates the manipulation of real images across selected semantic dimensions (i.e., a meaningful dimension in the latent manifold). This allows us to create a dimension involving a real person (here the infant's mother) by following previous implementations of image encoding (Bojanowski et al., 2018). Specifically, both the GAN output and the image to be encoded are projected into a common feature space, encoded by a perceptual model – an intermediate layer of the image classifier VGG16 (Simonyan and Zisserman, 2015). The latent codes, $d_{latents}$, are then optimised directly by gradient descent to try to minimise the perceptual loss. The perceptual loss is the difference between both projected images in the common feature space. Given the output of the feature space $F(I)$, where I is the image input, the perceptual loss can be given by the Euclidean distance between the two projections:

$$L_{percept}(I_1, I_2) = ||F(I_1) - F(I_2)||^2 \tag{4.1}$$

The resulting latent code that minimises the perceptual loss will be able to generate a similar recreation of the original photograph. As previous studies have demonstrated (da Costa et al., 2020, Radford et al., 2016), the latent space can be exploited to control for different facial features of the artificially generated faces. We can progressively change a given aspect of the face, while maintaining the other features mostly intact following the linear interpolation method described in section 2.1.2.3. Finally, the same linear interpolation can be applied to any two images generated from the GAN. By linearly manipulating two latent codes from one to the other, it is possible to obtain a continuous morphing of the faces, slowly changing their facial identity. These manipulations using the GAN latent space prove useful in neuroscientific research, as they allow for a flexibility in facial features, while maintaining realistic representations of the face. Instead of relying on categorical and discrete stimuli, these continuous variations can then be used for better assessment of how a given ERP varies.

These manipulations of artificial faces are done using Google's Colaboratory [†], an online Jupyter notebook environment that provides access to GPU computation, a requirement when doing inference with GANs. This module is run before the data collection, and it is responsible for generating the stimuli that will compose the space being analysed. In this module, the experimenter will define the dimensionality of the space, the semantic direction of each dimension and the original stimulus under analysis. The semantic direction can be chosen from a list of available options: gender, age, emotion, yaw, roll, pitch, lip ratio, nose-ratio, eye-ratio, eye distance, eye-to-eyebrow distance, nose-to-mouth distance, mouth open/closed, eyes open/closed, nose tip position and interpolation between two different faces (the option used in the present study). The present study interpolates between the faces of the participant's mother and a stranger. The original stimulus was an original photograph that is uploaded to the system; an alternative approach is to use a randomly generated artificial face. This configuration space builder is currently limited to facial stimuli, but it can be substituted by any stimuli that are predefined to model a configuration space to be analysed in real-time.

In this system, the configuration space is Euclidean, discrete and can be multi-dimensional. The number of dimensions, each of which should account for an independent variation of the stimuli, is only limited by the capability of the Bayesian optimisation to sample vast spaces, which is known to break down for spaces larger than 20 dimensions (Snoek et al., 2012). Each dimension should code a single and independent variation of the stimuli. This allows to disambiguate and eliminate confounders from the variation of the target metric along a given axis. The specific variations of stimuli being encoded per dimension should depend on the research questions being addressed.

In summary, in this proof-of-concept infant paradigm, there is a focus on brain functioning during visual processing of familiar and unfamiliar faces: i.e., the mother's face and a stranger's face. The stranger's face was the face of one of the researchers and was used across all infants. A stranger's face and a photograph of the mother of the infants were used as the extremes for the

[†] github.com/PedroFerreiradaCosta/NeuroadaptiveEEG/blob/main/FaceShiftBirkbeck.ipynb

mother-stranger continuum. In the photographs, the mother and stranger had a neutral expression, and the head was centred where the image was cropped at shoulder or clavicle height. Using the StyleGAN2, a continuous GAN latent space that represented the mother's face linearly changing into the stranger's face was generated. This resulted in 10 additional, realistic images of faces, bringing the total of possible sampled stimuli to 12 (see Figure 4.2).

4.2.2 Recording and real-time analysis of the neural responses to the stimuli

Participants were recruited via a database of families interested in research. Infants were aged between 5 and 9 months and excluded if they had a family or personal history of epilepsy, were born preterm ($<$ 31 weeks gestational age), had a clinical diagnosis or a sensory or motor impairment, or if they could not hold their head up without support. Information about the study was provided by email and informed consent was signed by the caregiver digitally (per Covid requirements). Mothers were asked to send a picture of themselves with a neutral expression before the visit. This picture was used to generate and prepare the face stimuli in advance. If a photograph was not received before the visit, a photo was taken in the lab with an iPad. Procedures were in accordance with the COVID-19 safety government regulations active at the time of data collection. The study and COVID-19 safety procedures were approved by the Department of Psychological Sciences ethics committee at Birkbeck (ref.no. 192001). Infants received a t- shirt as a thank-you.

Stimulus presentation. Stimuli were presented on a 24 inch diagonal screen (1080p; 1920 x 1200 pixels)) and controlled by a MacBook Pro (15-inch, 2018 with a 2.6 GHz Intel Core i7 processor) using Matlab (version R2018b). Stimulus presentation was controlled with Task Engine [‡] (Jones et al., 2019), Psychtoolbox 3.0.14, Gstreamer 1.14.4 for stimulus presentation, and a Lab Streaming Layer (LSL) to connect the EEG recording software to the Matlab software. A webcam (Logitech HD Pro Webcam C920) was placed on top of the screen. Open Broadcaster Software (OBS) was used to monitor the infants' looking behaviour during the session. An iPad (7th Generation) was used for taking photographs for generating the face stimuli and recording of EEG cap placement.

[‡] sites.google.com/site/taskenginedoc/

The experiment consisted of a series of blocks. At the start of each block, one of the face stimuli was selected to be sampled. This stimulus was repeatedly presented for a total of 12 trials per block; 12 trials were used in each block as this number of trials was 20% higher than the typical minimum trial number (10) used in infant Nc studies to allow for data loss (Munsters et al., 2019). Each block started with a red spiral to attract the infants' attention to the screen. When the infants were looking at the screen, the face stimuli were presented. This stimulus presentation was controlled with a key press by the researcher who monitored the infants' attention via the webcam. Each trial in the block started with a fixation cross presented on a grey screen for a duration of 1000ms for the first trial, and a jittered duration between 500 and 1000ms in subsequent 11 trials. The face stimulus was presented for 500ms. Immediately after, the next trial was presented. Whenever the infants were looking away, stimulus presentation was paused, and the red spiral was presented on the screen to re-attract the infants attention to the screen. A cartoon image of an object or animal was presented at the end of each block while the real-time EEG analysis and BO were performed.

EEG recording and analysis of the Nc response. During this procedure, EEG was continuously recorded using the Neuroelectrics Enobio with an 8-channel gel-based system (NE Neuroelectrics, Barcelona, Spain). The system was connected to the recording software Neuroelectrics NIC (v2.0.11.7, Barcelona, Spain) via wifi. CMS and DRL electrodes functioned as the system's reference. Data were recorded at a sampling rate of 500Hz. Since the Nc is most prominent at frontal-central sites (Courchesne et al., 1981) , EEG was recorded at six channels placed at locations FC1, Fz, FC2, C1, Cz, C2 in metal electrode holders in infant-sized caps (sizes K - 42cm, or KS - 46 cm). The two remaining electrodes were placed at locations P7, and Oz and functioned as the reference electrodes for re-referencing during the real-time EEG analysis. CMS and DRL electrodes were placed on the infants' mastoid using sticktrodes.

After the presentation of each block of stimuli, real-time EEG analysis started with reading in the markers and continuous EEG from the presented block. The continuous EEG data were filtered using an FIR digital band-pass filter from 1 to 20Hz with a Hanning window (as in Benedek et al. (2017), Webb et al. (2011)). Data were then segmented into trials based on the marker information from -100ms to 800ms after stimulus onset. Trials were baseline-corrected using the average amplitude across the -100ms to 0ms. Time series containing artefacts were identified on a channel by trial basis. Time series containing signals exceeding a threshold of -200μV or +200μV (Munsters et al., 2019), a range of 400μV, or showing a flat signal (absolute value below 0.0001μV) were marked and excluded from further analysis. Time series from all trials and channels of interest (FC1, Fz, FC2, C1, Cz, C2) were averaged together into one ERP. For the re-referencing, time series from all trials and the channels P7 and Oz were averaged together and subtracted from the ERP.

After the preprocessing, the Nc response was extracted from the ERP. The Nc response was defined as mean amplitude calculated across the time window from 300ms to 800ms. To measure data quality, the percentage of artefact-free trials included in the analysed ERP and number of trials included due to lack of threshold, range, or flat signal artefacts were also calculated. Both this data quality information and the ERP waveform were then displayed for visual inspection by the researchers.

Different criteria were explored on which this decision was based with different sessions (and different infants) and both automated and user-defined decisions. The preprocessed pipeline was fully automated. However, there is always the possibility of unforeseen interfering events or poor data quality that are not picked up by the automated pipeline. Interference may arise for a mother or experimenter accidentally disturbing the infant during a block, for example by talking and pointing at the screen. The signal of the EEG furthermore varies between individual infants where a certain threshold may be effective in one infant, but not pick up on the artefacts in another. In order to account for these possibilities, both automated and manual approaches to the decision were tried. In the automatic criterion, the block was included if the data quality was good. For the first infant, the following was implemented: Data quality was good if 1 or more channels were good (out of the 6 channels of interest). A channel was considered good if it contained 3 or more artefact-free trials (out of the 12 presented trials). In the remaining sessions, the manual criterion was implemented. This was a user-defined decision after each real time EEG block. In this researcher-based criterion, the researcher decided whether to continue, repeat or terminate the sampling. This subjective decision was based on the visual inspection of the ERP waveform, the percentage of data from the block included in the ERP.

4.2.3 Bayesian optimisation for sampling across the space

After each block, the relevant ERP metric (here, Nc mean amplitude) was passed to the Bayesian optimisation (BO) algorithm. The Bayesian optimisation sampling algorithm is run independently using python as the programming language.

Before the experiment, the user needs to define the number of burn-ins, where in the configuration space they are sampled, the maximum number of iterations run by the sampling algorithm and the level of desirable exploration (ξ). In the present study, the target EEG metric was the mean Nc amplitude. For four participants, it was optimised for its most negative value and for one participant it was optimised for its most positive value to verify that the resulting search pattern changed. It was defined that the first 4 iterations were burn-ins to Bayesian optimisation, to try to capture an initial model of the Nc amplitude's variation across the configuration space. The initial points sampled along the interpolation between the mother's face and stranger's face were in order, 100% mother's face, 100% stranger's face, $\frac{1}{3}$ mother's face and $\frac{2}{3}$ mother's face. EEG experiments in neurodevelopmental research are inherently limited regarding the number of iterations it can run for one participant in a given session. Because of the short-attention spans of infants, the method's parameters need to minimise the number of blocks the researcher must run. Towards this end, a ξ value of 0.1 was defined, which benefits exploitation of the identified maxima. One other reason for choosing a more exploitative ξ is the configuration space being relatively small with just one dimension of variation (i.e., mother-stranger interpolation). For larger spaces, higher values of ξ that benefit explorations should be considered. The maximum number of iterations was defined to be 15, which was identified to be an upper-bound for how many iterations the infants' managed to maintain attention (approximately 20 minutes). In the present study, the stopping criterion was always reached before the maximum number of iterations was surpassed.

4.2.4 Re-iterating steps 2 and 3 until an optimum target metric has been reached

The system is run iteratively in a closed loop. The acquisition function defines which stimulus to display next. The stimulus is presented to the participant and the target EEG metric (here the Nc amplitude) is collected after automatic processing of the ERPs. The acquisition function will progressively choose a stimulus to sample that is predicted to be closer to the predicted maximum until a stopping criterion is met or it has run a predetermined number of iterations. The early-stopping criterion objective is to finish the program if the BO algorithm is not capturing any new information on each block. The stopping criterion should then be a relevant proxy of the uncertainty present in the statistical model. This could be the mean standard deviation after each block, or the number of consecutive times a given stimulus has been sampled. If the sampling algorithm has identified its expected maximum (i.e., the image eliciting the strongest neural signature), then it will sample this point (i.e., present the selected image) until the predetermined number of iterations are run. In theory, increasing the number of blocks, i.e., obtaining the target EEG metric multiple times for the same stimulus, could be beneficial as it would allow the algorithm to average output across blocks and obtain a more reliable value. In practice, showing the same image to the infants would induce neural habituation and decrease the strength of the neural signal for the repeated stimulus; thus, these factors need to be balanced. To balance these considerations, in the present study a default stopping criterion of sampling the same image three times consecutively was used. If an image was sampled three times, it was assumed that the algorithm has converged to the unknown function's maximum.

4.3 Results

Good quality EEG data for analysis were collected from four infants (mean age = 6 months 4 days, range: 5 months 7 days - 7 months 22 days). One infant was excluded due to technical issues. Here, it is reported the results of the real-time analysis (4.3.1), a demonstration that the paradigm elicits expected effects when analysed in a traditional manner (4.3.2), and a discussion of two factors that need to be considered in the light of data collected – the balance between exploitation and exploration (4.3.3) and habituation (4.3.4).

4.3.1 Results of the real-time optimisation in infants

The method allows to build a statistical model of each participant's response across the configuration space by fitting their sampled responses with a Gaussian Process. In the problem presented here, the configuration space was a one-dimensional interpolation between the face of the infant's mother and a stranger. The signal was optimised towards the largest, i.e., most negative, amplitude of the Nc response (i.e., the minimum value obtained for the Nc). For all four participants the stopping criteria was met with a mean of 8.25 iterations and 1.64 of standard deviation, well below the number of possible stimuli. Furthermore, some images were sampled more than once, to allow the statistical model to assess if they indeed provided a larger Nc, as more samples of a given stimulus allow for a better estimate of the real elicited ERP. The results are displayed in

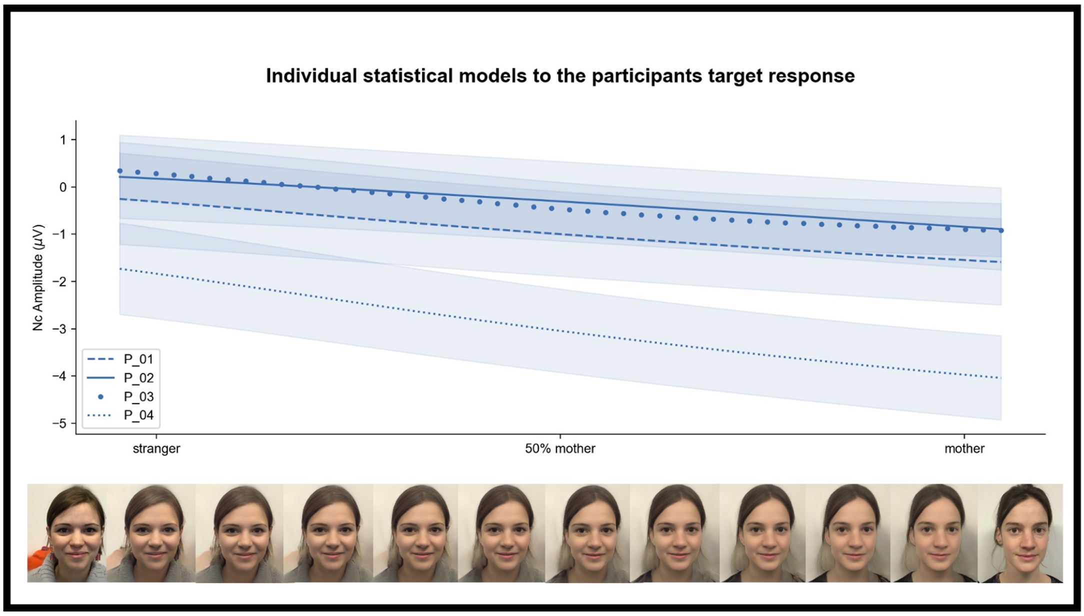

Figure 4.2: Individual model statistics for the 4 tested infants. *The BO sampling optimised the negative polarity of the Nc amplitude. The standard error for each participant's modelled response is displayed in shaded colour around each function.*

Figure 4.2. As hypothesised, the predicted measures of the Nc across the stimulus continuum show a negative slope for all four participants optimised towards the most negative Nc. Thus, the statistical model predicted that for these four participants, the image of their mother would produce the most negative Nc. This was not the case for the infant participant where the algorithm was run to optimise the most positive value. For this case, the model statistics failed to capture a variation of the Nc along the stimuli. Because Gaussian Processes inherit the properties of normal distributions, this method returns the standard deviation of the modelled function along the configuration space. This value works as a proxy of the model uncertainty. The high values of standard deviations obtained for the modelled responses are a consequence of brain metrics not being deterministic and being highly variant even when averaging across several trials.

4.3.2 Confirmation of expected mother-stranger effects using a traditional analysis

The grand average across four infants for the first two burn-in blocks (the mother and the stranger) are displayed in Figure 4.3.A. As expected, the neural response to the mother's face showed a more negative deflection than the response to the stranger face, most prominent during our time window of interest (300 - 800ms).

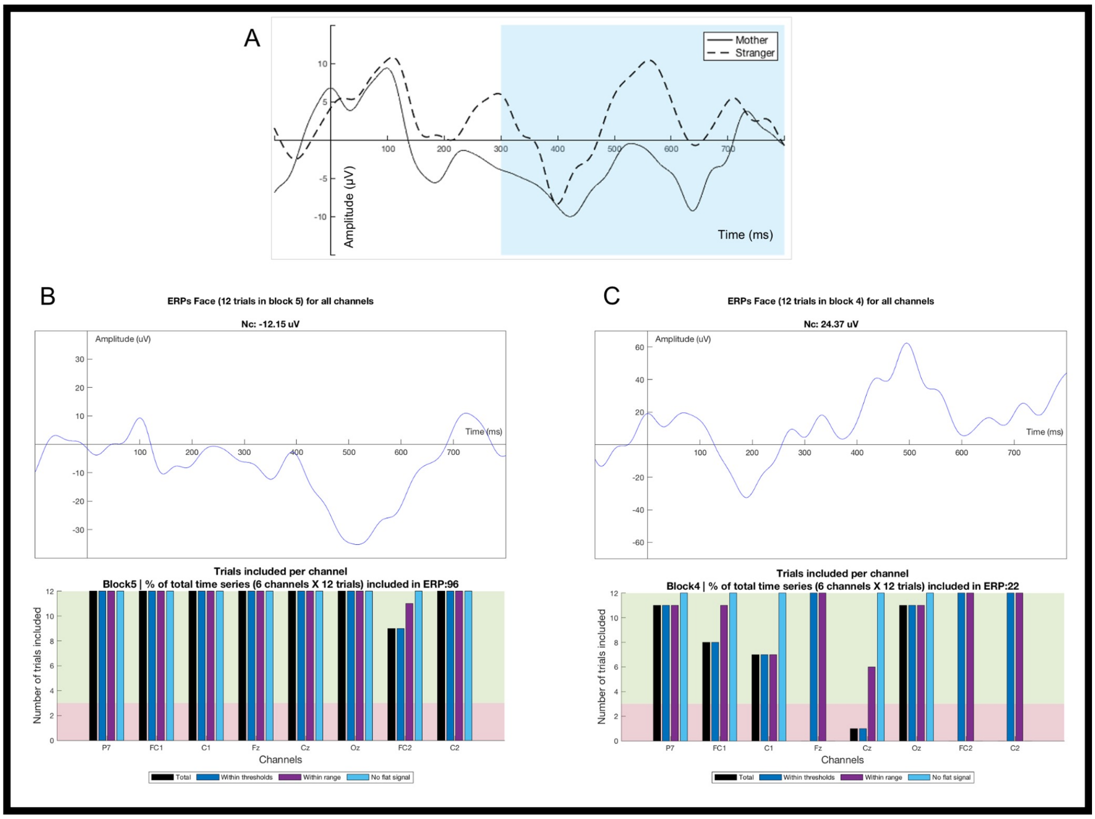

Figure 4.3: Real time EEG analysis. *A. Grand average for mother and stranger across the 4 infants with the shaded area reflecting the Nc time window of interest. B. ERP and data quality report for a good ERP; C. poor quality ERP with the ERP for the block in the top panel and the number of trials included for each channel in the bottom panel.*

Illustration of Real time metrics. Figure 4.3.B shows an example of the data visualisation display reviewed by the investigator after each block during data collection: an ERP with good data quality computed in real-time during data collection. In the ERP panel (top), there is a negative deflection during the time window of interest. The title of the ERP panel displays the value of the extracted ERP feature: Nc amplitude. The channel feedback panel (bottom) shows that the EEG signal, or time series, for all trials and most channels were included in the calculation of the ERP: a) the percentage of inclusion of time series for the channels of interest printed at the top is 96%, and b) the height of bars for channel FC2 indicates that 9 time series from this channel were without any artefacts (black bars), 9 were within the thresholds (blue bars), 11 within the range (purple bars), and 12 did not display a flat signal (light blue bars). The bars for the other channels indicate that all 12 time-series for the other channels were without any artefacts (thus, within the thresholds, within the range, and without a flat signal). Due to the high amount of clean time series, and a clear ERP waveform, the researchers decided to continue sampling after the data collection of this block.

Figure 4.3.C displays the real-time EEG feedback for a low-quality ERP for one of the infants that was excluded due to flat channels and excessive movement. In this low-quality ERP, the waveform does not show the typical shape and shows a positive deflection rather than a negative deflection during the time window of interest (ERP panel, top). More importantly, the channel feedback panel (bottom) shows that 22% of the time series from the channels of interest were excluded in the calculation of the ERP. For both channels P7 and Oz, 11 out of the 12 time-series are included into the calculation of the ERP. This suggests that the EEG data from the channels used as reference are relatively good quality. For channels Fz, FC2, and C2, there were no artefact-free trials (absence of black bars) due to flat signals (absence of light blue bars, whereas the height for the ‘within threshold/range’ bars are 12). For channels FC1, C1 and Cz, not all 12 presented trials were included either. This was mainly caused by the time series exceeding the thresholds and/or the range. In instances like these, the researcher would decide to repeat the block due to the low percentage of included time series and the low-quality ERP waveform. The researchers would first attempt to improve the EEG signal for the channels with low data quality (here, FC1, C1, Fz, Cz, FC2, and C2) by adjusting the cap and/or regelling the EEG electrodes before repeating the block.

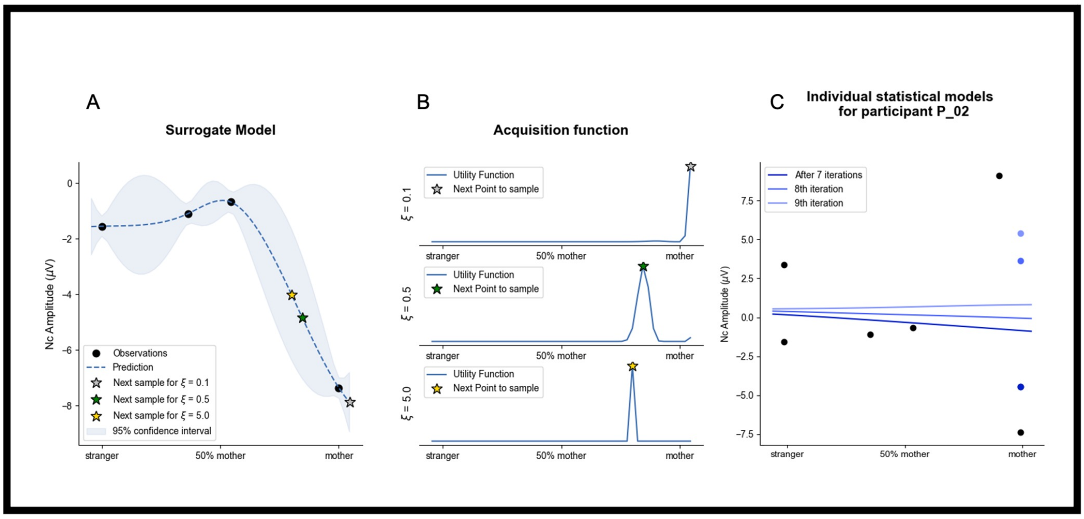

Figure 4.4: Case-study of exploration, exploitation and habituation on participant P_02. *A. Model statistics for participant P_02 after four iterations with the next points to sample from three different utility functions marked as stars. B. Acquisition function for the model statistics presented in A for three different values of ξ. The larger the value the more exploratory is the utility function and more will it privilege uncertainty over sampled maxima. The point to sample next is marked with a star. C. Example of habituation on participant P_02, where positive amplitudes of the Nc result in a change in the model statistics*

4.3.3 The exploration vs exploitation parameter

As mentioned in Section 4.2.3 the acquisition function, that guides the sampling algorithm, contains a hyperparameter ξ that controls the level of exploration of the space. In this proof-of-concept a conservative value of 0.1 was chosen, benefiting exploitation, as the number of blocks that could be run with infant participants was severely limited. Figure 4.4.A shows the surrogate model and the acquisition function for one of the participants after 4 burn-ins. The surrogate model's mean is a proxy for how the model interprets the Nc amplitude to change across the space and is displayed with a dashed line. The standard deviation of the surrogate model is interpreted as the uncertainty value across the fitted space. The uncertainty is minimal for the points already sampled and it increases the further it is from these points. This is due to the prior assumption that stimuli closer to each other in the configuration space elicit a similar response of the Nc ERP. The different acquisition function plots display the effect of different values of ξ for the next point to sample for this given participant – represented with a star. The higher the ξ parameter, the more exploratory the sampling behaviour and the more it will investigate sampling the regions where uncertainty is highest. At the extreme, the acquisition function will display active learning properties, sampling only to minimise the standard deviation of the space. The lower the ξ value, the more it will look into resampling what it found the highest value to be, disregarding uncertainty when ξ is 0. The small number of iterations it required to achieve the stopping criteria and the fact that the model captured the mother-stranger variation in Nc amplitude are good indications that a conservative value was the right choice for this specific paradigm.

4.3.4 The problem of habituation

Efficient sampling in EEG research and, more specifically in neurodevelopmental research, is fundamental to minimise problems of habituation or of short attention spans that plague the field (Snyder et al., 2008). This makes the case for algorithms that can predict individual response to a collection of stimuli, while only sampling a subset of them. Figure 4.4.C shows that, even while minimising the number of iterations, there are signs of habituation to the stimuli on P_02 that will shift how the model statistics of the space are predicted. When trying to predict an individual response to a set of stimuli, it is imperative to try to minimise signal variations that aren't directly related to the stimulus being displayed. To avoid the effect of the habituation to the stimulus on the model statistics for a given participant, the predicted model statistics before any habituation to the repeated stimulus had occurred was considered for participant P_02 in our *post hoc* analysis.

4.4 Discussion

This project presented the neuroadaptive EEG, a method that uses real-time data processing and machine learning algorithms to invert common research approaches by searching a large stimulus space to find the stimulus that maximally evokes a given neural response. In a proof of principle study, it was shows that neuroadaptive measurement of infant brain activity can locate a picture of the infant's mother from a one-dimensional face space. Here, it is discussed the advantages and limitations of this method; technical considerations for its use; and the potential of its application to a broad range of research questions across diverse fields.

4.4.1 Scientific robustness

Neuroadaptive EEG involves a fully closed-loop design. Thus, the neural feature targeted in the study and the stimulus search space must be predefined and hardcoded into the experiment itself. This completely removes analytic flexibility from the investigator since data is analysed during the experiment itself, solving one of the major challenges of current neuroimaging research (Head et al., 2015). Further, this approach requires brain signals to be reliable on an individual level within the experiment itself (if not, the search will not converge). This approach can be combined with other advances in robustness, such as external pre-registration of the selected EEG features and selected stimulus space to avoid the 'file drawer' problem. The approach is thus most suitable when the investigator has a known brain metric they are interested in investigating. However, the range of potential metrics to which the method can be applied is very broad; this could include connectivity between particular regions or at particular frequencies; activation in particular brain locations; or the speed or amplitude of well characterised event-related responses. In this way, the experimenter can build on the long history of stimulus-driven investigation of particular brain metrics whilst significantly extending our understanding by mapping new stimulus landscapes. Of note, this differentiates the present approach from more traditional brain-computer interface approaches (BCI) where the target metric is data-derived for each individual, making it hard to use as a tool for cumulative discovery. Finally, neuroadaptive methods support effective generalisation by computing the modulation of a metric across a large stimulus space. The importance of considering the limitations of inference to the specific stimuli selected in any given paradigm have recently been elegantly outlined (Yarkoni and Westfall, 2017). With neuroadaptive optimisation, the boundaries of generalisation can be objectively and efficiently probed. This approach is not only applicable to neuroscientific research as it can be used more broadly to multiverse analysis in general (Dafflon et al., 2020).

4.4.2 Efficiency of estimation across a space

This project's proof of principle case was the use of these methods in infants. Infants are a particularly challenging population because of their limited attention spans and inability to respond to verbal instruction, and this makes it fundamental to be able to capture dependencies between stimuli using as few iterations as possible. In this experiment a one-dimensional configuration space is used which interpolates between the stimulus of the participant's mother face and a stranger face. The mean amplitude of the Nc ERP is measured as the target EEG metric. Contrary to classic research paradigms, this method does not restrict the researcher to just two stimuli (i.e., mother and stranger), but instead the researcher is able to capture the continuous variation between these two images. This opens the possibility to explore not only if there are differences in the ERP amplitude, but also whether the signal variation is gradual or if it is abrupt, such that only the last image of the mother elicits a stronger Nc (as may be the case if face perception is categorical (Leopold and Rhodes, 2010, Moulson et al., 2011)). Here the individual model prediction is presented across the configuration space to show individual variation, but this method allows for the use of other data to account for intra-group variability; for example, the measure of uncertainty across the space can be useful to capture differences in signal reliability between participants. The path of exploration to exploitation used by the algorithm can also be relevant to discriminate between participants. The introduced method showed this capability in this proof-of-concept by predicting the variation of the signal across the stimuli while only sparsely sampling them. This method can be used in future neurodevelopmental research to test the reliability of individual ERPs as correlates of cognitive functions, and to understand commonality and individual differences in neural responses to visual stimuli.

4.4.3 Diverse populations

Neuroscience has suffered from a long history of collecting data in primarily White, neurotypical, English-speaking young adults, often from higher education institutions. The need to expand to broader and more diverse populations is well rehearsed but comes with significant challenges. Beyond the practicalities, we must recognise that selection of stimuli based on theories derived from our existing narrow samples may not be the most appropriate or efficient way to learn about brain function in more diverse cultures. Such an approach is highly prone to cultural bias (e.g., selection of White faces for experiments with diverse ethnicities) and a deficit-based perspective (identifying differences in how neurodivergent adults respond to stimuli selected by normative experimenters – such as examining diminished responses to faces in autism). The presented approach allows the investigator to move away from studying brain responses to the same stimulus in different groups, and towards studying how a comparable brain metric is modulated across a much broader stimulus space. Rather than asking why a young child with autism doesn't attend to faces, we can search for the type of stimulus that the child does find interesting. Of course, these approaches are complementary rather than exclusive, but together they may provide us with a richer suite of tools to study brain function across diverse populations and to devise individualised interventions that build on strengths.

These considerations are particularly important in studying the infant brain, where one of our core interests is the way in which mature cognitive topography emerges in development. For decades, researchers have debated whether the mapping between brain system and cognitive function is present at birth (nativist or modular accounts (Markram, 2006)) or whether brain regions have functions that change over developmental time (Johnson, 2011). One leading account proposes that brain regions progressively specialise through a process of interaction and competition (interactive specialisation (Johnson, 2011)). Testing such accounts is slow and challenging with traditional experimental paradigms, because stimulus selection is typically informed by adult cognitive models (e.g. since the fusiform face area or N170 component are face-selective in adults, we examine their response to faces vs objects in infants (Deen et al., 2017)). With a neuroadaptive approach, in principle investigators can move to mapping the modulation of the fusiform face area or N170 across a much broader stimulus space. Indeed, GANs can be used to create a large library of both faces and objects that are artificially manipulatable along a range of dimensions. In this way, we may generate new knowledge about the cognitive topography of the infant brain- seeing the world through their eyes.

4.4.4 Methodological considerations

The approach includes several parameters that can be adjusted depending on the research question that is being addressed. The presented method allows the researcher to navigate the exploration-exploitation boundaries of the configuration space by defining the ξ value, that allows the algorithm to balance between a more exploitative search of the maximum target metric. Experimental designs that are limited to a small number of iterations (because of the population studied), as was the case of the proof-of-concept with infants, should use a low value of ξ and benefit exploitation of the stimuli. Designs that explore larger configuration spaces, with many dimensions and stimuli, could utilise a larger value of ξ and exploration to better capture the modelled statistics of the space. As an example, if instead of just varying along one dimension (e.g., mother-stranger interpolation), the configuration space contained several semantic variations (e.g., a 4-dimensional space with faces varying across age, emotion, gender and eye-to-eye distance), a more explorative behaviour would be beneficial.

Selecting a robust target metric is important. The algorithm searches the space to identify the stimulus that elicits the maximum target metric we are measuring. For the four participants for whom the target EEG metric for the Bayesian optimisation was the most negative amplitude of the ERP, the algorithm is able to model a variation of the signal that maximised its negative amplitude for images of the mother's face. This was not the case for the participant in which the Bayesian Optimisation aimed instead to identify the image that produced the most positive Nc amplitude. There are several possible explanations for this; one may be that categorical face perception means that the infant is sensitive to varying degrees of proximity to the mother, but not varying degrees of 'strangerness' (M. De Haan and C. A. Nelson, 1997). Success of BO-based approaches are dependent on reliable and meaningful neural signatures to be used as target metrics. In the current proof-of-concept study, the parameters and measures for the ERP features were chosen based on the previous literature and offline analysis of pilot and existing datasets (Gui et al., 2021). The preprocessing methods and calculation of ERP features however vary between studies. It is possible to further examine how the process of optimisation would vary when using different EEG features or metrics. For example, in line with most previous work (Luyster et al., 2014, Webb et al., 2011) the mean amplitude from 300 to

800ms was used as an EEG metric. Selecting appropriate data processing parameters is also important. Researchers willing to use this method should conduct some preliminary offline analyses of individual-level ERPs to examine how varying recording and preprocessing methods affect the quality of the ERP waveform. Different thresholds or a larger number of trials within a block may improve the quality of the ERP waveform and decrease the signal to noise ratio. In this project, with infant data increasing the duration of the experiment and exposure to the same image within a block by including more trials does not improve the quality of the ERPs. Inspections of the individual infant ERPs obtained with a varying number of trials using existing and pilot datasets confirmed the choice that 12 trials were sufficient for our experiment. Another avenue is to develop an EEG metric that reflects the quality of the ERP waveform, such as standard deviation or area under the curve during the baseline, or a quantification of the shape of the ERP waveform and curve. One could implement this information into the BO algorithm by adding varying weights of the samples. Given that the present study aimed to test that BO could appropriately map neural signatures of attention engagement, it was essential to define a threshold for inclusion of good quality trials per block and channel. In case of poor EEG quality data, the block was excluded completely and the data collection was repeated for the same image. Here, the decision on how to proceed with the iterative optimisation process was dependent on the subjective inspection of the researchers. This always raises the risk of potential bias, particularly given the ERP and key metrics were displayed- although this was done to allow inspection of data quality, it raises the possibility that researchers could be biased towards rejection or selection based on the nature of the Nc response. Implementing and quantifying data quality measures would make this process more objective and therefore less potentially biased; however, in this pilot work automated quality measures weren't considered as effective at detecting poor data quality as a human researcher. Further work in building stronger automated quality measures for the signal are needed.

4.4.5 Limitations

The sampling algorithm and the real-time EEG can work with any configuration space setting, but the presented configuration space builder is limited to 16 pre-defined meaningful variations of faces (e.g., pitch change, age). Other meaningful variations of faces can be learned by following the approach described in Section 2.1.2.3. Non-facial stimuli are currently not supported by the configuration builder but could be integrated in future work. The sampling algorithm is further limited by the quality of the target EEG metric. Despite the algorithm being flexible to receive noisy input, if the averaging of the EEG signal per block fails to capture the dynamic of the ERP, the statistical model will not be able to capture the ERP dependency across the space. It is important to note that this work does not propose that the neuroadaptive method should replace traditional stimulus-driven approaches. Traditional methods will be important to further discovery of new neuroimaging metrics of interest. However, once such metrics have been identified and sufficiently well parameterised the neuroadaptive method allows the investigator to map the modulation of these metrics across a larger stimulus space, providing a complementary tool to further our understanding of the cognitive topography of the brain.

4.5 Conclusions

A core goal of functional neuroimaging is to study how the environment is processed and represented in the brain. The mainstream paradigm involves concurrently measuring a broad spectrum of brain responses to a small, preselected set of environmental features selected with reference to previous studies or a theoretical framework. As a complement, this project inverts this approach by allowing the investigator to record the modulation of a preselected brain response by a broad spectrum of environmental features. It was shown that online recording of the Nc infant brain engagement response can automatically identify the position of an individually salient face (the infant's mother) in a one-dimensional face space. The promise of this approach for studying the developing brain was demonstrated, where our theories based on adult brain function may fundamentally misrepresent the topography of infant cognition and where there are substantial practical challenges to data acquisition. This approach may also have significant potential in areas where theoretical frameworks or previous empirical data are impoverished or misleading, allowing us to tackle new questions and move beyond heteronormative undergraduate student populations. Furthermore, by using a prespecified closed-loop design the approach tackles fundamental challenges of reproducibility and generalisability in brain research. This approach has substantial potential in infancy research and beyond for accelerating our understanding of the cognitive topography of the brain.

5 | Multiverse analysis of processing pipelines

5.1 Introduction

Processing raw data and choosing statistical analysis methods are fundamental steps in hypothesis testing. Regardless of the field of research, many valid choices of processing steps can be taken. For each choice, a forking path is presented to the researcher resulting in a multitude of correct data processing paths that were not taken. Importantly, research has shown that the choice of data processing can directly impact the conclusions obtained by statistical analysis (Marek et al., 2022, Menkveld et al., 2021, Steegen et al., 2016). These results starkly contrast with current methodological best practices, where a single dataset analysis is sufficient to make research claims. This opens the door for p-hacking through assessing many processing paths and reporting only the most favourable outcome. Pre-registration of methodology and open-sourcing the code used for data processing are good practices but are blind to the uncertainty associated with the methods chosen and those not selected in the processing 'garden of forked paths'. Recent studies have pushed for a multiverse analysis, i.e., a survey of many possible processing paths, to analyse the robustness of outcomes and build better methodological gold standards (Wagenmakers et al., 2022). In this chapter, I study the relationship between the different possible paths and how the multiverse of analysis can be reduced to a Euclidean space with its organisation governed by the similarity between processing outcomes.

In the neuroscientific research literature, the research question and the hypothesis tend to be well defined (e.g., quantifying differences in functional connectivity measured with functional MRI (fMRI) between different groups or individuals), and the analysis pipeline tends to be less transparent. The analysis can vary depending on how functional MRI data is preprocessed to remove noise (e.g., which kernel, smoothing factor, or type of motion correction to use) or the type of summary metric of functional connectivity chosen (e.g., centrality, pairwise correlation, entropy). All these choices create a combinatorial panoply of valid analysis pipelines that can yield different conclusions from the same data (Bzdok and Yeo, 2017, Carp, 2012). In fledgling research areas, such as functional neuroimaging, where many ground truths are yet to be discovered, analytic exploration is an unavoidable aspect of the scientific process. Therefore, a central conceptual question facing the community is how to balance the data exploration needed for scientific progress with the analytical rigour necessary to minimise the number of such discoveries that are false positives. To highlight this issue (Botvinick et al., 2020) asked 70 independent teams to analyse the same dataset and test nine pre-defined hypotheses. Although all groups used different workflows to test these hypotheses and showed relatively high variability in the specific answers, a meta-analysis showed reasonable agreement among the overall results. The degree of consensus in such studies is essential because different approaches can yield broadly similar answers. This, in turn, provides confidence that these conclusions are not tied to a specific analytic approach (Steegen et al., 2016).

There is, however, a trade-off between the breadth of the initial exploratory approach and the sensitivity of subsequent scientific inferences. There are near limitless potential analysis approaches, and each additional analysis approach assessed reduces the sensitivity of statistical tests if appropriately corrected for the number of comparisons. As an example, if 1000 analysis approaches are tested for a specific hypothesis, then while hypothesis-testing, it is a correct methodological practice to correct p-values for the multiple comparisons performed (e.g., using a Bonferroni correction). This methodology dramatically reduces the statistical sensitivity and would require the corrected p-value to be under 0.00005 to be considered statistically significant at the default $p < 0.05$ level.

Moving to out-of-sample prediction rather than inferential statistics on the whole dataset does not avoid this trade-off. The robustness of a specific conclusion from exploratory neuroimaging studies can be evaluated by training models on a particular subset of a data set and testing it on an unseen portion of the data. While minimising overfitting problems from yielding false positive conclusions, this approach does not address reductions in sensitivity from performing many analyses on the same data.

In this project, a machine learning framework is established that maintains an exploration across analysis approaches and the sensitivity of predictive statistics and generalisability to out-of-sample data. This approach allows the researcher to explore many different features of the multiverse of pipelines and techniques, allowing many choices to be empirically compared without the need for exhaustive sampling. This is possible by building a low-dimensional space across different pipelines, efficiently mapping out using active learning (Settles, 2009).

The utility of this multiverse approach is illustrated in a research question in which the ground truth is entirely transparent - predicting age from functional connectivity obtained from functional MRI images of adolescent and young adult participants (Váša et al., 2020). In particular, this work focuses on graph theoretical analyses. The motivation for the choice of graph theoretical analysis is two-fold: (i) graph theory applied to fMRI data is a useful technique for exploring the interrelationships between brain regions (Bullmore and Sporns, 2009, van den Heuvel and Hulshoff Pol, 2010); (ii) such approaches have also been shown to be highly sensitive to preprocessing steps such as thresholding (Fornito et al., 2013, van den Heuvel et al., 2017). Moreover, there are dependencies between different types of graph theory measures (Rubinov, 2016), so the optimal analysis approach for any given dataset or question is typically unknown *a priori*. Indeed, previous work has demonstrated that variations in (structural) brain network construction and analysis pipelines substantially impact results (Phillips et al., 2015). Focusing on the graph theoretical analysis of functional brain networks to predict age allows us to evaluate the utility of the multiverse approach under conditions when the ground truth is known, but the ideal analytic approach is not. This approach could be applied more generally to many different types of neuroimaging problems (both functional and structural) or other types of data (e.g., univariate and multivariate analyses) and ultimately be applied to basic scientific and clinical research questions when the ground truth is less transparent.

5.2 Methods

5.2.1 Rationale and analytical workflow

While the proposed framework can be applied to any neuroimaging studies, here, its capabilities are shown in the context of predicting age. The framework consists of two main steps: (i) creating a low-dimensional continuous space of the different analysis approaches; (ii) using an active learning component that efficiently searches the space to find the optimal analysis pipeline (in this case, the pipeline which best predicts age) and produces estimates of other pipeline's performance. The proof-of-concept focuses on predicting brain age as it has been proposed as a valuable biomarker of neurological and psychiatric health (Cole and Franke, 2017, Kaufmann et al., 2019) and is predictive of a range of other factors, including mortality (Cole et al., 2017). More generally, predicting age is a valuable proof of principle for methodological demonstrations since the participant's age is known with certainty (Schulz et al., 2020), and a range of studies have shown that functional connectivity from resting-state correlates with age (e.g., Geerligs et al. (2015), Monti et al. (2020), Váša et al. (2020)). All code used for analyses and figure generation is available on GitHub* and can be run using Colab. The next section presents the data, followed by the range of analysis approaches considered, how the analysis space was constructed, and the active learning approach used to sample the space and so be able to estimate brain age prediction across the multiverse of analysis approaches without exhaustive testing.

5.2.2 Functional Connectivity data

The starting point is a functional MRI dataset of changes in functional connectivity across adolescence from Váša et al. (2020). This dataset consists of 520 scans from 298 neurologically healthy individuals (age 14-26, mean age=19.24, see Váša et al. (2020) for details). Here, only cross-sectional analyses were performed, so only the first scan for each individual was kept. The dataset was split into two parts: (i) 50 individuals, selected at random, were used to build the low-dimensional space; (ii) the remaining 248 individuals were subsequently used to perform search and validation on the space.

5.2.3 Analysis approaches

Many decisions are necessary to conduct a functional connectivity study, including choices regarding data acquisition, preprocessing, summary metrics and statistical models. For convenience, this project uses already acquired data which has been through extensive preprocessing pipelines to reduce many potentially confounding sources of non-neural artefacts. Usefully, two preprocessed datasets were shared by Váša et al. (2020) with two different types of correction for movement artefacts: (i) global signal regression and (ii) motion regression. The preprocessed fMRI time-series had been averaged within 346 regions of interest,

*https://github.com/Mind-the-Pineapple/into-the-multiverse including 330 cortical regions from the Human Connectome Project multi-modal parcellation (Glasser et al., 2016) (excluding 30 "dropout" regions with low signal intensity) and 16 subcortical regions from Freesurfer. The Pearson correlation coefficient was used to calculate functional connectivity (FC) between these regions. For further details regarding data preprocessing, see Váša et al. (2020). In future work, the approach outlined below could be applied to all of the image preprocessing steps, the choice of parcellation and the connectivity metric for a more extensive multiverse analysis.

To demonstrate the capabilities of the proposed method, I consider varying three distinct analysis pipeline choices. These are:

1. *The nature of regression:* data from two types of data regression, motion regression and global signal regression, were explored.

2. *The graph theoretical metric studied:* 16 distinct metrics covering both simple and higher-level metrics employed in previous neuroimaging studies were considered. Metrics were taken from the Python implementation of the Brain Connectivity Toolbox (Rubinov and Sporns, 2010); metrics included were those that produced nodal metrics. If prior community assignment information was required, I used the well-known Yeo network parcellation of the brain into seven networks (Yeo et al., 2011).

3. *The choice of threshold for estimated functional connectivity matrices:* 17 distinct threshold values ranging from 0.4 (resulting in highly sparse networks) to 0.01 (resulting in dense networks) were considered.

Every evaluated pipeline was built using one of the regression choices, one threshold and one graph theory metric. Therefore, the analysed multiverse space consisted of different analysis approaches. The full space of analysis options is presented in Table 5.1.

5.2.4 Constructing a low-dimensional space of analysis approaches

We need information about their general relationships to efficiently sample a large number of analysis approaches. This is achieved by building a low-dimensional space that quantifies the similarity between approaches in terms of a distance in the low-dimensional space (i.e., how similar is the obtained functional connectivity between the different approaches). To ensure utility to a range of questions, the space should capture the similarity between approaches across various problems (and potentially a range of different datasets, although that is not assessed here). For example, the same space can be used to ask questions about age. Still, it could be used to ask about other sources of individual variability (e.g., neuropsychiatric symptoms or cognitive ability). Below, one approach to building such a space is illustrated (precisely how this is done will vary depending on the data type and other factors).

All 544 analysis approaches were applied to 50 randomly selected participants' individual FC data to construct the low-dimensional space (Figure 5.1.A-C). The aim was to build a space to locate approaches in terms of how well they capture individual variability. Therefore, for each approach, the Euclidean distance matrix was calculated in terms of regional graph theory metrics between different participants (e.g., the Euclidean distance of betweenness centrality across all 346 regions for each pair of participants (Figure 5.1.D,E)). These were subsequently reshaped into a 2D matrix corresponding to between-participant distances (this led to a matrix of 1225 participant pairs by 544 analysis approaches). Finally, the low-dimensional space was constructed with established embedding algorithms. Five different algorithms were explored: local linear embedding (Roweis and Saul, 2000), spectral embedding (Belkin and Niyogi, 2003), t-distributed stochastic neighbour embedding (t-SNE) (van der Maaten and Hinton, 2008), multi-dimensional scaling (MDS) (Kruskal, 1964) and Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018). The objective of the embeddings was to create a space useful for active learning which could both: (i) capture similarity between approaches in terms of continuous distance in the space; as well as, (ii) distribute approaches relatively evenly across the space. Based on observations of the spaces resulting from the embedding mentioned above algorithms, MDS (see 2.1.1.2) was selected to create the space to apply active learning.

5.2.5 Searching the space

Using the low-dimensional space created with FC data from 50 participants, active learning was subsequently used with the remaining 248 participants to sparsely sample the space in order to: i) find the most successful approaches for predicting participant age based on FC; and (ii) estimate age prediction ability for all models, including the large majority of models which were not sampled. Active sampling is performed using closed-loop Bayesian optimisation with Gaussian processes (Shahriari et al., 2016). This loop involves: selecting a point in

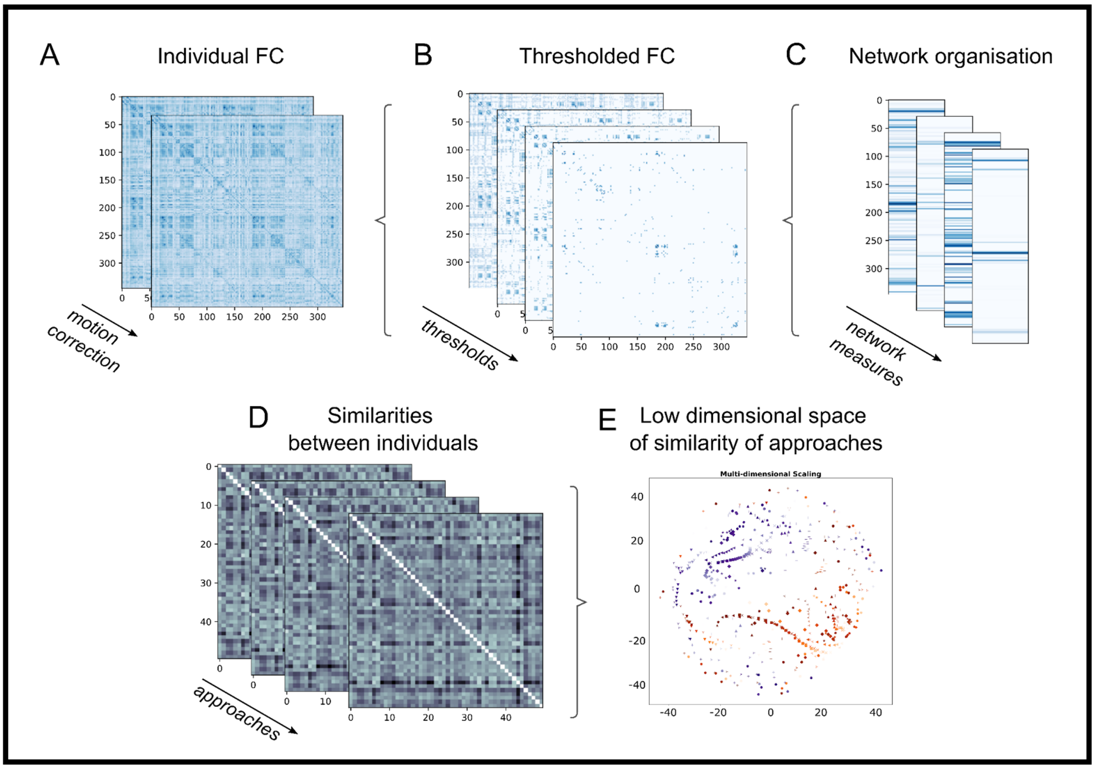
A
Individual FC
motion correction
B
Thresholded FC
thresholds
C
Network organisation
network measures
D
Similarities between individuals
approaches
E
Low dimensional space of similarity of approaches
Multi-dimensional Scaling

Figure 5.1: Creating a low-dimensional space to characterise the multiverse of different analysis approaches. *A low-dimensional space was constructed using 50 (randomly selected) participants' functional connectivity (FC) matrices. Each participants' data was analysed using 544 possible analysis approaches, which were composed by choosing from: two methods for motion correction (A), 17 different sparsity thresholds (B), and 16 different graph theory metrics (C). The different approaches were evaluated as to how well they capture individual variability by evaluating the pairwise cosine similarity for all the different analysis approaches (D). The distance matrices were then converted into a low-dimensional (2D) space summarising the similarity between approaches by using an embedding algorithm (such as multi-dimensional scaling (E)).*

the space to sample, evaluating it in terms of 5-fold cross-validated predictive accuracy, fitting a Gaussian process (GP) regression to the space, and evaluating an acquisition function using the GP regression to select the next point to sample.

When a point in the space is identified, the closest analysis approach to that point in the space is selected. Its predictive accuracy is evaluated using support vector regression for brain age prediction. It is essential to highlight that although the algorithm to predict age was kept constant, the input data varied depending on the selected analysis pipeline, which could have used different motion correction, thresholding or graph theory metrics. Predictive accuracy was calculated with 5-fold cross-validated negative mean absolute error.

For the examples presented in the results, there was an initial burn-in phase in which ten points in the space were randomly selected and evaluated before active learning began. Bayesian optimisation used the upper confidence bound (UCB) acquisition function (Shahriari et al., 2016). The Gaussian process regression model used a Matérn kernel combined with a white noise kernel, with kernel hyperparameters chosen in each iteration by maximising log-marginal-likelihood using the default optimiser.

5.3 Results

The first step was to construct a low-dimensional space of the analytic space. Six approaches were considered and are presented in Figure 5.2. All embedding algorithms demonstrate considerable structure in the position of the different approaches (e.g., similar types of motion correction, thresholding, graph metric are generally proximal). This suggests that the low-dimensional space captures the intended similarity between the approaches. A dissimilarity score was used to assess how much the different embedding algorithms preserved the topological information (i.e., similar analysis approaches should stay close after embedding). MDS, t-SNE and UMAP efficiently maintained the neighbourhood of the original space. In addition, MDS displayed a relatively even spread of approaches across the whole space, especially when contrasted with Local-Linear-Embedding and Spectral Entropy. An approximately even spread across the space is desirable for the subsequent active learning and Gaussian process regression. As such, MDS was used in subsequent analyses.

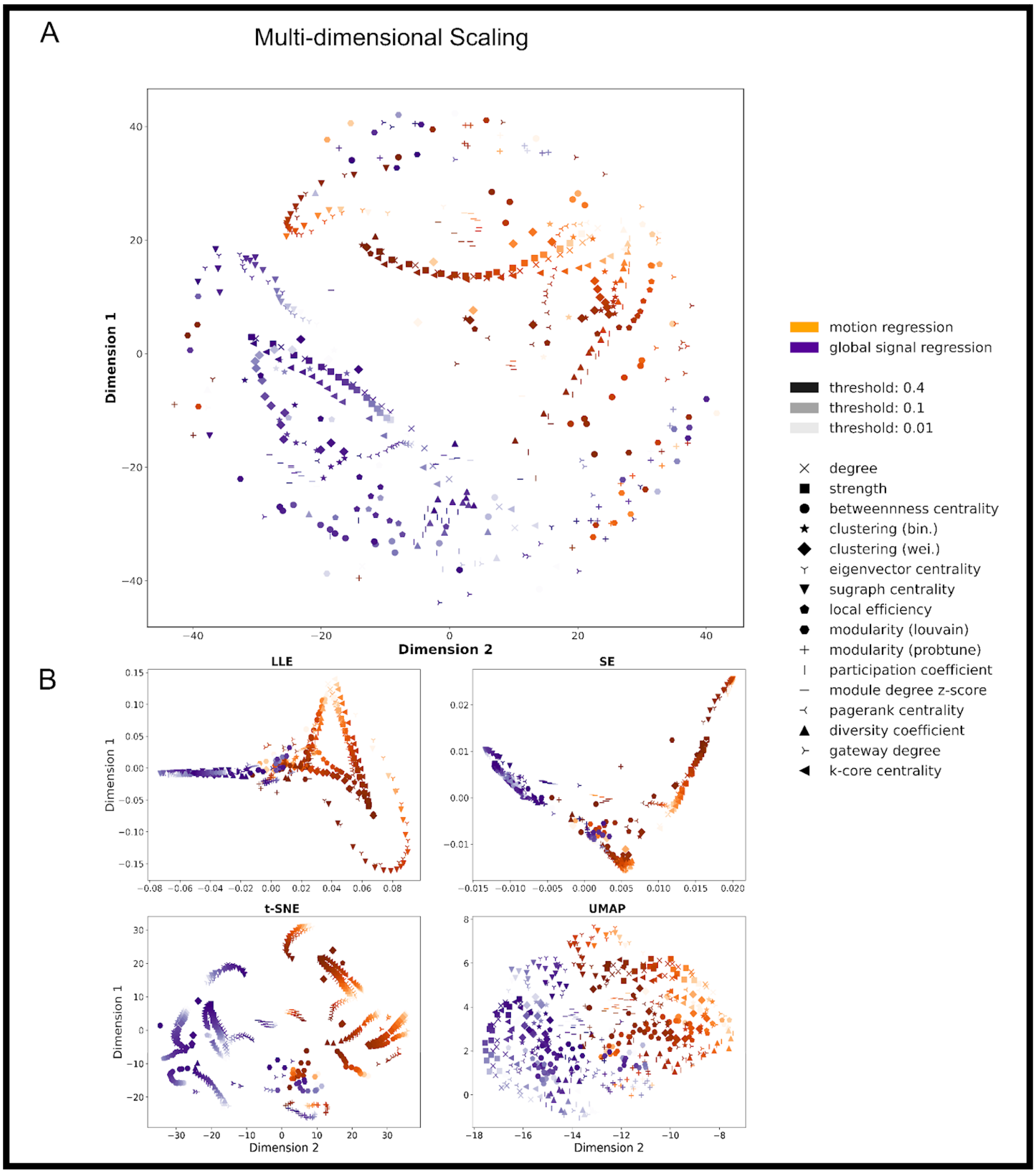

Figure 5.2: Low-dimensional embeddings of the different analysis approaches. *Each point represents a combination of data accounting for noise confounds, thresholding of connectivity weights and different graph theory metrics. In particular, the colours represent both motion correction methods used to preprocess the data (i.e., motion regression (orange) and global signal regression (blue)), the colour intensity represents the different thresholds used in each analysis and every graph theory metric is represented by a different symbol. A. Multi-dimensional scaling. B. Four other types of embedding: Local linear embedding (LLE), Spectral embedding (SE), t-Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP).*

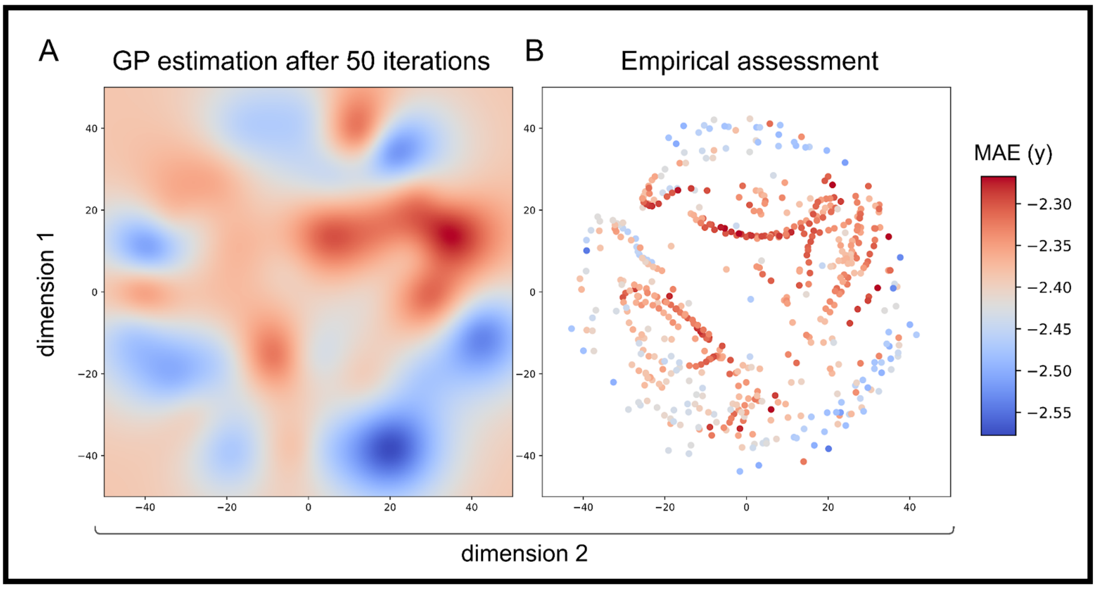

Figure 5.3: Learning the space and identification of the optimal analysis pipeline for age prediction. *A. After 50 iterations of Bayesian Optimisation the Gaussian Process (GP) closely estimates the empirical space. B. Empirical assessment of age prediction across the whole space. The colours correspond to negative mean absolute error of each model in years. Values closer to zero represent a more accurate prediction and are shown in red.*

There are two objectives for the use of active learning on the MDS-defined space of different analysis approaches: i) finding an approximately optimal analysis approach efficiently, controlling the number of multiple comparisons; while ii) approximately estimating performance on the multiverse of approaches without exhaustive sampling. These two objectives can be observed in Figures 5.3 and 5.5 where age-prediction models were trained and evaluated for different analysis approaches selected by active learning.

First, in Figure 5.3, the result of the Gaussian process regression after 50 iterations of active learning is shown. Based on the 50 different analysis approaches sampled, GP regression estimates performance across all 544 approaches (Fig.5.3.A); this identifies areas predicted to have higher age-prediction performance (in warm colours), including the optimum, as well as approaches which perform worse (in cooler colours). For comparison, the ground truth of performance across the space (from an exhaustive sampling of every approach) is presented in Fig.5.3.B. There is a generally good concordance between actual age prediction for each approach and the estimated prediction across the whole space (Spearman's $\rho = 0.61$, p < 0.0001).

The evolution of the active sampling and Gaussian process regression model is presented in Figure 5.4. A poor GP estimation of the space based on the first

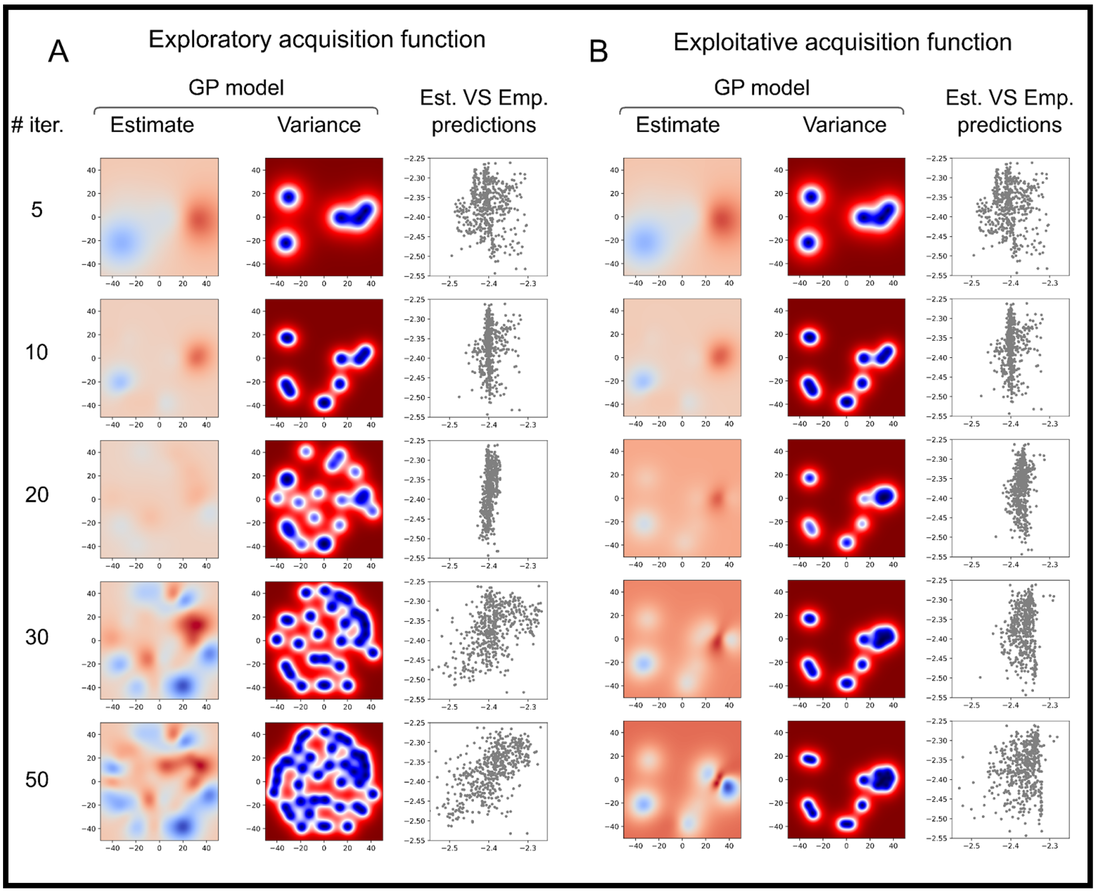

Figure 5.4:* *The evolution of the search across the space for: *A. a more exploratory acquisition function; and B) a more exploitative acquisition function. Within each panel, the first column is the estimated Gaussian Process (GP) model after different numbers of samples; the second column is the variance of the GP model across the space, indicating which points have been sampled; the third column is the estimated versus empirical predictions for all the analysis approaches in the space.*

ten random burn-in samples was initially observed. As the sampling increases, the space is progressively better estimated, achieving increasingly higher correlations between empirical and estimated spaces. Acquisition function parameters strongly affect the active sampling; to illustrate this, the parameter was varied to conduct both exploratory (10, Fig.5.4.A) and exploitative versions of active sampling (0.1, Fig.5.4.B). The exploratory version achieves a better estimation of the whole space, while the exploitative version focuses on an estimated optimum much more quickly, but the GP model changes much less subsequently resulting in a much lower correlation between estimated and empirical accuracies across the space.

To investigate the reliability of the active sampling, the process was repeated

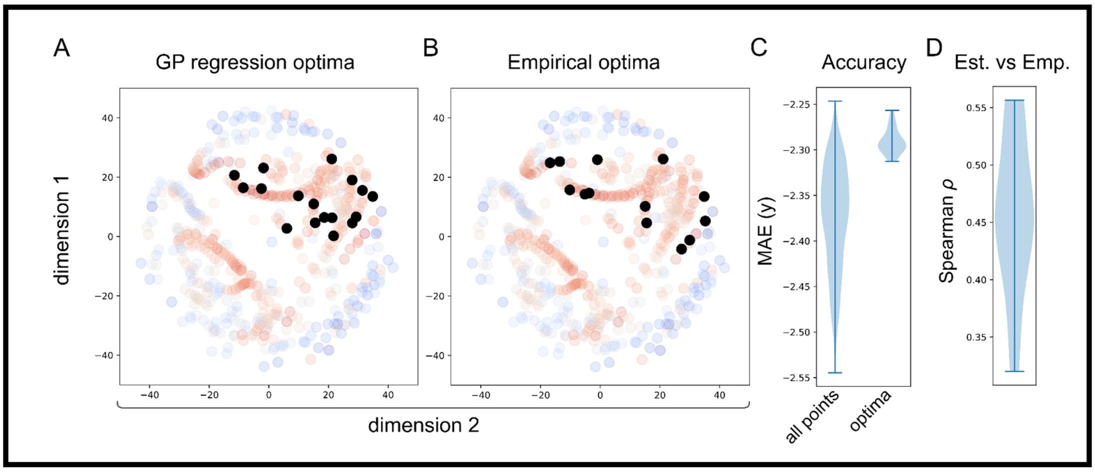

Figure 5.5: Performance of the optimisation across different random starting conditions. *For computational efficiency, only 20 iterations of active sampling were performed. Black dots represent optima of the 20 iterations based on A. the highest accuracy estimated using the GP model and B. the actual sampled points. C. Range of negative mean absolute error for the optima versus negative mean absolute errors across the whole space. D. correlations between actual and estimated accuracies across the whole space for the 20 replications.*

20 times (using the more exploratory 10) with different random seeds (and so different initial random burn-in samples). In Figure 5.5, the optima (i.e., model with the highest empirical accuracy) of the 20 repetitions are represented by the black dots, based both on the highest accuracy estimated using the GP model (Fig.5.5.A) and for the actual sampled points (Fig.5.5.B). Table 5.2 presents the optimal analysis approaches selected by each iteration. Many of the optima illustrated in Table 5.2 were obtained using the Betweenness centrality. This might suggest that this graph theory metric is more robust to using different preprocessing choices. The range of the mean absolute error for the different optima selected versus the full range of mean absolute errors across the whole space is presented in Fig.5.5.C, and the range of correlations between actual and estimated accuracies across the whole space for the 20 replications is shown in Fig.5.5.D. For inferential statistics, the optimal analysis approach selected for each of the 20 replications was assessed with a permutation test on cross-validated predictions (with 5000 random permutations), resulting in a range $p \sim= 0.004 - 0.044$ (Bonferroni corrected for 20 samples). Exhaustive sampling would result in correcting the best model for 544 comparisons rather than 20 (requiring an uncorrected p<0.000091 rather than p<0.0025).

5.4 Discussion

In this chapter, it was established that active sampling can be used to map out a low-dimensional space of the multiverse of analytic approaches allowing the processing pipelines with higher accuracy to be identified efficiently. This project focused on a question with a known ground truth, predicting brain age from resting state functional connectivity data. Since efficient exploratory research is critical for neuroimaging to become a mature scientific discipline, this multiverse approach is a crucial tool that balances the need for rapid discovery with analytic rigour in a highly cost-efficient manner.

The application of active sampling to predict age from functional connectivity is an illustrative example. A recently released dataset preprocessed using both motion regression and global signal regression (Váša et al., 2018) was used. The main aim was to showcase active sampling on a space of analysis approaches, rather than identify (the) optimal combination(s) of fMRI head motion correction, functional connectome threshold and graph theoretical method for age prediction. Nevertheless, it remains interesting to consider the approaches selected as optimal by the GP regression. By repeating the active learning method 20 times, there is substantial consistency in processing steps across the selected optima. The motion regression approach consistently outperforms the global signal regression; lower, but not the lowest sparsities, were also favoured using a range of simple and complex graph theoretical metrics, with betweenness centrality selected most frequently. These results are essential because global signal regression is one of the most debated fMRI processing steps, with many arguments for and against its inclusion in processing pipelines (Li et al., 2019, Murphy and Fox, 2017).

Regarding thresholding, it was observed that most of the optima had a higher threshold. This is in line with previous research that observed that connections with lower edge weights (i.e., correlation) are more likely to be spurious, suggesting that connectomes thresholded to lower densities might be less affected by noise (van den Heuvel et al., 2017, Váša et al., 2018). Finally, betweenness centrality had previously been found to perform well in network neuroimaging applications, including machine learning applications (Fagerholm et al., 2015). Our multiverse approach is highly generalisable and can easily be expanded to consider different analysis approaches and preprocessing techniques. The presented method is not limited to graph theory. It can be applied to any set of heterogeneous techniques in neuroimaging that can be evaluated by a standard target measure, as is the case of different machine learning pipelines that try to minimise a cost function. For example, a similar approach could be used with full preprocessing pipelines used in (f)MRI and potentially integrated with automated pipelines such as fMRIPrep (Esteban et al., 2019) to allow controlled, efficient exploration of a much more comprehensive range of analyses. Given that the analysis space is based on variability across individuals, it does not require that the analysis approach results in data of the same format. It is possible to combine univariate and multivariate analyses (e.g., single regions or every voxel or vertex measured) and even potentially different modalities, allowing the multiverse to cover a very heterogeneous collection of approaches.

In the current chapter, the analysis space was developed from a subset of the whole participant group; however, this need not be the case. A predefined space can be constructed using an existing dataset and subsequently applied to different datasets with minimal computational cost. For example, large open datasets such as the Human Connectome Project or UK Biobank could be used to define analysis spaces which can then be applied to smaller, e.g., clinical datasets. This would mirror the strategy taken with many deep learning approaches, which are computationally expensive to train but not to apply to new data. Performing multiverse analyses can increase the generalizability of results, similarly to other approaches (e.g., Baribault et al. (2018)). As recently revisited by Yarkoni (2019), when interpreting findings, we often go (both statistically and verbally) far beyond what is justified by the restricted nature of the data and analyses performed. Taking a multiverse approach explicitly tests the generalisability of the studies: indeed, the GP regression model quantifies the relationships between analysis approaches in the low-dimensional space. This can clarify how specific or general a given finding is across all approaches. The efficiency of the space sampling also ensures that the same data is only used a limited number of times, reducing the problems inherent in sequential analyses in terms of overfitting. In the extreme, to maximise generalisability, it is possible to perform each iteration of the active sampling on a different subset of participants who are not then reused; as such, each suggestion from the Bayesian optimisation for the next point to be sampled would involve out-of-sample prediction.

Similarly to previous work using Bayesian optimisation for the navigation of predefined experimental spaces (Lorenz et al., 2016, 2018), the method presented here can help improve the poor reproducibility present across much of (neuro)science. Sequential analysis, as applied here, is highly formalised, quantifiable and controllable, and as such, it can be readily combined with preregistration (Lorenz et al., 2017). Similarly, the route and samples taken by the analysis make it possible to deduce what the hypothesis (encoded as the target function of the optimisation algorithm) was at the time of testing. If a different target function was selected, the algorithm would have taken a different route through the analysis space. This means that questionable research practices such as SHARKing may be more challenging to pursue.

As with any analysis approach, using active sampling methodologies comes with inherent trade-offs. Most notably, for more exploitative problems, where the optimal analysis approach is known (or approximately known) a priori or highly theoretically constrained, then the additional costs (in terms of sequential analysis affecting statistical power and computational burden) are a serious limitation. The optimisation algorithm finding local minima resulting in poor overall performance is another potential limitation; this will depend heavily on the acquisition function, including the type used and hyperparameters controlling exploration and exploitation as well as decisions regarding the GP regression and types of kernels used to model the low-dimensional space. A related issue is the creation of the low-dimensional space itself; this will inevitably involve a trade-off between capturing relevant variance and creating a relatively simple search space, with few dimensions. It is shown here that the search space is coherent (in terms of the placement of similar analysis approaches near each other - Figure 5.2) and the GP regression can capture regularities in the space efficiently (Figure 5.3). However, building a compact search space may be more challenging for other problems, e.g., lower signal-to-noise, more heterogeneous variability across individuals, or more heterogeneous analysis approaches. Future work is needed to find the most useful acquisition function, GP regression and search spaces for applying active sampling approaches to multiverse analyses.

5.5 Appendix

Table 5.1: Different analysis approaches used to create the pipelines to be evaluated. *Every pipeline was composed of one type of regression data, one graph theory metric and a threshold leading to the creation of* $2 \times 16 \times 17 = 544$ *different analysis pipelines.*

Data	Graph Theory Metric	Threshold
Motion regression	Degree	0.4
Global signal regression	Strength	0.3
	Betweenness-centrality	0.25
	Binary clustering coefficient	0.2
	Weighted clustering coefficient	0.175
	Eigenvector-centrality	0.15
	Subgraph-centrality	0.125
	Local efficiency	0.1
	Modularity (Louvain)	0.09
	Modularity (ProbTune)	0.08
	Participation coefficient	0.07
	Module Degree ZScore	0.06
	Pagerank-centrality	0.05
	Diversity coefficient	0.04
	Gateway degree	0.03
	K-core centrality	0.02
		0.01

Table 5.2: List of the data, threshold, graph theory metric and obtained mean absolute error (MAE) for the empirical optima obtained for the 20 iterations.

Data	Threshold	Graph Theory Metric	MAE
Motion Regression	0.150	Betweenness Centrality	-2.342
Motion Regression	0.090	Betweenness Centrality	-2.299
Motion Regression	0.100	Betweenness Centrality	-2.326
Motion Regression	0.100	Betweenness Centrality	-2.299
Motion Regression	0.100	Betweenness Centrality	-2.401
Motion Regression	0.175	Betweenness Centrality	-2.274
Motion Regression	0.100	Betweenness Centrality	-2.395
Motion Regression	0.100	Betweenness Centrality	-2.331
Motion Regression	0.070	Modularity (louvain)	-2.280
Motion Regression	0.175	Gateway degree	-2.324
Motion Regression	0.175	Gateway degree	-2.274
Motion Regression	0.175	Gateway degree	-2.274
Motion Regression	0.175	Gateway degree	-2.333
Motion Regression	0.030	Pagerank Centrality	-2.401
Motion Regression	0.030	Pagerank Centrality	-2.369
Motion Regression	0.175	Eigenvector Centrality	-2.401
Motion Regression	0.175	Eigenvector Centrality	-2.432
Motion Regression	0.030	Degree	-2.300
Motion Regression	0.020	k-core centrality	-2.274
Motion Regression	0.040	k-core Centrality	-2.301

6 Multiverse analysis of data modelling pipelines using data-driven spaces

6.1 Introduction

The multiverse analysis framework introduced in Chapter 5 is applied to the specific problem of functional connectivity data, where different preprocessing parameters are considered. This framework is problem agnostic as it solely depends on a multiverse of choices existing, with each one having a tangible outcome that could be mapped in terms of similarity to other outcomes. In this chapter I generalise the framework for a large range of predictive modelling algorithms, that can be exploited for any dataset where we can map independent variables to a variable of interest.

The abundance of data generated and the development of powerful predictive model solutions have made machine learning (ML) ubiquitous in many industries and most research fields. One big limitation to the implementation of ML solutions is that there is no single model that outperforms all others for any given dataset (Wolpert and Macready, 1997). The choice of model will depend on the data linearity; its dimensionality; how missing values are inputted; etc. After choosing a set of methods, they will still depend on the careful tuning of their hyperparameters. Additionally, most raw data requires preprocessing for optimal model performance. For these reasons, ML solutions largely depend on experienced machine learning practitioners and on time-intensive brute force fine-tuning of successful configurations (Olson et al., 2016). To alleviate these limitations, the field of automated Machine Learning (autoML) has focused on building algorithms that create hands-free solutions for any given dataset, addressing the *Combined Algorithm Selection and Hyperparameter optimisation problem* (CASH) (Thornton et al., 2013). This is a hard problem because of the vast number of possible combinations of imputation methods, preprocessing algorithms and predictive models that make up a pipeline of data modelling. To address this challenge, this project organises the pipelines into a Euclidean configuration space, Θ, organised based on performance similarity of pairs of pipelines. The configuration space is built directly from the raw values of prediction generated by each pipeline on several datasets. This approach assumes that pipelines that perform similarly across several datasets are likely to perform similarly for new tasks presented.

This process benefits from being model agnostic as it takes no consideration of which is the pipeline that is generating the prediction, only its position in the configuration space. This allows the user to be better informed of the performance of different data modelling pipelines while only sparsely sampling the configuration space. From an explorative perspective, this Euclidean representation of data modelling pipelines into a configuration space allows for an efficient multiverse analysis of the different pipelines (i.e., how does the result change for different analysis) and a better understanding of the robustness of a given model of responses. From an exploitative perspective, this space allows to obtain an optimal data modelling pipeline while not requiring the user to try out all combinations of models, which would be time-consuming and prone to overfitting to the validation set. Finally, this approach opens up the possibility of pre-registering the exploration of pipeline choices by using an active learning mechanism to sample from the configuration space.

For this project, 20.000 instantiations of pipelines are randomly generated, comprising missing data imputation, preprocessing techniques and modelling algorithms, from a rich library of machine learning algorithms (see Table 6.1)and build a collection of prediction data on 64 diverse datasets from OpenML (Vanschoren et al., 2014), which we refer to as the prediction space P. Multidimensional Scaling (MDS), a dimensionality reduction technique that privileges local information in hyper-space, is used to create an Euclidean embedding of the pipelines that generated the predictions. This unsupervised approach naturally embeds in the space relevant meta-feature information that are imperceptible to evaluation metrics. As the spatial information encodes the inter-relationships between pipelines (i.e., pipelines that are close together are similar in their prediction compared to pipelines that are further away), this system presents a method to build ensembles that maximises prediction dissimilarity.

The main contributions of this project to the field of data modelling research are:

1. the creation of a space of pipelines organised by their prediction similarity - here named the configuration space;
2. a multi-purpose software that explores the space for any dataset with a target variable to both inform about the multiverse analysis of pipelines and optimising the pipeline-selection for a given classification task - here named ModelZoom.

I further demonstrate how it allows for an efficient optimisation of pipelines while testing memory and time constraints, by using a neighbouring approach when sampling the space. This novel approach is flexible to receive new settings as it can contain an open-ended library of algorithms, it can optimise for any evaluation metric that is a function of the prediction, and it allows to map the pipelines to any constraint. Finally, I show the applicability of this tool for data modelling in a proof-of-concept. This tool can be applied to problems of data modelling in any field of research, as the only existing constraint is that the problem that is being tackled is a classification task. One such field where there could be benefit for such a system is in neurodevelopment research. There has been an effort in this field to move from group-level to individual-level analysis in search for relevant biomarkers to identify elusive signs of psychiatric conditions (Baker and Kandasamy, 2022, Latal, 2009, De Ridder et al., 2020). One example is in the early diagnosis of Autism Spectrum Disorder (ASD), a condition known to cause very heterogeneous behavioural responses in patients, making it challenging to diagnose. Early signs of atypical disinterest in human faces (Jones and Klin, 2013, Maestro et al., 2002) and associated atypical neural responses to facial stimuli have been studied as a possible biomarker of ASD in the first year of life (de Haan, 2007). In Tye et al. (2022), this specific problem is considered by studying how a face processing event-related potential task in 8 months-old infants with (n=148) and without (n=68) older siblings with ASD can be explored as a predictor of ASD. I employ the same data using the introduced tool to explore the benefits and limitations of considering the multiverse of available pipelines when building individual prediction models.

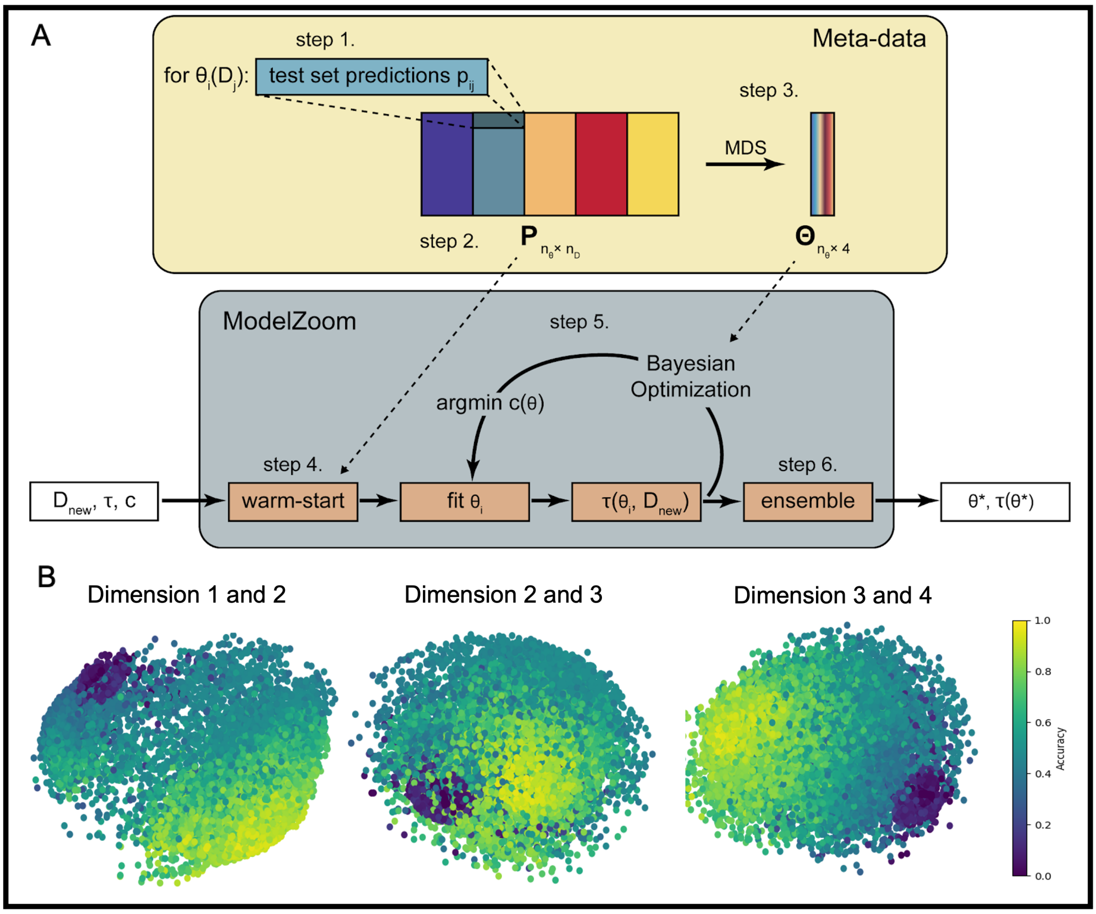

Figure 6.1: A visual description of the proposed method and space. *A. A configuration space Θ is created based on the predictions of 20.000 pipeline instantiations on a large set of datasets. Every individual pipeline θ is trained on a partition of each dataset $D_{j,train}$. They then predict a hold-out set, $D_{j,test}$ and the predictions $p_{i,j}$ are concatenated to build the prediction space P. Using multidimensional scaling, the high dimensional space is converted to a low-dimensional Euclidean embedding of the pipelines, our configuration space. The meta-data collected and the configuration space guide the warm-start module, the ensemble constructor and orient the sampling through Bayesian optimisation. When a new dataset is presented to the autoML system, it is trained and evaluated through cross-validation on 10 pipelines based on their meta-data performance (i.e., the warm-start module). Then the sampling of pipelines is guided by a Bayesian optimisation algorithm and, when the system requires constraining, it uses a neighbouring approach to sample the pipeline that minimises constraint for the sampled region. Lastly, the system uses the information provided by the sampled pipelines to build an ensemble of pipelines directed by performance and distance covered across the space. This ensemble is the solution provided by our system for any given dataset. B. Visual representation of the 4 dimensions of the configuration space colour coded by the mean accuracy across all datasets used for generating the space. Each point in the space represents one of the 20.000 data modelling pipelines.*

6.2 Methods

6.2.1 Defining the pipelines and generating the meta-data

This project's full model selection autoML system optimises over the choice of imputation method, preprocessing techniques, predictive models and their respective hyperparameters. Each pipeline's configuration is randomly selected from a pool of 5 imputation methods, 12 data preprocessing methods, distinguishing between numerical and categorical features, 18 classifiers and a total of 94 hyperparameters. The full list is printed in Table 6.1. A total of 20.000 pipelines were selected at random to populate the configuration space Θ. From this list, 1185 pipelines were unsuccessful and were discarded. The configuration space was built using prediction data, by training and evaluating each pipeline on 64 different datasets from OpenML (Vanschoren et al., 2014), including binary and multiclass datasets and datasets with missing values. The datasets were chosen solely based on their OpenML rating score. Each pipeline was trained with 5-fold cross-validation on 75% of the available data for a given dataset, and the trained pipeline's predictions on 25% of the data were stored to build the configuration space. Furthermore, other meta-data from the trained pipelines were saved (e.g., memory required, training time, testing time and accuracy).

6.2.2 The configuration space

This approach focuses on building a low-dimensional Euclidean configuration space that draws information solely from pipelines' predictions on a large number of datasets. It was assumed that if two pipelines have similar predictions in several different datasets, it is more likely that their prediction is similar for a new unseen dataset. This allows the algorithm to infer the predictions of pipelines neighbouring the sampled pipeline. This process, coupled with a Bayesian optimisation sampling mechanism, reduces the number of samples required to quantify pipeline performance across the full range of 20.000 configurations. Instead of focusing on an arbitrarily chosen evaluation metric to draw similarities between pipelines (Fusi et al., 2018), an Euclidean embedding of pipelines θ based on their raw predictions on hold-out sets was built. This approach does not limit the optimisation to one evaluation metric, instead generalising for any mapping of predictions. The organisation of the space Θ is obtained in an unsupervised fashion based on prediction but it approximately encapsulates other features in its structure such as evaluation metrics and sample-dependent performance of models.

Bayesian optimisation struggles to handle high dimensional spaces (Grünewälder et al., 2010), with many approaches seeking to search over a well-defined low-dimensional projection (e.g., Wang et al. (2013)). It follows that defining such a low-dimensional space is challenging. In this work, a low-dimensional Euclidean embedding $\Theta \in \mathbb{R}^{4n\theta}$ is constructed from the highly dimensional space of predictions, using metric multidimensional scaling (MDS) (Cox and Cox, 2000).

Thus, the configuration space was created based on dissimilarity measures, reducing the original space to $\Theta \in \mathbb{R}^{4n\theta}$. Besides allowing for better interpretation and visualisation of the configuration space, the reduced dimensionality facilitates the use of Bayesian optimisation to efficiently sample the space.

6.2.3 The sampling algorithm

The creation of a Euclidean embedding space of the pipelines' predicted output allows the system to infer the predictions of all pipelines as a function of the already evaluated pipelines. For any task t_{new}, our objective is to find the pipeline θ that generates the maximum of the underlying function f of a given evaluation metric τ that is dependent on a pipeline's prediction, $p_{i,new}$:

$$f(\theta(t_{new})^*) = \tau(p_{i,new}) = y^* \tag{6.1}$$

To find the maximum with a principled approach, Bayesian optimisation was used across the configuration space. It uses a surrogate function to map a probabilistic model of the evaluation metric across the configuration space based on the previously sampled pipelines' performance. Then it balances modelled predictions and uncertainty using an acquisition function to determine the utility of candidate pipelines to sample next. The suitor pipeline is evaluated through 5-fold cross-validation and its performance is used to update the surrogate function and define where to sample next. The kernel used to model the space covariance function was the Matérn kernel summed with a White kernel to account for random noise affecting pipeline performance. This choice of kernels accounts for a non-smooth assumption of the underlying space, as the configuration space order translates only an approximation of the evaluation metric. The selection of the next pipeline to evaluate is guided by the *Expected Improvement* (EI) acquisition function (Kandasamy et al., 2016) introduced in Section 2.2.

6.2.4 Warm-starting with meta-data

The Bayesian optimisation framework uses an acquisition function to guide where to sample next based on previously collected samples. When the system starts, as there are no pipelines sampled, the acquisition function cannot efficiently determine which pipeline to train next. This is known as the cold start problem. Typically, this issue is addressed with a burn-in phase, where trained pipelines are assessed (often selected at random) to populate the space before running the acquisition function. Instead of randomly drawing pipelines from the space, the initial sampling is guided by drawing pipelines which tend to present a good performance in a timely manner. Following approaches introduced by Misir and Sebag (2013) meta-data was collected by training the pipelines on a large number of datasets to build a rank-order scoring mechanism to the pipelines' mean and median regret and training time. each pipelines' rank orders were summed over and the 10 pipelines with the best performance on these criteria were chosen as the initial samples for the sampling algorithm.

6.2.5 Constraint mapping

In autoML, it is often the case that the user needs the system to minimise certain constraints unrelated to the evaluation metric, such as training time or memory resources. It can also be the case that the user intends to constrain the optimal pipeline to partake in certain conditions (e.g., containing a given class of models). Here, the constrained Bayesian optimisation literature was followed (Gelbart et al., 2014, Gramacy et al., 2012) to allow for these optimisation constraints. The underlying information of the configuration space was leveraged on the similarity of predictions p on neighbouring pipelines θ to build a principled constraining mechanism. This mechanism requires that a mapping of the expected constraint values for all pipelines be passed to the system, here named the constraint map c. Specifically, the pipeline sampled at each iteration of the BO algorithm will be chosen based on the minimisation of c for the 5 nearest neighbours of the sample chosen by the acquisition function. The collected meta-data was used to build constraint mappings evaluations constrained on training time and memory resources of the pipelines that populate the configuration space.

6.2.6 The ensemble constructor

Ensembles of models often outperform single models (Guyon et al., 2010, Lacoste et al., 2014). Ensembles are a powerful way to combat overfitting by weighting different predictions on the same data. For this reason, the predictors that compose the ensemble should be individually strong and diverse (Dietterich, 2000). This system is particularly well-suited for ensemble building, given that the configuration space is mapped based on prediction dissimilarity of each pipeline and the optimisation algorithm predicts how strong each pipeline is on a given evaluation metric. By measuring diversity on the basis of distinct predictions on prior experiments, the system can capture differences invisible to the common approach of minimising the correlation of the errors.

Following Feurer et al. (2015) analysis of different types of ensemble builders, the ensemble constructor uses ensemble selection (Caruana et al., 2004). The ensemble is built iteratively by taking the ordered 200 best pipelines considered by the surrogate function and, from this list, discarding the pipelines that lay closer than a 5% distance in the configuration space from the models already considered for the ensemble. A proto-ensemble with the added top pipeline is then trained with an accuracy-adjusted weighting, and if the performance improves on the previous best, the full ensemble is considered for the next iteration of models at a larger than 5% distance. If not, the pipeline is discarded.

6.3 Results

6.3.1 Space disposition analysis

Here it is considered how well the configuration space of 20.000 pipelines compactly represents the pipelines' predictions and how this allows optimisation of different evaluation metrics and different resource constraints.

The space order is evaluated by measuring the correlation of predictions of the pipelines for 10 different datasets. The Fisher transformation of Pearson's correlation is calculated between predictions on a hold-out set along the distance between pipelines in the configuration space. As expected, and denoted in Figure 6.2.B, there is an inverse relationship between pipelines' distances in space and the prediction similarity (Spearman's $\rho = -0.64, p < 0.0001$). This relationship allows the algorithm to explore the neighbouring pipelines with more desirable characteristics (e.g., less resource-intensive) but with similar prediction boundaries.

The configuration space order is again denoted by running a two-dimensional non-linear binary task of an artificially generated spiral dataset. Each pipeline's accuracy and f1-score on the spiral dataset is mapped onto the space in Figure

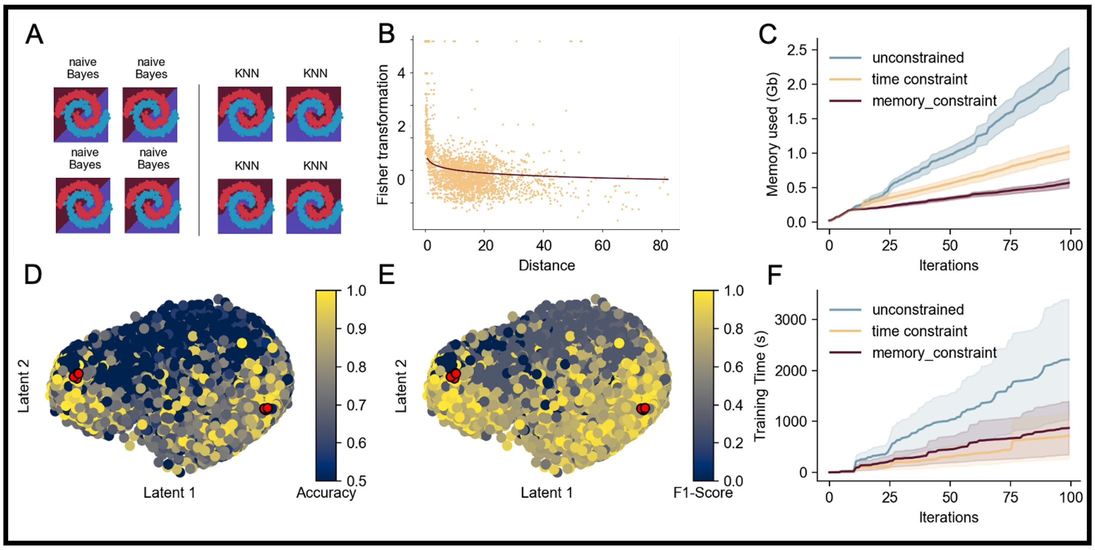

Figure 6.2: Study of the space organisation and constraint optimisation using a synthetic dataset. *A. Two examples of the decision boundary of neighbouring pipelines for the spiral task; B. Map of correlation of predictions vs. distance in the configuration space for the spiral dataset, fitted to an exponential function. C. Mean cumulative sum of the training time of the system across 100 iterations for an unconstrained system (blue), a system constrained to minimise computational resources (yellow) and a system constrained to minimise time resources (red) for 10 different datasets; D. Configuration space mapped to the accuracy of each pipeline for the spiral task; E. Configuration space mapped to the F1-Score of each pipeline for the spiral task; F. Mean cumulative sum of computational resource allocation across 100 iterations for an unconstrained system, a system constrained on time resources and a system constrained on memory resources for 10 different datasets.*

6.2.D,E. Different evaluation metrics will not necessarily optimise for the same pipelines nor the same region of the configuration space. Because the space is created based solely on prediction similarity, it will present structure and organisation for any specific objective related to predictions, even if they are not necessarily overlapping. For example, the f1-score, which is the harmonic mean of Precision and Recall, will better measure incorrectly classified cases than the accuracy metric. The desirable evaluation metric depends on the task being solved and on the context of the problem, but if it is a function of the pipeline prediction, it will find the coherence needed in the configuration space to run the sampling optimisation.

Two constraint mappings were created, which optimised for the mean training time and mean memory allocation to each pipeline in the configuration space using the meta-data collected when building the space. Three systems were tested in similar conditions, one unconstrained, one with training time constraint and one with memory constraint. Each was run for 100 iterations on 10 heterogeneous datasets obtained from OpenML (Vanschoren et al., 2014) that were not used to create the space (5 binary tasks, 5 multi-class tasks, 5 datasets with missing data) with accuracy as the evaluation metric. After 100 iterations the mean running time (in seconds) for the system and the standard error with time constraint were 717.9s (±448.2), for the system with memory constraint was 872.2s (±524.3) and for the unconstrained system was 2216.6s (±1184.5). The mean memory resources requested (in Gigabytes) for the system with memory constraints was 0.566Gb (± 0.069), for the system with time constraints was 1.017Gb (± 0.103) and for the unconstrained system was 2.230Gb (± 0.303). Despite the 68% decrease in running time compared to the unconstrained system, the time-constrained system presented a 0.009 (± 0.011) mean regret (i.e., difference for the best possible result) and, despite the 75% decrease in memory resources consumed compared to the unconstrained system, the memory-constrained system presented a 0.012 (± 0.020) mean regret. The unconstrained system presented a mean regret of 0.022 (± 0.033). The results of the three systems resource requirements across iterations are depicted in Figure 6.2.C and F.

6.3.2 Proof-of-concept exploring predictive models of diagnosis

In Tye et al. (2022), an ERP task is used to build predictive models of diagnosis of ASD. Specifically, the amplitude and latency of the P1, N290 and P400 ERPs of infant participants are measured for face stimuli with direct gaze, averted gaze, static faces and visual noise. The modelled variable of interest was the ASD diagnosis at 36 months of age which was evaluated using 10-fold cross validation. The Expectation Maximisation algorithm was used for imputation of missing values, a genetic algorithm was used for feature selection and a Support Vector Machine (SVM) with linear kernel was used as the modelling algorithm. An accuracy of 75.7% was obtained with this pipeline.

All pipelines are trained and evaluated on the presented dataset to build a ground-truth of the performance across the configuration space (Figure 6.3.A). This is expensive and time-consuming and in general a sub-optimal method to identify the best pipeline to use as it doesn't take into account the correlation of predictions between datasets. To explore the configuration space organisation, we use our optimisation algorithm, ModelZoom, to sparsely sample the space and train and evaluate the chosen pipelines. Here, the system was run twice for 50 iterations (i.e., it evaluated 50 pipelines per run), once with a more exploitative acquisition function ($\kappa = 0.01$) and once with a more explorative setting ($\kappa = 100$). The more exploitative setting is preferable for finding a pipeline with a good performance on the chosen metric. As such the optimal model chosen by the system, marked with a red star in Figure 6.3.B, obtained an accuracy of 77.3% using 10-fold cross validation. This pipeline was composed of a median value imputation method, a scaling of each feature by its maximum absolute value and a K-nearest neighbour (KNN) classifier with $k = 7$. Then, an ensemble was constructed using the prior information in the configuration space to avoid pipelines whose predictions would be the same - encoded as distance in the space - and an accuracy of 87.0% was obtained for a 4-pipeline ensemble using 10-fold cross-validation. The prediction of the system for the performance across the whole configuration space (presented in Figure 6.3.B) had a Pearson correlation of 0.37 against the ground-truth obtained by training all pipelines. A more explorative acquisition function will transform the problem from an optimisation of pipelines to an active learning of the space. By running the algorithm with an explorative acquisition function for 50 iterations, the prediction of the performance across the space (Figure 6.3.E) obtained a Pearson correlation of 0.47

against the ground-truth accuracy across the space. This was achieved while only sampling 0.25% of all available pipelines. The output of the acquisition function across the configuration space informs what are the prioritised regions to sample. Figure 6.3.C shows that the more exploitative run of the system leads to oversampling the region with higher performance, where Figure 6.3.F shows how the explorative run did a more homogeneous sampling across the space to gain a better understanding of all pipelines' performance.

6.4 Discussion

In this chapter, I presented a configuration space that was created by Euclidean embedding of a large compendium of pipeline's predictions across several datasets. I also presented a system for multiverse analysis and optimisation of data modelling pipelines that, for any given task, efficiently navigates the configuration space to find an adequate pipeline or better understand how the performance varies along pipelines. Each pipeline is composed of an imputation method, preprocessing techniques, a classifier model and each respective tuned hyperparameters. The system uses a Bayesian optimisation sampling method to efficiently navigate the organised low-dimensional Euclidean space and a warm-start method and ensemble constructor, following similar autoML literature (Feurer et al., 2015, Fusi et al., 2018). The configuration space presents a coherent organisation of pipelines across different unseen datasets and metrics as shown in the synthetic spiral dataset and for the proof-of-concept with real infant ERP data. It is also shown that ModelZoom performs well both on exploitation and exploration settings to obtain relevant pipelines. One explanation for the system's good performance is that the space can capture a higher detail of the pipeline prediction than it's possible for a space created on performance data. As an example, two pipelines can present the same accuracy of prediction but mislabelling different data points between the two. This would be imperceptible in the evaluation metric but not when comparing predictions directly. Furthermore, the ensemble can also draw information from the space that cannot be drawn from correlations between predictions. By building the space from a large compendium of predictions, two pipelines with the same prediction on the hold-out set and at a large distance in the configuration space are likely to have distinct decision boundaries, resulting in stronger ensembles. Because the space is organised on pipeline prediction, the only limitation for a pipeline to be included

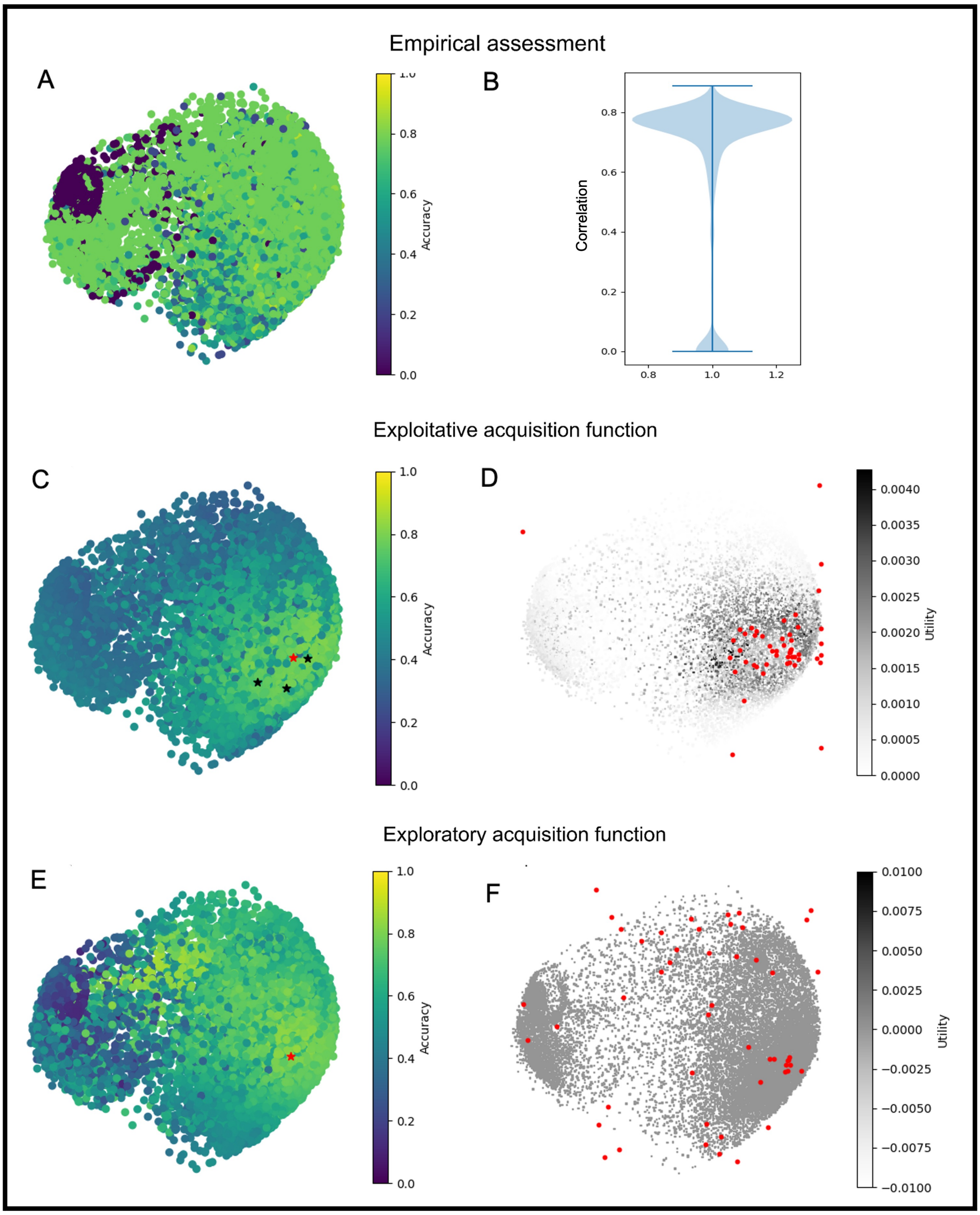

Figure 6.3: ***Evaluation of the proof-of-concept dataset on the space.*** *A. Configuration space (dimension 1 and 2) coloured by the accuracy obtained by each pipeline after training and evaluating on the ERP dataset; B. Violin plot of the distribution of accuracies from the 20.000 pipelines evaluated; C. Configuration space colour coded by the prediction of the ModelZoom run with an exploitative acquisition function. The suggested best pipeline is identified with a red star and the pipelines that make up the ensemble presented are marked with a black star; D. Acquisition function value across the configuration space for the exploitative run. The pipelines sampled by the algorithm are marked in red; E. Configuration space colour coded by the prediction of the ModelZoom run with an explorative acquisition function with the identified best model marked in red; F. Acquisition function across the configuration space for the explorative run, with the sampled pipelines marked in red.*

is for it to map any independent variable sample to a prediction of the dependent variable. Although only shallow classifiers were included in this study due to resource limitations, the same process would allow the inclusion of any deep learning architecture and any combination of hyperparameters. Another advantage of mapping directly from the prediction space of pipelines is that the system is not organised towards one evaluation metric but can generalise for optimising any function of prediction. This can be very useful when dealing with imbalanced class distributions or when the task context requires special attention to false negative or false positive rates, for which the accuracy metric is less reliable.

Another benefit of clustering pipelines with similar predictions is that it facilitates minimising constraints at a low cost on performance. It was shown how using a constraint mapping could minimise training time and memory expenditure at minimal cost in performance by using a neighbour-sampling approach, but competitive constraint mapping can also benefit from the space organisation. An interesting example that is left for future work is to test if a mapping of the interpretability of each pipeline in the space could allow for building more interpretable solutions at a minimal cost in performance.

This work shows the benefit of creating a low-dimensional configuration space that relates directly to the pipeline prediction instead of its performance. However, it does not imply that the dimensionality reduction approach is the optimal method for building the high-dimensional information in the prediction space, most likely it is not; future work could explore instead the decision boundaries of the fitted pipelines or further maximise the number of datasets used for generating meta-data or develop better ways to distil similarity between pipelines into a low-dimension configuration space.

6.5 Appendix

6.5.1 State-of-the-art comparison

The presented system in an exploitative setting was compared to existing full autoML systems. AutoWEKA (Thornton et al., 2013), auto-sklearn (Feurer et al., 2015) and auto-sklearn 2.0 (Feurer et al., 2020) are model-based derivative-free optimisation systems like ModelZoom. Specifically, the auto-sklearn approach follows a similar method to the one presented here as it is also composed of a warm-start module, Bayesian optimisation as the sampling algorithm and an ensemble constructor. We also considered the genetic programming autoML system

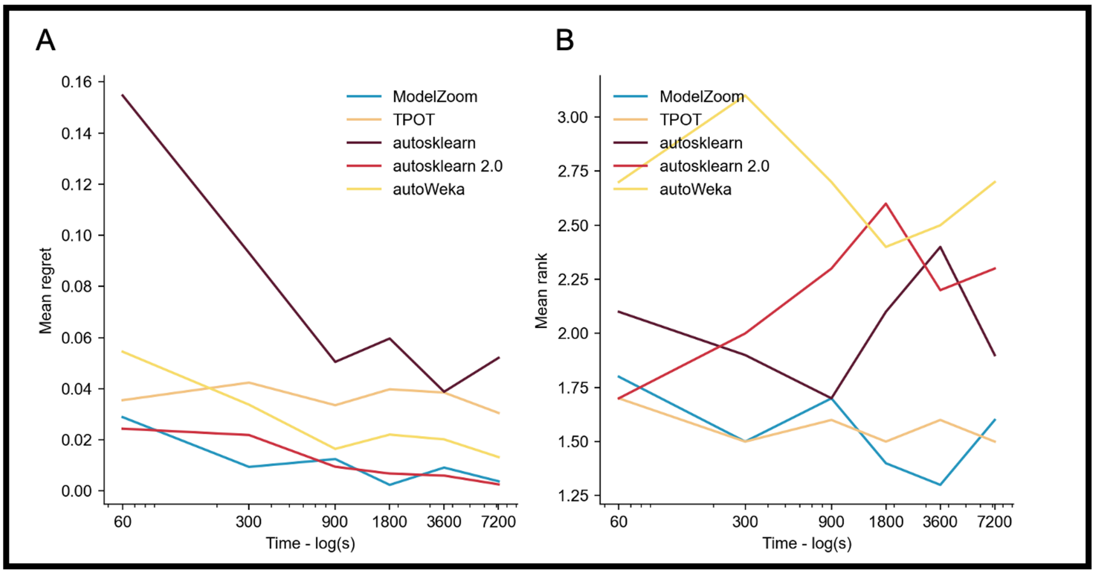

Figure 6.4: Performance against other state-of-the-art autoML softwares. *A. Mean regret for 4 classical autoML systems and our system (blue) for different time limits and 10 heterogeneous datasets; B. Mean rank-order for 4 classical autoML systems and our system (blue) for different time limits and 10 heterogeneous datasets*

TPOT (Olson et al., 2016). The 5 systems were run on 10 OpenML datasets that were not included in the creation of our configuration space. They were run for six different time limits: 1, 5, 15, 30, 60 and 120 minutes on a 28-CPU machine using accuracy as the evaluation metric. In Figure 6.4, we contrast the different systems' performance, demonstrating that ModelZoom performs on par with the common off-the-shelf autoML solutions, especially after shorter periods of training.

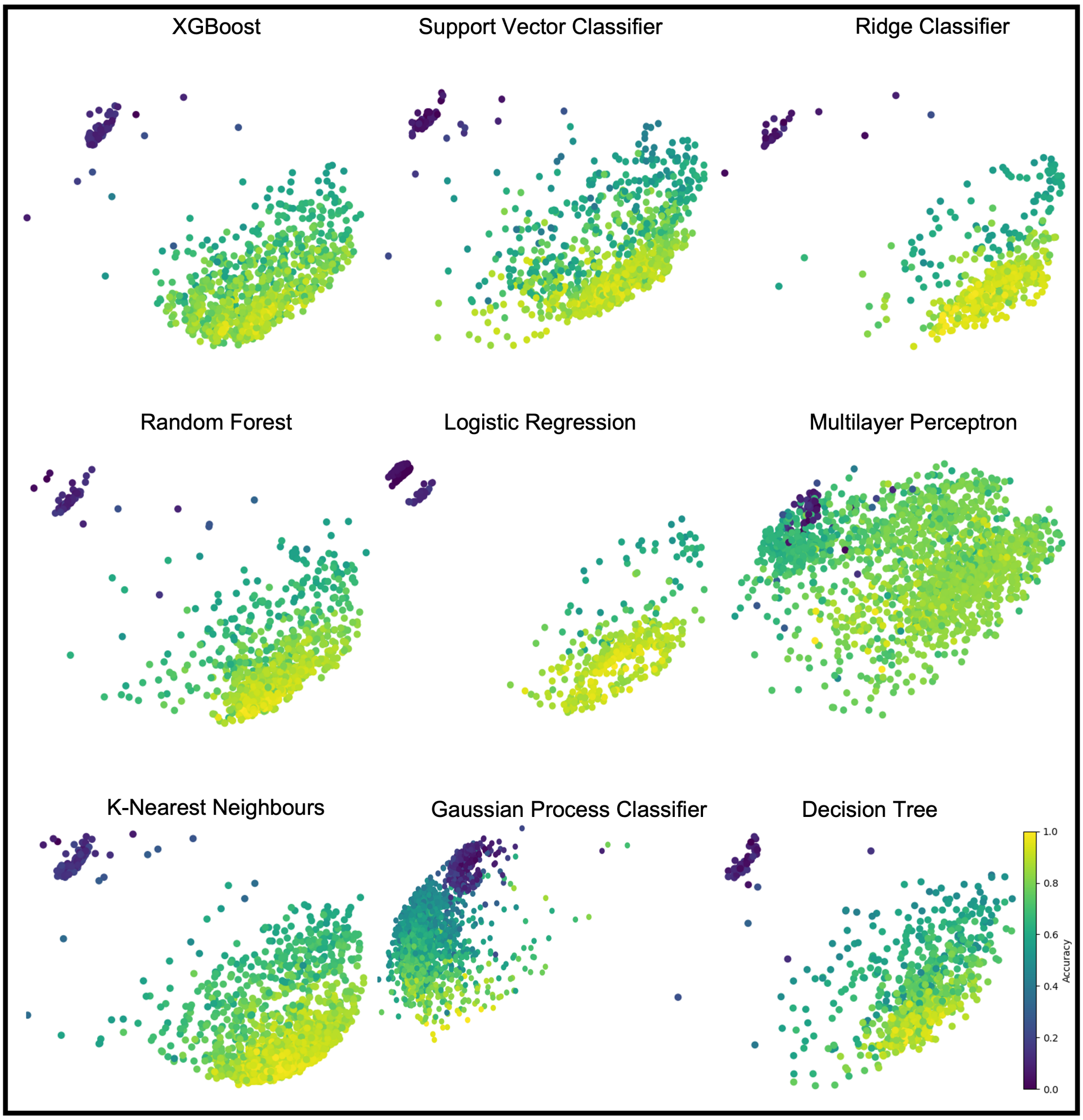

Figure 6.5: Pipelines across the configuration space mapped by their predictive model and colour coded by their mean accuracy in the datasets used to create the space. *From here it is clear how some predictive models greatly overlap in the same region of the configuration space, from where we can infer that their predictions across datasets are similar, and other predictive models, such as the Gaussian Process, seem to generate dissimilar prediction boundaries and are all allocated to a separate region of the configuration space. The dispersion across the configuration space present in most predictive models reveals how the choice of preprocessing algorithms and hyperparameters impacts the expressivity and outcome of the pipeline. Finally, the upper-left corner of the configuration space seems to account for failed pipelines, where the choice of algorithms fails to obtain any prediction.*

Table 6.1: *List of algorithms used and respective number of hyperparameters*

Methods	***Algorithm***	***Nr. of λ***	***Library***
Imputation	Mean substitution	0	scikit-learn
	Median substitution	0	
	Mode substitution	0	
	KNN substitution	2	
	Multivariate feature substitution	2	
Categorical Feature Processing	Ordinal Encoder	0	scikit-learn
	One-Hot Encoder	0	
Numerical Feature Preprocessing	Standard Scaling	0	scikit-learn
	Maximum Absolute Scaling	0	
	Robust Scaler	0	
	Power Transformer	1	
	Quantile Transformer	1	
	Normalization	1	
Discretization	K-bins Discretizer	3	scikit-learn
Dimensionality reduction	Fast-ICA	1	scikit-learn
	Feature Agglomeration	2	
	PCA	2	
	Variance Threshold	1	
Classifier	KNN	3	scikit-learn
	Gaussian Process	5	
	Gaussian Naive-Bayes	0	
	Multinomial Naive-Bayes	2	
	Complement Naive-Bayes	3	
	Bernoulli Naive-Bayes	2	
	Categorical Naive-Bayes	2	
	Decision Tree	7	
	Random Forest	7	
	AdaBoost-SAMME	3	
	Gradient Boosting	7	
	Ridge	1	
	Logistic Regression	5	
	Linear SVM	4	
	Support Vector Machine	2	
	Multi-layer Perceptron	6	
	XGBoosting	6	xgboost
	Relevance Vector Machine	1	sklearn-rvm
Total	36	82	3

7 | Transformer-based anomaly detection of early schizophrenia

7.1 Introduction

Schizophrenia is a chronic mental health disorder that causes a range of heterogenic psychological symptoms and significantly impairs the quality of life of millions of people worldwide. The current diagnosis gold-standard for schizophrenia, described in the Diagnostics and Statistical Manual of Mental Disorders (DSM-V) (American Psychiatric Association, 2013) relies on a professional inquiring the patient and evaluating the presence of three of the five main symptoms (i.e., delusions, hallucinations, disorganized or incoherent speaking, disorganized or unusual movements and negative symptoms). Issues of inter-rater reliability and ambiguous criteria descriptions (Welch et al., 2013) are factors that are pushing the field of psychiatry to search for more objective, operationalisable and personalised biomarkers of psychiatric conditions.

It has been shown that schizophrenia is associated with subtle brain abnormalities that can be detected with structural Magnetic Resonance Imaging (sMRI) data (Shenton et al., 2001). This has led the field to try to build reliable predictors of schizophrenia by employing machine learning algorithms. The most popular approach utilised supervised learning on structural data to build classifiers (Leonard et al., 1999, Squarcina et al., 2017). Despite presenting modest to good accuracies on the testing sets, most algorithms fail to generalize to the early stages of the condition and to cross-site validation (Pinaya et al., 2016, Vieira et al., 2019). These failures can be attributed to training on small and limited datasets that fail to capture the full distribution of patients and the population in general.

To address these limitations, there has been an effort to move to unsupervised learning techniques that focus instead on building normative models of the healthy brain that try to capture variations from normality as predictors of psychiatric conditions (Marquand et al., 2019) . The challenge of collecting large swathes of brain data becomes more manageable if only healthy participants' data are required. Furthermore, contrary to supervised algorithms, normative models can identify variations from any condition that markedly changes brain structure. Showcasing this idea Wolfers et al. (2018) built a normative model of Voxel-based morphometry and found significant variations between healthy and schizophrenic individuals. They also discuss how the interindividual differences between patients with schizophrenia mask group-level differences to healthy participants. Lv et al. (2021) built a normative model of 48 white tracts and 68 cortical regions and found that patients fell significantly outside of normative ranges. However, no tract accounted for the majority of deviations from normality, highlighting the heterogeneity in schizophrenia representation in brain abnormalities. Recently, researchers have proposed unsupervised anomaly detection algorithms to identify brain pathologies, such as brain lesions, from structural MRI (Baur et al., 2018, Chen et al., 2020). These methods are based on autoencoders to learn a latent representation of healthy brain data. After training, these models assess unknown examples to detect pathologies based on their deviation from normality. In this context, the current state of the art is held by variational autoencoder (VAE) based methods (Baur et al., 2020), which try to reconstruct a test image as the nearest sample on the learned normal manifold, using the reconstruction error to quantify the degree and spatial distribution of any anomaly (Pinaya et al., 2021). However, the success of this approach is limited by the fidelity of reconstructions from most VAE architectures (Dumoulin et al., 2016), and by unwanted reconstructions of pathological features not present in training data, which suggests a failure of the model to internalize complex relationships between distant imaging features (Pinaya et al., 2021). To address these issues, a recent study achieved the state of the art performance in unsupervised brain anomaly detection using an architecture based on transformers (Pinaya et al., 2021).

Transformers have revolutionized language modelling, becoming the primary choice for language-related tasks (Radford et al., 2019, Vaswani et al., 2017). They rely on attention mechanisms that capture the natural sequence of input data, completely dispensing the use of convolutions or recurrences. This mechanism allows modelling the dependencies of input data regardless of their distance, enabling the detection of complex long-range relationships. The robustness of transformers to map input data relationships, whose distances vary widely, makes them great candidates for neuroimaging tasks, especially anomaly detection (Graham et al., 2022, Pinaya et al., 2021).

Here, it is investigated if a normative model with an architecture based on transformers could be used to detect psychopathologies, such as schizophrenia, from brain 3D structural MRI and if it could be further used to study the local variations associated with the condition. The normative model was trained on 3D T1 images of neurotypical individuals (N=1,765). Then, the likelihood of neurotypical controls and psychiatric individuals with early-stage schizophrenia was obtained from an independent dataset (N=93) from the Human Connectome Project. Considering the mean likelihood of the scan as a proxy for a normative score, an AUROC of 0.82 was obtained when assessing the difference between controls and schizophrenic individuals using only unsupervised methods. The presented approach surpassed recent normative methods based on brain age and Gaussian Process, showing the promising use of deep generative models to help in individualised analyses.

7.2 Material and Methods

7.2.1 Datasets

In this study, T1-weighted volumes from healthy subjects were used to train our normative models of the brain. These volumes were from two datasets: the Human Connectome Project - Young Adult (HCP-YA) (Van Essen et al., 2013) and the Human Connectome Project - Development (HCP-D) (Somerville et al., 2018). From the HCP-YA dataset, 1,113 volumes were taken from the "1200 Subjects Data Release"[*]. From the HPC-D, 652 volumes were taken from the "Lifespan 2.0 Release"[†]. In total, 1,765 subjects were obtained (808 male and 957 female) with an age range from 5 to 37 years old (Avg. (SD) = 23.3(8.1) years old). To evaluate the presented method, I used the Human Connectome Project - Early Psychosis (HCP-EP) ("Release 1.1"[‡]), a study with the goal to acquire high quality imaging, behavioural, clinical, cognitive, and genetic data on an important cohort of early psychosis patients. Importantly, this study was performed at a different acquisition site than the data used for training the

[*] http://www.humanconnectome.org/documentation/S1200/

[†] https://www.humanconnectome.org/study/hcp-lifespan-development/document/hcp-development-20-release

[‡] https://www.humanconnectome.org/study/human-connectome-project-for-early-psychosis/document/hcp-ep

Table 7.1: Demographic information for the subjects from the Human Connectome Project Young Adults (HCP-YA), Human Connectome Project Development (HCP-D), and Human Connectome Project for Early Psychosis (HCP-EP). *For the HCPEP, we are presenting the data from the distinguish classes: Control (HP) and subjects with early psychosis (EP). We used Student's t test and Chi-square test to verify if age and gender, respectively, are significantly different in the HCP-EP dataset.*

	HCP-D (n=652)	HCP-YA (n=1,113)	HCP-EP (n=93)		
			HP (n=46)	EP (n=47)	stats
Age, y					$1.87_{p=0.07}$
$Mean_{\pm SD}$	$14.0_{\pm 4.1}$	$28.8_{\pm 3.7}$	$23.6_{\pm 2.8}$	$22.6_{\pm 2.6}$	
Range	5-21	22-37	16-30	19-31	
Sex, n					$2.07_{p=0.15}$
$\text{Men}_{(\%)}$	$301_{(46\%)}$	$507_{(45\%)}$	$29_{(63\%)}$	$37_{(79\%)}$	
$\text{Women}_{(\%)}$	$351_{(54\%)}$	$606_{(55\%)}$	$17_{(37\%)}$	$10_{(21\%)}$	

model. The HCP-EP focus on early psychosis (both affective and non-affective psychosis), within the first 3 years of the onset of psychotic symptoms. This is a critical time period when there are fewer confounds such as prolonged medication exposure and chronicity, and when early intervention strategies will be most effective. From the HCP-EP, the volumes of 46 healthy individuals and 47 volumes of subjects with early psychosis were used, specifically subjects with diagnosed with schizophrenia and the groups were statistically balanced for age and sex (see Table 7.1 for demographic details).

7.2.2 MRI processing

All images were corrected for intensity non-uniformity originating from the bias field using the function N4 bias-field correction (Tustison et al., 2010) from the Advanced Normalisations Tools (ANTs - version 2.3.4) (Avants et al., 2008). The images were also registered to a common space (MNI152NLin2009aSym) using rigid and affine transformation using the RegistrationSynQuick command from the ANTs. At the end, this project collected high-resolution volumes ($1mm^3$), where each volume had 192 x 224 x 192 voxels.

7.2.3 Normative model

The main component of the model is an autoregressive Transformer (Vaswani et al., 2017) that learns a mapping of probabilities of a given sequence of values, which is an approximation of the likelihood of the distribution. Transformers are able to capture highly complex dependencies across large distances due to their attention mechanism that weighs the linear transformation of the input with itself. The computational cost of the attention mechanism scales quadratically with the sequence length, making it unfeasible to sequence the original highly-dimensional brain data. In the presented method, the 3-dimensional sMRI brain data is encoded into a smaller discrete latent space before being fed as input to the autoregressive Transformer. Using a vector quantized variational autoencoder (VQVAE) (Van Den Oord et al., 2017), the dimensionality was reduced 512 times from more than 8 million voxels to 16,128 latent variables. This dimensionality-reduction step makes it computationally feasible for the Transformer to learn the probability distribution in the latent space. Both components are trained separately using only healthy participants. The architecture and methodological details of the VQVAE and the transformer are presented in section 2.1.2.5 and 2.1.2.6.

A decoder-only transformer architecture was used due to its autoregressive nature and because it outperforms other autoregressive models such as the Pixel-CNN (Pinaya et al., 2022). The autoregressive transformer receives as input 1D sequences which are unmasked sequentially so that the model is only informed by values it has already estimated. The 3D latent representation is flattened using the raster scan order before being given as input to the transformer. The categorical nature of the VQVAE latent representation allows the transformer to predict the likelihood of any of the available elements in the codebook. This is done through a softmax non-linear function as the transformer output.

7.2.4 Training the normative model

The VQVAE and the autoregressive Transformer are trained sequentially using the 3-dimensional T1-weighted brain scans from HCP-D and HCP-YA cohorts, composed solely of participants without a diagnosis of a neurological or psychiatric disorder, as training and validation data. The VQVAE is first optimized to correctly reconstruct the brain scans and create an efficient discrete latent representation as its bottleneck. The transformer is trained to predict the next element in the sequence from the latent representation obtained from the VQ-VAE and learns to estimate the likelihood of the next element in the sequence. To try to account for image variability caused by different scanners or acquisition parameters, data augmentation was performed during training by applying random shifts to the contrast and intensity of the brain data.

7.2.5 Evaluation and analysis

I estimate the probability of each index in the latent representation obtained by the VQVAE for each brain scan using the output of the trained transformer. The autoregressive transformer weighs the already evaluated latent indices to predict the conditional probability of the next index in the latent representation. This method will result in the model flagging indices that do not follow the sequencing observed from the trained data by assigning them low values of probability. This follows from cases where there are changes to the normal structuring of the brain as the transformer was trained on a healthy cohort. By summing the log-likelihood of all the elements of the latent representation, a log-likelihood estimation was obtained per individual in the evaluation set. The evaluation set is composed by the HCP-EP cohort, where both neurotypical participants (n=46), and participants with early-stage schizophrenia (n=47) are present. I hypothesise that images acquired from participants with schizophrenia will have a lower log-likelihood than controls due to subtle changes to their brain structure being captured by the transformer as unlikely. This is measured through a correlation analysis between the participants diagnosis and its log-likelihood. The efficiency of the method at identifying individuals with early stage schizophrenia is studied by measuring the Area Under the ROC curve (AUROC) of the log-likelihood estimation with the diagnosis as the target variable.

Finally, it was studied how likelihood measures vary locally for participants with schizophrenia by mapping the likelihood estimations across the latent representations to different regions in the brain space, using the Desikan-Killiany cortical atlas (aparc) and the automatic segmentation volume (aseg) (Desikan et al., 2006) originally provided by the Human Connectome Project. For each of the 113 measured brain regions, a correlation analysis between the region's median log-likelihood estimation and the participants' diagnosis is done and the effect sizes per region are measured using Cohen's d.

7.2.6 Baseline normative models

7.2.6.1 Voxel-wise Brain age prediction with deep neural networks

Brain age prediction consists in building predictive models of participants' age using only their brain data, generally T1-weighted MRI scans. It uses associations from changes in brain structure to the biological age, despite the ageing process not being uniform across a population. As symptoms of psychiatric disorders are exacerbated during ageing, the brain age predictor trained on healthy control participants has been suggested as a normative model for psychiatric conditions, by measuring prediction error (Cole et al., 2019). I used the state-of-the-art in brain age prediction, the Simple Fully Convolution Network (Peng et al., 2021) trained on the HCP-D and HCP-YA 3-dimensional brain data, as a normative model of the brain trained on predicting age. The model's prediction error on the HCP-EP data was used as a proxy measure for detection of participants with schizophrenia.

7.2.6.2 Region-wise Gaussian Process Regression

Following previous literature on phenotyping schizophrenia using normative models (Wolfers et al., 2018), a set of Gaussian process regressors (GPR) were trained to predict regional volumes of 183 brain regions using age and sex as covariables. In inference time, the trained models estimate the predicted brain volume and the confidence of prediction. The z-scores (i.e., the prediction error normalized by the prediction uncertainty) are used as proxies of anomaly by measuring the mean z-score over all regions for each participant. In this study, this is implemented through the Predictive Clinical Neuroscience (PCN) toolkit[§], that is optimised for normative modelling of clinical imaging data.

7.2.6.3 Region-wise Bayesian Linear Regression

As a third baseline, Bayesian linear regressive (BLR) models were used. They were trained to infer the same regional volumes from healthy control participants using age and sex as covariables (Huertas et al., 2017). Similar to this project's approach and unlike Gaussian process regressors, Bayesian linear regressors estimate likelihood-based statistics that can be used as proxies for normative models. They are able to model non-Gaussian predictive distributions, as is the case in this project's approach. As in the Gaussian process regression, I measure as

[§] github.com/amarquand/PCNtoolkit

Table 7.2: Performance of the different normative methods at identifying participants with schizophrenia as outliers. *All methods presented a significant difference between the distribution of controls and individuals with schizophrenia (i.e., p-value < 0.05) for their metrics metric of normality.*

Method	AUROC ↑	t-statistics ↑	P-value
GPR Marquand et al. (2016)	0.636	2.104	0.038
BLR Huertas et al. (2017)	0.678	3.013	0.003
Brain Age Peng et al. (2021)	0.732	8.901	<0.001
VQVAE + Transformer [**Ours**]	**0.828**	**6.033**	<0.001

proxy of anomaly detection the mean z-score over all regions measured for a given participant using the PCN toolkit. The larger the z-score, the more out of distribution the sample is.

7.3 Results

7.3.1 Image-wise detection of schizophrenia

This project's model, which is trained in a fully unsupervised manner without seeing examples of individuals with schizophrenia, successfully flags most cases of early-stage schizophrenia resulting in an AUROC of 0.828. A Pearson's correlation coefficient of 0.568 (p-value = 4.1e-9) was obtained when analysing the correlation between the brain log-likelihood and the diagnosis (subjects with schizophrenia = 1). As shown in Table 7.2 this project's model outperforms all baselines at identifying participants with schizophrenia from the HCP-EP dataset.

7.3.2 Region-level analysis

The normative scores of each cortical region and anatomical structure (from aseg+aparc parcellation) are estimated by calculating the median log-likelihood of the latent variables that are inside the region when upsampling them onto the original brain space. It was found that 16 regions out of 113 presented a different normative score between control and subjects with schizophrenia with a significance level below p=0.05, but none show significance once the result is corrected by the Bonferroni correction for multiple comparisons (Dunn, 1961). These results contrast with the global estimation where the difference between

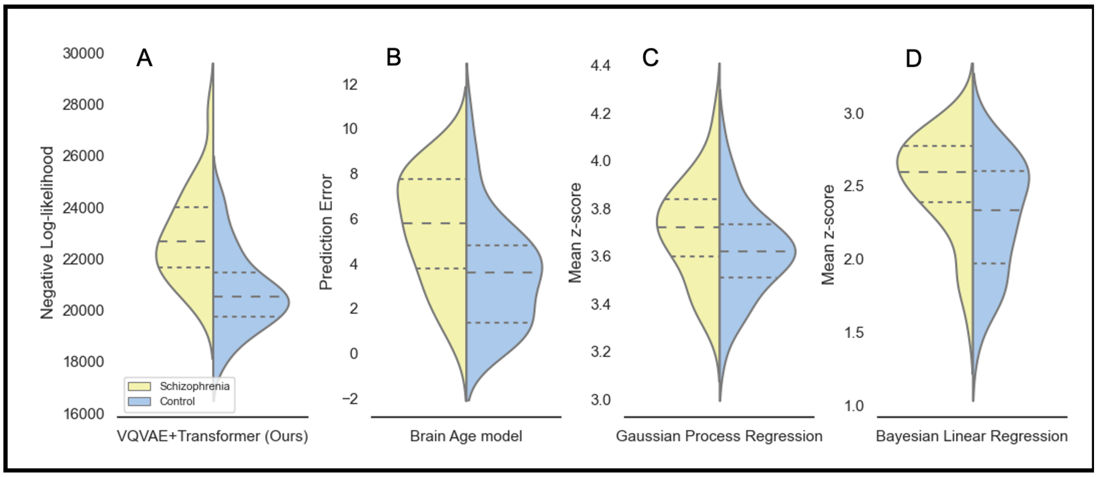

Figure 7.1: Violin plots of proxies used for anomaly detection. *A. presents our model's difference in distribution between control participants and individuals with early-stage schizophrenia using the estimated negative log likelihood per individual as a metric for detecting schizophrenia. B. presents the distributions obtained using the mean absolute error of a brain age model. C. presents the distribution between cohorts for the mean z-scores obtained using univariate Gaussian process regressions and D. presents the distributions of the mean z-scores using Bayesian linear regression. For each distribution, the dashed lines identify the median and the 25th and 75th quartiles.*

log-likelihoods of the two cohorts is significant (p-value < 0.001). The Bonferroni correction is quite conservative and assumes independence between estimations, which in statistical inferences in homotopic and adjacent regions is not a correct assumption. Instead, here we focus on the effect sizes of each statistical test. In Figure 7.2 and Table 7.3, I show the regions with the highest measured effect size (Cohen's d), varying between small and medium effect sizes (between 0.42 and 0.57). The regions with the highest values were present in the prefrontal cortex (i.e., the left precentral gyrus and the right pars orbitalis), the temporal cortex (i.e., the right and left fusiform gyri and the left transverse temporal gyrus), the right lateral ventricle, the anterior portion of the corpus callosum, and the left and right choroid plexus.

7.4 Discussion

In this study, a VQVAE and autoregressive transformers were applied to create a normative method to predict how likely it is for a sample to belong to the normative population. When applying it on subjects from the HCP-EP dataset, this

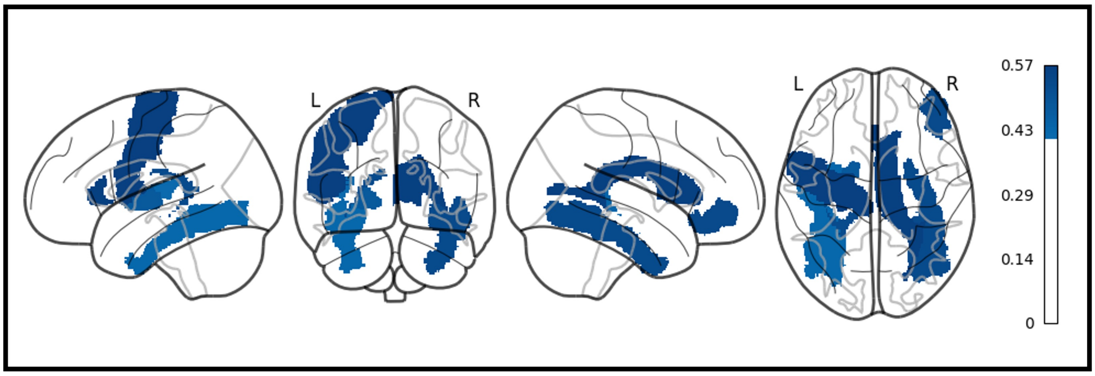

Figure 7.2: Top 10 brain regions with the highest effect size measured by Cohen's d.

Table 7.3: *Top 10 regions with highest effect sizes measured by Cohen's d.*

Rank	Region	Cohen's d ↑
1	Left precentral	0.574
2	Anterior corpus callosum	0.569
3	Right lateral ventricle	0.567
4	Right pars orbitalis	0.537
5	Right fusiform gyrus	0.532
6	Left choroid plexus	0.525
7	Left pallidum	0.479.
8	Right choroid plexus	0.432
9	Left fusiform gyrus	0.430
10	Left transverse temporal	0.420

method was shown to be able to distinguish between healthy controls and participants with early-stage schizophrenia with a AUCROC=0.828. This normative score was more robust compared to other baseline methods, such as the z-score from the fitted Gaussian process regression, the z-score from the fitted Bayesian linear regression or the prediction error of the SFCN brain age model. By comparing state-of-the-art methods, it was observed that the Gaussian process-based approach had the lowest performance. This might be because the latter is limited to modelling uncertainty, the most relevant metric for the normative model, as a Gaussian distribution (Rasmussen, 2004). This is not the case for the other three methods, where error prediction and likelihood estimation can have diverse distributions. This is visible in Figure 7.1 where even when considering the mean of the 113 Gaussian estimations, the GPR modelled uncertainty is the one that more closely follows a Gaussian distribution. The Bayesian linear regression had the second-worst performance, which can be attributed to building univariate normative models of local brain features, as is the case of the GPR. By fitting the confounding variables (i.e., age and sex) to the sole dependent variable (i.e., one specific local region) we are fitting as many models as there are local regions, which makes these two normative approaches incapable of retrieving information between brain regions when building uncertainty estimates. This goes against what is known about how schizophrenia affects the brain, as it is thought to be a condition that affects multiple areas simultaneously as well as the communication between these areas (Lv et al., 2021, Tsuang et al., 1990). Moreover, as we are fitting individual models per region for both BLR and GPR, it was necessary to obtain a low dimensionality representation of the structure of the brain, in this study, characterised by the volumes of the cortical regions and subcortical structures. Due to this limitation, the methods lost a lot of information about the brain of the investigated subject. This is not the case in both the brain age approach and our VQVAE+Transformer model, where the voxel-level data is used to train both models in a multivariate method, where deep neural networks' high non-linearity can extract high-level information from different regions at the same time in a fully data-driven manner. One explanation why the VQVAE+Transformer model outperforms the brain age approach is that learning to map the age of the individual through MRI brain data results in using the error as a proxy for identifying outliers which is an indirect task. The VQ-VAE+Transformer model learns to explicitly predict the likelihood distribution of brain data from healthy participants. Another advantage of our model against brain age modelling is that it outputs a regional estimation of likelihood, that can be used to analyse the local structures for biomarkers of a given anomaly. There is no direct local representation for the brain age model as it is predicting a single value, the age of the participant.

One important limitation of the presented model is that it does not account for demographic and other information about the participants, such as age and sex when building the normative model. This is relevant because these confounding variables will greatly impact what is considered normal (i.e., the normal brain structure at 15 years of age is very different than at 50) and other factors (e.g., smoking, BMI) may impact brain structure and differ systematically between groups. One avenue that is left for future work is to condition the transformer estimation based on context to have a likelihood estimation that is demographic dependent. Another limitation of the presented approach is the robustness of the model on data that do not follow the exact distribution of the training data (Molina et al., 2017). The model fails to identify healthy brains, marking all data with low likelihood. This shift in distribution can happen when using different acquisition settings in the scanner or different field strength. In this work, an external dataset is used (i.e., the HCP-EP) but it is under the umbrella of the Human Connectome Project as is the training data. One solution that was implemented here and can be further explored is data augmentation of the training data in order to simulate the outputs of these different settings and increasing the model's robustness to distribution shifts.

The difference in likelihood distribution between groups in individual cortical regions shows small effect sizes when measured through Cohen's d. One justification for larger effect sizes not being found can be that as the dataset only comprises early stages of schizophrenia, the data is not characterised by large scale pathology clearly visible on MRI (unlike many neurological disorders, e.g., focal stroke, Alzheimer's disease). Therefore, differences have to be relatively subtle and distributed across large areas of the brain, possibly with large individual variability. Relevantly, the global analysis of the likelihood presented significant difference between distributions, so there is no sole segmented region that is driving the correct prediction of participants with schizophrenia and taking the whole brain estimation benefits the prediction power. This assessment follows closely with what is known in the literature, as schizophrenia is known as a condition with heterogeneous brain structure profiles (Lv et al., 2021, Tsuang et al., 1990). The regions that presented largest effect sizes follow what is known from the literature about brain changes related to schizophrenia. The largest effect size was found in the left precentral (p = .0067; Cohen's d = .57), a region that has been associated with the mechanisms that underlie the onset of the psychiatric conditions (Rimol et al., 2010, Shepherd et al., 2012, Zhou et al., 2005). The right pars orbitalis (p = .0101; Cohen's d = -.54), a region associated with language processing (De Carli et al., 2007), is part of the inferior frontal lobe, which is associated with grey matter reduction in schizophrenic individuals (Shepherd et al., 2012). The left transverse temporal (p = .0458; Cohen's d = .42) and the right and left fusiform gyri (p = .0119; Cohen's d = .53 and p = .0406; Cohen's d = .43 respectively) are components of the temporal cortex, a region that has been associated with volume reduction (Shepherd et al., 2012) and positive symptoms (Walton et al., 2017) in participants with schizophrenia. The volume reduction in brain structures observed in patients with schizophrenia has been associated with substantial enlargements of the ventricles (Rimol et al., 2010). This can explain the observed effect size of the right lateral ventricle (p = .0074; Cohen's d = .57). The anterior corpus callosum (p = .0072; Cohen's d = .57), that connects the orbital, medial and lateral surfaces to the frontal lobe, is not commonly associated with schizophrenia in meta-analysis and reviews, but its roof forms the body of the lateral ventricle (Chaichana and Quiñones-Hinojosa, 2019) which can result in the likelihood estimation to confound the two regions. The left pallidum (p = .0230; Cohen's d = .48) sits in the basal ganglia-thalamocortical circuitry, that has been associated with a grey matter reduction in first-episode schizophrenia, but less so in chronic schizophrenia (Ellison-Wright et al., 2008). Importantly, the disruption of this circuitry has been studied to mediate executive functioning deficits in schizophrenia (Camchong et al., 2006). Finally, the right and left choroid plexi (p = .0400; Cohen's d = .43 and p = .0013; Cohen's d = .53 respectively) have also been found to have significant differences in previous work on normative modelling of psychiatric conditions (Pinaya et al., 2019). Taken together, these findings suggest that the presented model was sensitive to subtle neural changes that are directly associated with schizophrenia. The higher effect size in symmetric structures (i.e., the left and right choroid plexi and the left and right fusiform gyri) despite the transformer processing the brain structure sequentially and thus breaking its spatial information, is another example of how the algorithm is capturing real deviations from normality and not just spurious variations.

Future work will focus on further analysing local variations from normality at a finer grain (e.g., intra-region correlations), and evaluating how clinical scores from participants map onto deviations from normality in different regions. It is possible to measure if some regions' deviations in structure are more striking for positive or negative symptoms and if biomarkers can be associated with some specific phenotyping of the psychiatric condition.

7.5 Conclusion

The diagnosis of early psychosis is a challenging task, and several data-driven methods have been developed to help in this task. In this study, we demonstrated the potential of using deep generative models to assign the likelihood of MRI data belonging to a healthy population. The Transformer was better at identifying early psychosis patients compared to traditional volumetric based approaches and brain-age voxel-based approach. Furthermore, this approach allows the researcher to explore local maps of likelihood and retrieve information about how cortex regions are affected for each individual participant, allowing for new possibilities in personalised psychiatry.

Normative models benefit largely from not requiring any data from participants with conditions when training, and only learning from healthy controls, which are data that is more readily available. This approach can be extrapolated for rarer conditions where the absence of large cohorts tends to lead to unexplored research avenues. Due to fast pace of the development of deep generative models in other fields like computer vision, these models are a promising tool to help in psychiatry.

8 | Discussion

The research presented in this thesis tackles two crucial methodological challenges in neurodevelopmental and related research, the replication and generalisability crises. The main drivers of these crises are the inability to control the researcher's degrees of freedom in less constrained hypothesis settings and a lack of focus on the robustness of scientific results. Both elements can be considered as a lack of control of multidimensional problems, i.e., each methodological direction the researcher can take creates a new dimension of possible results; different recording equipment, different populations, and different settings hide multiple dimensions that are not considered when extrapolating results from data. To tackle this, new methodologies can: 1) focus on automating the experiment process, taking the researcher out of the loop and reducing the experimenter's degrees of freedom; 2) accept the variability present from the multiverse of possible choices and study the robustness of results in multiple-choice paths. For both these directions, there is a need for novel frameworks that can be considered new gold standards of research.

As such, this work focused on taking advantage of novel machine learning techniques to explore the underlying organisation of the multiverse of possible choices. The neuroadaptive optimisation framework was extended to optimise over learned stimuli manifolds and EEG paradigms. Furthermore, this work created novel frameworks using Bayesian optimisation to explore lower-dimensional spaces of methodological pipelines, both in neuroimaging preprocessing and in more general cases of predictive modelling of data. Finally, it extended a state-of-the-art outlier detection algorithm to efficiently learn the normative brain as a robust and unsupervised mechanism for detecting subtle changes to the brain structure. All these frameworks were accompanied by proof-of-concept examples in neuropsychiatry and neurodevelopment research as a clear demonstration of the benefit they pose in addressing the methodological challenges in our field and extending the researcher's toolkit when exploring new hypotheses.

In summary, the key contributions of this thesis were:

1. The extension of the neuroadaptive framework to explore a rich generative model-driven manifold of face stimuli in an automated and controlled manner. This framework extension can be considered for any object stimuli and is not limited to faces;
2. The extension of the neuroadaptive framework for EEG experiments, by automatically processing and evaluating ERP experiments and using Bayesian optimisation to guide the stimulus to be presented next that would maximise the measured signal or information, depending on the research goal;
3. The development of a navigable space of neuroimaging processing techniques that can be exploited to measure the robustness of the obtained results efficiently. This new framework is generalisable for any problem of multiverse analysis where each pipeline has a clear numeric output;
4. The extension of the latter framework to the more general case of predictive modelling. The space of modelling pipelines can be of use to any modelling problem where the robustness of the predictions matters for generalisation purposes;
5. The extension of the VQVAE + transformer algorithm to build unsupervised normative models of the human brain as a generalisable mechanism to detect subtle changes in brain structure caused by psychiatric conditions, such as early-stage schizophrenia.

8.1 Summary

The topics discussed in this thesis can be summarised as follows:

Chapter 1: The lack of reproducibility and generalisability of research results are impacting research fields such as neuropsychiatry and neurodevelopment. This chapter addresses how a blind spot to the challenges of the curse of dimensionality in methodological research can lead to such issues and how we can distil the relevant information down to lower dimensions to tackle the problems of robustness and reproducibility better. It also describes how machine learning has already introduced the algorithms needed to take the field in this direction.

Chapter 2: The algorithms used in this thesis for both building lower-dimensional representations of the data as well as efficiently and automatically navigating through these spaces are introduced in this chapter.

Chapter 3: Investigating the cognitive and neural mechanisms involved with face processing is a fundamental task in modern neuroscience and psychology. To date, the majority of such studies have focused on the use of pre-selected stimuli. The absence of personalised stimuli presents a serious limitation as it fails to account for how each individual face processing system is tuned to cultural embeddings or how it is disrupted in disease. The same stimulus can have different interpretations and elicit different results from different populations limiting the capacity for study replication in cognitive neuroscience. In this chapter, I introduced an extension to the neuroadaptive optimisation framework, which combines generative adversarial networks with Bayesian optimisation to identify individual response patterns to a rich set of faces. Formally, Bayesian optimisation is employed to efficiently search the manifold of faces learned by the generative models, with the aim to automatically generate novel faces to maximise an individual subject's response. I presented results from a web-based proof-of-principle study in self-recognition, where participants (n=30) rated images of themselves generated via performing Bayesian optimisation over the face-manifold of a generative model. The algorithm was able to efficiently locate an individual's optimal face while mapping out their responses across different semantic transformations of a face; inter-individual analyses suggest the approach can provide rich information about individual differences in face processing.

Chapter 4: A core goal of functional neuroimaging is to study how the environment is processed in the brain. The dominant experimental paradigm involves concurrently measuring a broad spectrum of brain responses to a small set of environmental features pre-selected with reference to previous studies or a theoretical framework. Recording a broad spectrum of metrics allows the researchers to choose the one that best fits their narratives. Here, I defined an approach where the researcher records the modulation of a single pre-selected brain response in a broad spectrum of environmental features. By using a pre-specified closed-loop design, the approach addressed fundamental challenges of reproducibility and generalisability in brain research. These conditions are particularly acute when studying the developing brain, where our theories based on adult brain function may fundamentally misrepresent the topography of infant cognition and where there are substantial practical challenges to data acquisition. This methodology employed machine learning to map the modulation of a neural feature across a space of experimental stimuli. The method collects, processes and analyses EEG brain data in real-time; and uses a neuro-adaptive Bayesian optimisation algorithm to adjust the stimulus presented depending on the prior samples of a given participant. In a mother-stranger paradigm proof-of-concept, I showed that the method could automatically identify the face of the infant's mother through an online recording of their Nc brain response to a face continuum. This method allows for the retrieval of model statistics of individualised responses for each participant, opening the door for early identification of atypical development. This approach has substantial potential in infancy research and beyond for improving the power and generalisability of mapping the individual cognitive topography of brain function.

Chapter 5: For most neuroimaging questions, the vast range of possible analytic choices leads to the possibility that conclusions from any single analytic approach may be misleading. Although it is possible to perform a multiverse analysis that evaluates all possible analytic choices, this can be computationally challenging and repeated sequential analyses on the same data can compromise inferential and predictive power. Here, I established how active learning in a low-dimensional space that captures the inter-relationships between analysis approaches could be used to approximate the whole multiverse of analyses efficiently. This approach balances the benefits of a multiverse analysis without the accompanying cost to statistical power, computational power and the integrity of inferences. I illustrated this approach with a functional MRI dataset of functional connectivity across adolescence, demonstrating how a multiverse of graph theoretic and simple preprocessing steps can be efficiently navigated using active learning. This chapter showed how this approach was able to identify the subset of analysis techniques (i.e., pipelines) which are best able to predict participants' ages and allow the performance of different approaches to be quantified.

Chapter 6: Although classical statistical analysis has been pushing forward our understanding of the human brain, the last decades have seen an exponential growth of studies employing machine learning to build predictive models of variables of interest. The prevalence of machine learning tools combined with the ever-growing number of predictive algorithms has resulted in the need for a large number of decisions to be taken when choosing how to train an algorithm and preprocess the data. In this chapter, I presented a framework solution that explores the prediction patterns of many ML algorithm pipelines on a large collection of data. I distilled the high-dimensional data to a low-dimensional configuration space that is efficiently sampled through Bayesian optimisation. I demonstrated how the automatically organised space captures information about the neighbouring pipelines for unseen datasets. I further demonstrated how the space organisation could be used for exploring the generalisability of predictions for the multiverse of pipeline options or exploiting the region in the space with the best performance while checking for the robustness of the results by evaluating the neighbouring pipelines. I showed the benefit of this framework with a proof-of-concept in neuropsychiatry, where an EEG dataset was used to build an accurate predictive model for classifying autism.

Chapter 7: Despite the impact of psychiatric disorders on clinical health, early-stage diagnosis remains a challenge. Classification approaches tend to be overly narrow, leading to challenges in generalising the model's performance for clinical practice. The overlap between conditions leads to high heterogeneity between participants that is not properly captured by classification models trained on an under-representative section of the population. To address these issues, normative approaches have surged in popularity as a robust alternative where pathologies are defined by their deviation from normality. In particular, transformer-based models showed great results as normative models to identify neurological lesions in the brain. However, neurological lesions usually are expressed in the data as significant changes in intensity, and experiments identifying subtle changes typically associated with psychiatric diseases are challenging. In this chapter, I evaluated the performance of transformer-based models to detect subtle changes expressed in adolescents and young adults. I trained a normative model on 3D T1 images of neurotypical individuals (N=1,765). Then, obtained the likelihood of neurotypical controls and psychiatric patients with early-stage schizophrenia from an independent dataset (N=93) from the Human Connectome Project. Using the mean likelihood of the scan as a proxy for a normative score, the model obtained an AUROC of 0.82 when assessing the difference between controls and schizophrenic individuals. This approach surpassed recent normative methods based on brain age and Gaussian processes, showing the promising use of deep generative models to help in individualised analyses.

8.2 Robustness at the individual level

Chapter 3 and Chapter 4 delve into the challenges of inter-individual differences in cohorts for generalisability and replication of results when doing group analysis. The same stimulus does not elicit the same brain response at the individual level in a given population, even when not considering random measurement errors. The disregard for the variability of brain responses is one factor that is stated as responsible for the generalisability crisis. This is ever more relevant in neurodevelopment research, where the brain is going through rapid changes which are not deterministic between individuals. In this thesis, I explored and developed frameworks that learn brain responses at an individual level across stimuli spaces in both EEG paradigms and self-reported questionnaires. These maps of brain responses can be used for building common grounds between individuals. For example, instead of relying on face stimuli selected for their optimal response in a small population, each participant can be assessed with face stimuli that are optimised for them individually. They can also be used for a richer group analysis where the individual variation across the stimuli space is considered. In conclusion, exploring these two frameworks is a step in the direction of addressing the impact of inter-individual variability in the lack of generalisability in neurodevelopment and neuropsychiatry research.

8.3 Robustness at the experiment level

Chapter 5 and Chapter 6 explore the challenges in reproducibility and robustness of experimental results that are derived from the abundance of data processing techniques and algorithms. The large space of valid choices can lead to different outcomes from the analysis of the same data. To tackle this problem, I developed a framework to study the multiverse of solutions and how they relate to each other. A better understanding and exposition of the impact of different valid data processing choices will strengthen an experiment's capability of being replicated at different laboratories. The framework worked for neuroimaging preprocessing techniques and predictive modelling algorithms but is extensible for any problem where a multiverse analysis is beneficial. Because it uses a fully data-driven approach to build the multiverse space, it can extract similarity information between techniques. This can facilitate exploring the multitude of options while only running a few experiments.

8.4 Robustness at the cohort level

Chapter 7 considers the challenge of the robustness of results in predictive modelling of small population cohorts. When building predictors of a psychiatric condition, the data privacy constraints and a small cohort of participants lead to the creation of models of prediction that fail to generalise for slightly different settings, recording equipment or ethnicities. To tackle this lack of robustness, I explored normative models using state-of-the-art machine learning algorithms for stronger predictions. A normative model only requires data from healthy controls, which are available in much larger quantities and have fewer issues related to data privacy. By using this approach, we can build more generalisable predictors of psychiatric conditions because these models are not trained for a specific condition but instead to identify outliers to the normal structuring of the brain. Models of normality present a good solution for building more robust and generalisable predictors of outliers, which is the main challenge before these algorithms can be used in a clinical setting.

8.5 Limitations and future directions

8.5.1 Dimensionality reduction

As all these frameworks deal with very large multidimensional data spaces, they all require that data are distilled down to be optimisable, or the frameworks would fail due to the curse of dimensionality. The generative model distils all possible pixel combinations in an image to only account for the face manifold (i.e., the region in the image space where images look like faces). The output of each processing pipeline considered in the multiverse space is concatenated with all others before the high-rank matrix is reduced to three or four dimensions using multidimensional scaling. The 3D brain image is originally more than 3 million voxels and is reduced three orders of magnitude using the VQVAE algorithm. In all these examples, information and data variance is lost. There is a trade-off between the manageability of the multidimensional space and the loss of variability within the data. I show that these algorithms efficiently capture much of the relevant information for each specific case, but future work should consider better methods of distilling information down while maximising retention of relevant information. Furthermore, when building spaces of stimuli or experiments in neuroadaptive optimisation paradigms, it is important to note that not all experiments translate well into this framework. The framework makes some assumptions when creating the space: it assumes the space of experiments or stimuli is continuous, and each axis controls a linear variation of the experiment (such as a continuous variation between an image of the mom and an image of a stranger); the space is bounded so there is a limit to where the optimisation algorithm is able to sample from. Future work could investigate the optimisation of discrete spaces using a similar optimisation framework or alternative adaptive algorithms.

8.5.2 Optimisation algorithm

All presented frameworks use an algorithm to extract information from the analysis space efficiently. For the neuroadaptive and multiverse analysis frameworks, the optimisation algorithm is Bayesian optimisation. Bayesian optimisation has some important limitations that curtail the applicability of these frameworks: 1) The optimisation process is sequential, where some analysis can be parallelised, Bayesian optimisation cannot as each sample depends on being informed on the result of the previous ones; 2) Although the algorithm performs global optimisation, it can get stuck in local maxima if using more exploitative acquisition functions; 3) if the optimal analysis/stimulus is largely known *a priori* or if the problem is highly constrained so that only a small region of the space is explorable, then Bayesian optimisation will be sub-optimal as it will require more resources (and potentially increase the chance of overfitting) then evaluating in an *ad hoc* manner; 4) if the measured signal variation is too low (i.e., if the effective brain response is drowned out by noise or there is limited variation between processing pipeline's outputs), then the surrogate model of the optimisation algorithm will fail to capture the space variation for a given individual or dataset. In traditional hypothesis testing, brain responses with high inherent noise are captured by averaging over many participants. Here, I tried a similar approach by averaging over 12 trials presented to one participant, with good results, but higher recording variability would prove challenging for the optimisation algorithm. Regarding the kernel chosen for each experiment, I little practical variation between the results obtained with a Mátern kernel and an RBF kernel, so both were a valid choice. Furthermore, many other available kernels could also been chosen, given that the kernel prior respects the type of data (e.g., for a cyclical signal, a cyclical kernel should be chosen). Similarly, the choice of acquisition function between UCB and EI did not impact the outcome of the experiments, but many other choices of acquisition function were available.

In the case of normative modelling, the algorithm that extracts information from the latent data is the transformer. Although very powerful at processing information and capturing the probabilistic distribution of the data, this deep learning algorithm has a large computational burden. This limitation can be addressed in future work by considering other probability estimator models, such as the diffusion models that present highly accurate results in computer vision with a fraction of the computation required by the transformer model.

8.5.3 Small proof-of-concepts

For each of the introduced frameworks, I explore a proof-of-concept that evaluates the framework on a neurodevelopment or neuropsychiatry problem. In exploring neuroadaptive optimisation for stimuli spaces, I study the individual variability of responses to a self-recognition task in a large stimuli space in 30 participants. In the extension of the neuroadaptive framework to EEG paradigms, I test out the mom-stranger paradigm in 4 infant participants, where I observe the optimisation to find the maximum ERP amplitude to follow the literature (i.e., the higher amplitude for images of the mums). In the multiverse analysis of preprocessing steps in functional connectivity data, I show how the framework can be applied to find the subset of analysis techniques that best process the data for optimising age prediction of adolescent participants. In the predictive modelling multiverse analysis, I show how space navigation can be used to optimise the prediction of autism participants over an ERP dataset. In the normative modelling framework, I present how it can be used to efficiently predict participants with early-stage schizophrenia as an outlier-identification task. These frameworks have applicability beyond neurodevelopment and neuropsychiatry research, but these are fields where new gold standards in methodology could present clear-cut benefits. Both fields are fundamental to our endeavour to understand the human brain better, but both rely on noisy data such as subjective questionnaires or jittery young participants' brain responses. An improvement in methodological gold standards could help push these fields forward by improving the reproducibility of the research outcomes. The presented proofs-of-concept are insufficient to present these frameworks as good practices as they are quite small and self-contained. Some studies, such as the neuroadaptive framework applied to EEG, were impacted by the COVID-19 pandemic, which limited the number of infant participants that could be tested. As such, although the projects are methodological sound, there is no evidence that these techniques generalise better. Future work will focus on broadening the scope of these frameworks to different tasks and pushing its boundaries in more complex problems to understand their limitations better. For example, the neuroadaptive EEG is currently being extended to contrast studies in adults and gaze vs emotion paradigms in infant research (Gui et al., 2022). The normative modelling framework is being applied to baby brain structures to study how pre-term variability from normality relates to future psychiatric conditions. To build evidence based examples of generalisation, we need to test different cohorts in different settings, using the proposed methods and evaluate how well the obtained results generalise for different populations.

8.6 Concluding remarks

This thesis focused on developing and expanding frameworks to address the challenges presented by the replication and generalisability crisis in neurodevelopment and neuropsychiatry. I associate these problems with the lack of robustness of research findings in different settings and populations and try to tackle them at the individual, experimental and cohort levels. At the individual level, I expand on the neuroadaptive optimisation framework for automatic optimisation of EEG paradigms and navigation of face spaces to capture variability that is hidden in group analysis. At the experiment level, I build and navigate Euclidean spaces that map the similarity of outcomes for processing steps in functional connectivity neuroimaging studies and in predictive modelling algorithms for efficiently exploring the multiverse of options when measuring experimental outcomes. At the cohort level, I extend a normative modelling framework that uses state-of-the-art deep learning algorithms and only trains on data from healthy controls to identify individuals with schizophrenia as outliers. This approach can be extended to other cohorts or clinical conditions. Altogether, these frameworks aim to improve the methodology of current research practices to make them more robust and reproducible.

Bibliography

Abdal, R., Qin, Y., and Wonka, P. (2019). Image2StyleGAN: How to embed images into the StyleGAN latent space? In *Proceedings of the IEEE International Conference on Computer Vision*.

Aggarwal, C. C. and Yu, P. S. (2001). Outlier detection for high dimensional data. *ACM SIGMOD Record*, 30(2):37–46.

Alon Halevy, Peter Norvig, and Fernando Pereira (2009). The Unreasonable Effectiveness of Data. *Expert Opinion - IEEE Computer Society*, pages 8–12.

American Psychiatric Association (2013). DSM-5 Diagnostic Classification. In *Diagnostic and Statistical Manual of Mental Disorders*.

Anderson, D. J. and Perona, P. (2014). Toward a Science of Computational Ethology. *Neuron*, 84(1):18–31.

Arbabshirani, M. R., Plis, S., Sui, J., and Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, 145(Pt B):137–165.

Arbuthnott, D. R. J. (1710). II. An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. By Dr. John Arbuthnott, Physitian in Ordinary to Her Majesty, and Fellow of the College of Physitians and the Royal Society. *Philosophical Transactions of the Royal Society of London*, 27(328):186–190.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein Generative Adversarial Networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR.

Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41.

Ayesha, S., Hanif, M. K., and Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, 59:44–58.

Baker, S. and Kandasamy, Y. (2022). Machine learning for understanding and predicting neurodevelopmental outcomes in premature infants: a systematic review. *Pediatric Research*.

Baptista, R. and Poloczek, M. (2018). Bayesian Optimization of Combinatorial Structures. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 462–471. PMLR.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White, C. N., De Boeck, P., and Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proc Natl Acad Sci U S A*, 115(11):2607–2612.

Barto, A. and Sutton, R. S. (1992). *Reinforcement Learning: An Introduction*. MIT Press.

Baur, C., Denner, S., Wiestler, B., Albarqouni, S., and Navab, N. (2020). Autoencoders for Unsupervised Anomaly Segmentation in Brain MR Images: A Comparative Study. 14(8):1–16.

Baur, C., Wiestler, B., Albarqouni, S., and Navab, N. (2018). Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. In *International MICCAI brainlesion workshop*, pages 161–169. Springer.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396.

Bell, G., Gray, J., and Szalay, A. (2005). Petascale Computational Systems: Balanced CyberInfrastructure in a Data-Centric World Computational Science and Data Exploration. *arXiv*.

Bell, S. J., Kampman, O. P., Dodge, J., and Lawrence, N. D. (2022). Modeling the Machine Learning Multiverse.

Bellman, R. (1957). *Dynamic Programming.* Princeton University Press.

Benedek, G., Horváth, G., Kéri, S., Braunitzer, G., and Janáky, M. (2017). The development and aging of the magnocellular and parvocellular visual pathways as indicated by vep recordings between 5 and 84 years of age. *Vision (Switzerland).*

Beyret, B., Hernández-Orallo, J., Cheke, L., Halina, M., Shanahan, M., and Crosby, M. (2019). The Animal-AI Environment: Training and Testing Animal-Like Artificial Cognition.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.*

Bojanowski, P., Joulin, A., Paz, D. L., and Szlam, A. (2018). Optimizing the latent space of generative networks. In *35th International Conference on Machine Learning, ICML 2018.*

Bortolon, C., Capdevielle, D., Altman, R., Macgregor, A., Attal, J., and Raffard, S. (2017). Mirror self-face perception in individuals with schizophrenia: Feelings of strangeness associated with one's own image. *Psychiatry Research*, 253:205–210.

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., and Kurth-Nelson, Z. (2020). Deep Reinforcement Learning and Its Neuroscientific Implications. *Neuron*, 107(4):603–616.

Box, J. F. (1978). *R. A. Fisher, the Life of a Scientist.*

Brochu, E., Cora, V. M., and de Freitas, N. (2010). A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. *arXiv.*

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.*, 10(3):186–198.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience 2013 14:5*, 14(5):365–376.

Bzdok, D. and Yeo, B. T. T. (2017). Inference in the age of big data: Future perspectives on neuroscience. *NeuroImage*, 155:549–564.

Camchong, J., Dyckman, K. A., Chapman, C. E., Yanasak, N. E., and McDowell, J. E. (2006). Basal ganglia-thalamocortical circuitry disruptions in schizophrenia during delayed response tasks. *Biological psychiatry*, 60(3):235–241.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

Carp, J. (2012). On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. *Frontiers in neuroscience*, 6:149.

Caruana, R., Niculescu-Mizil, A., Crew, G., and Ksikes, A. (2004). Ensemble selection from libraries of models. In *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*.

Chaichana, K. L. and Quiñones-Hinojosa, A. (2019). Preface. In Chaichana, K. and Quiñones-Hinojosa, A., editors, *Comprehensive Overview of Modern Surgical Approaches to Intrinsic Brain Tumors*, page xv. Academic Press.

Chen, X., You, S., Tezcan, K. C., and Konukoglu, E. (2020). Unsupervised lesion detection via image restoration with a normative prior. *Medical image analysis*, 64:101713.

Cohn, D., Ghahramani, Z., and Jordan, M. (1994). Active Learning with Statistical Models. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press.

Cole, J. H. and Franke, K. (2017). Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers. *Trends Neurosci.*, 40(12):681–690.

Cole, J. H., Marioni, R. E., Harris, S. E., and Deary, I. J. (2019). Brain age and other bodily ‘ages’: implications for neuropsychiatry. *Molecular Psychiatry*, 24(2):266–281.

Cole, J. H., Ritchie, S. J., Bastin, M. E., Hernández, M. C. V., Maniega, S. M., Royle, N., Corley, J., Pattie, A., and Harris, S. E. (2017). Brain age predicts mortality. *Nat. Publ. Gr.*, 23(5):1385–1392.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science (New York, N.Y.)*, 349(6251):aac4716.

Conte, S., Richards, J. E., Guy, M. W., Xie, W., and Roberts, J. E. (2020). Face-sensitive brain responses in the first year of life. *NeuroImage.*

Courchesne, E., Ganz, L., and Norcia, A. M. (1981). Event-related brain potentials to human faces in infants. *Child development.*

Cox, D. D. and John, S. (1992). A statistical method for global optimization. In *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics.*

Cox, T. and Cox, M. (2000). *Multidimensional Scaling, Second Edition.*

Creswell, A. and Bharath, A. A. (2019). Inverting the Generator of a Generative Adversarial Network. *IEEE Transactions on Neural Networks and Learning Systems.*

Cusack, R., Veldsman, M., Naci, L., Mitchell, D. J., and Linke, A. C. (2012). Seeing different objects in different ways: measuring ventral visual tuning to sensory and semantic features with dynamically adaptive imaging. *Human brain mapping*, 33(2):387–397.

da Costa, P. F., Dafflon, J., Mendes, S. L., Sato, J. R., Cardoso, M. J., Leech, R., Jones, E., and Pinaya, W. H. L. (2022). Transformer-based normative modelling for anomaly detection of early schizophrenia. In *Empowering Communities: A Participatory Approach to AI for Mental Health.*

da Costa, P. F., Haartsen, R., Throm, E., Mason, L., Gui, A., Leech, R., and Jones, E. J. H. (2021). Neuroadaptive electroencephalography: a proof-of-principle study in infants.

da Costa, P. F., Lorenz, R., Monti, R. P., Jones, E., and Leech, R. (2020). Bayesian Optimization for real-time, automatic design of face stimuli in human-centred research.

Dafflon, J., Costa, P. F. D., Váša, F., Monti, R. P., Bzdok, D., Hellyer, P. J., Turkheimer, F., Smallwood, J., Jones, E., and Leech, R. (2020). Neuroimaging: into the Multiverse. *bioRxiv.*

Dafflon, J., F. Da Costa, P., Váša, F., Monti, R. P., Bzdok, D., Hellyer, P. J., Turkheimer, F., Smallwood, J., Jones, E., and Leech, R. (2022). A guided multiverse study of neuroimaging analyses. *Nature Communications*, 13(1):3758.

De Carli, D., Garreffa, G., Colonnese, C., Giulietti, G., Labruna, L., Briselli, E., Ken, S., Macrì, M. A., and Maraviglia, B. (2007). Identification of activated regions during a language task. *Magnetic Resonance Imaging*, 25(6):933–938.

de Haan, M. (2007). *Infant EEG and Event-Related Potentials*. Psychology Press.

de Haan, M. and Nelson, C. A. (1999). Brain activity differentiates face and object processing in 6-month-old infants. *Developmental psychology*.

De Ridder, J., Lavanga, M., Verhelle, B., Vervisch, J., Lemmens, K., Kotulska, K., Moavero, R., Curatolo, P., Weschke, B., Riney, K., Feucht, M., Krsek, P., Nabbout, R., Jansen, A. C., Wojdan, K., Domanska-Pakieła, D., Kaczorowska-Frontczak, M., Hertzberg, C., Ferrier, C. H., Samueli, S., Benova, B., Aronica, E., Kwiatkowski, D. J., Jansen, F. E., Jóźwiak, S., Van Huffel, S., and Lagae, L. (2020). Prediction of Neurodevelopment in Infants With Tuberous Sclerosis Complex Using Early EEG Characteristics. *Frontiers in Neurology*, 11.

Deen, B., Richardson, H., Dilks, D. D., Takahashi, A., Keil, B., Wald, L. L., Kanwisher, N., and Saxe, R. (2017). Organization of high-level visual cortex in human infants. *Nature Communications*.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., and Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3):968–980.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., and Courville, A. (2016). Adversarially learned inference. *arXiv preprint arXiv:1606.00704*.

Dunn, O. J. (1961). Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56(293):52–64.

Efron, B. and Tibshirani, R. (1991). Statistical Data Analysis in the Computer Age. *Science*, 253(5018):390–395.

Eimer, M. (2012). The Face-Sensitive N170 Component of the Event-Related Brain Potential. In *Oxford Handbook of Face Perception.*

Ellison-Wright, I., Glahn, D. C., Laird, A. R., Thelen, S. M., and Bullmore, E. (2008). The Anatomy of First-Episode and Chronic Schizophrenia: An Anatomical Likelihood Estimation Meta-Analysis. *American Journal of Psychiatry*, 165(8):1015–1023.

Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., and Roberts, A. (2019). GANSynth: Adversarial Neural Audio Synthesis.

Esser, P., Rombach, R., and Ommer, B. (2020). Taming Transformers for High-Resolution Image Synthesis.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., Dupre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., and Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods*, 16(January).

Fagerholm, E. D., Hellyer, P. J., Scott, G., Leech, R., and Sharp, D. J. (2015). Disconnection of network hubs and cognitive impairment after traumatic brain injury. *Brain*, pages 1696–1709.

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., and Hutter, F. (2020). Auto-Sklearn 2.0: The Next Generation.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., and Hutter, F. (2015). Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems.*

Fisher, R. A. (1915). Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika.*

Fisher, R. A. (1992). *Statistical Methods for Research Workers*, pages 66–70. Springer New York, New York, NY.

Fornito, A., Zalesky, A., and Breakspear, M. (2013). Graph analysis of the human connectome: Promise, progress, and pitfalls. *Neuroimage*, 80:426–444.

Foster, E. D. and Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association : JMLA*, 105(2):203.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., and Yurovsky, D. (2017). A Collaborative Approach to Infant Research: Promoting Reproducibility, Best Practices, and Theory-Building. *Infancy : the official journal of the International Society on Infant Studies*, 22(4):421–435.

Fusi, N., Sheth, R., and Elibol, M. (2018). Probabilistic matrix factorization for automated machine learning. In *Advances in Neural Information Processing Systems*.

Geerligs, L., Renken, R. J., Saliasi, E., Maurits, N. M., and Lorist, M. M. (2015). A brain-wide study of age-related changes in functional connectivity. *Cerebral cortex*, 25(7):1987–1999.

Gelbart, M. A., Snoek, J., and Adams, R. P. (2014). Bayesian optimization with unknown constraints. In *Uncertainty in Artificial Intelligence - Proceedings of the 30th Conference, UAI 2014*.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., and Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, pages 1–11.

Golarai, G., Grill-Spector, K., and Reiss, A. L. (2006). Autism and the development of face processing. *Clinical Neuroscience Research*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Networks. *Corrosion*, page iii.

Graham, M. S., Tudosiu, P.-D., Wright, P., Pinaya, W. H. L., Jean-Marie, U., Mah, Y., Teo, J., Jäger, R. H., Werring, D., Nachev, P., and Others (2022). Transformer-based out-of-distribution detection for clinically safe segmentation. *arXiv preprint arXiv:2205.10650*.

Gramacy, R. B., Lee, H. K., Holmes, C., and Osborne, M. (2012). Optimization Under Unknown Constraints. In *Bayesian Statistics 9*.

Grünewälder, S., Audibert, J. Y., Opper, M., and Shawe-Taylor, J. (2010). Regret bounds for Gaussian process bandit problems. In *Journal of Machine Learning Research*.

Gui, A., Bussu, G., Tye, C., Elsabbagh, M., Pasco, G., Charman, T., Johnson, M. H., and Jones, E. J. (2021). Attentive brain states in infants with and without later autism. *Translational Psychiatry*.

Gui, A., Throm, E., da Costa, P., Haartsen, R., Leech, R., and Jones, E. (2022). Proving and improving the reliability of infant research with neuroadaptive bayesian optimization. *INFANT AND CHILD DEVELOPMENT*. Publisher Copyright: © 2022 John Wiley Sons Ltd.

Guyon, I., Saffari, A., Dror, G., and Cawley, G. (2010). Model selection: Beyond the bayesian/frequentist divide. *J. Mach. Learn. Res.*, 11:61–87.

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, 95(2):245–258.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). The Extent and Consequences of P-Hacking in Science. *PLOS Biology*, 13(3):e1002106.

Huertas, I., Oldehinkel, M., van Oort, E. S. B., Garcia-Solis, D., Mir, P., Beckmann, C. F., and Marquand, A. F. (2017). A Bayesian spatial model for neuroimaging data based on biologically informed basis functions. *NeuroImage*, 161:134–148.

Ioannidis, J. P. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, 2(8):e124.

Ioannidis, J. P. (2014). How to Make More Published Research True. *PLoS Medicine*.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23(5):524–532.

Johnson, M. H. (2011). Interactive Specialization: A domain-general framework for human functional brain development?

Jones, E. J., Mason, L., Begum Ali, J., van den Boomen, C., Braukmann, R., Cauvet, E., Demurie, E., Hessels, R. S., Ward, E. K., Hunnius, S., Bolte, S., Tomalski, P., Kemner, C., Warreyn, P., Roeyers, H., Buitelaar, J., Falck-Ytter, T., Charman, T., and Johnson, M. H. (2019). Eurosibs: Towards robust measurement of infant neurocognitive predictors of autism across Europe. *Infant Behavior and Development*.

Jones, J. P. and Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, 58(6):1187–1211.

Jones, W. and Klin, A. (2013). Attention to eyes is present but in decline in 2-6-month-old infants later diagnosed with autism. *Nature*, 504(7480):427–431.

Kandasamy, K., Dasarathy, G., Oliva, J., Schneider, J., and Póczos, B. (2016). Gaussian process bandit optimisation with multi-fidelity evaluations. *Adv. Neural Inf. Process. Syst.*, (Nips):1000–1008.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*.

Kanwisher, N. and Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive Growing of GANs for Improved Quality, Stability, and Variation.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Karras, T., Laine, S., and Aila, T. (2020). A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kaufmann, T., Meer, D. V. D., Doan, N. T., Schwarz, E., Lund, M. J., Agartz, I., Alnæs, D., Barch, D. M., Baur-streubel, R., Tsolaki, M., Ulrichsen, K. M., Vellas, B., Wang, L., Westman, E., and Westlye, L. T. (2019). Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nature Neuroscience.*

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):1–9.

Kerr, N. L. (1998). HARKing: hypothesizing after the results are known. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc*, 2(3):196–217.

Konyushkova, K., Sznitman, R., and Fua, P. (2017). Learning Active Learning from Data. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Kriegeskorte, N., Formisano, E., Sorger, B., and Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America.*

Kropotov, J. D. and Kropotov, J. D. (2016). Chapter 1.6 – Event-Related Potentials. *Functional Neuromarkers for Psychiatry.*

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129.

Lacoste, A., Larochelle, H., Marchand, M., and Laviolette, F. (2014). Agnostic Bayesian learning of ensembles. In *31st International Conference on Machine Learning, ICML 2014.*

Lai, J. W., Ang, C. K. E., Rajendra Acharya, U., and Cheong, K. H. (2021). Schizophrenia: A Survey of Artificial Intelligence Techniques Applied to Detection and Classification. *International Journal of Environmental Research and Public Health 2021, Vol. 18, Page 6099*, 18(11):6099.

Latal, B. (2009). Prediction of Neurodevelopmental Outcome After Preterm Birth. *Pediatric Neurology*, 40(6):413–419.

Leonard, C. M., Kuldau, J. M., Breier, J. I., Zuffante, P. A., Gautier, E. R., Heron, D. C., Lavery, E. M., Packing, J., Williams, S. A., and DeBose, C. A. (1999). Cumulative effect of anatomical risk factors for schizophrenia: an MRI study. *Biological psychiatry*, 46(3):374–382.

Leopold, D. A. and Rhodes, G. (2010). A Comparative view of face perception. *Journal of Comparative Psychology.*

Li, J., Kong, R., Orban, C., Tan, Y., Sun, N., Holmes, A. J., Sabuncu, M. R., Ge, T., and Yeo, B. T. T. (2019). Global signal regression strengthens association between resting-state functional connectivity and behavior. *Neuroimage*, 196(April):126–141.

Lindsay, G. (2020). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, pages 1–15.

Livezey, J. A., Bouchard, K. E., and Chang, E. F. (2019). Deep learning as a tool for neural data analysis: Speech classification and cross-frequency coupling in human sensorimotor cortex. *PLOS Computational Biology*, 15(9):e1007091.

Lorenz, R., Hampshire, A., and Leech, R. (2017). Neuroadaptive Bayesian Optimization and Hypothesis Testing. *Trends in Cognitive Sciences.*

Lorenz, R., Johal, M., Dick, F., Hampshire, A., Leech, R., and Geranmayeh, F. (2021). A Bayesian optimisation approach for rapidly mapping residual network function in stroke. *Brain.*

Lorenz, R., Monti, R. P., Violante, I. R., Anagnostopoulos, C., Faisal, A. A., Montana, G., and Leech, R. (2016). The Automatic Neuroscientist: A framework for optimizing experimental design with closed-loop real-time fMRI. *NeuroImage.*

Lorenz, R., Monti, R. P., Violante, I. R., Faisal, A. A., Anagnostopoulos, C., Leech, R., and Montana, G. (2015). Stopping criteria for boosting automatic experimental design using real-time fMRI with Bayesian optimization.

Lorenz, R., Violante, I. R., Monti, R. P., Montana, G., Hampshire, A., and Leech, R. (2018). Dissociating frontoparietal brain networks with neuroadaptive Bayesian optimization. *Nature Communications.*

Lundervold, A. S. and Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, 29(2):102–127.

Luyster, R. J., Powell, C., Tager-Flusberg, H., and Nelson, C. A. (2014). Neural measures of social attention across the first years of life: Characterizing typical development and markers of autism risk. *Developmental Cognitive Neuroscience.*

Lv, J., Di Biase, M., Cash, R. F. H., Cocchi, L., Cropley, V. L., Klauser, P., Tian, Y., Bayer, J., Schmaal, L., Cetin-Karayumak, S., Rathi, Y., Pasternak, O., Bousman, C., Pantelis, C., Calamante, F., and Zalesky, A. (2021). Individual deviations from normative models of brain structure in a large cross-sectional schizophrenia cohort. *Molecular Psychiatry*, 26(7):3512–3523.

M. De Haan and C. A. Nelson (1997). Recognition of the mother's face by six-month-old infants: A neurobehavioural study. *Child Development*, 68(2):187–210.

Maestro, S., Muratori, F., Cavallaro, M. C., Pei, F., Stern, D., Golse, B., and Palacio-Espasa, F. (2002). Attentional Skills During the First 6 Months of Age in Autism Spectrum Disorder. *Journal of the American Academy of Child Adolescent Psychiatry*, 41(10):1239–1245.

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., Moore, L. A., Conan, G. M., Uriarte, J., Snider, K., Lynch, B. J., Wilgenbusch, J. C., Pengo, T., Tam, A., Chen, J., Newbold, D. J., Zheng, A., Seider, N. A., Van, A. N., Metoki, A., Chauvin, R. J., Laumann, T. O., Greene, D. J., Petersen, S. E., Garavan, H., Thompson, W. K., Nichols, T. E., Yeo, B. T. T., Barch, D. M., Luna, B., Fair, D. A., and Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902):654–660.

Markram, H. (2006). The Blue Brain Project. *Nature Reviews Neuroscience.*

Marquand, A. F., Kia, S. M., Zabihi, M., Wolfers, T., Buitelaar, J. K., and Beckmann, C. F. (2019). Conceptualizing mental disorders as deviations from normative functioning. *Molecular Psychiatry*, 24(10):1415–1424.

Marquand, A. F., Wolfers, T., Mennes, M., Buitelaar, J., and Beckmann, C. F. (2016). Beyond Lumping and Splitting: A Review of Computational Approaches for Stratifying Psychiatric Disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 1(5):433–447.

McCartney, M., Haeringer, M., and Polifke, W. (2020). Comparison of Machine Learning Algorithms in the Interpolation and Extrapolation of Flame Describing Functions. *Journal of Engineering for Gas Turbines and Power*, 142(6).

McInnes, L., Healy, J., Saul, N., and Grossberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software*, 3(29):861.

Menkveld, A. J., Dreber, A., Holzmeister, F., Huber, J., Johanneson, M., Kirchler, M., Razen, M., Weitzel, U., Abad, D., Abudy, M. M., Adrian, T., Ait-Sahalia, Y., Akmansoy, O., Alcock, J., Alexeev, V., Aloosh, A., Amato, L., Amaya, D., Angel, J. J., Bach, A., Baidoo, E., Bakalli, G., Barbon, A., Bashchenko, O., Bindra, P. C., Bjonnes, G. H., Black, J. R., Black, B. S., Bohorquez, S., Bondarenko, O., Bos, C. S., Bosch-Rosa, C., Bouri, E., Brownlees, C. T., Calamia, A., Cao, V. N., Capelle-Blancard, G., Capera, L., Caporin, M., Carrion, A., Caskurlu, T., Chakrabarty, B., Chernov, M., Cheung, W. M. Y., Chincarini, L. B., Chordia, T., Chow, S. C., Clapham, B., Colliard, J.-E., Comerton-Forde, C., Curran, E., Dao, T., Dare, W., Davies, R. J., De Blasis, R., De Nard, G., Declerck, F., Deev, O., Degryse, H., Deku, S., Desagre, C., Van Dijk, M. A., Dim, C., Dimpfl, T., Dong, Y. J., Drummond, P., Dudda, T., Dumitrescu, A., Dyakov, T., Dyhrberg, A. H., Dzieliński, M., Eksi, A., El Kalak, I., ter Ellen, S., Eugster, N., Evans, M. D., Farrell, M., Félez-Viñas, E., Ferrara, G., FERROUHI, E. M., Flori, A., Fluharty-Jaidee, J., Foley, S., Fong, K. Y. L., Foucault, T., Franus, T., Franzoni, F. A., Frijns, B., Frömmel, M., Fu, S., Füllbrunn, S., Gan, B., Gehrig, T., Gerritsen, D., Gil-Bazo, J., Glosten, L. R., Gomez, T., Gorbenko, A., Güçbilmez, U., Grammig, J., Gregoire, V., Hagströmer, B., Hambuckers, J., Hapnes, E., Harris, J. H., Harris, L., Hartmann, S., Hasse, J.-B., Hautsch, N., He, X.-Z. T., Heath, D., Hediger, S., Hendershott, T. J., Hibbert, A. M., Hjalmarsson, E., Hoelscher, S., Hoffmann, P., Holden, C. W., Horenstein, A. R., Huang, W., Huang, D., Hurlin, C., Ivashchenko, A., Iyer, S. R., Jahanshahloo, H., Jalkh, N., Jones, C. M., Jurkatis, S., Jylha, P., Kaeck, A., Kaiser, G., Karam, A., Karmaziene, E., Kassner, B., Kaustia, M., Kazak, E., Kearney, F., van Kervel, V., Khan, S., Khomyn, M., Klein, T., Klein, O., Klos, A., Koetter, M., Krahnen, J. P., Kolokolov, A., Korajczyk, R. A., Kozhan, R., Kwan, A., Lajaunie, Q., Lam, F. Y. E. C., Lambert, M., Langlois, H., Lausen, J., Lauter, T., Leippold, M., Levin, V., Li, Y., Li, M. H., Liew, C. Y., Lindner, T., Linton, O. B., Liu, J., Liu, A., Llorente-Alvarez, J.-G., Lof, M., Lohr, A., Longstaff, F. A., Lopez-Lira, A., Mankad, S., Mano, N., Marchal, A., Martineau, C., Mazzola, F., Meloso, D. C., Mihet, R., Mohan, V., Moinas, S., Moore, D., Mu, L., Muravyev, D., Murphy, D., Neszveda, G., Neumeier, C., Nielsson, U., Nimalendran, M., Nolte, S., Nordén, L. L., O'Neill, P., Obaid, K., Ødegaard, B. A., Östberg, P., Painter, M., Palan, S., Palit, I., Park, A., Pascual Gascó, R., Pasquariello, P., Pastor, L., Patel, V., Patton, A. J., Pearson, N. D., Pelizzon, L., Pelster, M., Pérignon, C., Pfiffer, C., Philip, R., Plíhal, T., Prakash, P., Press, O.-A., Prodromou, T., Putnins, T. J., Raizada, G., Rakowski, D. A., Ranaldo, A., Regis, L., Reitz, S., Renault, T., Wang, R., Renò, R., Riddiough, S., Rinne, K., Rintamäki, P., Riordan, R., RITTMANNSBERGER, T., Rodríguez Longarela, I., Rösch, D., Rognone, L., Roseman, B., Rosu, I., Roy, S., Rudolf, N., Rush, S., Rzayev, K., Rzeźnik, A., Sanford, A., Sankaran, H., Sarkar, A., Sarno, L., Scaillet, O., Scharnowski, S., Schenk-Hoppé, K. R., Schertler, A., Schneider, M., Schroeder, F., Schürhoff, N., Schuster, P., Schwarz, M. A., Seasholes, M. S., Seeger, N., Shachar, O., Shkilko, A., Shui, J., Sikic, M., Simion, G., Smales, L. A., Söderlind, P., Sojli, E., Sokolov, K., Spokeviciute, L., Stefanova, D., Subrahmanyam, M. G., Neusüss, S., Szaszi, B., Talavera, O., Tang, Y., Taylor, N., Tham, W. W., Theissen, E., Thimme, J., Tonks, I., Tran, H., Trapin, L., Trolle, A. B., Vaduva, M., Valente, G., Van Ness, R. A., Vasquez, A., Verousis, T., Verwijmeren, P., Vilhelmsson, A., Vilkov, G., Vladimirov, V., Vogel, S., Voigt, S., Wagner, W., Walther, T., Weiss, P., van der Wel, M., Werner, I. M., Westerholm, P. J., Westheide, C., Wipplinger, E., Wolf, M., Wolff, C. C. P., Wolk, L., Wong, W. K., Wrampelmeyer, J., Wu, Z.-X., Xia, S., Xiu, D., Xu, K., Xu, C., Yadav, P. K., Yagüe, J., Yan, C., Yang, A., Yoo, W., Yu, W., Yu, S., Yueshen, B. Z., Yuferova, D., Zamojski, M., Zareei, A., Zeisberger, S., Zhang, S., Zhang, X., Zhong, Z., Zhou, Z. I., Zhou, C., Zhu, X., Zoican, M., Zwinkels, R. C., Chen, J., Duevski, T., Gao, G., Gemayel, R., Gilder, D., Kuhle, P., Pagnotta, E., Pelli, M., Souml;nksen, J., Zhang, L., Ilczuk, K., Bogoev, D., Qian, Y., Wika, H. C., Yu, Y., Zhao, L., Mi, M., and Bao, L. (2021). Non-Standard Errors. *SSRN Electronic Journal.*

Minka, T. P. (2000). Automatic Choice of Dimensionality for PCA. In *NIPS.*

Misir, M. and Sebag, M. (2013). Algorithm Selection as a Collaborative Filtering Problem. *Res. Rep.*, (December):40.

Molina, D., Pérez-Beteta, J., Martínez-González, A., Martino, J., Velasquez, C., Arana, E., and Pérez-García, V. M. (2017). Lack of robustness of textural measures obtained from 3D brain tumor MRIs impose a need for standardization. *PloS one*, 12(6):e0178843.

Monti, R. P., Gibberd, A., Roy, S., Nunes, M., Lorenz, R., Leech, R., Ogawa, T., Kawanabe, M., and Hyvärinen, A. (2020). Interpretable brain age prediction using linear latent variable models of functional connectivity. *Plos one*, 15(6):e0232296.

Moulson, M. C., Balas, B., Nelson, C., and Sinha, P. (2011). EEG correlates of categorical and graded face perception. *Neuropsychologia*.

Munafó, M. R. (2009). Reliability and replicability of genetic association studies. *Addiction*, 104(9):1439–1440.

Munsters, N. M., van Ravenswaaij, H., van den Boomen, C., and Kemner, C. (2019). Test-retest reliability of infant event related potentials evoked by faces. *Neuropsychologia*.

Murphy, K. and Fox, M. D. (2017). Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *Neuroimage*, 154(November 2016):169–173.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T. A., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Paluck, E. L., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E. J., Wilson, R., and Yarkoni, T. (2015). Promoting an open research culture. *Science*.

Olson, R. S., Bartley, N., Urbanowicz, R. J., and Moore, J. H. (2016). Evaluation of a tree-based pipeline optimization tool for automating data science. In *GECCO 2016 - Proceedings of the 2016 Genetic and Evolutionary Computation Conference*.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science: Open Science Collobaration. *Science*.

Pang, R., Lansdell, B. J., and Fairhall, A. L. (2016). Dimensionality reduction in neuroscience. *Current Biology*, 26(14):R656–R660.

Pascalis, O., De Martin de Viviés, X., Anzures, G., Quinn, P. C., Slater, A. M., Tanaka, J. W., and Lee, K. (2011). Development of face processing. *Wiley Interdisciplinary Reviews: Cognitive Science*.

Pashler, H. and Wagenmakers, E. J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, 7(6):528–530.

Peng, H., Gong, W., Beckmann, C. F., Vedaldi, A., and Smith, S. M. (2021). Accurate brain age prediction with lightweight deep neural networks. *Medical Image Analysis*, 68:101871.

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, 45(1):S199–S209.

Phillipou, A., Abel, L. A., Castle, D. J., Hughes, M. E., Gurvich, C., Nibbs, R. G., and Rossell, S. L. (2015). Self perception and facial emotion perception of others in anorexia nervosa. *Frontiers in Psychology*, 6.

Phillips, D. J., Mcglaughlin, A., Ruth, D., Jager, L. R., and Soldan, A. (2015). NeuroImage : Clinical Graph theoretic analysis of structural connectivity across the spectrum of Alzheimer 3 s disease : The importance of graph creation methods. *NeuroImage Clin.*, 7:377–390.

Pinaya, W. H., Tudosiu, P.-D., Gray, R., Rees, G., Nachev, P., Ourselin, S., and Cardoso, M. J. (2022). Unsupervised brain imaging 3D anomaly detection and segmentation with transformers. *Medical Image Analysis*, 79:102475.

Pinaya, W. H. L., Gadelha, A., Doyle, O. M., Noto, C., Zugman, A., Cordeiro, Q., Jackowski, A. P., Bressan, R. A., and Sato, J. R. (2016). Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Scientific Reports*, 6(1):38897.

Pinaya, W. H. L., Mechelli, A., and Sato, J. R. (2019). Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study. *Human brain mapping*, 40(3):944–954.

Pinaya, W. H. L., Tudosiu, P.-D., Gray, R., Rees, G., Nachev, P., Ourselin, S., and Cardoso, M. J. (2021). Unsupervised brain anomaly detection and segmentation with transformers. *arXiv preprint arXiv:2102.11650.*

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., and Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2):115–126.

Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery 2011 10:9*, 10(9):712–712.

Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., and Moreno-Noguer, F. (2018). GANimation: Anatomically-aware facial animation from a single image. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).*

Radford, A., Metz, L., and Chintala, S. (2016). DCGAN. *4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.*

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., and Others (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents.

Rasmussen, C. E. (2004). Gaussian Processes in machine learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).*

Rasmussen, C. E. and Williams, C. K. I. (2018). *Gaussian Processes for Machine Learning.*

Richards, J. E., Reynolds, G. D., and Courage, M. L. (2010). The neural bases of infant attention. *Current Directions in Psychological Science.*

Rimol, L. M., Hartberg, C. B., Nesvåg, R., Fennema-Notestine, C., Hagler, D. J. J., Pung, C. J., Jennings, R. G., Haukvik, U. K., Lange, E., Nakstad, P. H., Melle, I., Andreassen, O. A., Dale, A. M., and Agartz, I. (2010). Cortical thickness and subcortical volumes in schizophrenia and bipolar disorder. *Biological psychiatry*, 68(1):41–50.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

Rubinov, M. (2016). Constraints and spandrels of interareal connectomes. *Nat. Commun.*, 7:1–11.

Rubinov, M. and Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *Neuroimage*, 52(3):1059–1069.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science 2021 2:3*, 2(3):1–21.

Schulz, M.-A., Yeo, B. T. T., Vogelstein, J. T., Mourao-Miranada, J., Kather, J. N., Kording, K., Richards, B., and Bzdok, D. (2020). Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nat Commun*, 11(1):4238.

Settles, B. (2009). *Active learning literature survey.* University of Wisconsin-Madison Department of Computer Sciences.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and Freitas, N. D. (2016). Taking the Human Out of the Loop : A Review of Bayesian Optimization. *Proc. IEEE*, 104(1):148–175.

Shenton, M. E., Dickey, C. C., Frumin, M., and McCarley, R. W. (2001). A review of MRI findings in schizophrenia. *Schizophrenia Research*, 49(1-2):1–52.

Shepherd, A. M., Laurens, K. R., Matheson, S. L., Carr, V. J., and Green, M. J. (2012). Systematic meta-review and quality assessment of the structural brain alterations in schizophrenia. *Neuroscience Biobehavioral Reviews*, 36(4):1342–1356.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.*

Singer, J. D. and Willett, J. B. (2009). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.*

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*.

Snyder, K. A., Blank, M. P., and Marsolek, C. J. (2008). What form of memory underlies novelty preferences? *Psychonomic Bulletin and Review*.

Somerville, L. H., Bookheimer, S. Y., Buckner, R. L., Burgess, G. C., Curtiss, S. W., Dapretto, M., Elam, J. S., Gaffrey, M. S., Harms, M. P., Hodge, C., and Others (2018). The Lifespan Human Connectome Project in Development: A large-scale study of brain connectivity development in 5–21 year olds. *Neuroimage*, 183:456–468.

Squarcina, L., Castellani, U., Bellani, M., Perlini, C., Lasalvia, A., Dusi, N., Bonetto, C., Cristofalo, D., Tosato, S., Rambaldelli, G., Alessandrini, F., Zoccatelli, G., Pozzi-Mucelli, R., Lamonaca, D., Ceccato, E., Pileggi, F., Mazzi, F., Santonastaso, P., Ruggeri, M., and Brambilla, P. (2017). Classification of first-episode psychosis in a large cohort of patients using support vector machine and multiple kernel learning techniques. *NeuroImage*, 145:238–245.

Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspect. Psychol. Sci.*, 11(5):702–712.

Strauss, E. and Kaplan, E. (1980). Lateralized Asymmetries in Self-Perception. *Cortex*, 16(2):289–293.

Sussillo, D., Stavisky, S. D., Kao, J. C., Ryu, S. I., and Shenoy, K. V. (2016). Making brain–machine interfaces robust to future neural variability. *Nature Communications*, 7:13749.

Thornton, C., Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2013). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Tim Salimans, Andrej Karpathy, Xi Chen, Diederik P. Kingma, Y. B. (2017). Pixelcnn++: A Pixelcnn Implementation With Discretized Logistic Mixture Likelihood and Other Modifications. *ICLR*.

Tomczak, J. M. (2022). *Deep Generative Modeling*. Springer.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., and Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. Supporting Online Material. *Science (New York, N.Y.).*

Tsao, D. Y. and Livingstone, M. S. (2008). Mechanisms of Face Perception. *Annual Review of Neuroscience.*

Tsuang, M. T., Lyons, M. J., and Faraone, S. V. (1990). Heterogeneity of Schizophrenia: Conceptual Models and Analytic Strategies. *British Journal of Psychiatry*, 156(1):17–26.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320.

Tye, C., Bussu, G., Gliga, T., Elsabbagh, M., Pasco, G., Johnsen, K., Charman, T., Jones, E. J. H., Buitelaar, J., and Johnson, M. H. (2022). Understanding the nature of face processing in early autism: A prospective study. *Journal of psychopathology and clinical science*, 131(6):542–555.

van den Heuvel, M. P., de Lange, S. C., Zalesky, A., Seguin, C., Yeo, B. T. T., and Schmidt, R. (2017). Proportional thresholding in resting-state fMRI functional connectivity networks and consequences for patient-control connectome studies: Issues and recommendations. *Neuroimage*, 152:437–449.

van den Heuvel, M. P. and Hulshoff Pol, H. E. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity. *Eur. Neuropsychopharmacol.*, 20(8):519–534.

Van Den Oord, A., Vinyals, O., and Kavukcuoglu, K. (2017). Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):6307–6316.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., Ugurbil, K., Consortium, W.-M. H. C. P., and Others (2013). The WU-Minn human connectome project: an overview. *Neuroimage*, 80:62–79.

Vanschoren, J., van Rijn, J. N., Bischl, B., and Torgo, L. (2014). OpenML: networked science in machine learning. 15(2):49–60.

Váša, F., Bullmore, E. T., and Patel, A. X. (2018). Probabilistic thresholding of functional connectomes: Application to schizophrenia. *Neuroimage*, 172(May 2018):326–340.

Váša, F., Romero-Garcia, R., Kitzbichler, M. G., Seidlitz, J., Whitaker, K. J., Vaghi, M. M., Kundu, P., Patel, A. X., Fonagy, P., Dolan, R. J., Jones, P. B., Goodyer, I. M., the NSPN Consortium, Vértes, P. E., and Bullmore, E. T. (2020). Conservative and disruptive modes of adolescent change in human brain functional connectivity. *Proc. Natl. Acad. Sci. U. S. A.*, 117 (6):3248–3253.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All you Need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vieira, S., Gong, Q. Y., Pinaya, W. H., Scarpazza, C., Tognin, S., Crespo-Facorro, B., Tordesillas-Gutierrez, D., Ortiz-García, V., Setien-Suero, E., Scheepers, F. E., van Haren, N. E., Marques, T. R., Murray, R. M., David, A., Dazzan, P., McGuire, P., and Mechelli, A. (2020). Using Machine Learning and Structural Neuroimaging to Detect First Episode Psychosis: Reconsidering the Evidence. *Schizophrenia Bulletin*, 46(1):17–26.

Vieira, S., Gong, Q.-y., Pinaya, W. H. L., Scarpazza, C., Tognin, S., Crespo-Facorro, B., Tordesillas-Gutierrez, D., Ortiz-García, V., Setien-Suero, E., Scheepers, F. E., Van Haren, N. E. M., Marques, T. R., Murray, R. M., David, A., Dazzan, P., McGuire, P., and Mechelli, A. (2019). Using Machine Learning and Structural Neuroimaging to Detect First Episode Psychosis: Reconsidering the Evidence. *Schizophrenia Bulletin*, 46(1):17–26.

Vieira, S., Pinaya, W. H., and Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience Biobehavioral Reviews*, 74:58–75.

Visser, I., Bergmann, C., Byers-Heinlein, K., Dal Ben, R., Duch, W., Forbes, S., Franchin, L., Frank, M. C., Geraci, A., Hamlin, J. K., and et al. (2022). Improving the generalizability of infant psychological research: The manybabies model. *Behavioral and Brain Sciences*, 45:e35.

Vu, M. A. T., Adalı, T., Ba, D., Buzsáki, G., Carlson, D., Heller, K., Liston, C., Rudin, C., Sohal, V. S., Widge, A. S., Mayberg, H. S., Sapiro, G., and Dzirasa, K. (2018). A Shared Vision for Machine Learning in Neuroscience. *The Journal of Neuroscience*, 38(7):1601.

Wagenmakers, E.-J., Sarafoglou, A., and Aczel, B. (2022). One statistical analysis must not rule them all. *Nature*, pages 605, 423–425.

Wakefield, J. C. (2013). DSM-5: An Overview of Changes and Controversies. *Clinical Social Work Journal 2013 41:2*, 41(2):139–154.

Walton, E., Hibar, D. P., van Erp, T. G. M., Potkin, S. G., Roiz-Santiañez, R., Crespo-Facorro, B., Suarez-Pinilla, P., Van Haren, N. E. M., de Zwarte, S. M. C., Kahn, R. S., Cahn, W., Doan, N. T., Jørgensen, K. N., Gurholt, T. P., Agartz, I., Andreassen, O. A., Westlye, L. T., Melle, I., Berg, A. O., Mørch-Johnsen, L., Faerden, A., Flyckt, L., Fatouros-Bergman, H., Jönsson, E. G., Hashimoto, R., Yamamori, H., Fukunaga, M., Preda, A., De Rossi, P., Piras, F., Banaj, N., Ciullo, V., Spalletta, G., Gur, R. E., Gur, R. C., Wolf, D. H., Satterthwaite, T. D., Beard, L. M., Sommer, I. E., Koops, S., Gruber, O., Richter, A., Krämer, B., Kelly, S., Donohoe, G., McDonald, C., Cannon, D. M., Corvin, A., Gill, M., Di Giorgio, A., Bertolino, A., Lawrie, S., Nickson, T., Whalley, H. C., Neilson, E., Calhoun, V. D., Thompson, P. M., Turner, J. A., and Ehrlich, S. (2017). Positive symptoms associate with cortical thinning in the superior temporal gyrus via the ENIGMA Schizophrenia consortium. *Acta psychiatrica Scandinavica*, 135(5):439–447.

Wang, Y., Wu, C., Herranz, L., van de Weijer, J., Gonzalez-Garcia, A., and Raducanu, B. (2018). Transferring GANs: Generating Images from Limited Data. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, pages 220–236, Cham. Springer International Publishing.

Wang, Z., Zoghiy, M., Hutterz, F., Matheson, D., and De Freitas, N. (2013). Bayesian optimization in high dimensions via random embeddings. In *IJCAI International Joint Conference on Artificial Intelligence*.

Webb, S. J., Jones, E. J., Merkle, K., Venema, K., Greenson, J., Murias, M., and Dawson, G. (2011). Developmental Change in the ERP Responses to Familiar Faces in Toddlers With Autism Spectrum Disorders Versus Typical Development. *Child Development*.

Welch, S., Klassen, C., Borisova, O., and Clothier, H. (2013). The DSM-5 controversies: How should psychologists respond? *Canadian Psychology*, 54(3):166–175.

Westfall, J., Nichols, T. E., and Yarkoni, T. (2016). Fixing the stimulus-as-fixed-effect fallacy in task fMRI. *Wellcome open research*, 1:23.

Westfall, J., Nichols, T. E., and Yarkoni, T. (2017). Fixing the stimulus-as-fixed-effect fallacy in task fMRI. *Wellcome Open Research*.

Wolfers, T., Doan, N. T., Kaufmann, T., Alnæs, D., Moberget, T., Agartz, I., Buitelaar, J. K., Ueland, T., Melle, I., Franke, B., Andreassen, O. A., Beckmann, C. F., Westlye, L. T., and Marquand, A. F. (2018). Mapping the Heterogeneous Phenotype of Schizophrenia and Bipolar Disorder Using Normative Models. *JAMA psychiatry*, 75(11):1146–1155.

Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*.

Woodman, G. F. (2010). A Brief Introduction to the Use of Event-Related Potentials (ERPs) in Studies of Perception and Attention. *Attention, perception psychophysics*, 72(8):2031–2046.

Yarkoni, T. (2019). The generalizability crisis. *PsyArXiv*.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, 45.

Yarkoni, T. and Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, 12(6):1100–1122.

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., and Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*.

Zhang, C. and Ma, Y. (2012). *Ensemble Machine Learning - Methods and Applications*. Springer.

Zhou, S.-Y., Suzuki, M., Hagino, H., Takahashi, T., Kawasaki, Y., Matsui, M., Seto, H., and Kurachi, M. (2005). Volumetric analysis of sulci/gyri-defined in vivo frontal lobe regions in schizophrenia: Precentral gyrus, cingulate gyrus, and prefrontal region. *Psychiatry Research: Neuroimaging*, 139(2):127–139.

Zimek, A., Schubert, E., and Kriegel, H. P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):363–387.