Exploring and mitigating analytical variability in fMRI results using representation learning

Élodie Germani

COLLEGE MATHS, TELECOMS
DOCTORAL INFORMATIQUE, SIGNAL
BRETAGNE SYSTEMES, ELECTRONIQUE

Université de Rennes

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES

ÉCOLE DOCTORALE N° 601
Mathématiques, Télécommunications, Informatique, Signal, Systèmes, Électronique
Spécialité : *Informatique*

Par
Elodie GERMANI

Exploring and mitigating analytical variability in fMRI results using representation learning

Thèse présentée et soutenue à Rennes, le 16 Septembre 2024
Unité de recherche : Empenn

Rapporteurs avant soutenance :

Ninon BURGOS Chargée de recherche, CNRS
Karim LEKADIR Professeur, Universitat de Barcelona

Composition du Jury :

Attention, en cas d'absence d'un des membres du Jury le jour de la soutenance, la composition du jury doit être revue pour s'assurer qu'elle est conforme et devra être répercutée sur la couverture de thèse

Président :	Mathieu ACHER	Professeur	Université de Rennes
Examinateurs :	Ninon BURGOS	Chargée de recherche	CNRS
	Carole LARTIZIEN	Directrice de recherche	CNRS
	Karim LEKADIR	Professeur	Universitat de Barcelona
Dir. de thèse :	Camille MAUMET	Chargée de recherche	Inria
Co-dir. de thèse :	Elisa FROMONT	Professeure	Université de Rennes

Acknowledgement

Tout d'abord, j'aimerais remercier les rapporteur.euse.s Dr. Ninon Burgos et Prof. Karim Lekadir d'avoir accepté de relire ma thèse et de faire partie du jury de ma thèse. Merci également aux deux autres membres du jury : Dr. Carole Lartizien et Prof. Mathieu Acher d'avoir fait partie de mon CSI pendant ces trois années.

Je souhaite également remercier mes deux directrices de thèse, Dr. Camille Maumet et Prof. Elisa Fromont pour leur soutien pendant ces trois années. Merci d'avoir su me rassurer quand j'en avais besoin, me conseiller quand c'était nécessaire et surtout m'écouter et me faire confiance. Merci de m'avoir poussée et d'avoir cru en moi. Merci pour votre bonne humeur à chaque réunion et pour tous vos conseils. Bref, MERCI POUR CES 3 ANNÉES !

Je voudrais remercier Prof. Tristan Glatard et Prof. Jean-Baptiste Poline pour m'avoir accueilli dans leurs laboratoires à Montréal pendant 4 mois, et de m'avoir fait découvrir tant de choses, que ce soit professionnellement ou humainement. Une pensée à tous les collègues avec qui j'ai collaboré au cours de ce séjour, et à ceux avec qui je n'ai pas collaboré, mais avec qui j'ai passé de superbes moments. Un merci tout particulier à Rémi Gau, merci d'avoir été un super compatriote accro au travail, pour ces week-ends de hardworking au Leaves House Cafe.

Merci à tous mes collègues de chez Empenn et LACODAM, pour les afterworks, les séminaires d'équipe, les pauses cafés et votre bonne humeur de manière générale. Je souhaite aussi remercier tous les collègues plus lointains que j'ai pu rencontrer en conférence ou en école d'été, et qui sont parfois devenu des amis. Grâce à vous, les conférences avaient une tout autre saveur et j'ai déjà hâte des prochaines pour vous retrouver !

Enfin, merci à ma famille, mes amis, mon conjoint et surtout à mon chat Doudou d'avoir supporté les aléas d'avoir une patronne thésarde, et qui a été forcé de prendre l'avion avec moi pour aller vivre 4 mois au Canada.

Table of Contents

Préambule en français **xix**

Preamble **2**

List of publications **12**

Open Science **16**

I Context 19

1 Functional Magnetic Resonance Imaging (fMRI) 20

1.1 From brain activity to fMRI raw data . . . 20

1.1.1 Brain imaging and fMRI . . . 20

1.1.2 Principle of BOLD fMRI . . . 21

1.1.3 Experimental design and protocols . . . 22

1.2 From fMRI raw data to fMRI results . . . 22

1.2.1 Preprocessing . . . 23

1.2.2 First-level analysis . . . 30

1.2.3 Second-level analysis . . . 37

1.2.4 Statistical inference . . . 38

2 Analytical variability 42

2.1 Different sources of variability . . . 42

2.1.1 Intra-individual variability . . . 43

2.1.2 Inter-individual variability . . . 43

2.1.3 Technical variability . . . 44

2.1.4 Analytical variability . . . 44

2.2 Focus: Analytical variability . . . 44

2.2.1 A large analytical space . . . 45

2.2.2 Effect of analytical variability at different levels . . . 47

2.2.3 Challenges related to analytical variability 50

II How to facilitate data re-use with deep representation learning? 56

3 Deep learning for medical imaging 57

3.1 Foundations of deep learning 58
3.1.1 From artificial intelligence to deep learning 58
3.1.2 Different learning processes 60
3.1.3 Neural Networks 61
3.2 Deep learning in medical imaging 65
3.3 Challenges related to medical imaging 67
3.3.1 Challenges related to data 68
3.3.2 Challenges related to models 70
3.4 Solutions using deep learning 71
3.4.1 Transfer learning 72
3.4.2 Image-to-image transition and style transfer 76

4 Leveraging variability in fMRI results with self-taught learning 86

4.1 Introduction 86
4.2 Materials and Methods 88
4.2.1 Overview of the datasets 90
4.2.2 Preprocessing 92
4.2.3 Model architectures 93
4.2.4 Convolutional AutoEncoder (CAE) training 93
4.2.5 Convolutional Neural Network (CNN) training 94
4.2.6 Benefits of self-taught learning and impact of different factors . . . 95
4.2.7 Explainability 99
4.3 Results 99
4.3.1 Convolutional AutoEncoder (CAE) performance 99
4.3.2 Hyperparameters optimisation for Convolutional Neural Network (CNN) 100
4.3.3 Benefits of self-taught learning on a homogeneous dataset 102
4.3.4 Benefits of self-taught learning on a heterogeneous dataset 105

4.3.5 How do we explain these benefits? . . . 106
4.4 Discussion . . . 112
4.4.1 Summary . . . 112
4.4.2 Limitations . . . 113

5 Mitigating analytical variability in fMRI results with style transfer 117
5.1 Introduction . . . 117
5.2 Materials and Methods . . . 119
5.2.1 Dataset . . . 119
5.2.2 Generative Adversarial Network (GAN) frameworks . . . 120
5.2.3 Denoising Diffusion Probabilistic Model (DDPM) frameworks . . . 120
5.2.4 Evaluation of performance . . . 123
5.3 Results . . . 124
5.3.1 Generative Adversarial Network (GAN) frameworks . . . 124
5.3.2 Denoising Diffusion Probabilistic Model (DDPM) frameworks . . . 125
5.3.3 Impact of multi-target images . . . 129
5.4 Discussion . . . 129

III How to explore the fMRI analytical space? 132

6 The HCP multi-pipeline dataset 133
6.1 Introduction . . . 133
6.2 Methods . . . 135
6.2.1 Raw Data: the Human Connectome Project . . . 135
6.2.2 Analyses pipelines . . . 136
6.3 Data Records . . . 139
6.4 Technical Validation . . . 139
6.5 Discussion . . . 144

7 Uncovering communities of pipelines 146
7.1 Introduction . . . 146
7.2 Materials and Methods . . . 148
7.2.1 Dataset . . . 148
7.2.2 Data processing . . . 148
7.2.3 Graph computation and community detection . . . 149

7.2.4 Communities statistic maps . . . 150
7.3 Results . . . 151
7.3.1 Communities for the contrast *right-hand* . . . 151
7.3.2 Communities for the contrast *right-foot* . . . 153
7.4 Discussion . . . 156

8 Mega-analyses using data processed with different pipelines 159
8.1 Introduction . . . 159
8.2 Materials and Methods . . . 161
8.2.1 Dataset . . . 162
8.2.2 Between-group analyses . . . 162
8.2.3 False Positive Rates Estimation . . . 164
8.2.4 Statistical distributions and P-P plots . . . 164
8.3 Results . . . 165
8.3.1 Analyses using the same pipeline (baseline) . . . 165
8.3.2 Analyses using pipelines with different parameters . . . 165
8.3.3 Analyses using pipelines with different software packages . . . 176
8.4 Discussions . . . 177

Conclusion and perspectives 182

Appendix - Contributions on reproducibility 190

A State-of-the-art on reproducibility 190

B Reproduction and replication of a study: Predicting Parkinson's disease trajectory using clinical and functional MRI features 203

C Reproduction of analysis pipelines: the NARPS Open Pipelines project 244

Appendix - Supplementary materials 256

D Supplementary materials for Chapter 4 256

E Supplementary materials for Chapter 5 260

F Supplementary materials for Chapter 7 **262**

G Supplementary Materials for Chapter 8 **267**

Bibliography **269**

List of abbreviations

AE AutoEncoder

ALFF Amplitude at Low Frequency Fluctuation

BIDS Brain Imaging Data Structure

BOLD Bold Oxygen Level Dependent

CAE Convolutional AutoEncoder

cGAN conditional Generative Adversarial Network

CNN Convolutional Neural Network

CSF CerebroSpinal Fluid

CT Computed Tomography

DDPM Denoising Diffusion Probabilistic Model

DUA Data Usage Agreement

EEG Electroencephalography

EPI Echo Planar Imaging

FDR False Discovery Rate

fALFF Fractional Amplitude at Low Frequency Fluctuation

fMRI Functional Magnetic Resonance Imaging

FWE Family-Wise Error

FWHM Full-width at Half-Maximum

GAN Generative Adversarial Network

GDPR General Data Protection Regulation

GLM General Linear Model

HCP Human Connectome Project

HRF Haemodynamic Response Function

I2I Image-to-image transition

ICA Independant Component Analysis

IRM Imagerie par Résonance Magnétique

IRMf Imagerie par Résonance Magnétique fonctionnelle

MDS-UPDRS Movement Disorder Society-Unified Parkinson's Disease Rating Scale

MEG Magnetoencephalography

MNI Montreal Neurological Institute

MRI Magnetic Resonance Imaging

NARPS Neuroimaging Analysis Replication and Prediction Study

OS Operating System

PD Parkinson's Disease

PET Positon Emission Tomography

PPMI Parkinson's Progression Markers Initiative

R2 R-squared - coefficient of determination

ReHo Regional Homogeneity

RMSE Root Mean Squared Error

ROI Region of Interest

rs-fMRI resting-state functional Magnetic Resonance Imaging

VAE Variational AutoEncoder

WM White Matter

List of Figures

1 Cartes statistiques de niveau groupe pour le paradigme *main droite* xx
2 Concept d'apprentissage de représentation dans les réseaux de neurones convolutifs xxiv
3 Group-level statistic maps for the paradigm *right-hand* 3
4 Concept of representation learning in Convolutional Neural Network (CNN) 6

1.1 Illustration of common fMRI protocols 22
1.2 Example of a standard fMRI pipeline 24
1.3 Characteristics of the Haemodynamic Response Function (HRF) 32
1.4 Convolutions of the Haemodynamic Response Function (HRF) 33
1.5 Modelling variance at the second-level of statistical analysis in fMRI 39

2.1 Different sources of variability in neuroimaging studies 43
2.2 Overview of analytical flexibility in fMRI pipelines 45
2.3 Fraction of teams reporting a significant result during Botvinik-Nezer et al., 2020 49

3.1 Venn diagram of artificial intelligence and deep learning 59
3.2 Main learning processes in deep learning 60
3.3 Comparison of Convolutional Neural Network (CNN) and AutoEncoder (AE) 63
3.4 Summary of deep learning papers in medical imaging, from Litjens et al., 2017 65
3.5 Main applications of deep learning in medical imaging 66
3.6 Evolution of accuracy in Alzheimer's disease diagnosis with deep learning, from Varoquaux et al., 2022 69
3.7 Different types of transfer learning based on the context 73
3.8 Paired and unpaired training datasets 78
3.9 Schematic representation of the learning process of Pix2Pix (Isola et al., 2017) 80

3.10 Schematic representation of the learning process of CycleGAN (Zhu et al., 2017) . . . 81
3.11 Schematic representation of the learning process of StarGAN (Choi et al., 2018) . . . 82
3.12 Schematic representation of the learning process of conditional Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2021) . . . 83

4.1 Flow diagram of the self-taught learning methodology . . . 89
4.2 Overview of the process used to split the datasets for cross-validation and performance evaluation . . . 97
4.3 Examples of reconstruction using Convolutional AutoEncoder (CAE) . . . 101
4.4 Performance of models on contrast classification with the HCP dataset . . . 103
4.5 Performance of models on task classification with the HCP dataset . . . 104
4.6 Performance of models on one-contrast task classification with the HCP dataset . . . 104
4.7 Performance of models on classification of mental concepts with the BrainPedia dataset . . . 106
4.8 Features learned by the different Convolutional Neural Network (CNN) . . . 107
4.9 Similarities of features learned by the different Convolutional Neural Network (CNN) . . . 109

5.1 Flow diagram of the Denoising Diffusion Probabilistic Model (DDPM)-based frameworks . . . 121
5.2 Source, target and generated images for Generative Adversarial Network (GAN)-based frameworks . . . 124
5.3 Source, target and generated images for Denoising Diffusion Probabilistic Model (DDPM)-based frameworks . . . 126
5.4 Features learned by the pipeline classifier . . . 128

6.1 Example of subject-level and group-level statistic maps of the HCP-multi-pipeline dataset . . . 140
6.2 Workflow of technical validation of statistic maps in the HCP multi-pipelines dataset . . . 141
6.3 Distribution of mean Percentage of Activation inside the Primary Motor Cortex . . . 142

6.4 Thresholded statistic maps for contrasts *right foot* and *right hand* for group-level analysis of group 3 with pipeline spm-5-0-0 . . . 143
6.5 Distribution of Percentage of Activation inside the Primary Motor Cortex for pipeline spm-5-0-0 . . . 143

7.1 Workflow of community detection in the pipeline space across different groups of participants and contrasts . . . 149
7.2 Adjacency matrix representing the number of times each pair of pipelines belong to the same community across different group-level statistic maps of the contrast *right-hand* . . . 151
7.3 Mean statistic map for the contrast *right-hand* across groups (of participants) for a representative pipeline in each community . . . 152
7.4 Adjacency matrix representing the number of times each pair of pipelines belong to the same community across different group-level statistic maps of the contrast *right-foot*. . . . 154
7.5 Mean correlations (across groups) between statistic maps for each pair of pipelines with the contrast *right-foot* . . . 155
7.6 Mean statistic map for the contrast *right-foot* across groups (of participants) for a representative pipeline in each community . . . 155

8.1 Overview of the methodology . . . 161
8.2 False positive rates for pipelines with different parameters within SPM and FSL . . . 166
8.3 Bland-Altman P-P plots for pipelines with different parameters and with the same parameters within SPM . . . 167
8.4 Bland-Altman P-P plots for pipelines with two different parameters and with the same parameters within SPM . . . 168
8.5 Bland-Altman P-P plots for pipelines with different parameters and with the same parameters within FSL . . . 169
8.6 Bland-Altman P-P plots for pipelines with two different parameters and with the same parameters within FSL . . . 170
8.7 Distribution of statistical values for multiple between-group analyses under SPM, compared to the expected distribution . . . 171
8.8 Distribution of statistical values for multiple between-group analyses under FSL, compared to the expected distribution . . . 172

8.9 False positive rates for pipelines with different software packages 177
8.10 Bland-Altman P-P plots for pipelines with different software packages . . . 178
8.11 Distribution of statistical values for between-software analyses, compared to the expected distribution. 178

A.1 Reproducible research defined by The Turing way 192

B.1 Summary of the different workflows implemented to reproduce the results of Nguyen et al., 2021 and explore their robustness to different analytic conditions 209
B.2 Workflow of model selection and performance evaluation 220
B.3 Distribution of MDS-UPDRS scores reported in the original paper's cohort, the replication cohort and the closest-to-original cohort 227
B.4 Performance of models trained for prediction at each time point, using Fractional Amplitude at Low Frequency Fluctuation (fALFF) or Regional Homogeneity (ReHo), with variations in the workflow 233
B.5 Predictive features learned by the best performing models to predict MDS-UPDRS score at each time point for the original study (extracted from Nguyen et al., 2021) and the *default workflow* using Regional Homogeneity (ReHo) . . 236
B.6 Predictive features learned by the best performing models to predict MDS-UPDRS score at each time point for the original study (extracted from Nguyen et al., 2021) and the *default workflow* using Fractional Amplitude at Low Frequency Fluctuation (fALFF) 237

C.1 Workflow used to study the impact of sample size on vibration of effects . . 250
C.2 Thresholded and unthresholded statistic maps obtained with the different pipelines from the NARPS Open Pipelines project and different sample sizes 252
C.3 Maximum statistical value inside the ROI of the ventromedial prefrontal cortex depending on sample size for the 8 reproduced pipelines from the NARPS Open Pipelines Project 253
C.4 Ratio largest/smallest maximum statistical value inside ROI depending on sample size 253

D.1 Schematic visualisation of the architectures of the Convolutional AutoEncoder (CAE) and Convolutional Neural Network (CNN) with 4 layers . . . 257

D.2 Schematic visualisation of the architectures of the Convolutional AutoEncoder (CAE) and Convolutional Neural Network (CNN) with 5 layers . . . 258

E.1 Histogram of the number of shared participants between two groups for each pair of groups across the whole dataset 261

F.1 Adjacency matrix representing the number of times each pair pipelines belong to the same community across different group-level statistic maps of the contrast *left-hand* 262
F.2 Adjacency matrix representing the number of times each pair pipelines belong to the same community across different group-level statistic maps of the contrast *left-foot* 263
F.3 Adjacency matrix representing the number of times each pair pipelines belong to the same community across different group-level statistic maps of the contrast *tongue* 264
F.4 Mean statistic map for the contrast *right-hand* across groups (of participants) for a representative pipeline of each community 265
F.5 Mean statistic map for the contrast *right-foot* across groups (of participants) for a representative pipeline of each community 265
F.6 Mean thresholded statistic map for the contrast *right-hand* across groups (of participants) for a representative pipeline of each community 266
F.7 Mean thresholded statistic map for the contrast *right-foot* across groups (of participants) for a representative pipeline of each community 266

G.1 Bland-Altman P-P plots for pipelines with two different parameters and with the same parameters within SPM 267

List of Tables

4.1 Overview of the datasets . . . 90
4.2 Reconstruction performance of the Convolutional AutoEncoder (CAE) . . 100
4.3 Hyperparameters chosen for each dataset and corresponding performance of the classifier on the validation set of the dataset. . . . 100
4.4 Classification performance of the models with the HCP dataset . . . 102
4.5 Classification performance of the models on BrainPedia datasets . . . 105
4.6 Per-class accuracies for classification of contrasts with the HCP dataset . . 108
4.7 Classification performance with different numbers of transferred layers . . . 110
4.8 Classification performance with different numbers of frozen layers . . . 111

5.1 Description of Generative Adversarial Network (GAN)-based frameworks . 120
5.2 Description of Denoising Diffusion Probabilistic Model (DDPM)-based frameworks . . . 121
5.3 Performance associated with four transfers for Generative Adversarial Network (GAN)-based frameworks . . . 124
5.4 Performance associated with four transfers for Denoising Diffusion Probabilistic Model (DDPM)-based frameworks . . . 126
5.5 Mean correlations between features maps learned at each layers for each pair of pipelines . . . 127
5.6 Performance associated with four transfers with Denoising Diffusion Probabilistic Model (DDPM)-based frameworks with different implementation . 129

7.1 Mean number of activated voxels in the thresholded mean statistic maps of the representative pipeline of each community and inside the ROI . . . 153

8.1 False positive rates for between-groups analyses with the same pipeline in both groups, with SPM and FSL and for all possible sets of parameters . . 165

A.1 Principles of statistical testing. . . . 198

B.1 Summary of cohort selection procedure . . . 224

B.2 Demographic and clinical variables for the different cohorts 225
B.3 Predictive performance achieved for each MDS-UPDRS time point and each imaging feature type, computed through leave-one-out cross-validation 230
B.4 Performance reported using different model selection and evaluation methods 235

C.1 Criteria for validating the reproduction of pipelines in the NARPS Open Pipelines project 247

E.1 Performance associated with four transfers on out of distribution data (different paradigm) 261

G.1 False positive rates for between-groups analyses using contrast maps without post-processing with the same pipeline in both groups, with SPM and FSL and for all possible sets of parameters 268

PRÉAMBULE EN FRANÇAIS

Introduction

Imagerie par Résonance Magnétique Fonctionnelle

Au cours des dernières décennies, le développement de l'imagerie cérébrale a considérablement enrichi notre compréhension du cerveau et de ses pathologies, ouvrant la voie à de nouvelles approches diagnostiques et thérapeutiques. En particulier, les techniques d'imagerie cérébrale telles que l'Imagerie par Résonance Magnétique (IRM) ont permis de mieux comprendre la structure du cerveau avec une grande précision, tout en restant non invasives. Aujourd'hui, la question de la compréhension des fonctions cérébrales occupe une place importante dans de nombreux domaines de recherche, allant de la médecine et de la psychologie à l'intelligence artificielle et à la philosophie. Démêler la complexité du cerveau et déchiffrer la manière dont les différentes régions cérébrales interagissent est un défi scientifique qui captive les chercheurs. L'Imagerie par Résonance Magnétique fonctionnelle (IRMf) est une technique d'imagerie cérébrale qui permet aux chercheurs d'étudier l'activité cérébrale des individus pendant qu'ils effectuent des tâches prédéfinies. Le nombre d'études publiées utilisant cette modalité a explosé au cours des dix dernières années : en 2018, plus d'un millier d'études enregistrées sur le site web `clinicaltrial.gov` utilisaient l'IRMf comme mesure de résultat (Sadraee et al., 2021).

Dans les études traditionnelles d'IRMf, un ensemble de participants est choisi sur la base de différents critères, en fonction de l'objectif de l'expérience (participants sains, stade de la maladie, etc.). Les études sont conçues pour répondre à une ou plusieurs hypothèses de recherche, par exemple sur l'activation d'une zone cérébrale au cours d'une tâche ou sur la présence de différences entre la force d'activation chez deux groupes d'individus. Les participants sont soumis à une acquisition d'IRMf, qui consiste en une séquence de tâches constituant le paradigme d'activation. Ce paradigme est composé d'une tâche de référence, généralement le repos, et d'une tâche d'intérêt dont la seule différence avec la référence correspond au processus cognitif à explorer. Les chercheurs récupèrent les

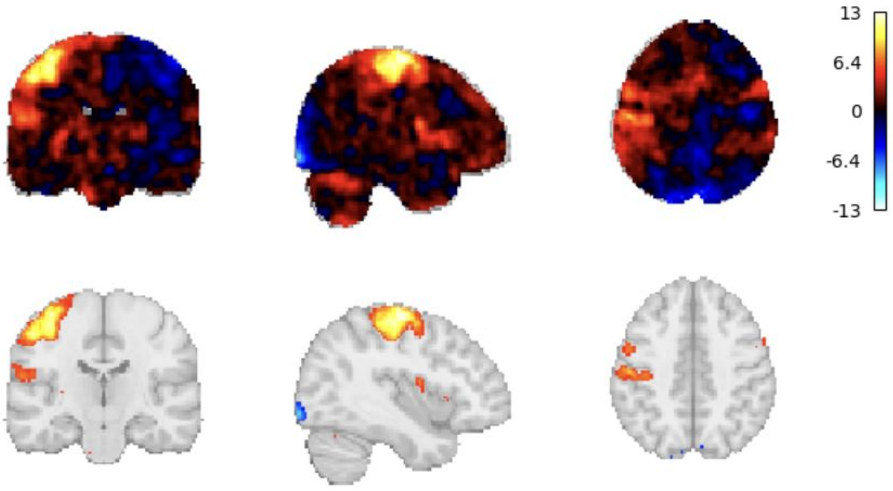

FIGURE 1 – Cartes statistiques de niveau groupe pour le paradigme *main droite*, sans seuillage (en haut) et avec seuillage (en bas) en utilisant un seuil de $p < 0,05$, corrigé par Bonferroni.

données brutes de l'IRMf sous la forme de matrices à 4 dimensions pour tous les participants et appliquent une séquence d'étapes de prétraitement et d'analyse statistique, appelée "chaîne de traitement". À la fin de cette chaîne de traitement, les résultats sont présentés sous la forme de cartes statistiques tridimensionnelles montrant l'activation du cerveau pour chaque participant et chaque question de recherche, et pour l'ensemble des participants si un deuxième niveau d'analyse a été effectué. Ces cartes sont généralement seuillées par inférence statistique afin d'identifier les régions du cerveau qui présentent une activation significative. Ce processus s'appuie sur des tests statistiques pour déterminer si les différences observées ou les relations entre les variables sont probablement dues au hasard ou si elles représentent des effets réels.

Les études d'imagerie cérébrale, et en particulier l'IRMf, sont soumises à différentes sources de variabilité. Par source de variabilité, nous entendons tout facteur dont les variations entraînent des modifications des résultats finaux. Par exemple, selon l'heure de la journée ou la présence ou non de médication, le même participant peut avoir des activations différentes, ce qui représente une source de variabilité, appelée variabilité intra-individuelle (Chen et al., 2016). De même, si les données sont acquises pour le même participant, mais avec des paramètres d'acquisition différents, cela pourrait entraîner des variations dans les résultats (Wittens et al., 2021). Naturellement, il semble logique que deux participants aient des résultats différents, mais il est difficile pour les études de

représenter l'ensemble des variations inter-individuelles (Valizadeh et al., 2018). Les études d'imagerie cérébrale sont généralement de petite taille (∼30 participants en 2015 (Poldrack et al., 2017)) et composées de données acquises dans un seul centre, ce qui entraîne une faible généralisabilité des résultats.

Variabilité analytique

Un autre exemple de source de variabilité est liée aux variations des résultats finaux causées par les différentes implémentations de la chaîne de traitement utilisée pour traiter et analyser les données brutes. Dans une méta-analyse de plus de 200 articles sur l'IRMf, Carp, 2012b a montré que ces chaînes de traitements sont très flexibles, laissant aux chercheurs de nombreux choix à faire pour analyser leurs données. Cette déclaration a soulevé des questions sur les effets potentiels de ces différentes implémentations sur les résultats. Récemment, dans le cadre d'une étude multi-analystes (Botvinik-Nezer et al., 2020), 70 équipes de recherche ont été chargées d'analyser le même ensemble de données d'IRMf de tâche avec leur méthode habituelle. Dans l'ensemble, il n'y avait pas de chaînes de traitements identiques et les résultats finaux étaient relativement variables d'une équipe à l'autre. Ce phénomène, également connu sous le nom de "variabilité analytique", peut résulter de différents niveaux de variation dans le processus d'analyse :

— À chaque étape, lorsque l'on change l'algorithme à utiliser ou la valeur d'un paramètre.

— Lorsque l'on modifie le logiciel utilisé pour mettre en oeuvre la chaîne de traitement.

— À un niveau inférieur, lors de la modification de l'environnement de calcul.

Aujourd'hui, l'idée que différentes approches analytiques peuvent conduire à des résultats différents est acceptée par la communauté. Les chercheurs recherchent désormais des solutions pour relever les différents défis liés à la variabilité analytique (Botvinik-Nezer et al., 2023). La flexibilité des chaînes de traitements est particulièrement difficile à gérer pour les chercheurs, car il n'y a pas de vérité terrain qui puisse être utilisée pour comparer et mesurer les performances de chaînes de traitements concurrentes. Ainsi, il n'existe que peu d'accords sur les bonnes pratiques pour guider le choix d'une chaîne de traitement. Dans la pratique, les chercheurs explorent couramment de multiples alternatives analytiques valides, mais ne rapportent souvent que les résultats d'une seule d'entre elles (ou d'un petit nombre de variantes). Cette publication sélective peut entraîner une augmen-

tation des résultats faux positifs (Ioannidis, 2008a ; Simmons et al., 2011 ; Gelman et al., 2019). Comme solution potentielle, les chercheurs peuvent utiliser des analyses multiverselles (Steegen et al., 2016) pour comparer et rapporter les résultats de plusieurs approches analytiques. Cependant, une étude systématique de l'espace des chaînes de traitements n'est pas réaliste en raison du grand nombre de combinaisons possibles. Dans les deux cas, une meilleure connaissance de l'espace des chaînes de traitements serait utile pour identifier les principaux facteurs de cette variabilité des résultats et, par exemple, faciliter la sélection d'un ensemble représentatif de chaînes de traitements.

Une autre question liée à la variabilité en IRMf est la réutilisation des données. Avec l'émergence des pratiques de partage des données (Niso et al., 2022), il est possible d'augmenter la taille des échantillons des études d'imagerie cérébrale en réutilisant les données disponibles publiquement. L'utilisation d'échantillons plus importants et plus diversifiés contribuerait à améliorer la reproductibilité et la généralisabilité des résultats et offrirait une plus grande souplesse quant aux questions de recherche à étudier. Dans la pratique, les études avec réutilisation de données sont généralement effectuées avec des données brutes provenant de différentes études, qui sont ensuite réanalysées avec la même chaîne de traitement. Une autre solution consiste à utiliser des données dérivées (déjà traitées). Cette solution est plus optimale, non seulement parce que le partage des cartes statistiques n'est pas aussi difficile que le partage des données brutes en raison des contraintes réduites en matière de protection de la vie privée, mais aussi parce qu'elle évite d'avoir à effectuer de nouveaux calculs coûteux. Toutefois, il a été démontré que les données dérivées provenant de différentes sources devraient être combinées avec soin dans les études statistiques pour éviter d'augmenter le nombre de faux positifs (Rolland et al., 2022). En outre, les données partagées sur des bases de données publiques telles que NeuroVault manquent généralement d'annotations (Gorgolewski et al., 2015), ce qui rend leur réutilisation difficile. Il est donc nécessaire de trouver des solutions pour tirer parti de cette grande quantité de données dérivées partagées sur les bases de données publiques.

Apprentissage de représentations

L'apprentissage statistique, un sous-domaine de l'intelligence artificielle, se concentre sur le développement d'algorithmes capables d'apprendre à partir de données et de faire des prédictions ou de prendre des décisions sans qu'on leur indique explicitement la manière de procéder. En particulier, l'"apprentissage de représentation" désigne le processus

par lequel des caractéristiques significatives sont extraites des données pour créer des représentations plus faciles à comprendre et à traiter. Grâce à leur capacité à modéliser des relations complexes, les réseaux de neurones, utilisés dans le cadre de l'apprentissage profond (LeCun et al., 2015), ont montré des performances prometteuses pour cette tâche dans de nombreux domaines de recherche. Dans ce contexte, les représentations des données sont apprises de manière hiérarchique par les réseaux de neurones et contiennent des caractéristiques significatives pour la tâche sous-jacente à laquelle le réseau a été formé.

Dans le domaine de la vision par ordinateur, les chercheurs utilisent les réseaux de neurones convolutifs en raison de leur capacité à extraire des caractéristiques visuelles à l'aide d'opérations de convolution. Au fur et à mesure que les données d'entrée passent par des couches successives, ces réseaux apprennent des caractéristiques de plus en plus abstraites et complexes. Les couches inférieures capturent des caractéristiques de base telles que les bords et les textures, tandis que les couches supérieures représentent des caractéristiques plus significatives sur le plan sémantique et pertinentes pour la tâche à accomplir. Ces représentations sont ensuite utilisées pour produire des résultats pour cette tâche, mais elles peuvent également être manipulées et transférées entre différents tâches ou entre différentes données afin d'améliorer les performances des réseaux. L'"apprentissage par transfert", un cas d'utilisation de l'apprentissage de représentations, s'appuie sur des réseaux pré-entraînés dont les connaissances (paramètres des couches entraînées) sont transférées à un autre réseaux qui sera appliqué à des données provenant d'un autre domaine ou à une autre tâche. De même, les représentations peuvent être manipulées pour transférer des attributs entre les données. Ce cas d'utilisation est également connu sous le nom de "transfert de style" (Gatys et al., 2016) et utilise des modèles génératifs, dans lesquels les réseaux apprennent à modéliser la distribution des données, à partir de laquelle de nouvelles données peuvent être échantillonnées ou des données existantes peuvent être transférées.

Ces techniques sont prometteuses pour les problèmes décrits précédemment, car elles permettraient de construire une représentation complète des résultats d'IRMf et de leurs sources de variabilité. Toutefois, l'apprentissage d'une représentation utile et efficace nécessite une grande quantité de données d'entraînement pour représenter la diversité des données cibles potentielles (Ricci Lara et al., 2022). Ce problème est particulièrement important dans le domaine de l'imagerie cérébrale, où les études sont généralement réalisées

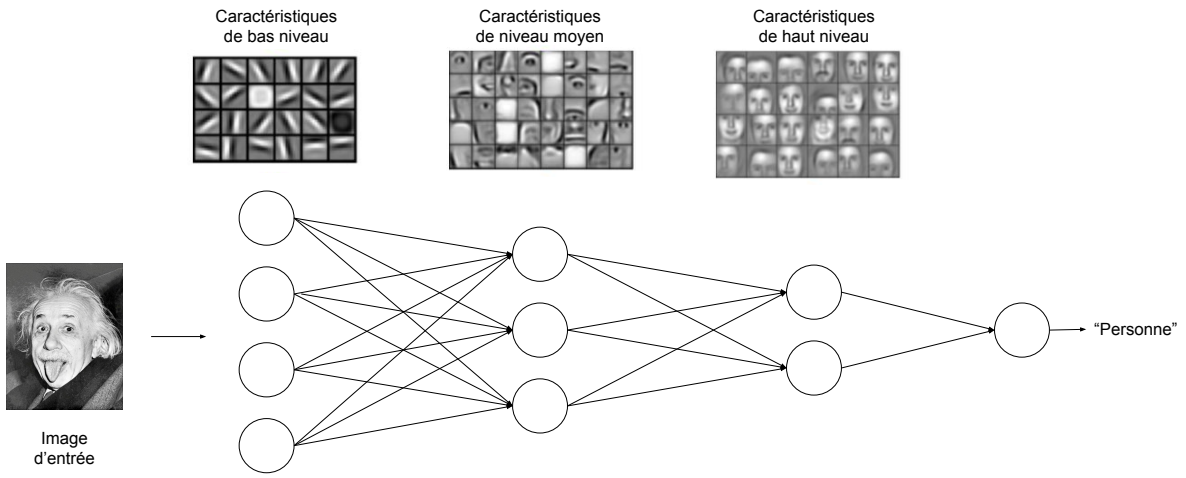

FIGURE 2 – Concept d'apprentissage de représentation dans la vision par ordinateur à l'aide de réseaux de neurones convolutifs. Les caractéristiques de niveau inférieur sont extraites dans les premières couches, tandis que les caractéristiques de niveau supérieur sont apprises par la suite.

sur des échantillons petits et homogènes. Comme indiqué ci-dessus, les plateformes de partage de données contiennent au contraire un grand nombre de données, provenant de différentes sources, et présentent donc un bon niveau de variabilité en termes de protocoles d'acquisition, de machines, de sites et de chaînes de traitements. L'utilisation de ces données nécessite le recours à des méthodologies particulières, car elles ne sont généralement pas étiquetées ou ne font pas l'objet d'un processus d'étiquetage standardisé (Poldrack et al., 2011b). En outre, les cartes statistiques d'IRMf ont des propriétés particulières qui nécessitent une adaptation des méthodes traditionnelles d'apprentissage de représentations. Elles contiennent des informations quantitatives (valeurs statistiques dans notre contexte), la localisation spatiale est une information cruciale (la même activation dans différentes régions du cerveau conduit à une interprétation complètement différente) et la dimensionnalité des images médicales est beaucoup plus grande (une carte statistique d'IRMf contient des dizaines de milliers de dimensions).

Contributions

La variabilité analytique des résultats d'IRMf pose des problèmes aux chercheurs qui conçoivent une nouvelle étude et à ceux qui tentent de réutiliser les données issues d'autres études. Dans ce contexte, la construction de représentations compréhensibles et significa-

tives de la diversité des données d'IRMf aiderait à obtenir une meilleure connaissance de l'espace analytique, mais aussi à fournir des solutions aux chercheurs qui souhaitent bénéficier des données dérivées partagées par la communauté sur des plateformes publiques. Toutefois, cela nécessite l'utilisation de grandes quantités de données et des méthodologies adaptées pour traiter les spécificités des données d'IRMf.

Dans la première série de contributions de cette thèse, nous proposons deux solutions concrètes pour apprendre et manipuler des représentations basse dimension des résultats d'IRMf afin de faciliter la réutilisation de la grande quantité de données dérivées partagées. Premièrement, nous tirons parti de la base de données NeuroVault (Gorgolewski et al., 2015) - une grande base de données publique de neuro-imagerie qui a été construite collaborativement - pour apprendre une représentation non supervisée des cartes statistiques d'IRMf, pouvant être transférée dans un cadre d'apprentissage "autodidacte" pour aider à résoudre de nouvelles tâches (par exemple, décodage du cerveau). Ce travail a donné lieu à la publication d'un article dans *Gigascience*, et les modèles pré-entraînés ont été partagés avec la communauté en vue d'une réutilisation ultérieure. Deuxièmement, nous avons supposé que les chaînes de traitements pouvaient être considérés comme un composant de style des cartes statistiques d'IRMf et nous avons proposé d'utiliser des méthodes de transfert de style pour convertir les cartes statistiques entre les chaînes de traitements. Dans cette contribution, nous avons développé une méthode basée sur des modèles de diffusion qui utilise une représentation latente des cartes statistiques dans laquelle les données sont structurées sur la base des caractéristiques les plus importantes qui les distinguent à travers les chaînes de traitement. Cette contribution a fait l'objet d'un article, disponible en préprint et bientôt soumis à *Human Brain Mapping*.

Dans une deuxième série de contributions, nous explorons les caractéristiques de l'espace des chaînes de traitements en IRMf. Pour ce faire, nous avons d'abord construit un ensemble de données multi-chaînes de traitements composé d'un grand nombre de participants et qui représente une partie non exhaustive mais contrôlée de l'espace analytique. Nous avons publié un article de présentation des données en tant que préprint (bientôt soumis à Scientific Data) et nous sommes en train de partager le jeu de données avec la communauté sur Public-nEUro. Dans cet ensemble de données, nous avons utilisé des algorithmes de détection de communautés (apprentissage de représentation pour identifier les groupes de données sur les graphes) pour explorer l'espace analytique et évaluer

la stabilité des relations entre les résultats de différentes chaînes de traitements à travers différents groupes de participants et de tâches. Ce travail a donné lieu à la rédaction d'un article, accepté à la conférence *ICIP 2024* et disponible sous forme de préprint. Enfin, nous explorons le potentiel de réutilisation des données et étudions la validité des analyses statistiques qui combinent des données traitées avec différentes chaînes de traitements (par exemple avec différents algorithmes, valeurs de paramètres ou logiciels). Un préprint est disponible pour cette contribution, en co-auteur avec Xavier Rolland, et a été soumis à *Imaging Neuroscience*.

Ce manuscrit est composé de trois parties. La première partie est consacrée à l'introduction de l'analyse des données d'IRMf et de la variabilité analytique.

(i) Dans le Chapitre 1, nous présentons le champ d'application de cette thèse, l'IRMf. Nous introduisons les grands principes des études d'IRMf, en décrivant le parcours de l'activité cérébrale aux données brutes d'IRMf. Ensuite, nous exposons le processus de transformation de ces données brutes en résultats finaux et décrivons les principales étapes de traitement incluses dans les chaînes d'analyse d'IRMf traditionnelles.

(ii) Dans le Chapitre 2, nous donnons un aperçu des différentes sources de variabilité qui peuvent être observées dans les études d'IRMf. Après une brève description de chaque source, nous nous concentrons sur la variabilité analytique et montrons que des changements peuvent être apportés à différents niveaux de la chaîne de traitement, entraînant des variations dans les résultats. Nous décrivons également les principaux défis liés à la variabilité analytique et les solutions développées pour les relever.

Dans la deuxième partie du manuscrit, nous exposons notre première série de contributions dans lesquelles nous avons utilisé l'apprentissage de représentations profondes pour atténuer la variabilité (analytique) des résultats d'IRMf et faciliter la réutilisation des données.

(i) Le Chapitre 3 présente le contexte de l'apprentissage de représentations profondes et son application au domaine de l'imagerie médicale. Nous détaillons les concepts

fondamentaux de l'apprentissage de représentations profondes, les défis liés à l'imagerie médicale et décrivons deux cas particuliers d'apprentissage de représentations, à savoir l'apprentissage par transfert et le transfert de style, dans lesquels les représentations sont transférées entre les tâches et/ou les données.

(ii) Dans le Chapitre 4, nous proposons une première solution pour faciliter la réutilisation des données et nous exploitons la base de données NeuroVault dans un cadre d'apprentissage autodidacte, un type spécifique d'apprentissage par transfert. Pour ce faire, nous apprenons une représentation agnostique des cartes statistiques d'IRMf à l'aide d'un autoencodeur convolutif, puis nous l'adaptons à diverses tâches.

(iii) Alors que les représentations ont été construites et transférées entre les tâches dans le chapitre précédent, nous proposons une autre solution pour réutiliser les données dérivées partagées et manipuler les représentations pour convertir les données entre les différentes chaînes de traitements dans le Chapitre 5. Dans ce contexte, nous proposons aussi une nouvelle méthode qui utilise un réseaux de neurones convolutif entraîné à distinguer les cartes statistiques entre les chaînes de traitements et à extraire les caractéristiques de haut niveau des données pour faciliter la transition entre les chaînes de traitements avec des modèles de diffusion.

L'utilisation concrète de ces solutions repose sur l'identification des principaux facteurs de variation pour trouver les cas critiques où l'atténuation de la variabilité analytique est nécessaire et appropriée. Dans la troisième partie, nous explorons l'espace analytique d'IRMf pour mieux comprendre les relations entre les résultats des chaînes de traitements et identifier certains défis particuliers liés à la réutilisation des données et à la variabilité analytique.

(i) Une étude complète de l'espace des chaînes de traitements n'est pas pratique car il est particulièrement grand, nous proposons en revanche d'explorer une plus petite partie de cet espace dans différents contextes : un grand nombre de participants et de groupes, et plusieurs tâches pour 24 chaînes de traitements. Dans le chapitre 6, nous décrivons l'ensemble de données "HCP multi-pipeline" que nous avons construit et partagé avec la communauté pour faciliter l'étude de la variabilité analytique dans différents contextes.

(ii) Dans le Chapitre 7, nous utilisons l'ensemble de données HCP multi-pipelines pour explorer l'espace des chaînes de traitements et évaluer la stabilité des relations entre les résultats des chaînes de traitements au sein de différents groupes de participants et de tâches. Nous avons utilisé des algorithmes de détection de communautés, *i.e.* un type d'apprentissage de représentation qui permet d'identifier des groupes ou des communautés de noeuds sur des graphes, et de tirer des conclusions quant aux paramètres des chaînes de traitements qui donnent le plus souvent des résultats similaires dans différents contextes.

(iii) Les relations entre les chaînes de traitements peuvent être évaluées en termes de similitudes de leurs résultats, mais aussi en termes de compatibilité de ces résultats. Au Chapitre 8, nous explorons la validité des méga-analyses d'IRMf (c'est-à-dire la combinaison de cartes statistiques au niveau du sujet provenant de différentes études) qui combinent des données traitées différemment au niveau du sujet. Nous montrons que certains cas sont plus critiques que d'autres, ce qui entraîne un plus grand nombre de faux positifs.

Enfin, dans une quatrième partie, nous présentons quelques perspectives et travaux futurs.

En Annexe, nous présentons également d'autres travaux relatifs à l'impact de la variabilité analytique sur la crise de la reproductibilité. Dans un premier temps (Annexe A), nous examinons le contexte de la crise de la reproductibilité dans la recherche expérimentale. Ensuite, nous présentons deux études dans lesquelles nous avons exploré la relation entre la variabilité analytique et la reproductibilité en IRMf :

(A-i) Dans l'Annexe B, nous explorons l'impact de plusieurs variations dans les chaînes de traitements sur la performance des biomarqueurs de la maladie de Parkinson dérivés de l'IRMf de repos.

(A-ii) Dans l'Annexe C, nous présentons notre travail sur le projet *NARPS Open Pipelines*, une base de code reproduisant les 70 chaînes de traitements utilisés dans une étude multi-analystes (Botvinik-Nezer et al., 2020).

Au début de chaque chapitre, afin de reproduire les expériences et les figures, nous fournissons les différents codes développés pour le chapitre et un lien vers les publications

ou preprints associés. Pour plus de clarté, les détails de la mise en oeuvre de chaque expérience (architecture du modèle, hyperparamètres, etc.) sont disponibles dans une section dédiée de l'annexe.

Contexte de la thèse

Cette thèse a été réalisée au sein des équipes Empenn et LACODAM (LArge COllaborative DAta Mining) du laboratoire IRISA (Institut de Recherche en Informatique et Systèmes Aléatoires), unité mixte de recherche (UMR 6074) issue d'une collaboration entre neuf établissements pluridisciplinaires : CentraleSupélec, CNRS, ENS Rennes, IMT Atlantique, Inria, INSA Rennes, Inserm, Université Bretagne Sud, Université de Rennes. Les équipes de recherche Empenn (ERL U1228) et LACODAM sont conjointement affiliées à l'Inria, et Empenn est également affilié à l'Inserm (Institut National de la Santé et de la Recherche Scientifique).

Ce travail a été partiellement financé par la Région Bretagne (ARED MAPIS) et l'Agence Nationale pour la Recherche pour le programme de contrats doctoraux en intelligence artificielle (projet ANR-20-THIA-0018).

Ce projet de recherche a également fait l'objet d'une collaboration avec le laboratoire Big Data for NeuroInformatics du Dr. Tristan Glatard à l'Université Concordia et le laboratoire ORIGAMI du Dr. Jean-Baptiste Poline à l'Université McGill, toutes deux basées à Montréal, Canada. Le stage de mobilité internationale réalisé dans ce cadre a été financé par une bourse de recherche Mitacs Globalink (IT34055) et par une bourse du Collège Doctoral de Bretagne et de Rennes Métropole.

PREAMBLE

Introduction

Functional Magnetic Resonance Imaging

Over the last decades, the development of brain imaging has considerably enriched our understanding of the brain and its pathologies, paving the way for new diagnostic and therapeutic approaches. In particular, brain imaging techniques such as Magnetic Resonance Imaging (MRI) provided insights on the brain structure with high precision, while remaining non invasive. Nowadays, the question of understanding brain functions has an important place in many research fields ranging from medicine and psychology to artificial intelligence and philosophy. Unraveling the complexity of the brain and deciphering how different brain regions interact is a scientific challenge that captivates researchers. Functional Magnetic Resonance Imaging (fMRI) is a brain imaging technique that allows researchers to study brain activity of individuals while they perform predefined tasks. The number of published studies making use of this modality exploded in the last ten years: in 2018, more than one thousand studies registered in the website `clinicaltrial.gov` were using fMRI as an outcome measure (Sadraee et al., 2021).

In traditional fMRI studies, a set of participants is chosen based on different criteria, depending on the purpose of the experiment (*e.g.* healthy controls, disease stage, etc.). Studies are built to answer one or multiple research hypotheses, *e.g.* on the activation of a brain area during a task or on the presence of differences between activation strength between two groups of individual. Participants undergo an fMRI acquisition, which consists in a sequence of tasks constituting the activation paradigm. This paradigm is composed of a reference task, usually rest, and a task of interest whose only difference with the reference corresponds to the cognitive process to explore. Investigators recovers the raw fMRI data in the form of 4-dimensional matrices for all participants and apply a sequence of preprocessing and statistical analysis steps, called a "pipeline". At the end of a pipeline, results are output in the form of 3-dimensional statistic maps showing the activation of

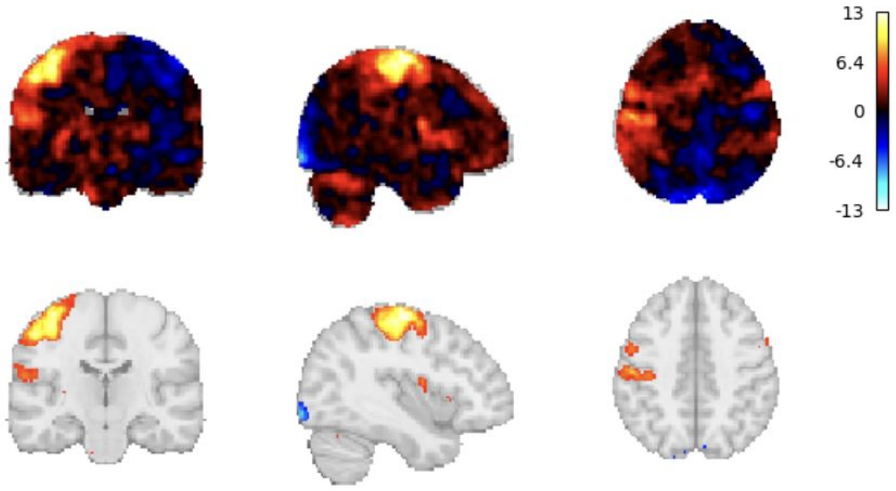

Figure 3 – Group-level statistic maps for the paradigm *right-hand*, unthresholded (upper) and thresholded (lower) using a threshold of $p < 0.05$, Bonferroni corrected.

the brain for each participant and each research question, and across the set of participants if a second-level of analysis was performed. These maps are usually thresholded with statistical inference to identify the brain regions which display significant activation. This process rely on statistical testing to determine whether observed differences or relationships between variables are likely due to random chance or if they represent true effects.

Brain imaging studies, and in particular fMRI, are subject to different sources of variability. By sources of variability, we refer to any factor whose variations lead to changes in the final results. For instance, depending on the time of the day, or on the medication state, the same participant could have different set of activations, representing a source of variability, called intra-individual variability (Chen et al., 2016). Similarly, if data are acquired for the same participant, but with different acquisition parameters, this could lead to variations in the results (Wittens et al., 2021). Naturally, it seems logical that two participants would have different results, but studies might fail to represent the whole set of inter-individual variations (Valizadeh et al., 2018). Brain imaging studies are usually small (∼30 participants in 2015 (Poldrack et al., 2017)) and composed of data acquired at a single center, leading to a poor generalizability of findings.

Analytical variability

Another source of variability relates to the variations in the final results caused by different implementations of the pipeline used to process and analyze raw data. In a meta-analysis of more than 200 papers about fMRI, Carp, 2012b showed that these pipelines are highly flexible, leaving researchers with many choices to make to analyze their data. This statement raised questions about the potential effects of these different implementations on the results. Recently, in a many-analyst study (Botvinik-Nezer et al., 2020), 70 research teams were tasked to analyze the same task-fMRI dataset with their usual method. Overall, there were no identical pipeline and the final results were relatively variable across teams. This phenomenon, also known as "analytical variability", can arise from different levels of variations in the analysis pipeline:

— At each step, when changing the algorithm to use or a parameter value.

— When varying the software package used to implement the pipeline.

— At a lower level, when varying computing conditions.

Nowadays, the idea that different analytical approaches can lead to different results is accepted in the community. Researchers now look for solutions to face the different challenges related to analytical variability (Botvinik-Nezer et al., 2023). The flexibility of analysis pipelines is particularly challenging for researchers as there is no ground truth that can be used to compare and measure the performance of competing pipelines. Thus, there is only few agreements on the good practices to guide the choice of pipeline. In practice, researchers commonly explore multiple valid analytic alternatives, but often report their results relative only to a single pipeline (or to a few set of variants). This selective reporting can result in an increase of false positive findings (Ioannidis, 2008a; Simmons et al., 2011; Gelman et al., 2019). As a potential solution, researchers can use multiverse analyses (Steegen et al., 2016) to compare and report the results of multiple analytical approaches. But, a systematic investigation of the pipeline space is impractical due to the high number of possible pipelines. In both cases, a better knowledge of the pipeline space would be useful to identify the main drivers of this variability in the results and for instance, facilitate the selection of a representative set of pipelines.

Another issue related to the variability in fMRI is the reusability of data. With the emergence of data sharing practices (Niso et al., 2022), there is an opportunity to increase

sample sizes of brain imaging studies by re-using shared data. The use of larger and more diverse samples would help to improve the reproducibility and generalizability of results and provide more flexibility as to which research question can be investigated. In practice, data re-use is usually performed with raw data coming from different studies, that are then re-analyzed with the same pipeline. Another solution is to use derived data (*i.e.* already processed). This is more optimal, not only because sharing statistic maps is not as difficult as sharing raw data due to reduced privacy constraints, but also because it avoids having to perform costly re-computations. However, there have been some evidence that derived data coming from different sources should be combined carefully in statistical studies to avoid increasing the number of false positives (Rolland et al., 2022). Moreover, there is usually a lack of annotations on data shared on public databases such as NeuroVault (Gorgolewski et al., 2015), making it challenging to re-use them. Thus, there is a need for solutions to benefit from this large amount of derived data shared on public databases.

Representation learning

Machine learning, a subfield of artificial intelligence, consists in providing real world data to a model, which will learn patterns in these data to answer a problem at hand. In particular, "representation learning" (Bengio et al., 2013) refer to the process where meaningful features are extracted from raw data to create representations that are easier to understand and to process. With their ability to model complex relationships, neural networks, used in deep learning frameworks (LeCun et al., 2015), showed promising performance for this task in many research fields. In these frameworks, representations of data are learned in a hierarchical manner by neural networks and contain meaningful features for the underlying task for which the network was trained.

In computer vision, researchers make use of Convolutional Neural Network (CNN) due to their ability to extract visual features with convolution operations. As the input data passes through successive layers, these networks learn increasingly abstract and complex features. Lower layers capture basic features like edges and textures, while higher layers represent more semantically meaningful features relevant to the task at hand. These representations are then used to output results for this task, but they can also be manipulated and transferred between models or between data to improve other models performance. "Transfer learning" (Pan et al., 2010), a use case of representation

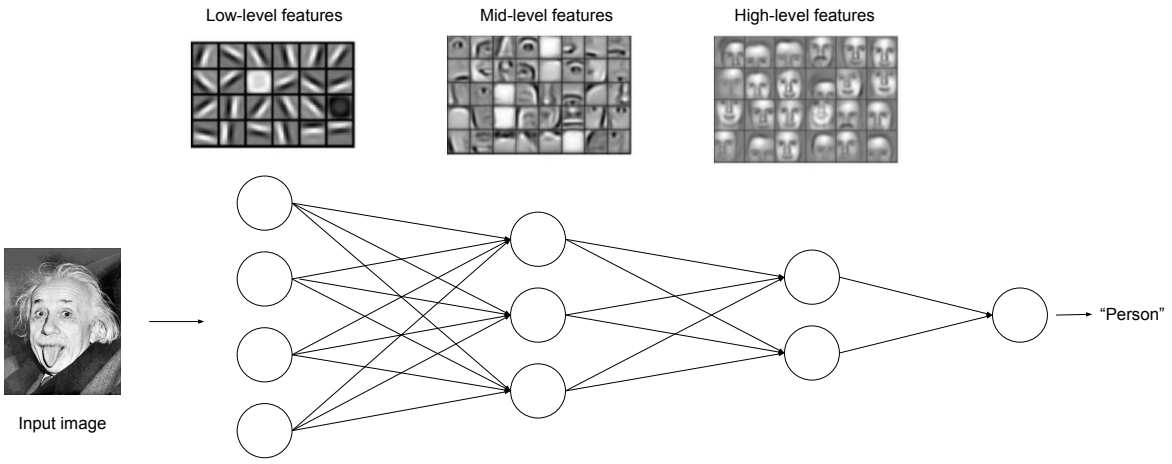

Figure 4 – Concept of representation learning in computer vision using Convolutional Neural Network (CNN). Lower-level features are extracted at the first layers, while higher-level features are are learned after.

learning, leverage pretrained models whose knowledge (*i.e.* parameters of trained layers) is transferred to another model that will by applied on data from another domain or for another task. Similarly, representations can be manipulated to transfer attributes between data. This use case is also known as "style transfer" (Gatys et al., 2016) and make use of generative models, in which networks learn to model the distribution of training data, from which new data can be sampled or existing data can be transferred.

These techniques are promising for the problems outlined earlier, as these would allow to build a comprehensive representation of fMRI results and of their sources of variability. However, learning a useful and efficient representation requires a large amount of training data to represent the diversity of the potential target data (Ricci Lara et al., 2022). This issue is of main importance in brain imaging, where studies are usually made on small and homogeneous samples. As stated above, data sharing platforms, on the contrary, contain a large number of data, coming from different sources, and thus display a good level of variability in terms of acquisition protocols, machines, sites and analysis pipelines. Using these data necessitate the use of particular methodologies, as these are usually not labeled or does not have a standardized labelling process (Poldrack et al., 2011b). Moreover, fMRI statistic maps have particular properties, which require adaptation of traditional representation learning frameworks. They contain quantitative information (*i.e.* statistical values in our context), spatial localisation is crucial information (*i.e.* the same activation

in different regions of the brain leads to a completely different interpretation) and the dimensionality of medical images is much larger (*i.e.* an fMRI statistic map contains tens of thousands of dimensions).

Contributions

Analytical variability in fMRI results lead to challenges, for researchers that design a new study, and for researchers who try to re-use data from other studies. In this context, building comprehensible and meaningful representations of the diversity in fMRI data would help to get a better knowledge of the analytical space, but also to provide solutions for researchers that want to benefit from derived data shared by the community on public platforms. However, this requires the use of large amount of data and adapted methodologies to deal with the specificities of fMRI data.

In the first series of contributions of this thesis, we propose two concrete solutions to learn and manipulate lower-dimensional representations of fMRI results to facilitate the re-use of the large amount of shared derived data. First, we leverage the NeuroVault database (Gorgolewski et al., 2015) – a large public neuroimaging database that was built collaboratively – to learn an unsupervised representation of fMRI statistic maps, that can be transferred in a self-taught learning framework to help solve new tasks (*e.g.* brain decoding). This work led to the publication of a journal paper in *Gigascience*, and pretrained models were shared with the community for further re-use. Secondly, we made the assumption that pipelines could be seen as a style component of fMRI statistic maps and proposed to use style transfer frameworks to convert statistic maps between pipelines. In this contribution, we adapted several state-of-the-art frameworks for Image-to-image transition (I2I) to our task and developed a framework based on Denoising Diffusion Probabilistic Model (DDPM). This framework makes use of a latent representation of statistic maps in which data are structured based on the most important features that distinguish them across pipelines. This contribution was the subject of a paper, available as preprint and soon submitted to *Human Brain Mapping*.

In a second series of contributions, we explore the characteristics of the fMRI pipeline space. To do so, we first built a multi-pipeline dataset composed of a large number of participants and that represents a non-exhaustive but controlled part of the pipeline

space. We published a data paper as a preprint, we plan to submit it to *Scientific Data* and to share the dataset with the community on *Public-nEUro*. Within this dataset, we used community detection algorithms (*i.e.* representation learning to identify clusters on graphs) to explore the pipeline space and assess the stability of relationships between pipeline results across different groups of participants and tasks. This work led to a paper that was accepted to the conference *ICIP 2024* and that is also available as preprint. Finally, we explore the potential of data re-use and study the validity of statistical analyses that combine data processed differently (*e.g.* with different algorithms, parameters values or software packages). A preprint is available for this contribution, in co-authorship with Xavier Rolland, and was submitted to *Imaging Neuroscience*.

This manuscript is composed of three parts. The first part is dedicated to the introduction of fMRI data analysis and of analytical variability.

(i) In Chapter 1, we present the field of application of this thesis, fMRI. We introduce the main principles of fMRI studies, outlining the journey from brain activity to fMRI raw data. Afterwards, we expose the process of translating these raw data into final results and describe the main processing steps included in traditional fMRI analysis pipelines.

(ii) In Chapter 2, we give an overview of the different sources of variability that can be observed in fMRI studies. After a brief description of each source, we focus on analytical variability and show that changes can be made at different levels of the pipeline, leading to variations in the results. We also describe the main challenges related to analytical variability and the solutions developed to tackle these.

In the second part of the manuscript, we expose our first series of contributions in which we used deep learning to mitigate (analytical) variability in fMRI results and facilitate data re-use.

(i) Chapter 3 poses the context of deep learning and its application to the field of medical imaging. We detail the fundamental concepts of representation learning and in particular, deep learning, the challenges related to medical imaging and

describe two particular cases of deep learning, namely transfer learning and style transfer, in which representations are transferred between tasks and/or data.

(ii) In Chapter 4, we propose a first solution to facilitate data re-use and we leverage the NeuroVault database in a self-taught learning framework, a specific type of transfer learning. To do so, we learn an agnostic representation of fMRI statistic maps using a Convolutional AutoEncoder (CAE) and then fine-tune it towards a variety of tasks.

(iii) While representations were built and transferred between tasks in the previous chapter, in Chapter 5, we propose another solution to re-use shared derived data and manipulate representations to convert data between different pipelines. We propose a new framework that makes use of a Convolutional Neural Network (CNN) trained to distinguish statistic maps between pipelines and extract the higher-level features of data to help transition between pipelines with diffusion models.

The concrete use of these solutions rely on the identification of the main drivers of variations to find critical cases where mitigation of analytical variability is required and appropriate. In the third part, we explore the fMRI analytical space to better understand relationships between pipelines results and identify some particular challenges related to data re-use and analytical variability.

(i) A full investigation of the pipeline space is impractical as it is particularly large, we propose in contrast to explore a smaller part of this pipeline space across different contexts: a large number of participants and groups, and several tasks for 24 pipelines. In Chapter 6, we describe the HCP multi-pipeline dataset that we built and shared with the community to facilitate the study of analytical variability across different contexts.

(ii) In Chapter 7, we make use of the HCP multi-pipeline dataset to explore the pipeline space and assess the stability of relationships between pipeline results across different groups of participants and tasks. We used community detection algorithms, *i.e.* a type of representation learning that allows to identify clusters or communities of nodes on graphs, and derive conclusions as to which pipelines parameters mostly give similar results across contexts.

(iii) Relationships between pipelines can be seen in terms of similarities of their results, but also in terms of compatibility of these results. In Chapter 8, we explore the validity of fMRI mega-analyses (*i.e.* combining subject-level statistic maps from different studies) that combine data processed differently at the subject-level. We show that some cases are more critical than other, leading to a higher number of false positives.

Finally, in a fourth part, we present some perspectives and future works.

In Appendix, we also share additional works related to the impact of analytical variability on the reproducibility crisis. In a first Chapter (A), we discuss the context of the reproducibility crisis in experimental research. Then, we present two studies in which we explored the relationship between analytical variability and reproducibility:

(A-i) In Appendix B, we explore the impact of several variations in the workflow on the performance of resting state fMRI derived Parkinson's disease biomarkers.

(A-ii) Then, in Appendix C, we present our work on the *NARPS Open Pipelines* project: a codebase reproducing the 70 pipelines used in a many-analyst study (Botvinik-Nezer et al., 2020).

At the beginning of each chapter, in order to reproduce the experiments and the figures, we provide the different source codes developed for the experiments and link to the associated publications or preprints.

Context of the thesis

This thesis was carried out in Empenn and LACODAM (LArge COllaborative DAta Mining) teams at the IRISA (Institut de Recherche en Informatique et Systèmes Aléatoires) laboratory, a joint research unit (UMR 6074) resulting from a collaborative effort between nine multi-disciplinary establishments: CentraleSupélec, CNRS, ENS Rennes, IMT Atlantique, Inria, INSA Rennes, Inserm, Université Bretagne Sud, Université de Rennes. Both Empenn (ERL U1228) and LACODAM research teams are jointly affiliated with Inria, and Empenn is also affiliated with Inserm (National Institute of Health and Scientific Research).

This work was partially funded by Region Bretagne (ARED MAPIS) and Agence Nationale pour la Recherche for the program of doctoral contracts in artificial intelligence (project ANR-20-THIA-0018).

This research project was also the subject of a collaboration with the Big Data for NeuroInformatics lab of Dr. Tristan Glatard at Concordia University and ORIGAMI lab of Dr. Jean-Baptiste Poline at McGill University, Montreal Canada. The mobility internship realized in this context was funded by a Mitacs Globalink Research Award (IT34055) and by a scholarship from the College Doctoral de Bretagne and Rennes Metropole.

List of publications

In neuroimaging, the first authors are the people who carried out the work, and the last authors are the supervisors.
Journal ranks come from Scimago, conference ranks come from Conference rank.

* means co-authors, with equal contributions.

Journal papers

On the benefits of self-taught learning for brain decoding
Germani, E., Fromont, E., Maumet, C.
GigaScience, 2023, Vol. 12, pp.1-17. DOI: 10.1093/gigascience/giad029.
[Paper] [HAL] [Code] [Derived data]

Conference papers

Uncovering communities of pipelines in the task-fMRI analytical space
Germani, E., Fromont, E., Maumet, C. DOI: arXiv:2312.06231.
Accepted at ICIP 2024 (Rank A-B).
[Preprint] [Code]

Submitted papers

On the validity of fMRI studies including subject-level data processed with different pipelines
Germani, E.*, Rolland, X.*, Maurel, P., Maumet, C. DOI: arXiv:2402.12900v1.
** Co-first authors. In revision at Imaging Neurosciences.*
[HAL] [Code]

Predicting Parkinson's disease trajectory using clinical and functional MRI features: a reproduction and replication study
Germani, E., Baghwat, N., Dugré, M., Gau, R., Montillo, A. M., Nguyen, K. P., Sokolowski, A., Sharp, M., Poline, JB., Glatard, T. DOI: arXiv:2403.15405v1.
In revision at PLOS ONE.
[Preprint] [Code]

Preprints

Mitigating analytical variability of fMRI results with style transfer
Germani, E., Fromont, E., Maumet, C. DOI: arXiv:2404.03703v1.
Target journal: Human Brain Mapping.
[Preprint] [Code]

The HCP multi-pipeline dataset: an opportunity to investigate analytical variability in fMRI data analysis
Germani, E., Fromont, E., Maurel, P., Maumet, C. DOI: arXiv:2312.14493.
Target journal: Scientific Data. Publication of the dataset in Public nEUro.

***Data sharing challenges**: discussion with Public-nEUro and our data protection officers to share the dataset in compliance with the European General Data Protection Regulation (GDPR), discussion with the BIDS maintainers on the best way to organize our dataset into Brain Imaging Data Structure (BIDS) format, which does not contain specification for sharing statistic maps.*
[Preprint] [Code]

Conference posters

National conference

Prédire l'évolution de la maladie de Parkinson à l'aide de données cliniques et d'IRM fonctionnelles: reproduction et robustesse d'une étude
Germani, E., Baghwat, N., Dugré, M., Gau, R., Sokolowski, A., Sharp, M., Poline, JB., Glatard, T.
IABM 2024 - 2ème édition du Colloque Français d'Intelligence Artificielle en Imagerie Biomédicale , Mar 2024, Grenoble, France
🏆 ***This poster was awarded with the Best poster in medical imaging price during the conference.***
[Poster]

Representation learning for more reproducible fMRI data analyses
Germani, E., Fromont, E., Maumet, C.
IABM 2023 - Colloque Français d'Intelligence Artificielle en Imagerie Biomédical, Mar 2023, Paris, France. pp.1
[Poster]

International conference

Exploring variability patterns in the task-fMRI analytical space
Germani, E., Fromont, E., Maumet, C.
OHBM 2023 - 29th Annual Meeting of the Organization for Human Brain Mapping, Jul 2023, Montreal, Canada
[Poster]

FMRI data analysis: How does analytical variability vary with sample size?
Germani, E., Maumet, C.
OHBM 2022 - 28th Annual Meeting of the Organization for Human Brain Mapping, Jun 2022, Glasgow, United Kingdom. pp.1-5
[Poster] [Code]

Conference or Workshop Summaries

Proceedings of the OHBM Brainhack 2022
Moia, S., Wang, H.-T., Heinsfeld, A. S., Jarecka, D., Yang, Y. F., Heunis, S., Svanera, M., De Leener, B., Gondova, A., Kim, S., Basavaraj, A., Bayer, J. M. M., Bayrak, R. G., Bazin, P.-L., Bilgin, I. P., Bollmann, S., Borek, D., Borghesani, V., Cao, T., Chen, G., De La Vega, A., Dresbach, S., Ehses, P., Ernsting, J., Esteves, I., Ferrante, O., Garner, K. G., Gau, R., Germani, E., Ghafari, T., Ghosh, S. S., Goodale, S. E., Gould Van Praag, C. D., Guay, S., Gulban O. F., Halchenko, Y. O., Hanke, M., Herholz, P., Heuer, K., Hoffstaedter, F., Huang, R., Huber, R., Jensen, O., Keeratimahat, K., Kosciessa, J. Q., Lukic, S., Magielse, N., Markiewicz, C. J., Martin, C. G., Maumet, C., Menacher, A., Mentch, J., Monch, C., More, S., Muller, L., Muller-Rodriguez, Leonardo, Nastase, Samuel A., Nicolaisen-Sobesky, E., Nielson, D. M., Nolan, C. R., Paugam, F., Pinheiro-Chagas, P., Pinho, A. L., Pizzuti, A., Poldrack, B., Poser, B. A., Rocca, R., Sanz-Robinson, J., Sarink, K., Sitek, K. R, Spychala, N., Stirnberg, R., Szczepanik, M., Torabi, M., Toro, R., Urchs, S. G. W., Valk, S. L., Wagner, A. S., Waite, L. K., Waite, A. Q., Waller, L., Wishard, T. J., Wu, J., Zhou, Y., Bijsterbosch, J. D.
Aperture Neuro, 2024, Vol. 4. DOI: 10.52294/001c.92760.
[Paper]

Scientific Popularization

Brain Awareness Week. Editions 2022 and 2024.
[Website][Slides]

L Codent, L Créent. Editions 2021-2022 and 2022-2023. Talk at the colloquium Femmes&Sciences 2022.
[Website] [Slides]

TISSAGE. Quiz: “Mieux comprendre l’intelligence artificielle en étudiant l’intelligence humaine”.
[Slides]

Open Science

During my thesis, I had the chance to discover and dive into the Open Science community. My journey took the form of several principles that I tried to respect throughout my research and other scientific contributions.

Opening my research

At the beginning of each project, I create a code repository using Git (Chacon et al., 2014) to version my code and GitHub or GitLab to make it available publicly online. Even if the project ends or is left in standby, this allows to let an imprint of my work online, which can be useful to help future researchers who would like to continue this project or start a new one. For published or finished works, I use Software Heritage (Cosmo et al., 2017) to archive this code and preserve it in the long term.

My projects are sometimes associated with derived data. These data can take the form of pretrained models, which was the case for the works described in Chapter 4 and 5. I always share these models with the community, for instance on Zenodo (European Organization For Nuclear Research et al., 2013), to facilitate their re-use by other researchers.

For several of my works, I use a dataset of statistic maps processed with different pipelines that I built using publicly available data (see Chapter 6). These data are currently in the process of being shared with the community, as these might help researchers to perform their own analyses.

When a project lead to a written paper, I also take care to publish the preliminary versions as preprint on HAL [1] and arXiv [2].

1. `https://inria.hal.science/`
2. `https://arxiv.org/`

Making my research more reproducible

When I share my work, I try to make my experiments as reproducible as possible. This starts by writing comprehensive and interactive README files for my code repositories, to help users that would like to re-run the project. In these repositories, I also use Jupyter Notebooks to facilitate reproducibility and comprehension of my code. As a further step towards reproducibility, during some of my projects, I developed several Docker (Merkel, 2014) and Singularity (Kurtzer et al., 2017) containers that I also shared with the community and that could be directly used to re-launch my experiments.

In all my projects, I use open datasets (*e.g.* Human Connectome Project (Van Essen et al., 2013), NeuroVault (Gorgolewski et al., 2015), etc.), which first allows me to make my work easily reproducible. These datasets are very useful as initial 'pilot' data for methods development and experimentations, as they reduce the time and cost of acquiring new data. Due to their large size (more than 1TB for Human Connectome Project (HCP)), I had to find solutions to store these data and analyze them easily.

Using such data also allows me to work with common data representations, in my case the Brain Imaging Data Structure (BIDS) (Gorgolewski et al., 2016). Some of the datasets that I used were not initially in BIDS format (*e.g.* raw data from the Human Connectome Project (HCP)) and I used conversion software, such as HeudiConv (Halchenko et al., 2024), to convert the dataset in BIDS and facilitate the re-launching of my code on other datasets in BIDS format.

Collaborating with other researchers

These previously detailed efforts are also to me a solution to facilitate collaborations with other researchers. During the first year of my thesis, I attended the OHBM Brainhack (in Glasgow, 2022) and realized that collaborating was leading to new knowledges, new ideas and thus, a better research. During the three days of this Brainhack, I had the chance to present the *Narps Open Pipelines* project (detailed in Appendix C) that I started during my master's internship, and that is now at a far higher level thanks to the collaborations initiated at this event. We recently published the Proceedings of this Brainhack in *Aperture Neuro* (see). I also participated to the OHBM Brainhack in Montreal, 2023 and to local hackathons which I organized in the ORIGAMI team during my mobility internship, and in the Empenn team. This project led to several fundings

which allowed to hire a research engineer for 18 months to continue the project (Boris Clenet, research engineer in the Empenn team).

During my thesis, I also had the opportunity to collaborate with international researchers through an international mobility in Montreal, Canada. I spent four months in Tristan Glatard's and JB Poline's labs to work on the LivingPark project, co-hosted by their labs. This project showed me the importance of international collaborations and allowed me to discover new techniques to facilitate open science and reproducibility.

Communicating to the general public

To my opinion, opening science also means opening our research to the general public through scientific popularization. During my thesis, I participated to several actions to transmit knowledge to the public. In particular, as part of the *L Codent, L Créent* action, I animated educational sessions (8 × 45min) of creative programming for middle school girls (12-13yo) during the Editions 2021-2022 and 2022-2023. The goal was to promote computer science and demystify coding.

I also took part in the organization of an event for the Brain Awareness Week [3] (Editions 2022 and 2024) and was volunteer for the organization of the festival Pint of Science 2024 [4]. In order to improve my skills, I also followed the training to scientific popularization as part of the TISSAGE project (TrIptyque Science Société pour AGir Ensemble) [5] led by the Ministry of Higher Education, Research and Innovation.

3. `https://www.semaineducerveau.fr/`
4. `https://www.pintofscience.fr`
5. `https://www.univ-rennes.fr/saps-tissage`

Part I

Context

Chapter 1

Functional Magnetic Resonance Imaging (fMRI)

Functional Magnetic Resonance Imaging (fMRI) is a brain imaging technique used to explore brain activity and the functional connections between different brain regions. The first section of this chapter outlines the journey from brain activity to the acquisition of fMRI raw data, we provide: an overview of the objectives of brain imaging and the primary techniques employed in this field, with a specific focus on fMRI (1.1.1), explanations about the main physiological process behind fMRI (1.1.2) and descriptions of the fundamental concepts involved in traditional fMRI acquisition (1.1.3). In a second section, we delve into the process of translating fMRI raw data to fMRI results, with a description of the multiple processing steps that can be used to analyze fMRI data, starting from preprocessing (1.2.1) to statistical analysis (1.2.2 and 1.2.3) and inference (1.2.4).

1.1 From brain activity to fMRI raw data

1.1.1 Brain imaging and fMRI

Brain imaging - also known as neuroimaging - provides the opportunity to capture rich and descriptive information about the structure and functional architecture of the brain non-invasively. Nowadays, brain imaging techniques are commonly used to acquire raw data that are processed to answer questions about the healthy and pathological brain, in medicine and in psychology. Depending on the research question, different brain imaging techniques can be used, involving different physical and biological processes: radiation (X-ray emission, detection of injected radioactive products), measurement of electrical activity or magnetic fields. We mainly distinguish two types of brain imaging techniques:

- **Structural imaging** that explores the anatomy of the brain, for instance with Computed Tomography (CT) scan (based on X-rays) and Magnetic Resonance

Imaging (MRI) (based on magnetic fields).

- **Functional imaging** in which the activity of brain areas is studied during different tasks, using for instance Positon Emission Tomography (PET) scan (based on the injection of a radioactive tracer), Electroencephalography (EEG) and Magnetoencephalography (MEG) (respectively measuring the activity of neurones using electrodes that measure the electrical potential and the magnetic fields at the surface of the brain) and Functional Magnetic Resonance Imaging (fMRI) (which measures variations in the local oxygenation of blood, which in turn reflects the amount of local brain activity.).

For the remainder of the manuscript, we will focus on fMRI, and particularly on task-fMRI. There are two main types of fMRI: resting-state and task-fMRI. In resting-state fMRI, brain activity is recorded when participants are at rest, *i.e.* when they are not performing any task supposedly underlying any cognitive process. This allows to investigate the synchronicity of activations between different brain regions and thus, to identify resting-state networks. In task-fMRI , participants are asked to perform specific tasks during the acquisition, *e.g.* movement, speaking, etc. This allows to measure variations in the recorded signal at the time of the expected response and thus, to detect activations in specific areas of the brain related to the task. The physiological processes involved for the acquisition of resting-state and task-fMRI are the same, and relates to haemodynamic changes in the brain. However, the preprocessing and statistical analysis used to analyse the data acquired varies between the two, with common preprocessing steps.

1.1.2 Principle of BOLD fMRI

The most common method used for fMRI takes advantage of the Bold Oxygen Level Dependent (BOLD) signal, which rely on the fact that increased neuronal activity in a region of the brain correlates with increased blood flow to that specific region. This amount of blood sent to the active neurons is larger than what is needed to oxygenate neurons, leading to a relative surplus in oxygenated blood. These changes in levels of oxygenated or deoxygenated blood can be detected on the basis of their differential magnetic susceptibility, and can be compared to the expected haemodynamic response for each task to estimate brain activity (Poldrack et al., 2011a).

1.1.3 Experimental design and protocols

In research settings, to explore brain activity under different conditions, task-fMRI studies usually involve multiple participants for which a temporal sequence of brain volumes is acquired while they perform a set of tasks. During the acquisition, images are acquired with a regular time interval between two consecutive images (called repetition time or TR) during a specific time. A protocol is developed to explore a cognitive process and participants are asked to perform a set of tasks (*e.g.* raising hands and raising foots) with predefined onsets, durations and sometimes, intensities. This set of task and the associated stimuli are called a paradigm (*e.g.* motor paradigm). Paradigms are designed to manipulate specific mental processes, in order to better understand how they relate to brain activity. Two main experimental designs exists: block designs and event-related designs. Block designs involve long-lasting stimuli, while event-related designs involve brief stimuli leading to short neural responses. Event-related designs usually offer greater flexibility, but may have lower signal-to-noise ratios compared to block designs (Liu, 2012; Petersen et al., 2012).

1.2 From fMRI raw data to fMRI results

Figure 1.1 – Illustration of common fMRI protocols

For each participant, several data are acquired during a session: functional data corresponding to 4-dimensional matrices containing a time-series of concatenated brain volumes and additional files used to process them, for instance structural data (*i.e.* with more precise anatomical information) or field maps (*i.e.* measuring field inhomogeneity). Stimulus time-series, which give information about the set of tasks performed by

the participant, is also recorded for further analysis. Brain imaging data are composed of voxels (3-dimensional version of a pixel, *i.e.* a value on a regular grid in 3-dimensional space) that represent the intensity of the signal in the corresponding part of the brain. In functional data, there is a supplementary time dimension, meaning that each voxel is associated with a time course (see Figure 1.1). A session of functional acquisition is also called a "run". Depending on the study protocol, multiple runs can be acquired, resulting in multiple functional data for a participant for the same paradigm. The data obtained after acquisition are called the "raw data".

The sequence of steps applied to the raw data to obtain the final results is called a "pipeline". In fMRI data analysis, a pipeline is traditionally split into three main steps, itself composed of multiple sub-steps:

1. Preprocessing: cleaning and preparation of data for further analyses, usually involving additional 3-dimensional files such as structural data or field maps.
2. First-level analysis: at the run or at the subject-level, to analyze each voxel's time course to identify changes in the BOLD signal in response to some manipulation.
3. Second-level analysis: referring to the combination of run-level results in a subject-level analysis or to the combination of subject-level results at the group-level.

To facilitate comprehension, in the following, we will refer to subject-level analysis for both run-level and subject-level analyses.

1.2.1 Preprocessing

An fMRI pipeline usually starts by the preprocessing of the raw data. This step is fundamental to perform further analyses, due to the high number of artifacts in the data and due to the variations in the shape of the brain across participants. Using a sequence of several image and signal processings, the goal is to increase signal-to-noise ratio and to prepare data for group-level analyses with standardization steps that aligns data between participants so that a voxel's coordinate in the brain of participant A corresponds to the same location in the brain of participant B.

The preprocessing part of the pipeline is composed of several steps, that have an effect at the temporal level (*i.e.* involving operations that filter or affect the properties of data across the time dimension) or at the spatial-level (*i.e.* involving operations that filter or affect the spatial properties of data, such as spatial orientation, resolution, and shape).

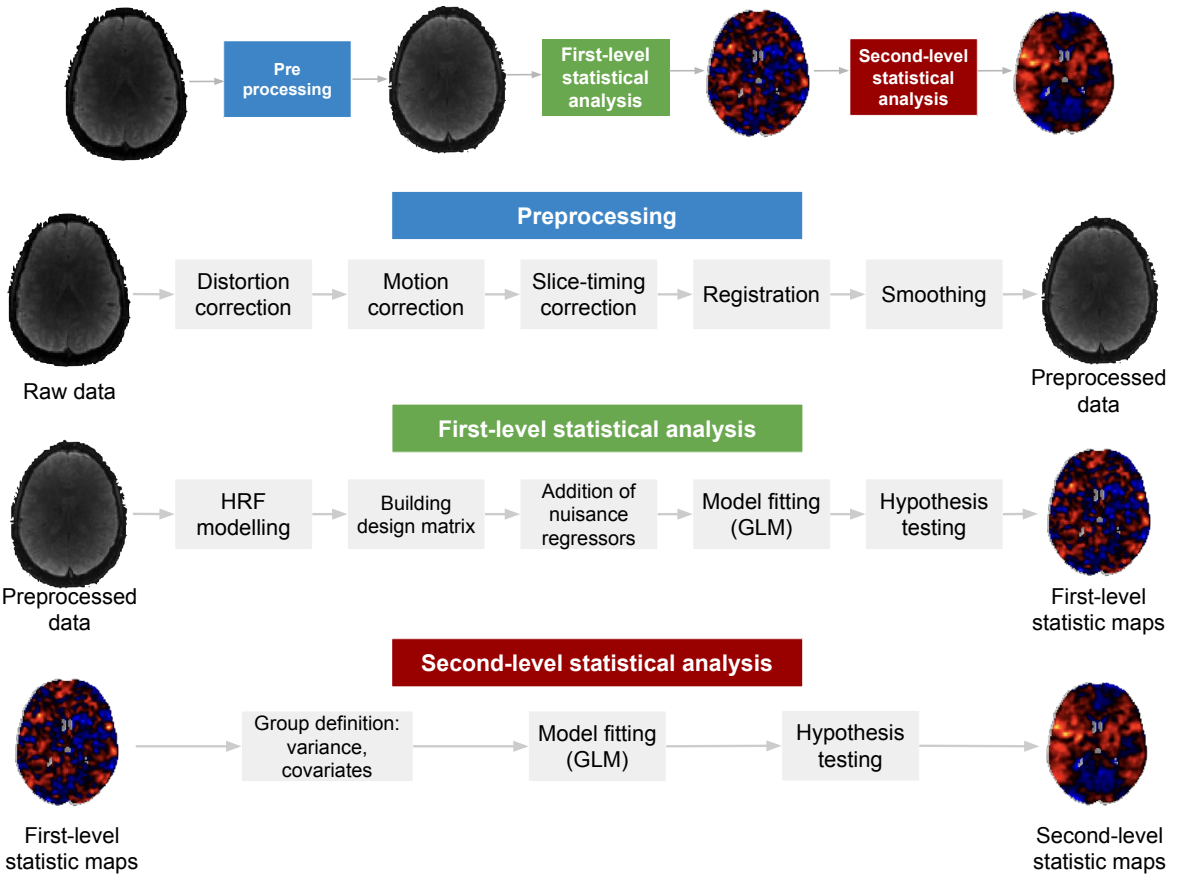

Figure 1.2 – Example of a standard fMRI pipeline: pre-processing, first and second-level statistical analyses.

Example of a standard preprocessing workflow is presented in Figure 1.2, with an example of functional raw data before and after preprocessing.

In the following subsections, we describe several preprocessing steps that can be used to clean and align raw data between participants. These steps can be performed on both resting-state and task-fMRI data, but some of these are more commonly used during resting-state fMRI data analysis, as signal-to-noise ratio in the data is usually lower.

1.2.1.1 Distortion correction

Echo Planar Imaging (EPI), the technique used to acquire fMRI BOLD data, is very sensitive to magnetic field inhomogeneity, causing geometric distortion in the images. This phenomenon particularly affects regions where there is an air-tissue interface, *i.e.* where the magnetic field varies, causing dropouts and distortions. Different techniques

have been developed to correct for image distortions. We will refer to "unwarping" for the non field map based technique and to "undistortion" for the one that involves the use of field maps representing the field inhomogeneities (Hutton et al., 2002).

Unwarping is based on the susceptibility-by-movement assumption (Andersson et al., 2001). After realignment (motion correction, explained in 1.2.1.2), there is residual movement-related artefacts caused by the object having different shape at different time points. The unwarping technique uses these remaining artefacts and the movements parameters computed during realignment to estimate how the distortions change with participant movement. However, this corrects images to some "average" distortion and does not actually remove the "static" distortions.

Undistortion, the field map based technique, attempt to correct for the "static" component of the geometric distortion, *i.e.* not related to motion. This technique can be used in complement to unwarping to improve anatomical fidelity but it requires the acquisition of supplementary images to build the field map.

1.2.1.2 Motion correction

The movement of a participant during the acquisition can impact the analysis of the resulting images. Indeed, if we look at the signal at a specific voxel coordinate, the same signal may changes coordinates across time due to movement. Moreover, the brain signal may vary because of the movement and not because of the paradigm of interest.

The first step to correct for movement is to perform a rigid-body transformation to realign data to a reference scan: often, the first or mean volume of functional raw data. Translations and rotations of the brain on the x, y and z axes are compared to the reference and differences are computed to obtain 6 movement regressors. In some cases, in addition to the computation of the movement regressors, the image is also "realigned", *i.e.* modified to permanently apply the computed transformations to the image.

However, realignment does not solve all movement-related issues, in particular due to the interaction of movement with the inhomogeneity of the field. This can cause distortions of voxels and thus, non-rigid movements. The remaining motion in the image can be mitigated at two other steps: 1) movement regressors computed at this step can be regressed out from the signal to further remove movement-related artifacts (Friston et al., 1996) (see 1.2.2.3) and 2) unwarping, which tries to estimate the effects of interactions between field inhomogeneity and movement and correct for them (see 1.2.1.1).

1.2.1.3 Slice-timing correction

During acquisition in two dimensions, each 2-dimensional slice of a volume is acquired at a distinct time point, since images are collected in discrete slices. However, when we model the data at each voxel for statistical analysis, we assume that all of the slices were acquired simultaneously, which can be a problem when modelling rapid events. To correct for this, the time series of each slice can be adjusted to make it appears like all slices were acquired at the same time. This step may not be beneficial for all acquisitions, this depends on the acquisition parameters and on the experimental protocol. Slice-timing correction might also interact with other processing steps (Parker et al., 2019), for instance more accurate motion estimate can be obtained if motion correction is performed before slice-timing.

1.2.1.4 Co-registration and standardization

To further correct for head movement and obtain comparable images between participants, two steps are usually performed: co-registration and standardization (also known as normalization).

Co-registration Co-registration corresponds to the alignment between two acquisitions: a structural image and the functional images. Usually, the structural image is realigned to the functional ones, using the same method as for realignment (rigid-body transformations, see 1.2.1.2). This step is performed before normalization and allows to compute normalization parameters on the structural image, which has higher spatial resolution and fewer artefacts, and after, apply these parameters to the functional images afterwards. This step can also be bypassed in favor of direct transformation into the standard template coordinate system (Calhoun et al., 2017).

Normalization / Standardization Normalization also corresponds to the alignment between images, but, contrary to co-registration, it aligns functional data of different participants into a common standard template. Indeed, participants have different brain shapes and to allow for group-level analyses, it is important that each voxel of the brain is located at the same coordinate between different participants. Similarly to realignment and co-registration, linear transformations are applied in a first step: translations and rotations, plus zooms and shears. These transforms are often complemented by non-linear registration using deformation fields to further reduce distortions. It can also incorporate

regularization (*i.e.*, imposing penalties for excessive distance between the parameters and their expected values) or segmentation(Ashburner et al., 2005) (*i.e.*, separating gray and white matter) to obtain more robust results.

The most widely used standard template is the one of the Montreal Neurological Institute: MNI152 (Fonov et al., 2009), but this template is specific to a certain category of the population and may not fit some specific studies. For instance, studies on infants or on a specific demographic category of the population might require specific templates.

1.2.1.5 Spatial smoothing

Spatial smoothing consists in averaging the signals of neighboring brain voxels, which can be justified by the correlation of their function and blood supply. This step helps improving signal-to-noise ratio but decreases spatial resolution and blurs the image. Depending on the features that need to be extracted from the raw data, in particular for resting-state fMRI, this step can be deprecated. However, it is commonly performed for task-fMRI. The standard method implies convolution of the raw data with a Gaussian kernel that multiply the signals of close neighboring voxels with a high weight and from more distant voxels with a lower weight. The optimal kernel size is variable, but in practice, Full-width at Half-Maximum (FWHM) value of the Gaussian kernel is typically set to 4 to 6 mm for participant-level studies and to 6 to 8 mm for group-level analyses.

1.2.1.6 Temporal filtering

Functional data suffer from temporal noise, which refers to changes in signal over time due to factors that are not related to brain activity. It can arise from the scanner (physical noise) but also from the participant (physiological noise, such as motion, breathing and cardiac pulsation). This temporal noise can be corrected with different steps during preprocessing. Smoothing, which is known to reduce spatial noise, is also beneficial for temporal noise as it cleans time courses by reinforcing signals and cancelling noise.

Detrending The origin of the linear trend of fMRI signal is still discussed in the community. Two hypotheses are discussed: some believe that it arises from scanner instability (Huettel et al., 2004), while others believe that it may have other meaning, at least in resting-state fMRI (Wang et al., 2014). However, the linear trend may be problematic when trying to estimate brain activity. Linear detrending consists of modelling the voxel's time-series using a General Linear Model (GLM) and subtracting the linear component

from the original signal (Bandettini et al., 1993). If the data does have a trend, detrending forces its mean to zero and reduces overall signal variations.

High-pass filtering To further remove temporal noise and in particular linear trend and scanner drifts, high-pass filtering can also be performed. It also filters out linear trends, so adding linear trend removal is redundant but often described as two different steps in studies. Noise is particularly expressed in the low-frequencies in fMRI signal, so high-pass filtering can help to remove this low-frequency noise.

Low-pass filtering Highest temporal frequencies can also be filtered, allowing only the low frequencies to pass and limit the impact of physiological noise such as respiratory or cardiac noise, which are associated with high signal frequencies. One way to do this is to smooth the time-series with a Gaussian kernel over time, similarly to spatial smoothing but instead of computing the weighted average of neighboring voxel intensities at the same time, temporal smoothing computes those averages over time, using neighboring time points. This step is used in resting-state fMRI preprocessing pipelines, but usually not applied in task-fMRI.

1.2.1.7 Regression of nuisance signals

To further remove any non-neural activity-related process, several nuisance signals are often regressed out from data using multiple linear regression. Indeed, even if some of the noise can be remove by high-pass and low-pass temporal filtering, high-frequency confounds from breathing, heart beat and movement may still remain in the signal. Time series of physiological noise can be included as noise regressors into a GLM to remove the part of the signal explained by the nuisance regressor from the residuals. Confounds such as motion regressors, CerebroSpinal Fluid (CSF) or White Matter (WM) signals can also be regressed out, with few consensus on which ones to use. Indeed, while it might sound interesting to remove as much noise as possible from the signal, a high number of regressors in a GLM might lead to a more conservative significance testing of the model due to a lower number of degrees of freedom. This is of particular importance in task-fMRI, but less taken into account for resting-state which is especially vulnerable to physiological artifacts. The use of such regressors in statistical analysis for task-fMRI is described in Section 1.2.2.3.

There are multiple ways to compute these confounds. Motion-related regressors are

often computed during the realignment step of the preprocessing, sometimes enriched with the squares, derivatives, and squares of derivatives of the six original parameters to obtain 24 movement regressors (Yan et al., 2013a). Regarding breathing and cardiac noise, it is possible to record these values during the acquisition, but this requires a specific setup and the recording can contain artifacts too. A common procedure is to extract physiological noise using the time series of voxels located in white matter (WM) or ventricles (CSF) since these signals are of no interest (Weissenbacher et al., 2009).

Other techniques make use of dimensionality reduction to identify specific parts of the signal that correspond to noise. Component Based Noise Correction Method (CompCor) (Behzadi et al., 2007) derives significant principal components of noise from regions-of-interest in which the signal is unlikely to be modulated by brain activity. These component can then be included as confounds in the GLM, similarly to other nuisance regressors. Independant Component Analysis (ICA) techniques decompose the data in a set of spatial components and their associated time-course, with the intention of regressing out the components representing noise. However, to identify these components, one need to manually annotate them or train a model to identify these components, which might be delicate and very specific depending on the study.

Regression of global signal is a debated topic. It might be beneficial as it reduces the impact of motion but it also removes some signal of interest (Yan et al., 2013a; Satterthwaite et al., 2013). Studies also showed that regression of global signal can add anti-correlation and alter connectivity structure (Yan et al., 2013b; Weissenbacher et al., 2009).

1.2.1.8 Data cleaning: scrubbing, despiking

Even after all these preprocessing steps, it may remain some large intensity increase in the signal, called "spikes", which are caused by scanner instability or high level movements (coughing for instance). These spikes cannot be properly removed with temporal filtering and require a specific processing.

To deal with these spikes, several methods can be used. The first one is despiking, in which the signal of abnormally high voxels will be made lower artificially. This method allows to modify the signal while keeping all volumes and time points. The second method consists in identifying the time points where large movements occur and adding this information as a nuisance regressor or removing the identified volumes from the data. This technique is called "scrubbing" and the identification of outlier volumes rely on

metrics such as framewise displacement (FD) or the DVARS, representing the spatial root mean square of the data after temporal differentiation (Power et al., 2012; Power et al., 2013; Afyouni et al., 2018). However, this technique mostly relies on motion parameters computed during realignment, which may not be perfect and requires to choose a threshold for the different metrics to identify outliers. This outlier identification can also lead to the elimination of participants due to a large number of outlier volumes.

1.2.2 First-level analysis

After preprocessing, functional data are cleaned and ready for further feature extraction using different analyses. In task-fMRI, the goal is to explore brain activity and to measure which part of the signal recorded during acquisition is related to the task performed by the participant. Since we have access to the time-series of the stimuli and since we approximately know how the signal should vary in response to a stimuli, we can model the expected brain response in case of brain activation for each task and compare this to the real signal in the raw data. To do so, we use a General Linear Model (GLM) to fit the fMRI signal present in functional data at each position of the brain to regressors computed to represent the different tasks of interest. This step can be performed at the run-level (each run analyzed separately) or at the subject-level (concatenation of runs). Example of a standard first-level statistical analysis workflow is presented in Figure 1.2, with an example of first-level statistic maps.

In the following section, we explain how GLM works and in particular, multiple linear regression, which is used for statistical analysis at the first and second-level. This corresponds to a summary of the description provided in Appendix A of Poldrack et al., 2011a. Then, we explain how the haemodynamic response is modeled to estimate the parameters of the GLM and present several options that can be used to build the design matrix. Finally, we briefly describe the principles of hypothesis testing and how it can be used at the first-level of fMRI data analyses.

1.2.2.1 General Linear Model

The purpose of GLM is to explain a vector Y with a sum of weighted vectors X and an error term ϵ. The model is the following:

$$Y = X\beta + \epsilon \tag{1.1}$$

with:

- Y, a matrix of the data we want to fit to the model.
- X, a design matrix with $p + 1$ columns with p independant explanatory variables and one constant term. The columns of X are called "regressors".
- β, a vector of parameter values associated to each regressor, these parameters are what we want to estimate. Each β_i associated to a variable X_i is interpreted as the effect of X_i controlling for all other variables in the model.
- ϵ, a vector of random variables that constitute noise in the data.

To explain this vector Y and thus, fit the data to the design matrix, we must estimate $\hat{\beta}$, the set of parameter values β that best explains the data Y in function of the regressors in X, while minimizing the residual noise. Since X is not a square matrix, we cannot directly solve the model equation, but we can multiply both sides by X': $X'Y = X'X\beta$. This leads to the following equation $\hat{\beta} = X'Y \times (X'X)^{-1}$, for which any β that satisfies the equation minimize the sum of squares of the residuals. This equation assumes that $X'X$ is invertible: *i.e.* X must have full column rank, and thus regressors must not be linear combinations of other regressors in the design matrix. If this is not the case, $\hat{\beta}$ could take multiple possible values to minimize the sum of squares of residuals and the estimation would be highly unstable.

In the first-level of task-fMRI data analysis, multiple linear regression is used on each voxel of the preprocessed functional data to estimate parameter values that explain the regressors that are modeled to correspond to the tasks performed. If we go back to Equation 1.1, Y represents the time-serie of the voxels after preprocessing and X the design matrix modeling the expected response depending on the tasks performed, with potentially other regressors included to represent noise (see Nuisance regressors).

1.2.2.2 Modelling the expected response

To build the design matrix X, we must model the expected response of the brain depending on the task. The BOLD signal measures variations of the haemodynamic response (see Section 1.1.2), whose time course is a low-pass-filtered expression of the total neural activity (Logothetis et al., 2001). This response starts by an increase shortly after the neuronal activity (1-2 seconds), called the initial dip. It reaches a peak 4 to 6 seconds after the stimulus and then starts decreasing until 12 to 20 seconds. We observe a post-stimulus undershoot, which is relatively small compared to the positive amplitude.

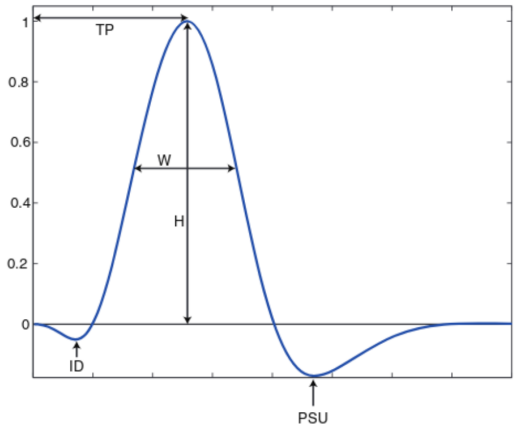

Figure 1.3 – Characteristics of the Haemodynamic Response Function (HRF). ID corresponds to the initial dip (which is sometimes not represented), TP corresponds to the time from the stimulus until peak, H to the height of response, W to the width of the HRF at half of the height and PSU to the post stimulus undershoot. Figure extracted from Poldrack et al., 2011a.

This haemodynamic response can be modeled using a function called the Haemodynamic Response Function (HRF). Figure 1.3 shows the characteristics of the HRF.

To model the specific haemodynamic response of the brain to the tasks performed, the HRF is convolved with the stimulus time-series (Cohen, 1997). This can be done thanks to two properties of the haemodynamic response in function of the neuronal activation: linearity and time invariance. These two properties state that:

- Same scaling factor: the amplitude of the haemodynamic response is proportional to the amplitude of the neuronal response,
- Additivity: the haemodynamic response for a sum of activations is equal to the sum of the response for each independant activation,
- Time invariance: if a stimulus is shifted by t seconds, the haemodynamic response will also be shifted by t seconds.

Canonical HRF In Handwerker et al., 2004, a study of the haemodynamic response shape showed that both time until peak and width of the haemodynamic response varied within-subjects (across regions of the brain) and between-subjects, with a larger inter-subject variability. Choosing an appropriate HRF to model the haemodynamic response is important to capture the shape as best as possible and to ensure a good fit during the GLM. Multiple possibilities exist in terms of modeling with differences in assumptions and

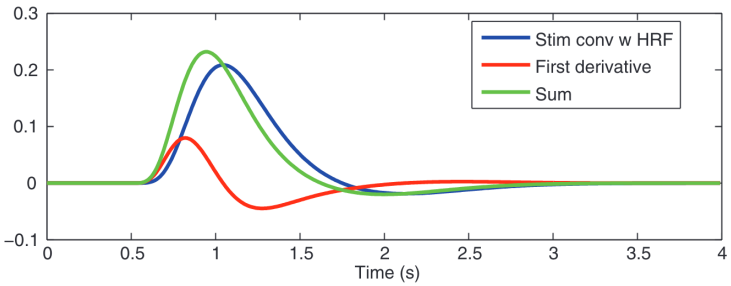

Figure 1.4 – The stimulus convolved with the Haemodynamic Response Function (HRF) (blue), its derivative (red), and the sum of the two (green), illustrating that including a derivative term in your linear model can adjust for small shifts in the timing of the stimulus. Figure extracted from Poldrack et al., 2011a.

model complexity (Lindquist et al., 2009). The optimal shape of the HRF was estimated by Friston et al., 1994; Lange et al., 2002 using deconvolution and found that in general, it could be approximately described by a gamma function. However, a single gamma function does not model the post-stimulus undershoot. Thus, a double-gamma HRF was adopted as the canonical HRF (Friston et al., 1998) (*i.e.* the default one) by multiple researchers, based on the combination of two gamma functions, one modelling the shape of the initial stimulus response and the second the undershoot.

Beyond the canonical HRF When using the canonical HRF to model the response, we are biased to only find responses that are similar to that function. Researchers tried to use more complicated models, that allows more flexibility in the shape of the HRF with more parameters, but this lead to more variability in the estimate. The goal is to find a tradeoff between bias and variance.

To build more flexible HRF, a popular approach is to use a set of HRF basis functions, that will be convolved to the stimulus onset to fit the signal instead of just convolving a single HRF. The most commonly used basis set is the canonical HRF + derivatives (temporal +/- dispersion) (Handwerker et al., 2004). Adding the temporal derivative allows for small offsets in the time to peak of the HRF and the dispersion allows for variations in the width of the HRF. Figure 1.4 shows a standard regressor (stimulus convolved with the canonical HRF), its temporal derivative and the sum of the two. We can see that the addition of the two leads to slight shift to the left and a small increase in peak height.

Another option is to use a Finite Impulse Response model (FIR), in which we do not

give any indication on the shape of the expected HRF. We only make a supposition of the signal length and we decide on a number of points to estimate, which allows for a subject-specific modeling of the HRF and thus, a high variability (Goutte et al., 2000). Between high bias with canonical HRF and high variability with the FIR, researchers built what is known as the "constrained basis sets" (Woolrich et al., 2004). This set of basis functions can be built by first generating a set of reasonable HRF shapes with varying parameters, and then by applying a principal components analysis (PCA) to extract a set of the most representative basis functions. It leads to a more flexible and less biased estimation, but also to more variability when using a high number of representative functions.

1.2.2.3 Building the design matrix

The stimulus time series is convolved with the modeled HRF to obtain the regressors of the design matrix. The choice of these regressors and their position in the design matrix can impact the parameters estimate:

- Depending on the study protocol, one might want to add a parametric modulation to a regressor (see Parametric modulation) or to add the response time as a regressor (see Modelling response time),
- In specific cases, to remove correlation between regressors, one might want to apply orthogonalization (see Orthogonalization),
- To correct for remaining artifacts, one might also want to add nuisance regressors to the model (see Nuisance regressors).

Parametric modulation In complex and specific studies, stimuli can be parametrically varied (*e.g.* contrast of a visual stimulus, volume of an auditory stimulus, etc.) and we can expect that the strength of the neuronal response will reflect these variations. We can thus add an additional parametric regressor to the design matrix, which will model these variations. To create a parametric regressor, the onsets of each stimulus are modified to have a height that reflect the variations. Adding this parametric regressor does not prevent from including an unmodulated regressor, but the height values of the parametric one must be demeaned to avoid any correlation between the modulated and the unmodulated regressor.

Modelling response time During acquisition, participant's response times might be different across participants and trials, causing variations in the neuronal response. In-

deed, longer stimuli lead to a higher haemodynamic response and thus, participants might exhibit a greater activation simply due to the duration of the task, rather than to any difference in neuronal response. To include this response time in the model, two options are possible. First, the regressors of the models can be created using the exact trial duration and not a fixed duration across trials and participants. However, this decreases the sensitivity for responses that are constant across trials. Thus, the second option is preferable and consists in creating a primary regressor with constant duration across trials and including an additional parametric regressor that varies with response time. Effects of response time are thus removed from the model and we can separate the constant effects and the effects that vary with response time.

In practice, incorporating response time regressors in fMRI analysis is strongly recommended to accurately model the relationship between neural activity and the BOLD signal. This approach helps mitigate the response time paradox by accounting for the temporal overlap in hemodynamic responses that can arise for tasks with long time response (Mumford et al., 2024).

Orthogonalization The regressors included in the design matrix are usually correlated to each other, for instance, the time response regressor will be correlated to the primary regressor (see 1.2.2.3). The variability described by two regressors X_1 and X_2 has three components: the one that is unique to X_1, the one that is unique to X_2 and the one who is shared by X_1 and X_2. When regressors are highly correlated, this shared variability is high and the portion of variability explained by each regressor independantly is small. This leads to instabilities of the parameters estimates for these regressors, since the variability of the signal explained by one regressor can easily shift to the other.

A solution to remove the correlation between regressors is called "orthogonalization". It consists in removing the shared variability from one of the two regressor. However, the remaining regressor does not represent the same portion of explained variability anymore and should be interpreted carefully. Moreover, one must choose which regressor to orthogonalize to which, as the portion of variability common to both can be attributed either to the first or to the second regressor. Thus, orthogonalization should be applied only in specific cases where variables are clearly having a supplementary role only (*e.g.* derivatives of the HRF, time response, etc.).

Nuisance regressors Nuisance signals like motion can cause artifacts in the data even after applying correction with realignment (see 1.2.1.2). These signals, including motion estimates, can be added as regressors to the model to reduce error variance and improve detection power. While adding nuisance regressors is strongly recommended in practice, this step should also be taken carefully as these nuisance signals can be correlated with the stimuli and thus, including them in the model might decrease sensitivity. A detailed explanation of the different nuisance signals that can be added and the methods used to compute them was done in Section 1.2.1.7.

1.2.2.4 Hypothesis testing at the first-level

Now that we modeled the signal and estimated the parameters β of the GLM, the brain activity related to each task can be estimated using hypothesis tests. To perform hypothesis testing, we must define a hypothesis H_0, called *null hypothesis*, and we will try to see if the information contained in our data give us enough confidence to reject this hypothesis. Usually, this null hypothesis is about the absence of effect or no difference between two elements. The opposite hypothesis of H_0, is called *alternative hypothesis*, H_1 and consists in the presence of an effect or the presence of a difference.

In our case, we have to define our hypothesis as a contrast that will be tested, this contrast is a vector with length equal to the number of regressors of the GLM and consist in a linear combination of parameters estimates. For instance, if our model was composed of four regressors with associated parameters $[\beta_0, \beta_1, \beta_2, \beta_3]$, the contrast that tests the effect of the first regressor (*i.e.* that tests if the first parameter β_0 is different from 0, $H_0 : \beta_0 = 0$) would be $c = [1, 0, 0, 0]$ since $c\beta = \beta_0$. To test whether two parameters are different from each other or if one is superior to another, the contrast would be for instance $c = [0, -1, 1, 0]$ for $H_0 : \beta_2 = \beta_3$. Since each regressor X_i correspond to a specific task or stimulus, testing if its corresponding estimated parameter β_i is different from 0 is like testing if the task lead to a significant neural response.

For instance, to test a single contrast, we can use a t-statistic, which under the null hypothesis, is distributed as a Student distribution. To test for multiple contrasts at a time, F-tests can also be performed, for instance to test for $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$. After computing the statistic of the test, we can estimate the p-value, corresponding to the probability under the null hypothesis of having a test statistic larger than the one actually observed. The analysis of these p-values is described after in Section 1.2.4.

The main outputs of first-level analyses are the following:

— A matrix with the different parameters estimates for each regressor at each voxel,

— 3-dimensional statistic maps corresponding to the results of the hypothesis tests: for each contrast tested, a 3-dimensional volume of the brain is output with each voxels value corresponding to the statistic of the test performed,

— 3-dimensional contrast maps corresponding to the results of the hypothesis tests in terms of percent BOLD change: for each contrast tested, a 3-dimensional volume of the brain is output with each voxels value corresponding to the combination of estimated parameters in the tested contrast.

— 3-dimensional variance contrast maps corresponding to the expected variance of the estimated contrast maps.

Hypothesis testing is also performed at the second-level to test for mean activations inside a group of participants or to compare activations between groups of participants. We will describe this in the following section.

1.2.3 Second-level analysis

During the first-level analysis, estimates of contrasts and variance have been obtained for each voxel for several participants. If the first-level was performed at the run-level, the second-level analysis consist in a single-subject analysis that combines the different run-level contrasts of a participant to obtain subject-level statistic and contrast maps. Typically, when mentioning second-level analyses, we refer to group-level analyses, in which subject-level contrast maps can be combined to test for the mean effect of a regressor within a group or to compare this effect between groups.

1.2.3.1 Hypothesis testing at the second-level

In both cases, as in the first-level, we use a GLM (see 1.2.2.1), with Y corresponding to the list of contrasts maps for the participants of the group (or runs of a participants). At the group-level, additional regressors can be added with informations regarding the participants, these can be quantitative (*e.g.* age) or qualitative (*e.g.* sex or gender). Similarly to the first-level, once the parameters are estimated, we define a contrast, consisting in a linear combination of parameters, and perform hypothesis test. For instance, if we have two groups of participants and want to test for any difference between the two, the contrast would be $c = [-1, 1]$.

1.2.3.2 Modelling variance

At the group level, we must take into account multiple sources of variance: the contrast variance estimated at the first-level (*i.e.* within-subject) and the between-subject variance, that needs to be estimated (Mumford et al., 2006). Two main models exist to model variance: mixed-effects and fixed-effects. In the fixed-effects models, typically used when the second-level analysis combines the different runs for a single participant, only the within-subject variance is taken into account. In such case, a mixed-effects model is impractical due to the limited number of runs per participant, making it difficult to properly estimate the between-run variance. For group-level analyses, the mixed-effects models is used. It assumes that the total variance is composed of both between-subject and within-subject variance. The goal here is to estimate the between-subject variance, while taking into account within-subject variance. This is usually done iteratively by successively computing the between-subject mean and variance while incorporating participants to the model (Worsley et al., 2002). In a simpler case, we can assume that within-subject variance is equal for all participants of the group or that it is negligible compared to between-subject variance (random effect).

1.2.4 Statistical inference

After estimating the contrasts and performing hypothesis testing, the goal is to determine whether or not the detected effect is significant. This is done by applying statistical inference on statistic maps, resulting in 3D thresholded statistic maps. Multiple thresholding methods can be applied on statistic maps to identify significantly activated voxels: at the voxel-level and at the cluster-level.

1.2.4.1 Voxel-wise inference

Statistic maps resulting from the hypothesis tests are composed of voxels with associated statistic values. Thus, it might seems logical that one should test each voxel individually, by comparing their associated value to a threshold, to test if the effect is significant or not. Such method allows to make very specific inferences, in particular on small areas of the brain. At this level, voxels are all analyzed independently and spatial information is not taken into account. This "naive" approach is also known as uncorrected voxel-wise inference, note that in practice a correction for multiple comparison should be applied (see 1.2.4.3).

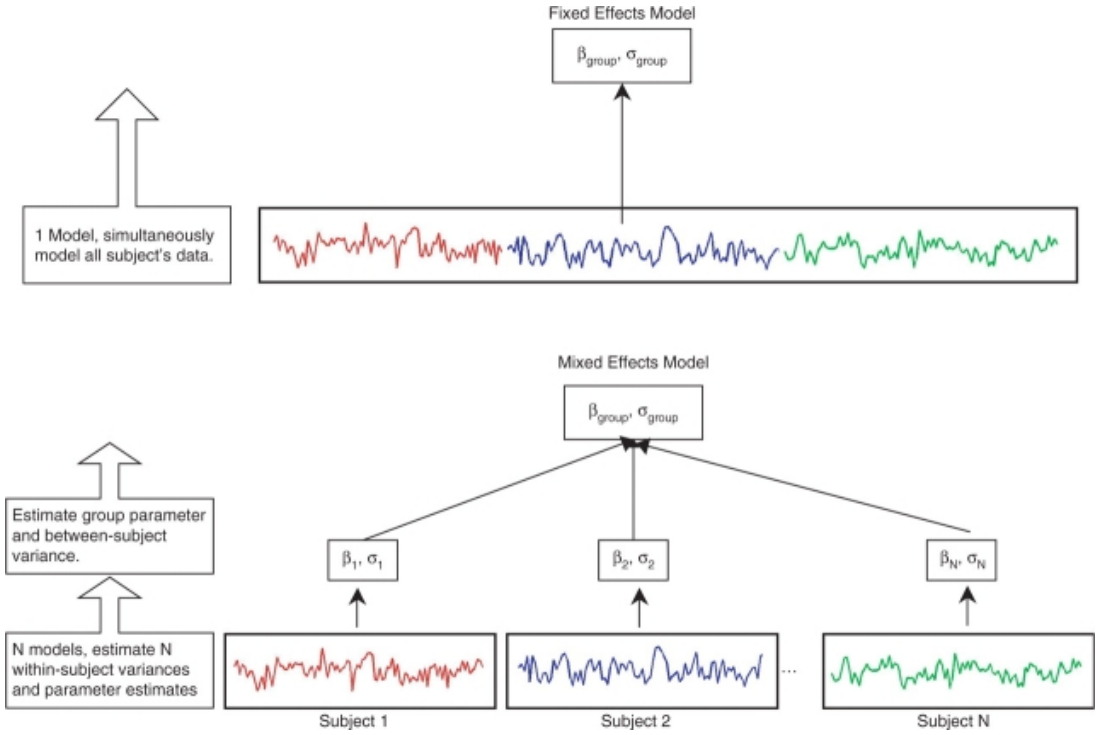

Figure 1.5 – Modelling variance at the second-level. Top panel: fixed effects analysis where all subject's data are combined into a single model with only one source of variability. Bottom panel: two-stage summary statistics mixed model, each subject's time series is first analyzed individually, supplying within-subject parameter estimates and variances, and the second stage uses the first stage parameter estimates and variances and estimates the between-subject variance and group parameter estimate. Extracted from Mumford et al., 2006.

1.2.4.2 Cluster-wise inference

At the cluster-level, we use spatial information in the image, such as the fact that significantly activated voxels might be located in close areas of the brain. Indeed, brain regions activated during the tasks are usually larger than the size of a single voxel (around $2mm^3$) and data are often spatially smoothed during preprocessing, leading to a spreading of the signal across many voxels of the image. Cluster-wise inference is usually done in two steps: first, a cluster-forming threshold is applied to the statistic map, and groups of contiguous voxels above the threshold are defined as clusters. Neighboring voxels must be defined before the thresholding step to decide if 6 (voxels sharing a face), 18 (voxels sharing face + edge) or 26 (voxels sharing face + edge + corner) are taken into account. Then, the size of each cluster is used to determine its significance, by comparing it to a

critical cluster size that must also be defined a priori.

1.2.4.3 Correction for multiple testing

In standard hypothesis tests, we have a control on the level of false positive risk with the appropriate selection of α, usually set to 0.05. However, this is only valid if a single test is performed. In the hypothesis testings performed in fMRI, we test all the voxels of the image simultaneously, which mean that if we set $\alpha = 0.05$, 5% of the voxels of the image will be false positives. This problem is also known as *multiple testing problem* and must be corrected. Two main measures of false positive risk were defined: Family-Wise Error (FWE) and False Discovery Rate (FDR). FWE corresponds to the chance that across all voxels, one or more is a false positive, meaning that if we set $\alpha_{FWE} = 0.05$, on average there will be one or more false positive voxels in the thresholded map 5% of the time. To control for FWE, multiple methods are available:

- **Bonferroni correction**, which consists in defining a threshold $\alpha = \alpha_{FWE}/V$ with V being the number of tests (here, voxels of the image). This correction usually shows highly conservative results as it is optimal when voxels values are independent, which is not the case for fMRI statistic maps. In practice, this correction technique is not commonly used.
- **Random Field Theory**, which takes into account the intrinsic smoothness of the data, *i.e.* the one present in all imaging data and the one applied during preprocessing.
- **Non parametric approaches**, in which no assumption is made about the independence of the data. These approaches make use of the data themselves to estimate the appropriate threshold to use. The most widely used methods are permutation tests and bootstrap.

FWE methods were the first available for researchers, but were criticized due to the few number of results that were left after correction. A more lenient alternative to FWE is the control of the false discovery portion, the fraction of detected voxels that are false positives, through FDR procedures. The FDR corresponds to the chance that voxels identified as significant are false positives, *i.e.* an FDR of 5% means that, among all voxels detected as significant, on average 5% of these are false positives.

Take-home Message

- Functional Magnetic Resonance Imaging is a brain imaging technique in which brain activity is studied during the realisation of predefined tasks. This technique is based on the Bold Oxygen Level Dependent (BOLD) signal.
- After acquisition, raw data are processed and analyzed using a sequence of steps called a "pipeline".
- These pipelines are composed of multiple steps that aim to clean and prepare data for further analysis, and to identify changes in the BOLD signal in response to the task.
- At each step of a pipeline, multiple options are available and researchers have to make choices to build their pipeline.

Chapter 2

Analytical variability

In the previous chapter, we presented the main steps of an fMRI analysis pipeline, from raw data to final results. In fMRI data analysis, pipelines are highly flexible (Carp, 2012b), leaving researchers with many choices to make (*e.g.* software package, algorithm, parameters value, etc.), also known as *researchers degrees of freedom*. In the past few years, multiple studies have shown that different choices could have a large impact on the results. This *analytical variability*, induced by different protocols and methods applied on the data, lead to a multiplicity of possible results for a given study, called a *vibration of effects* (Ioannidis, 2005).

Other sources of variability exist in neuroimaging studies, for instance across participants (inter-individual variability) or acquisition parameters (technical variability). In the following chapter, we will first describe the different sources of variability that can be observed in neuroimaging studies. Then, we will focus on analytical variability with a description of the variations in the analytical protocol that can lead to different results, the main studies that explored this topic and the challenges related to analytical variability.

2.1 Different sources of variability

To build a neuroimaging study and in particular, an fMRI study, researchers have many choices to make, from the definition of the study to the analysis of the results. At each step, different sources of variability must be taken into account. Figure 2.1 illustrates these different sources of variability: inter-individual variability (between participants), intra-individual variability (longitudinal comparison or test-retest variability), technical variability (relative to differences during acquisition) and analytical variability (relative to data processing and analysis).

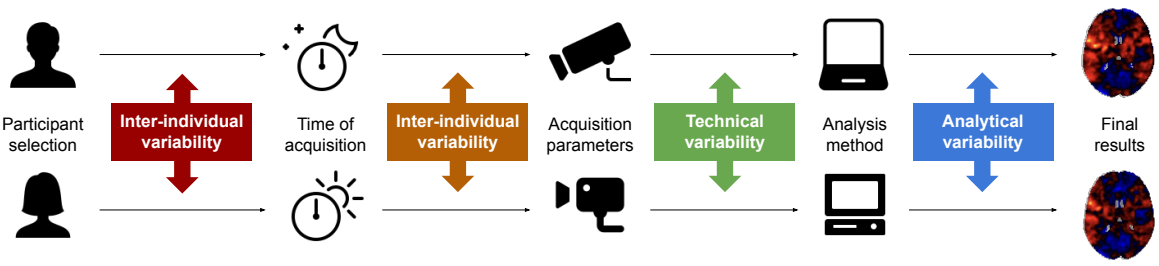

Figure 2.1 – Different sources of variability in neuroimaging studies

2.1.1 Intra-individual variability

For a single participant, variability in the results can arise when repeating the analysis with the exact same protocol (same acquisition and instrument, same method) (Chen et al., 2016). This type of variability, also know as *intra-individual variability*, relates to changes within a participant across time. Researchers usually assess the effect of this type of variability by measuring intra-class correlations (Weir, 2005) between measures obtained from a single participant at different time points. This is of particular importance for disease biomarkers identification as it might be difficult to detect true longitudinal experimental effects if intra-individual variability is large. This variability can also put into question the reliability of fMRI studies due to the amount of uncertainty between two supposedly similar measurements (Noble et al., 2019; Elliott et al., 2020). Aron et al., 2006 explored the long-term test-retest reliability of fMRI-based measurements, showing their potential as biomarkers for brain development and neurodegenerative diseases.

2.1.2 Inter-individual variability

Like with fingerprints, each brain is different (Valizadeh et al., 2018). In particular, environmental and genetic factors shape the brain structures and functions. Thus, the results obtained when analyzing the data from two participants using the same methods in an fMRI study can be really different. This phenomenon is known as *inter-individual variability* and was widely studied in the literature. In 1.2.1, we saw that several preprocessing and analysis steps applied to fMRI raw data are used to mitigate and take into account inter-individual variability. First, participants have differences regarding brain morphology (Rademacher et al., 1993; Thompson et al., 1996). A standardization step (see 1.2.1.4) is included in most neuroimaging pipelines to be able to compare participants

and to combine them in group-level analyses. Participants can also be different in terms of functional activation (Lebreton et al., 2019). This variability is taken into account at the second-level using mixed-effect modelling (see 1.2.3.2).

2.1.3 Technical variability

In neuroimaging studies, finding a sufficient number of participants might be difficult, in particular for rare pathologies. To tackle this issue and increase sample sizes, multicenter studies started to develop. However, these studies are subject to another type of variability due to the use of several (and different) acquisition sites. Multiple studies have shown that differences in MRI intensities (*i.e.* voxel values in raw data) between scanning parameters can be larger than the biological differences observed in these images (Wittens et al., 2021; Mackin et al., 2015). This led researchers to explore the role of various factors to explain the impact of this variability in the results, for instance, how different acquisition could change the smoothness of the image (Friedman et al., 2006). They also developed new methods to reduce these differences, and thus enhance multicenter reproducibility (Fortin et al., 2016).

2.1.4 Analytical variability

As stated at the beginning of this chapter, the exact choice of protocols and methods applied on the data can have a non-neglectable impact on the results. This phenomenon, also known as *analytical variability*, can be induced by different levels of variations including: different software environments, different software packages, different sets of parameters, different algorithms, etc. Compared to other sources of variability, this one is less understood and there is no established method to correct for it.

2.2 Focus: Analytical variability

In the following section, we will focus on analytical variability. We will show how specific choices in pipeline definition can lead to variations in the results. Then, we explain why analytical variability is of particular importance in neuroimaging and we present the main studies that tried to assess and mitigate it. Finally, we describe the remaining challenges regarding analytical variability and the ones we tackle in this manuscript.

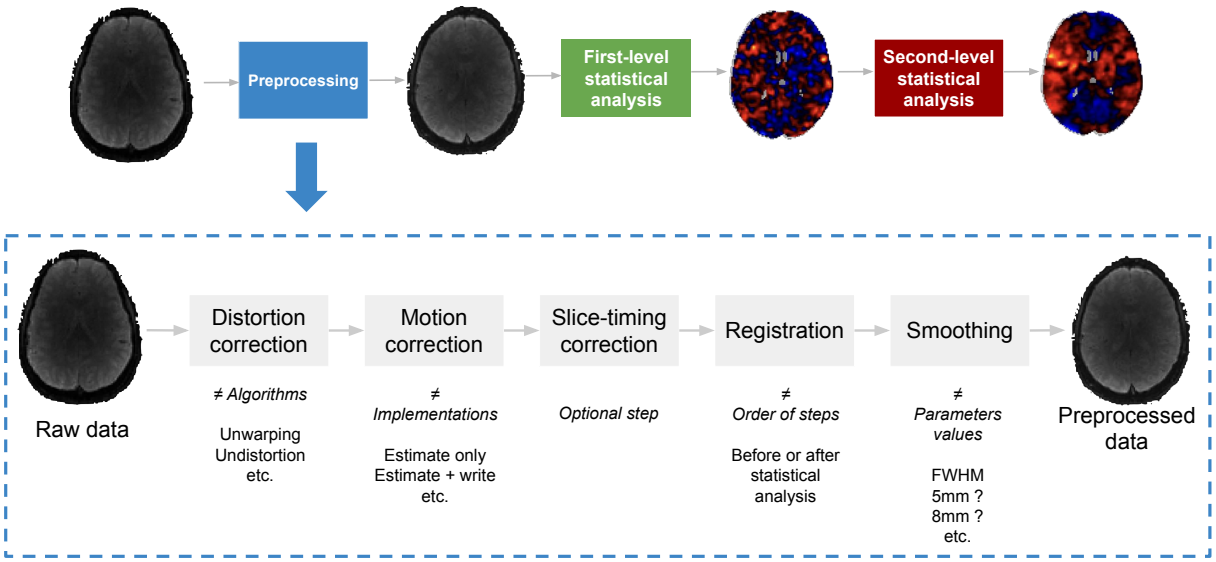

Figure 2.2 – Overview of possible choices to make during a standard fMRI preprocessing pipeline.

2.2.1 A large analytical space

In the previous chapter (see 1.2), we detailed the different steps that may - or must - be included in a standard fMRI pipeline. Analytical variability includes the variations in the results that arise when deciding to perform or not a processing step, when changing the order of operation, or even when modifying the value of a parameter. It also includes the variability in the results induced by different computing conditions such as the operating system and its version. Figure 2.2 illustrates several options, from which the researcher has to choose, during a standard preprocessing.

We distinguish three main types of variations:

— **Parameters variations**, which arise from changes in the choice of algorithm to use, the values of parameters or the order of operations.

— **Software variations**, which arise from the different implementations of a pipeline between different software packages.

— **Variations in computing conditions**, which arise from changes in computing environment.

During the analysis, researchers can modify their pipeline in different ways. The first possible choice is whether to include or not some processing steps in the pipeline. For instance, the use or not of slice-timing correction in fMRI pipelines is still a debated

topic, and depends on other characteristics of the study (Parker et al., 2019), such as the repetition time (see 1.2.1.3). Researchers can also choose to change the type of algorithm to use. As an example, if they want to perform distortion correction, they have the choice between using field-map based or non field-map based techniques, depending on its preferences, but also on the presence of field map in the dataset. Parameters can also be modified inside an algorithm. For instance, smoothing is usually applied by convolving images with a Gaussian kernel, but it can be applied with different intensity levels, defined by the FWHM of the kernel. There is no best practice regarding the right smoothing kernel to use, studies have shown that the decision of using a large or a small kernel size should be taken based on other study parameters (in particular due to its interaction with statistical inference (Hayasaka et al., 2003)). Here, we provided examples on variations of preprocessing steps but choices also have to be made during statistical analysis. These include the choice of HRF (see 1.2.2.2) between classical or double gamma functions, as well as Finite Impulse Response Models or Constrained Basis Sets. The design matrix can also be customized (see 1.2.2.3) to add nuisance regressors or to use HRF derivatives.

In practice, researchers usually do not make all these choices. They use software packages that implement a default pipeline, with only minimal user input required. Multiple software packages were developed to analyze fMRI data, the three most used being SPM (Penny et al., 2011), FSL (Jenkinson et al., 2012) and AFNI (Cox, 1996), which represented 80% of the published studies in 2012 (Carp, 2012b). Note that other software packages were developed since, and are now widely used in the community, for instance fMRIprep (Esteban et al., 2019). In these software packages, the default pipelines usually implement similar steps, but are built differently from one software to another. The main difference between SPM and FSL default pipelines is the order of operation, in particular for the registration. In FSL, registration parameters are computed during preprocessing, but only applied after first-level statistical analysis, on contrast and statistic maps directly. In SPM, these parameters are computed and applied during preprocessing, and statistical analysis is thus performed on standardized data. Some pipeline steps can be modified to align standard pipelines between software packages, but some remain very specific to a software (*e.g.* percent BOLD change estimation). Software packages can also be implemented in different programming language (*e.g.* Matlab for SPM, Python, C and other programming languages for FSL). Each programming language comes with a set of predefined functions, with differences that can impact the results.

Inside each software package, a well-known issue also relates to changes in software

version. During the development process of a software package, new versions are issued regularly, fixing known bugs and improving existing tools and/or adding new ones. These changes are usually reported, but can lead to modifications of a pipeline implementation, and thus to variations in the results obtained between two software versions. Another related question is whether differences in the results may arise due to different releases of the operating system (OS). This phenomenon can be related to differences in the way different systems handle floating point values (Glatard et al., 2015). This usually induces variability in the results of each step, accumulating towards the whole pipeline.

2.2.2 Effect of analytical variability at different levels

We showed that variations in the pipeline can arise at different levels: inside a pipeline, between pipeline implementations and at a lower-level with variability between computing environments. Here, we present the main studies that have shown the impact of such variations in neuroimaging results and their conclusions. These studies explore different types of neuroimaging data and analyses, not only task-based fMRI, and differences induces by analytical variability are observed across modalities.

2.2.2.1 Exploring analytical variability

In task-based fMRI, there is usually no ground-truth to evaluate the behavior of a pipeline (*i.e.* if the pipeline behaves correctly or not). This is also known as the "oracle problem" in software engineering and several approaches were developed to test it (Barr et al., 2015). Often, multiple comparable pipelines are run and results are compared to identify the most impacting parameters and potential discrepancies.

Assessing the impact of pipeline variations in neuroimaging results allows researchers to better visualize the effect of different choices, and guide them to build their pipeline. In practice, the main goal is usually to optimize the pipeline with metrics closely linked to the research and diagnostic questions addressed at the end of the pipeline (Strother et al., 2004). In several studies, ground-truth values were used to benchmark pipeline results and select the most suited one for the study at hand (Klein et al., 2009; Dafflon et al., 2022). In other cases, reproducibility metrics were used to assess the performance of the pipeline (LaConte et al., 2003), with for instance the NPAIRS framework developed by Strother et al., 2002.

2.2.2.2 Variations in pipeline parameters

Several studies investigated the impact of changes at a specific step by comparing the outputs of this step. For instance, Klein et al., 2009 evaluated the performance of fourteen nonlinear deformation algorithms for brain MRI registration and found substantial variations in the outputs of this step, but also on the accuracy (*i.e.* the final performance) of the methods. In particular, they found a correlation between the number of degrees of freedom of the deformation and registration accuracy.

Usually, these studies were also exploring how a single change could impact the final results of the analysis (Oakes et al., 2005; Nørgaard et al., 2020; Bhagwat et al., 2021; Carp, 2012a). In Oakes et al., 2005, different motion correction algorithms were compared in the context of task-fMRI. The goal was to see if performance of algorithms, quantifiable using chosen metrics, were different and if these could be related to differences in the final results. In the end, they found that the performance of the different methods could not predict any difference in final results. Bhagwat et al., 2021 explored variations in cortical surface analyses using different parcellation methods and showed that these variations had a large impact on the results of several tasks, such as age prediction or statistical analysis using a GLM.

The multiplicity of options at each step result in a very important number of potential pipelines, with multiple variations from one to another. Nørgaard et al., 2020 explored preprocessing in general and computed different pipeline variations to process PET-scan data. One of the largest study exploring the variability in the results obtained from different pipelines is the one by Carp, 2012a. In this study, authors estimated the variability of fMRI methods across ten preprocessing and model estimation steps. For each step, he proposed two or more variations, yielding 6,912 individual combination of parameters. He showed that there were large method-related variations in the results regarding activation strength, location and extent. Some results were shown to be stable across different analytical conditions, mostly the quantitative ones, but others like the size and localization of the activation peak were highly unstable.

2.2.2.3 Variations in software packages

Each software package implement different algorithms or have different default parameters values. In Bowring et al., 2019, authors explored the results of the three main fMRI software packages: SPM, FSL and AFNI. Across the three studies analyzed, variations

were found in the results obtained with the different software packages in terms of size and shape of detected clusters. In a follow-up study, Bowring et al., 2022 tried to identify which stage of the pipelines were producing the greater variations. They found that variations were mostly related to changes during the first-level statistical analysis, but the exact largest source of variation was different between studies. Other studies explored the impact of software package during specific steps such as MRI segmentation (Palumbo et al., 2019), parcellation (Bhagwat et al., 2021) or for a full preprocessing (Li et al., 2021; Kharabian Masouleh et al., 2020).

These studies usually use a fixed set of pipelines with predefined variations, leading to a constrained pipeline space. To explore the pipeline space from which researchers actually chose their pipeline from, Botvinik-Nezer et al., 2020 built a many-analyst study. They provided the same fMRI dataset to 70 research teams and tasked them to analyze it using their usual processing pipeline. Research teams had to answer to 9 binary hypotheses and they had provide the corresponded unthresholded and thresholded statistic maps. In the end, there were no identical pipelines across the different teams and results showed substantial variations. Distances between statistic maps revealed some clusters of pipelines that were giving similar results, but others were highly different, even leading to contradictory answers to binary hypotheses.

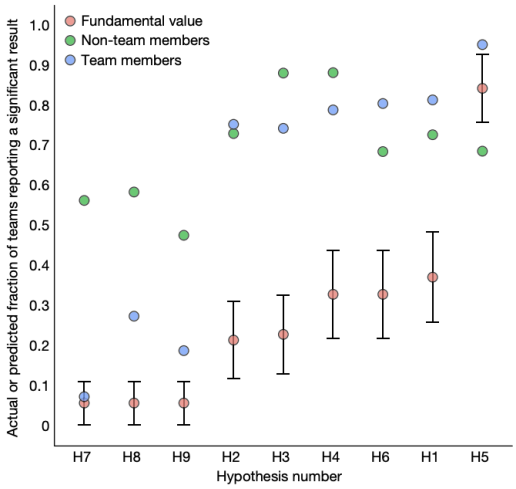

Figure 2.3 – Fraction of teams reporting a significant result during the many-analyst study for each binary hypothesis. Extracted from Botvinik-Nezer et al., 2020. Consortium was reach for H7, H8, H9 and H5.

2.2.2.4 Variations in technical conditions

Several studies explored the impact of computing variability, related to changes in software package versions or in operating systems and their versions. Gronenschild et al., 2012 showed the effect of using different versions of FreeSurfer (Fischl, 2012), different workstation types and operating system versions. They found significant differences in the results between different versions of the software package, in particular between version 5.0 and earlier versions. This suppose that a major change happened between this version, leading to large variations in the results. In another study, Glatard et al., 2015 quantified the differences between results of pipelines computed on different computing platforms. Differences were found to be related to single-precision floating point arithmetic used in certain algorithms and whose implementation evolve between different operating systems and their versions. At a single step, these variations have a small impact on the results, but their accumulation across the high number of steps of a pipeline lead to sometimes large changes in the results.

2.2.3 Challenges related to analytical variability

We showed that 1) to build neuroimaging pipelines, and in particular fMRI pipelines, researchers have access to a broad range of experimental design and data analytic strategies, and 2) these different strategies yield different results. At first, researchers explored and measured the flexibility of research outcomes across analytical conditions. In a second time, they tried to find some solutions to the challenges related to analytical variability. In this section, we will explore these challenges: what to do with analytical variability, how to deal with it when building a pipeline and how it impacts the validity of research findings. In the end, we will expose some open questions and drive towards some of my contributions to these questions.

2.2.3.1 How to interprete this variability?

As shown in 2.2.2.4, variations in low-level features like floating point arithmetic can change the results of a pipeline. While the impact of these variations seem small compared to those induced by different software packages or parameters, it actually points the lack of robustness of the original results. To test the robustness of a pipeline, Kiar et al., 2021 proposed a method in which small variations are added in the input data, and at different steps of the analysis. If these small perturbations lead to large variations in the results,

then one could consider that the pipeline is not robust.

In other circumstances, observing variations (or not) in the results due to changes in analytical conditions can inform on the research question. In Botvinik-Nezer et al., 2020; Carp, 2012a, while there were some strong variations regarding certain aspects of the results, others showed relative consistency, providing confidence that these conclusions are not tied to a specific analytic approach. Such consensus can be obtained using specific analyses, called *multiverse analyses* (Steegen et al., 2016). These allow to systematically explore and integrate pipeline variation on the results. Researchers (and readers of the study) can thus have an idea of how much the conclusions change because of arbitrary choices and which choices have the largest impact on the results.

2.2.3.2 How to take this variability into account when choosing a pipeline?

To limit the impact of analytical variability, researchers tried to optimize their pipelines to improve the quality of the results. While this was supposed to limit the number of options and to reduce the effect of analytical variability, it also led to new processing possibilities. Multiple challenges appear when considering the optimization of pipelines as a way to reduce analytical variability. As described in 1.2 and in 2.2.1, it is not yet clear whether some choices for an analytical step would be better than others. There is no ground-truth to benchmark pipeline results and to assess the superiority of a method compared to another. Moreover, the optimal processing choices may vary depending on the dataset and the analysis, *e.g.* slice-timing correction is more useful for studies with large TR acquisitions. Some studies still proposed solutions to identify optimal pipelines with respect to a predefined criterion (*e.g.* predicting brain age (Dafflon et al., 2022), segmentation tasks (Vanderbecq et al., 2020)).

A large number of possible choices can also be necessary when building a pipeline. Each step of an analytic pipeline is the implementation of a method that comes with some assumptions. For example, during statistical analysis, the GLM comes with the assumption that regressors are independant, which might not be the case when using an additional regressor for trial response time (Mumford et al., 2024). Sometimes, there are no consequences to these assumption violations, but these can sometimes lead to failure of the method, and thus invalidity of the results (Eklund et al., 2016). Pipelines are composed of multiple steps, each characterized by their own assumptions, leading to a pyramid of assumptions. Assessing all of these might be very difficult, but can also help researchers to make appropriate choices between methods for which the does not break

the assumptions and methods that are robust to assumption violations (Mumford et al., 2009).

Usually, large neuroimaging consortia like the Human Connectome Project (Van Essen et al., 2013) or the UK Biobank (Sudlow et al., 2015) develop their own preprocessing pipelines. This allows researchers that want to use these datasets to apply these pipelines and to minimize the potential analytical variability related to studies involving these datasets. However, since these pipelines are optimized for particular data acquisition protocols, they might not be applicable to other datasets. As a proposition to solve this issue, Esteban et al., 2019 developed fMRIprep, a preprocessing pipeline for task-based and resting-state fMRI data that is robust to idiosyncracies in the dataset and that requires minimal inputs from the user.

2.2.3.3 How does it impact the validity of research findings?

A direct consequence of analytical variability is the risk of analytical flexibility. Ioannidis, 2005 showed with a mathematical model of bias in scientific studies that the number of false positives in published research findings rises with the flexibility of research results. In practice, when performing their analysis, researchers commonly explore multiple valid analytic alternatives, but often report their results relative only to a single pipeline (or to a few set of variants). This selective reporting can result in an increase of false positive findings (Ioannidis, 2008a; Simmons et al., 2011; Gelman et al., 2019). Some of the solutions exposed in the above sections, such as the use of multiverse analyses or of a standard pipeline can help to reduce this effect.

2.2.3.4 Open questions

While some solutions were proposed to tackle and mitigate the effect of analytical variability, some questions remain open.

Reusing data Over the past few years, concerns have been raised regarding the lack of reproducibility of neuroimaging findings (Button et al., 2013; Poldrack et al., 2017; Botvinik-Nezer et al., 2023). In particular, the low statistical power of studies was criticised, as effectively leading to low probabilities of identifying true effects but also to high probabilities of reporting false positive findings in the literature (Button et al., 2013). Researchers proposed different approaches to increase sample sizes, and thus statistical power, for instance with the development of large-scale studies (Sudlow et al., 2015; Van

Essen et al., 2013). However, acquiring such amount of data is costly and due to the challenge of finding participants, these studies often contain a few number of data per participant. In fMRI, these datasets usually cover a limited subset of brain functions, limiting the flexibility of research questions to explore. A potential solution to increase sample size while avoiding these challenges, is to re-use the data already acquired in other studies into meta- or mega-analyses (Salimi-Khorshidi et al., 2009; Costafreda, 2009).

With the emergence of the FAIR principles (Wilkinson et al., 2016), and the development of standards for sharing brain imaging data, the process of sharing data became easier and now, more and more established in the community. Data sharing platforms (Markiewicz et al., 2021; Gorgolewski et al., 2015) were developed to facilitate the re-use of raw data but also of derived data. These can be used to increase sample sizes of studies, in meta- and mega-analyses (Costafreda, 2009), or to train more powerful machine learning models. In neuroimaging, derived data coming from different studies can be impacted by the many sources of variability arising during the experiment. This put into question the validity of experiments performed with data coming from different sources (*e.g.* derived data obtained with different pipelines in mega-analyses (Rolland et al., 2022)), but also the generalizability of the results obtained from one study to another (Sun et al., 2022).

While it has been shown that adding more variability to the data would lead to more reproducible and generalizable results (Tang et al., 2021; Raviv et al., 2022), the practical application of this paradigm is not always straightforward. In a recent thesis, Rolland, 2022 proposed a method to correct for differences in processing pipelines to perform more valid mega-analyses. However, this method was limited to situations where the proportions of data processed with each pipeline within each group was reasonable (limited to 70/30% or 80/20%). Moreover, labeled databases are not always available in neuroimaging, and if they are, the unconstrained annotations and the heterogeneity of tasks and studies make them difficult to use to train supervised machine learning models.

In the second part of this manuscript, we will present two practical solutions that can be used to facilitate data re-use in two cases: to increase sample sizes and build larger and valid mega-analyses while using shared derived data from different pipelines and to leverage large unlabeled databases in an agnostic manner and then fine-tune towards a variety of problems. Both methods make use of deep learning for their ability to model complex nonlinear relationships in the data.

Exploring the analytical space The neuroimaging community has now realized different pipelines lead to different results, and that a better understanding of the variability induced by alternative analytical paths is crucial. A systematic investigation of the pipeline space is impractical due to the high number of possible pipelines. To improve our knowledge of the pipeline space, it is necessary to find a way to measure distances between different analysis methods. Such relationship measurements can facilitate the selection of a set of pipeline parameters that are the main drivers of variability in the result space. The definition of this pre-defined set of pipelines to test would improve the quality of the results of a multiverse analysis (Steegen et al., 2016), but also decrease the computational time required for such experiment.

Investigating the pipeline space can also help in understanding the homogeneity (*i.e.*, pipelines that give similar results) but also the heterogeneity (*i.e.* pipelines that have a different behavior) of the pipeline space. Rolland et al., 2022 recently showed the problems arising when combining subject-level results obtained from different pipelines for group-level analyses. As we can suppose that such issue is exacerbated for pipelines presenting more distant results, their identification using dedicated measurements would be a first step to help improving generalizability by increasing sample sizes through data reuse.

Due to the high computational cost of storing and analyzing task-fMRI data, recent studies investigating analytical variability in neuroimaging focused on a restricted number of participants and cognitive tasks. One open question is whether patterns observed across pipelines are stable across different contexts: group of subjects, cognitive paradigm, acquisition parameters, etc. In Chapter 5, we propose a method to combine results from different pipelines by converting them between pipelines using style transfer. Style transfer frameworks aim at learning a mapping between two domains and at applying this mapping to data. If the mapping is different between contexts (*e.g.* different cognitive tasks), a framework trained to transfer statistic maps of a particular paradigm would not be applicable to other statistic maps. Exploring the stability of the relationships between pipeline results is thus of particular importance to assess the potential of our solution, and beyond of any solution that aims at being generalizable across different set of participants or fMRI cognitive tasks.

In the third part of this manuscript, we focus on the exploration of the fMRI analytical space. Our contributions are three-fold, 1/ we propose a new dataset called "HCP multi-pipelines" to explore analytical variability and present two use cases: 2/ a study of pipeline relationships, and whether patterns observed across pipelines are stable across different

contexts (group of subjects, cognitive paradigm, acquisition parameters, etc) and 3/ a study of the validity of analysis combining data from different pipelines as a follow-up of the work of Rolland, 2022.

Take-home Message

- fMRI studies are subject to numerous sources of variability, at the participant-level, or at the study-level.
- In particular, analytical variability is the phenomenon by which variations in the results arise due to changes in pipelines.
- These variations can be induced at different levels: software environment, software packages, parameters-level, etc.
- This analytical variability comes with challenges as it leads to difficulties to interpret these variations, but also when building a pipeline. This also puts into question the validity of research findings.
- While some solutions were developed to limit these challenges, some questions remain open regarding data re-use and relationships between pipelines in the analytical space.

Part II

How to facilitate data re-use with deep representation learning?

Chapter 3

Deep learning for medical imaging

Since the emergence of computer vision in the 1950s, researchers tried to build more and more performing systems for automated medical image analysis. At first, image processing was done with the sequential application of mathematical transforms. Then, researchers started to gather large amounts of data and developed techniques, using pattern recognition or machine learning approaches (Fradkov, 2020). These techniques required to design a feature extractor that would convert the raw data (such as the pixel values of an image) into an appropriate representation (also known as feature vector), that would be given as input to an algorithm for a learning task.

More recently, researchers developed new systems in which computers learn the features that optimally represent the data for the problem at hand, solving the issue of complex and time-consuming feature extraction. This particular machine learning process, also known as "representation learning" (Bengio et al., 2013), consists in the extraction of features that capture the underlying structure or characteristics of the data. Deep learning is a particular type of representation learning, which focuses on learning hierarchical representations of data through the use of deep neural networks with multiple layers. These techniques gained prominence due to their ability to automatically learn complex features from raw data, leading to state-of-the-art performance in various domains such as computer vision (LeCun et al., 2015).

In this chapter, we will first position the concepts of representation learning, machine learning and deep learning in the field of artificial intelligence. We will then explain the main learning techniques and models used in deep learning. We further focus on our main application cases and on the challenges that researchers face when using deep learning techniques in the field of medical imaging. Finally, we explore two applications of deep representation learning, namely transfer learning and style transfer, which contribute to the extraction, adaptation, and manipulation of meaningful representations from data. In Chapters 4 and 5, we will present two studies in which we used these techniques to mitigate the variability of fMRI results, and in particular analytical variability.

3.1 Foundations of deep learning

3.1.1 From artificial intelligence to deep learning

Artificial intelligence is a field of computer science in which machines, *i.e.* computers, simulate human intelligence and use their capabilities to answer complex tasks. These can consist in tasks that are intellectually difficult for human beings but relatively straightforward for computers (*i.e.* easily converted to a list of formal rules) or in tasks that are easy for human being to perform, but complex to describe formally (for instance recognizing elements in an image).

The former category of tasks is usually solved by techniques known as *knowledge-base approaches*, which consist in the encoding of statements in formal language, from which the computer can reason using logical inference rules. For instance, Lenat et al., 1990 developed Cyc, an inference engine based on a database of statements and formal rules that were supposed to accurately describe the world.

The later category of tasks makes use of *machine learning* algorithms, which consist in providing real world data to the model, which will learn patterns in these data to answer a problem at hand. For instance, logistic regression (Berkson, 1944), a simple machine learning algorithm, makes use of logistic functions to predict the probability of a binary outcome.

In machine learning, the data given as input to the algorithms are chosen to best represent the observations of the real world, while being understandable by a machine. For instance, if a model is taught to predict the weather for the next day, the real world data could be represented by measures of the temperature, air pressure or precipitation rates. We refer to such measures as a *representation* of the data, each piece of information included in this representation being known as a *feature*.

While many tasks can be solved by manually designing and extracting the right features from data for a task, then giving these as input to a simple algorithm, for many tasks, the feature extraction strategy to adopt is not straightforward. For such case, an approach called *representation learning* can be used, and consist in using machine learning to discover the most important patterns for the task at hand, but also to find the best representation of the data for this task.

Deep learning (LeCun et al., 2015; Goodfellow et al., 2016) is a particular type of representation learning, which focuses on learning hierarchical representations of data through the use of deep neural networks, *i.e.* networks with multiple layers. Figure 3.1

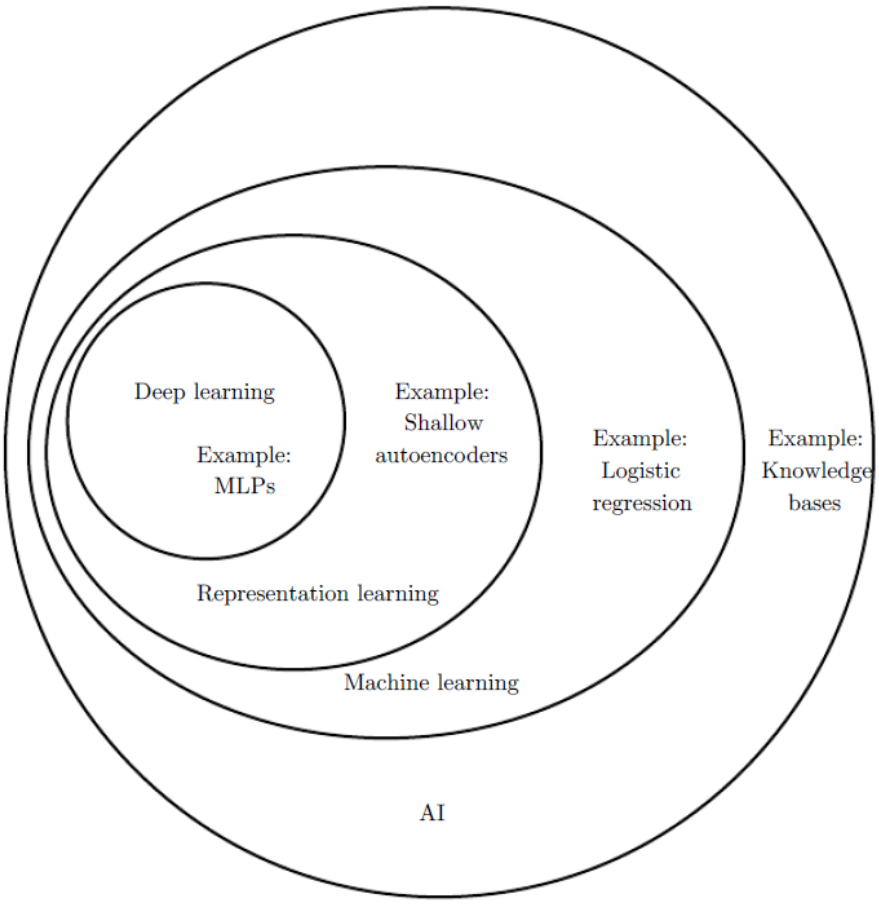

Figure 3.1 – Venn diagram showing how deep learning is a kind of representation learning, which is in turn a kind of machine learning, which is used for many but not all approaches to AI. Extracted from Goodfellow et al., 2016.

illustrates the relationships between these different disciplines from artificial intelligence.

In the following, we will focus on deep learning, and in particular on the deep representations of data that are learned, also known as *deep representation learning*. We will refer to deep learning for the process of learning a deep representation of data. We will describe the main foundations of deep learning and show its potential to extract meaningful representations of data for different applications. We detail the different learning concepts (3.1.2) and the models used to learn deep representations of the data (3.1.3).

3.1.2 Different learning processes

The idea behind machine learning is that algorithms can learn by observing data. This stems from the observation that humans and animals learn from experience, exposure to stimuli, and feedback from the environment. Machine learning algorithms can be divided in several categories depending on the type of experience and feedback they have during the learning process.

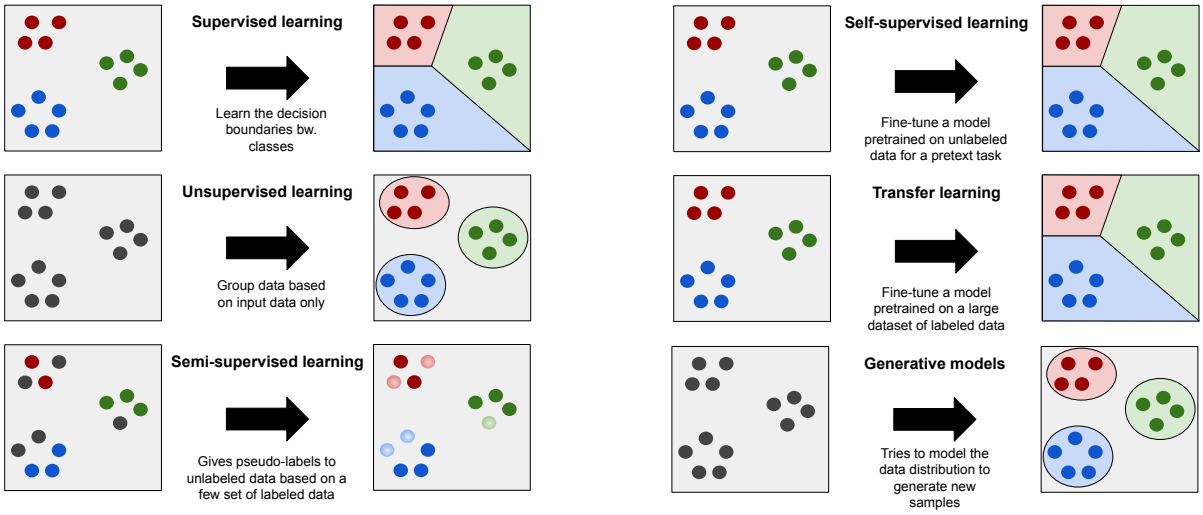

Figure 3.2 – Main learning processes in deep learning

Supervised learning involves training a model on a labeled dataset, where each input data point is associated with a corresponding target output. Representations are learned to answer tasks like classification (assigning input data to predefined categories) or regression (predicting continuous values). **Unsupervised learning** involves training a model on an unlabeled dataset, *i.e.* the model aims to discover patterns, structures, or representations within the data without explicit human guidance. Common tasks include clustering (grouping similar data points together) and dimensionality reduction (reducing the number of features while preserving important information). In both cases, a representation of data is built to answer the task at hand, *i.e.* a representation in which the features associated with the data are the most relevant for the task.

With the difficulty of gathering labeled data and the challenges related to unsupervised learning, methods like **semi-supervised learning** emerged by combining elements of both supervised and unsupervised learning. The model learns from the labeled examples

while also leveraging the additional information present in the unlabeled data to improve performance.

As an attempt to reach supervised learning performance without any labeled data, researchers also proposed **self-supervised learning** (Doersch et al., 2017). In this specific form of representation learning, the model is trained to produce meaningful representations using labeled data whose data have been derived from the data itself without human intervention. For instance, the model can be trained to predict missing parts of an image (image inpainting) or predicting the next word in a sentence given previous words (language modeling). The resulting learned representations can then be transferred to downstream tasks, usually in supervised settings with few labeled data.

Similarly, **transfer learning** (Pan et al., 2010) leverages knowledge learned by pretraining a model on a large-scale dataset and fine-tunes it to a target task with limited data. We will explore this technique in more details in section 3.4.1, and Chapter 4.

To tackle this issue of lack of data, researchers also proposed techniques for data augmentation and in particular, using **generative models**. Generative models learn to generate realistic data samples by capturing the underlying structure and distribution of the training data, enabling them to generate new samples that resemble the original data.

3.1.3 Neural Networks

Perceptrons In deep learning, the extraction of meaningful and hierarchical representations from data is performed by deep neural networks. Introduced by Frank Rosenblatt in the late 1950s, perceptrons (Rosenblatt, 1958) were the initial type of neural networks. However, their inability to process data that are not linearly separable caused a reduction in their use for several years (Minsky et al., 1969). A perceptron consists in a single neuron characterized by parameters W, B, with W indicating the neurons weights and B its biases and a non linear activation function a. Neuron inputs x_i and parameters are linearly combined as a weighted sum and then passed through the activation function (*e.g.* sigmoid, hyperbolic tangent, or softmax functions) to produce the output y.

$$y_i = a(W \cdot x_i + B) \tag{3.1}$$

Multi-Layer Perceptrons Multi-Layer Perceptrons (MLP) (Haykin, 1999) or feedforward neural network is a stack of multiple layers with different numbers of neurons, which are perceptrons. These are composed of at least three layers: an input layer, one or more

hidden layer, and an output layer. Every neuron uses a non-linear activation function, and each neuron in one layer connects with a certain weight W_{ij} to every neuron in the following layer. We call these layers *fully-connected layers*. The learning process of perceptrons works by changing the weights of neurons after having seen a batch of data. Indeed, these parameters decide on how the values of the input vector affect the output. The training process updates these weights and biases so that they can transform the input data to their corresponding target values. Thus, the network learns how to distinguish certain similarities and patterns among the features of the input data. The process used to update the weights is known as backpropagation (Lecun, 1985; LeCun et al., 1989).

Convolutional Neural Networks From the basis of MLP, multiple architectures of neural networks emerged, the most widely used in (medical) image analysis being the Convolutional Neural Network (CNN) (Lecun et al., 1998). A CNN is defined as any neural network that includes at least one convolutional layer. In contrast to fully-connected layers, convolutional layers makes use of a kernel (matrix, smaller than input data) which slides across the input data, performing a dot product with the corresponding part of the data and producing a feature map that highlights specific patterns or features in the input. At each position, a feature map is output and the final output of a convolutional layer is the concatenation of the feature maps at the different positions. The first convolutional layers (*i.e.* lower layers) typically learn basic features like edges or textures (called low-level features), while the highest layers learn more semantic features relevant to the task at hand (called higher-level features). In traditional neural network, the size of the feature maps extracted from the data is descending, meaning that layers are composed of descending numbers of neurons. In CNN, downsampling can be performed by using strided convolution or by incorporing pooling layers, where pixel values of neighborhoods are aggregated using a permutation invariant function, typically the max or mean operation.

The first notable CNN architectures were proposed by Lecun et al., 1998 with LeNet, and by Krizhevsky et al., 2012 with AlexNet. These two are very similar in terms of architecture, with two to five convolutional layers associated with fully-connected layers at the end for classification. After 2012, the trend was to build far deeper models, with the emergence of VGG (Visual Geometry Group) models (Simonyan et al., 2015), like VGG-19 with 19 layers. Nowadays, neural networks are usually composed of a sequence of complex blocks of neurons, called building blocks. These blocks improve the efficiency of

the training procedure and reduce the amount of parameters, with for instance Inception blocks (Inception models) (Szegedy et al., 2015) or Residual blocks (ResNet models) (He et al., 2016).

AutoEncoders In a particular type of neural network, this sequence of downsampling layers is followed by the opposite sequence of upsampling layers, leading to an architecture called AutoEncoder (AE) (see Figure 3.3). These models consists of an encoder network that maps input data to a latent representation and a decoder network that upsample this representation to output data with the same size as the input. These architectures are widely used to learn a lower-dimensional representation of data in unsupervised settings. In traditional AE, the output is a reconstruction of the input and the loss function is the error between the original input and its reconstruction. By doing so, AE learn to capture the most relevant and informative features of the input data in the latent space. This process encourages AE to discover meaningful representations that are useful for reconstructing the input data accurately. To avoid learning identity functions, the latent representation is usually much smaller than the original data dimension and can be learned with other constraints. The loss to minimize can be changed to adapt to other tasks, for instance Variational AutoEncoder (VAE) (Kingma et al., 2022) also minimize a regularization term that encourages the latent space to follow a predefined distribution (typically Gaussian) and then, allows to generate new data by sampling from this latent space.

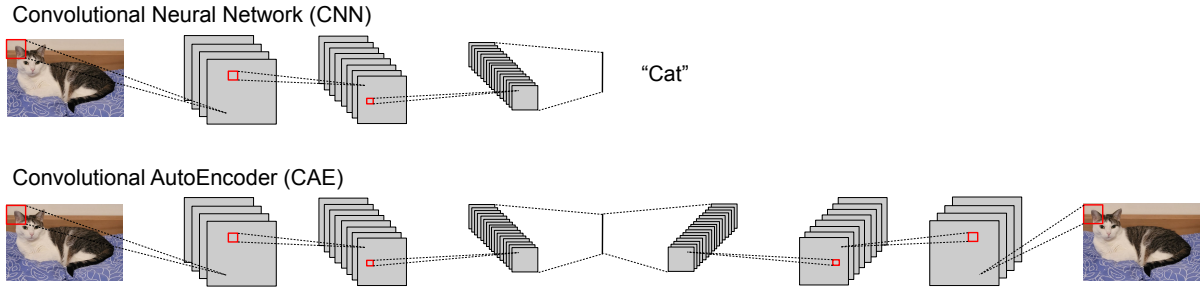

Figure 3.3 – Comparison of architectures between traditional Convolutional Neural Network (CNN) and AutoEncoder (AE).

Variational AutoEncoder (VAE) are representatives of a specific type of models called generative models. These models aim to learn and approximate the distribution of the samples of a dataset to generate new samples. We distinguish multiple categories of

generative models, namely function-based, energy-based and score-based models. Representatives of function-based models are VAE (Kingma et al., 2022) and Generative Adversarial Network (GAN) (Goodfellow et al., 2014). Boltzmann Machines (Hinton et al., 1983), Restricted Boltzmann Machines (Smolensky, 1986) and Deep Belief Networks (Hinton, 2009) are examples of energy-based models, and Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) are score-based models. Here, we will describe the main principles of GAN, which were the state-of-the-art models for medical image synthesis for years, and DDPM, that are their main competitors since their appearance in 2020.

Generative Adversarial Networks GAN (Goodfellow et al., 2014) are composed of two networks that play a two-player minimax game: a generator G that learns to model the data distribution and a discriminator D that learns to distinguish between samples coming from the training data rather than from G. In their original design, generators learn to generate new samples from a random noise variable z, the mapping from z to the data space is then represented by $G(z;\theta_g)$, θ_g being the learnable parameters of G. The distriminator D sees samples generated by G and samples from the training data and is trained to distinguish between the two. Thus, the job of the generator G is to fool the discriminator, for which it will be increasingly difficult to distinguish false images from real ones. Both D and G can be any type of neural networks, and are trained to minimize the following adversarial loss:

$$\min_{\mathrm{G}}\max_{\mathrm{D}} V(D,G) = \mathbb{E}_{x\sim p_{data}}(x)[logD(x)] + \mathbb{E}_{z\sim p_z}(z)[log(1-D(G(z)))]. \tag{3.2}$$

Denoising Diffusion Probabilistic Models More recently, diffusion models (Ho et al., 2020) appeared as new competitors to GAN for image generation. These models work by successively adding noise to the training data, and then learn to reverse the process to construct desired data samples from the noise. In the forward diffusion process, the source image X_0 is subjected to t steps of gradual noise ϵ addition to generate intermediate noisy versions of the image $\{X_0, X_1, ..., X_t\}$. In Ho et al., 2020, the $t-th$ version of the image is expressed as:

$$X_t = \sqrt{\bar{\alpha}_t} * X_0 + \sqrt{1-\bar{\alpha}_t}\cdot\epsilon \quad \text{with} \quad \epsilon \sim N(0,I) \tag{3.3}$$

where α_t corresponds to fixed hyper-parameters between 0 and 1 related to the variance and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_t$.

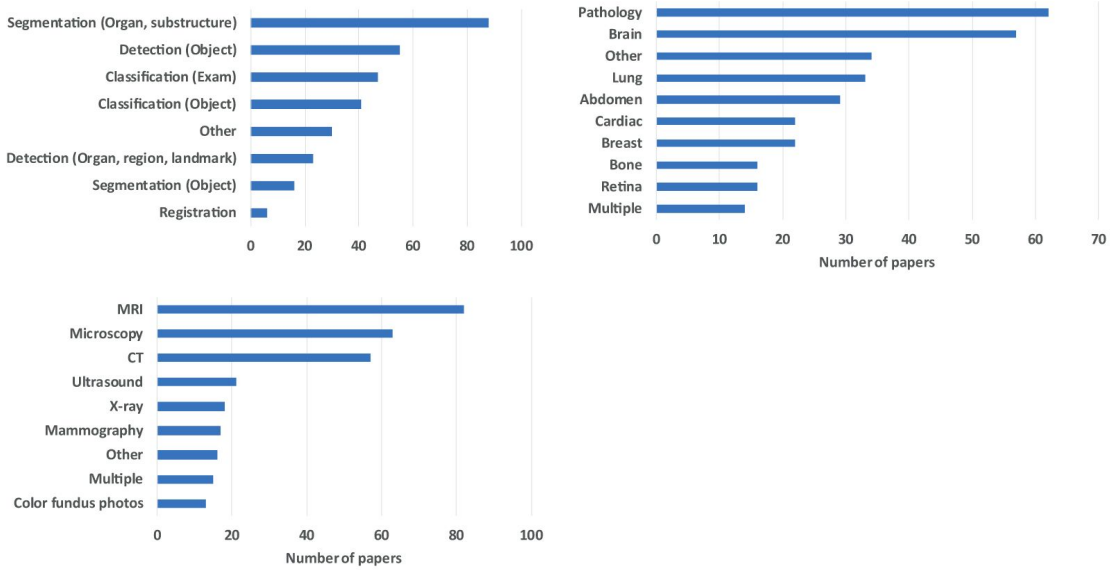

Figure 3.4 – Summary of the papers included in Litjens et al., 2017 in terms of task, modality and organs. Extracted from Litjens et al., 2017.

The reverse diffusion process uses a neural network trained to predict the noise added to the image $\widehat{\epsilon_t} = \epsilon_\theta(X_t, t, C)$ at each time step t given X_t the noisy version of X_0 and t the corresponding time step. Starting from X_t and using the predicted noise, the image X_{t-1} from the previous step can be reconstructed using the following equation and we can reconstruct X_0 by repeating this process for t times:

$$\widehat{X}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \cdot (X_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \cdot \widehat{\epsilon_t}) + \sqrt{1-\bar{\alpha}_{t-1}} \cdot z \quad \text{where} \quad z \sim N(0, I) \tag{3.4}$$

The equation is extracted from Ho et al., 2020. The network $\epsilon_\theta(X_t, t, C)$ is trained using a Mean Squared Error loss, $\mathcal{L}_{MSE} := \mathbb{E}_{X,C,t,\epsilon\sim N(0,1)}[\| \epsilon_t - \widehat{\epsilon_t} \|_2^2]$.

Note that other methods have been proposed since 2020, for example Nichol et al., 2021; Song et al., 2021. In Chapter 5, we use DDPM as described in Ho et al., 2020.

3.2 Deep learning in medical imaging

Deep learning is used to extract meaningful features in medical images in a wide range of applications types and areas. In a survey, Litjens et al., 2017 analyzed 300 papers about deep learning in medical imaging and showed the diversity of medical contexts

where these methods have been integrated successfully (see 3.4).

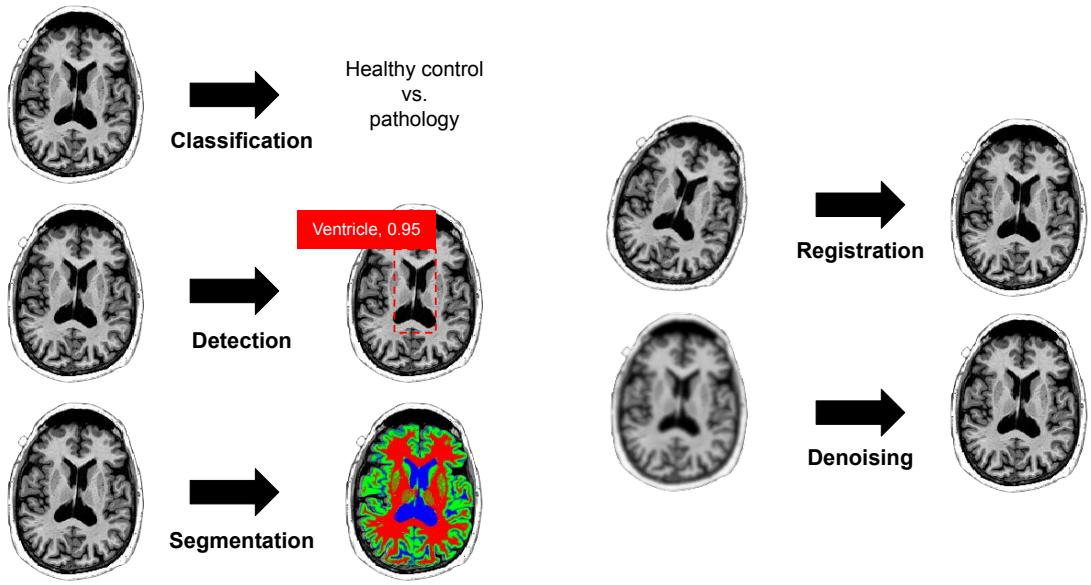

Figure 3.5 – Main applications of deep learning in medical imaging

Classification was one of the first task for which deep learning made a major contribution in medical imaging. This task consist in predicting a categorical variable from features extracted from one or multiple image per individual. These can be used for different purposes, such as disease diagnosis or prediction (Yin et al., 2022). For fMRI data, classification algorithms can also be used for brain decoding (*i.e.* identifying stimuli and cognitive states from brain activities) (Firat et al., 2014). Close to this task, regression tasks consist in predicting a quantitative variable from images, for instance predicting a clinical score (Hou et al., 2016) or a physiological age (*e.g.* brain age models (Baecker et al., 2021) which measures the effects of ageing on the brain).

Another task in which deep learning can be used in medical imaging is object classification, which usually focuses on the classification of a small (previously identified) part of the medical image into two or more classes (*e.g.* nodule classification in chest CT (Shen et al., 2015)). In such task, accurate classification necessitates that the learned representation contains both information on the appearance and localization of lesions. Generic deep learning frameworks often do not support this integration, necessitating the adoption of approaches like multi-stream architectures (Tu et al., 2018).

In image and object classification, objects are identified based on all the pixels of the image. Object detection and segmentation (Wang et al., 2022; Yang et al., 2021) consist,

on the other hand, of the individual classification of each pixel of the image, where the objects in the image are located. In other words, while an object classification will only give an output based on the presence or not of the object, an object detection task will give the position of the object in the input image. In medical imaging, these objects can correspond to anatomical structures (*e.g.* organs) or anomalies (*e.g.* pulmonary nodules).

Image registration (Chen et al., 2023) (also known as spatial alignment) is the process of aligning two or more images based on image appearances. In the medical imaging field, this process seeks to find an optimal spatial transformation that best aligns the underlying anatomical structures of two images. There are two strategies in the literature: using neural networks to estimate a similarity measure for two images to drive an iterative optimization strategy (Yang et al., 2016), and direct prediction of the transformation parameters using deep regression networks (Miao et al., 2016).

Medical images are acquired with several imaging techniques and are thus susceptible to noise and artifacts (Mohd Sagheer et al., 2020). Several types of noise can occur in the image: random noise, white noise characterized by a uniform frequency distribution, or noise that depends on the frequency, usually coming from the acquisition or from image processing techniques. This noise can blur the image, or add artifacts that may lead to difficulties for image analysis. Some types of models, such as denoising AE (Vincent et al., 2010) or GAN (Wang et al., 2023), were built on purpose to learn to reconstruct an image with higher resolution, and with reduced noise.

3.3 Challenges related to medical imaging

Learning efficient deep representations of medical images comes with difficulties due to the particular properties of the data. We refer to challenge to describe these difficulties, which limit the performance of deep learning models in medical imaging. Although the lack of available training data is frequently cited as the primary barrier, it is not the only challenge that may arise in this context. In this section, we describe the main challenges related to deep learning for medical imaging in two categories: the challenges related to the data (3.3.1) and the challenges related to the models (3.3.2).

3.3.1 Challenges related to data

In most medical image analysis competitions, CNN and their derivatives are almost always the top performers, with few differences in performance between different CNN architectures. Today, the most determining factor to achieve the best performance is rather related to the way data are treated and handled (Bengio et al., 2013).

The hope of large datasets One of the major breakthrough in deep learning for natural image processing was the appearance of large labeled datasets, such as ImageNet (Deng et al., 2009) or MNIST (Lecun et al., 1998). These datasets are composed of respectively 3 millions and 70,000 images, with a diversity of classes and images inside each class. Such datasets were built under the paradigm that we must present as many examples as possible during training to instill robustness by learning what is unnecessary, or what represents noise. In opposite, medical imaging datasets are usually smaller (around hundreds or thousands of participants), with few classes and few variations inside each class. These low sample sizes of medical imaging datasets can be related to ethical and privacy constraints, but also to issues related to the cost and difficulties of annotation.

Contrary to natural images, sample size of medical imaging datasets is expressed in number of participants, with sometimes multiple images per participants (*e.g.* different modalities, or time points). These data are usually high-dimensional with sometimes hundreds of thousands of values for 3D data. In relation with the large number of trainable parameters in deep learning models (Cho et al., 2016), this makes it particularly difficult to build fair and generalizable deep learning models for medical imaging (Ricci Lara et al., 2022).

In Willemink et al., 2020, a list of sixteen large medical imaging datasets was shared and sizes were ranging from 267 to 65,000 participants. Even with larger datasets, evidence showed that the increase in dataset sample sizes did not come with better performance of models. Varoquaux et al., 2022 performed a meta-analysis of 478 studies from six reviews on Alzheimer's disease diagnosis or subtypes identification. Figure 3.6 shows the results of this meta-analysis, with a downward trend in performance as sample sizes increase.

A lack of annotations This stagnation of performance with increasing sample size is mostly related to the lack of labeled samples for these large datasets, in particular for complex tasks such as segmentation. Labeling a medical imaging dataset requires some

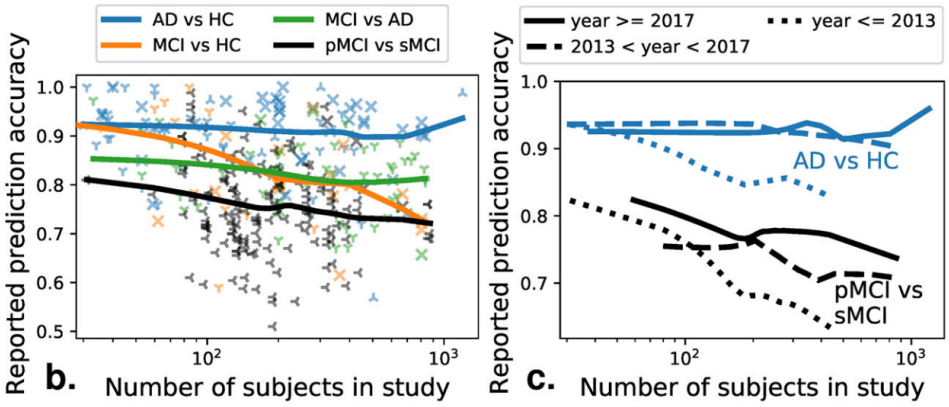

Figure 3.6 – Evolution of reported accuracy for Alzheimer's disease diagnosis and subtype identification in a meta-analysis of 478 studies. Extracted from Varoquaux et al., 2022.

domain expertise, usually from radiologists or pathologists, and is highly time consuming. For example, for segmentation tasks, datasets are often composed of 3D data and require slice-by-slice annotations. Even when datasets are annotated by experts, these experts might disagree on some annotations, leading to *label noise*. This inter-observer variability among experts necessitates to define consensus labels or proper methods of aggregating the labels from multiple experts (Nir et al., 2018; Bridge et al., 2016; Karimi et al., 2020). Another solution is the use of unsupervised or self-supervised methods to limit the need of annotated data (Cheplygina et al., 2019).

Dataset bias and heterogeneity The low diversity of medical imaging datasets, caused by opportunistic data collection, leads to biases and thus, poor generalizability of models (Chekroud et al., 2024). Biases in datasets arise when the distribution of the training data, which is used to build the decision model, differs from the distribution of the test data, where the model is actually employed (Dockès et al., 2021). For instance, such bias has been demonstrated in medical imaging for chest X-ray analysis (Larrazabal et al., 2020), where researchers showed that models trained on data from men participants had a large performance drop when applied on women data. Such issue has also been showed for brain imaging by Wachinger et al., 2021 who showed that simply pooling scans from distinct studies can introduce substantial biases due to differences in sampling strategies, data acquisition, etc. While we usually discuss biases related to population sampling, it must be noted that machine or method related artifacts can also produce biases (Moskal et al., 2022; Li et al., 2023; Korbmacher et al., 2024). Oakden-Rayner et

al., 2020 showed that deep learning models for pneumothorax diagnosis could be biased against images with a chest drain, which is a treatment for pneumothorax. A checklist for bias evaluation on computer vision datasets is presented in Zendel et al., 2017.

3.3.2 Challenges related to models

While most challenges in deep learning for medical imaging are related to data scarcity, the exact model choice and the particular properties of these models are also an issue to the development of deep learning in clinical practice.

Beyond biased data For years, a prevalent belief was that bias in deep learning models reflected unfairness of the dataset, and that the algorithm itself did not contribute to harm. In many cases, these biases are dealt with data augmentation or resampling, while in fact, the overall bias is a caused by interactions between the data and model design choices. We define algorithm bias as the way the model learns underrepresented features in data (Hooker, 2021). For instance, Jiang et al., 2021 showed that underrepresented features, usually more challenging to learn, are learned later in the training process and that the choice of the learning rate and of the training length has an impact on model bias. In another work, Bagdasaryan et al., 2019 showed that differential privacy techniques such as gradient clipping and noise injection could lead to a decrease in performance on certain subsamples of the test population, here, dark-skinned faces.

A need for more robust models Another well-known topic in deep learning is the issue of adversarial attacks. These attacks highlight the vulnerabilities of models by showing how a small change in the inputs can completely alter the outputs, causing the model to confidently answer a problem with wrong conclusions. This issue has been demonstrated for almost all application fields and all types of algorithms, from logistic regression to deep neural network (Biggio et al., 2018). However, such issue is even more complex in medical settings due to the often-competing interests within healthcare, but also the dramatic consequences that a wrong diagnosis or wrong treatment planning made by deep learning models could have. In Finlayson et al., 2019, adversarial attacks were executed against three highly accurate medical image classifiers and were found successful, showing the need for solutions to fight against these attacks.

A lack of interpretability In order to build trustworthy models according to the current guidelines in medical settings, it is essential that models be fair, robust against attacks, but also transparent (Lekadir et al., 2023). This latter is probably one of the biggest criticism that is made to deep learning models, in particular for medical imaging. Deep learning models frequently earn the label of "black-boxes", because their inner workings are not as easily interpreted as those of conventional models such as nearest neighbors algorithms or decision trees. Usually, deep learning models provide an output (*e.g.* "Healthy control") with a probability or confidence strength, whereas information on why and how this decision was make are hidden (Durán et al., 2021). Legally speaking, the European's General Data Protection Regulation (GDPR) law requires that any algorithm utilized for patient care should provides a clear explanation of its decision making process (Temme, 2017). Additionally, the usefulness of a black-box model in healthcare is constrained because it fails to reveal its reasoning, limitations, and biases. Making deep learning models interpretable not only exposes potential errors in the algorithms, but also facilitates the identification of significant details in imaging data that might otherwise remain hidden (Salahuddin et al., 2022).

3.4 Solutions using deep learning

The challenges exposed in the previous section are well-known in the community, and researchers have already proposed some solutions to tackle them. Lots of these solutions try to work around the requirement of large datasets for training, using deep learning techniques that do not necessitate labeled data (*e.g.* unsupervised learning, or self-supervised learning) or that allows to make use of more diverse datasets without privacy constraints (Rehman et al., 2023). Other solutions were proposed to tackle the issues of model related challenges, such as the lack of interpretability or to defend from adversarial attacks, but these will not be discussed here. In this section, we will outline two solutions making use of representation learning that researchers use to overpass the lack of diverse training data: transfer learning and data augmentation using generative models, in particular for image-to-image transition and style transfer. In these two concepts, learned representations are manipulated and used to transfer from one context to another: respectively, to transfer knowledge from one task or domain to another, and to transfer the style of data from one domain to another while preserving the content.

3.4.1 Transfer learning

Transfer learning came as a solution for researchers to overcome data scarcity issues (see 3.3.1). The fundamental motivation for transfer learning was first evoked at *NeurIPS-95*, at a workshop on Learning to Learn, leading researchers to put more and more attention to this field. In 2005, the Broad Agency Announcement (BAA) 05-29 of Defense Advanced Research Projects Agency (DARPA)'s Information Processing Technology Office (IPTO) proposed a new definition for transfer learning: "transfer learning aims to extract the knowledge from one or more source tasks and applies the knowledge to a target task". In this chapter, we use the definitions from Pan et al., 2010.

Definitions

A **domain** $\mathcal{D}$ consists of two components: a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$ with $X = \{x_1, x_2...\} \in \mathcal{X}$.

A **task** $\mathcal{T}$ consists of two components: a label space $\mathcal{Y}$ and an objective predictive function $f(\cdot)$, which can be learned from the training data. Training data consists of pairs $\{x_i, y_i\}$ where $x_i \in X$ and $y_i \in \mathcal{Y}$.

Given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T$ and learning task $\mathcal{T}_T$, **transfer learning** aims to help improve the learning of the target predictive function $f_t(\cdot)$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, with $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$.

These definitions suggest that two domains can be different because feature spaces are different, or because marginal distributions of the feature spaces are different. Two tasks can also be different if their label spaces are different or if the conditional probability distributions are different. Such variations lead to several types of transfer learning, explained in the next section (3.4.1.1). Transfer learning can also be performed using diverse approaches, described in 3.4.1.2. The particularities of transfer learning with neural networks are explained in 3.4.1.3. Finally, in 3.4.1.4, we will expose some studies that made use of transfer learning for deep learning in medical imaging, in particular with fMRI data.

3.4.1.1 Types of transfer learning

Figure 3.7 represents the different types of transfer learning. These are defined based on the presence of labeled data in the source and target domains, but also on the context, *i.e.* whether domains are different, tasks are different or both are different.

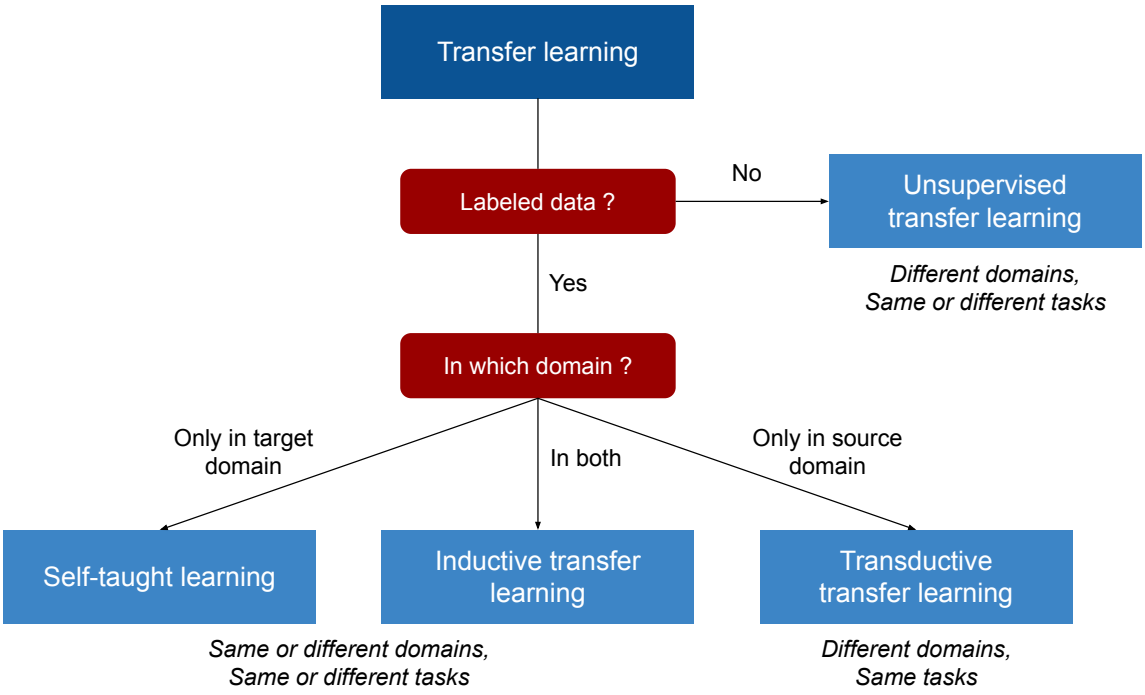

Figure 3.7 – Different types of transfer learning based on the context

Inductive transfer learning In this setting, we should have access to labeled data in both source and target domains. There are multiple cases in which inductive transfer learning could be used.

- First, we could have access to a large source dataset with specific labels (*e.g.* anatomical segmentations) and a smaller target dataset with different labels (*e.g.* lesion segmentations), thus $\mathcal{D}_S = \mathcal{D}_T$ and $\mathcal{T}_S \neq \mathcal{T}_T$. The knowledge learned by training on the source domain for anatomical segmentations could then be transferred to the task of lesion segmentations in the target domain. The supposition is that representations of data learned to segment organs would help for the task of segmenting lesions.

- We could also have the same labels in both domains, but differences between domains (*e.g.* lesion segmentation in T1 MRI and in CT Scan), in such case $\mathcal{D}_S \neq \mathcal{D}_T$ and $\mathcal{T}_S = \mathcal{T}_T$.
- Lastly, both domains and tasks could be different (*e.g.* anatomical segmentation in T1 MRI and lesion segmentation in CT Scan), $\mathcal{D}_S \neq \mathcal{D}_T$ and $\mathcal{T}_S \neq \mathcal{T}_T$. In each case, both datasets have access to labels, whatever the type of labels.

Self-taught learning A subtype of inductive transfer learning concerns the situation where no labels are available in the source dataset. In self-taught learning (Raina et al., 2007), domains can be different or similar. The main point is that we have access to a large unlabeled source dataset (*e.g.* T1 MRI) and a smaller labeled target dataset (*e.g.* lesion segmentation in CT Scan). The knowledge learned on an unsupervised task with T1 MRI could thus be transferred to improve training of the lesion segmentation model in CT Scan.

Transductive transfer learning In the opposite case, we could have access to a large labeled source dataset and a smaller unlabeled target dataset. In this setting, the source and target tasks are the same, while the source and target domains are different. For instance, we could learn a feature mapping from T1 to CT images while optimizing to segment lesion in CT.

Unsupervised transfer learning Finally, when we have no labeled data in both datasets, we could use unsupervised transfer learning. In this setting, the target task can be different from but must be related to the source task, however, it is only possible to solve unsupervised learning tasks in the target domain, such as clustering, dimensionality reduction, and density estimation (Wang et al., 2008).

3.4.1.2 Common approaches to transfer

In Pan et al., 2010, authors define four types of approaches for transfer, at different levels: instance-level, feature-level, parameter-level and relationship-level. In the former, we assume that some parts of the source data can be reused to learn in the target dataset using re-weighting. At the feature-level, the goal is to find a feature representations that would minimize domain divergence and model error. This feature representation can be

built using supervised or unsupervised methods. In the parameter-level approach, we assume that models trained for related tasks share some parameters or prior distributions of hyperparameters. The target task could thus benefit from the parameters learned by the model on the source task. Finally, in some cases where data have a specific representation, such as network data, relationships between data can be used to transfer knowledge between source and target domain. In this context, statistical relational learning techniques are proposed to solve these problems.

3.4.1.3 Particularities of deep learning

The principles of deep learning and the architecture of networks make them highly suitable for two approaches in particular: at the feature-level and at the parameter-level. At the feature-level, neural networks are first trained on the source domain for a source task. The convolutional layers are then extracted and weights and biases are frozen. These layers are then used to extract features of the target dataset and directly input them to another model for the target task. At this level, the lower-level representation of data learned for the source task on the source domain is used for the target tasks, implying that both tasks and domains should be close.

At the parameter-level, models are also trained on the source domain for a source task, layers are also extracted, but not frozen. The weights and biases of these layers are used to initialize another model that will be trained on the target task. This technique is also known as "fine-tuning" and suggests that representations learned during the first training phase are useful for the target task or domain, but remain too different to be used directly.

3.4.1.4 Applications in medical imaging

The potential of transfer learning to deal with data challenges in medical imaging led researchers to a massive use of these techniques. A search on PubMed for transfer learning on medical imaging led to more than 20,000 papers, with an ascending tendency since 2015 [1]. The lack of large public datasets has led to the widespread adoption of transfer learning from ImageNet (Deng et al., 2009), a famous natural images dataset, to improve performance on medical imaging tasks (Bengio, 2012). It might seems surprising that such transfer from natural images to the medical domain gives good performance, due to

1. PubMed was queried on May, 7th 2024.

the large difference between the two domains (Raghu et al., 2019), and thus potentially on data representations learned by neural networks. In a recent paper, Matsoukas et al., 2022 showed that such transfer was more beneficial when the target dataset had a small size and was close to the source dataset (*i.e.* to natural images). These benefits are mostly related to the feature extraction that is similar between the two domains, as also demonstrated in Kim et al., 2022.

In some cases, medical imaging data might be too far from natural images, or might have different properties, making it difficult to take advantage of large natural image datasets. This is the case for fMRI data, which are usually 4-dimensional for raw data, or 3-dimensional when using statistic maps, contrary to natural image datasets which are composed of 2-dimensional images. Moreover, statistic maps are composed of voxels, whose value does not represent pixel intensities between 0 and 255, but statistical values that can take a wider range of value (positive or negative). Transfer learning in this setting might require the use of another dataset, closer to the target data, or some data adaptation to remain close to natural images properties. For instance, Thomas et al., 2023 pretrained two deep learning classifiers on a large, public fMRI dataset of raw data, fine-tuned them and evaluated their performance on another task on the same dataset and on a fully independent dataset. In another study, Y. Gao et al., 2019 used the ImageNet dataset (Deng et al., 2009) to pretrain a model and fine-tuned it to classify 2-dimensional fMRI data. This database was also used in Malik et al., 2022 for pretraining a 2-dimensional structural MRI classifier. In the same paper, the Kinetics dataset (Kay et al., 2017) was also used to evaluate the transfer learning process with 3-dimensional images. In a recent work, Thomas et al., 2022 used self-supervised learning frameworks to pretrain brain decoding models across a broad fMRI dataset, comprising many individuals, experimental domains, and acquisition sites. These studies showed improved classification accuracies as well as quicker learning and less training data required.

3.4.2 Image-to-image transition and style transfer

The term neural style transfer was first employed by Gatys et al., 2016 to define the separation and recombination of the image content and style using neural networks. In the algorithm, features of content and style of images are matched in the convolutional layers of a CNN. Despite the results showed in the paper, the principle of neural style transfer remained unclear. Li et al., 2017 thus theoretically showed that neural style transfer could be seen from a domain adaptation point of view and that matching the Gram matrices

of the features was equivalent to minimizing the Maximum Mean Discrepancy between the features. Such findings were promising, and lead researchers to apply neural style transfer for a wide range of tasks, including domain adaptation (Li et al., 2017) and data augmentation (Zheng et al., 2019). In this section, we will focus on neural style transfer with Image-to-image transition (I2I), as opposed to text-to-image or other forms of style transfer. This technique was found particularly successful in medical imaging to overcome the data scarcity issues, and we will present some examples at the end of the section.

3.4.2.1 Foundations of image-to-image transition

Following the definitions exposed in 3.4.1, the goal of I2I is to convert an input image x_A from a source domain $\mathcal{A}$ to a target domain $\mathcal{B}$ with the intrinsic source content preserved and the extrinsic target style transferred. This means that we need to learn the mapping $G_{A\to B}$ that would generate $x_{AB} \in \mathcal{B}$, with the content of $x_A \in \mathcal{A}$ and the style of $x_B \in \mathcal{B}$.

I2I frameworks can be categorized using several criteria. First, we distinguish supervised and unsupervised I2I. In supervised settings, we have access to paired datasets, meaning that we have a dataset $X_A = \{x_{A_1}, x_{A_2}, ...\} \in \mathcal{A}$ and a dataset $X_B = \{x_{B_1}, x_{B_2}, ...\} \in \mathcal{B}$, with $X_{B_i} = G_{A\to B}(X_{A_i})$. In other words, we should have access to the ground-truth, *i.e.* the exact version of each image of the domain $\mathcal{A}$ in the domain $\mathcal{B}$. In unsupervised settings, we only have access to unpaired datasets, meaning that we have a dataset $X_A = \{x_{A_1}, x_{A_2}, ...\} \in \mathcal{A}$ and a dataset $X_B = \{x_{B_1}, x_{B_2}, ...\} \in \mathcal{B}$, but this time, we do not have any ground-truth, *i.e.* data are not matched and supposedly, there is no equivalent of $x_{AB} \in \mathcal{B}$ in the dataset X_B. Figure 3.8 illustrates the notion of paired and unpaired datasets. Other types of I2I exists, for instance semi-supervised I2I or few-shot I2I, that will not be detailed here.

We also distinguish I2I frameworks according to the fact that only two domains are involved, or multiple ones, *i.e.* two-domains and multi-domains frameworks. In the former, only one transfer is learned at a time, and datasets are only composed of data from two different domains. If we take the example of facial attributes modification, an application of style transfer, this means that each framework will learn to transfer a single facial attribute (*e.g.* hair color, age, etc.).

If datasets are composed of n models, such frameworks would require to learn $n \cdot (n-1)$ models to learn all possible mappings. Such training is highly time consuming and limited since models cannot use the global information available in the whole dataset

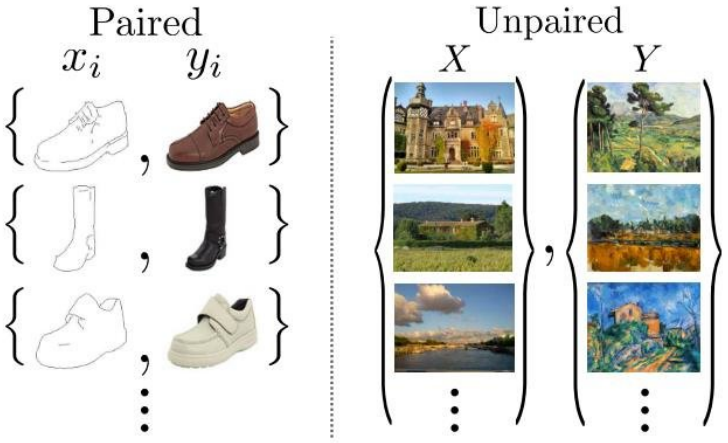

Figure 3.8 – Paired training data (left) consists of training examples x_i, y_i, where the correspondence between x_i and y_i exists (Isola et al., 2017). Unpaired training data (right) consists of a source set x_i and a target set y_j, with no information provided as to which x_i matches which y_j. Figure extracted from Zhu et al., 2017.

and the mappings between other domains. Thus, researchers studied the multi-domains I2I problem. These frameworks are composed of single unified model in which different outputs might contain different style modifications. For the facial attributes, this means that a single model would be able to learn to transfer both hair color and age. This is done in practice by encoding a precise query in the model, for instance using conditioning.

3.4.2.2 Models and architectures

In computer vision, recent advances gave rise to performing deep generative models such as GAN (Goodfellow et al., 2014) and DDPM (Ho et al., 2020). These models produce high quality results for generating new images from a known distribution, and in the task of I2I using their conditional versions (Isola et al., 2017; Saharia et al., 2022). Architectures and learning strategies of GAN and DDPM were described in 3.1.3. Here, we present their conditional versions and expose several frameworks developed for I2I.

Conditional Generative Adversarial Networks. Conditional versions of GAN (Mirza et al., 2014) - conditional Generative Adversarial Network (cGAN) - can be constructed by inputting the data, y, we wish to condition the generation on, to both the generator G and the discriminator D. The condition y can be any kind of information, from class labels to images, and is usually set as input to the networks by concatenation with the

data. In such case, the adversarial loss remain similar to Equation 3.2, with both D and G conditioned on y:

$$\min_{G}\max_{D} V(D,G) = \mathbb{E}_{x\sim p_{data}}(x)[logD(x|y)] + \mathbb{E}_{z\sim p_z}(z)[log(1-D(G(z|y)))]. \quad (3.5)$$

In Odena et al., 2017, conditioning is further improved by training the discriminator to differentiate between real and fake images, but also to correctly classify real and fake image in the target class label. The adversarial loss is thus combined with another non-adversarial losses, dedicated to classification.

These models were first developed to generate new samples that follow the training distribution, in particular for data augmentation. The sampling process (*i.e.* the generation of new images after training) starts with the initialization of a random vector z from which images are sampled. The cGAN provided an opportunity to perform conditional image generation, but the absence of conditioning of the (noise) input variable z prevent them to directly perform style transfer. In I2I, one starts from a source image x_A that is given as input to the framework and modified using conditioning. In the following, we will describe several frameworks developed for I2I:

Age-cGAN Antipov et al., 2017 created Age-cGAN, a conditional GAN coupled with an encoder to approximate an initial latent vector that would preserve the person's identity. This allows to conditionally generate images by constraining on a target age y_{target} and to perform face aging on a specific image using the approximate latent vector.

Pix2Pix Isola et al., 2017 introduced Pix2Pix, a framework to tackle supervised two-domains I2I problems. The generator receives as input an image from the input domain $\mathcal{A}$ and learns to convert it to the target domain $\mathcal{B}$ by minimizing a reconstruction error (Mean Squared Error - MSE or Mean Absolute Error - MAE loss) in addition to the adversarial loss. The discriminator learns to differentiate between the fake output $G(x_A)$ and the desired ground truth output image x_B.

CycleGAN Due to the difficulty of building paired datasets for training, researchers developed methods to perform I2I using unpaired datasets for training. The state-of-the-art for unsupervised image-to-image transition is CycleGAN (Zhu et al., 2017). In this framework, two generators and two discriminators are trained:

- $G_{A\to B}$ learns to map data from $\mathcal{A}$ to $\mathcal{B}$

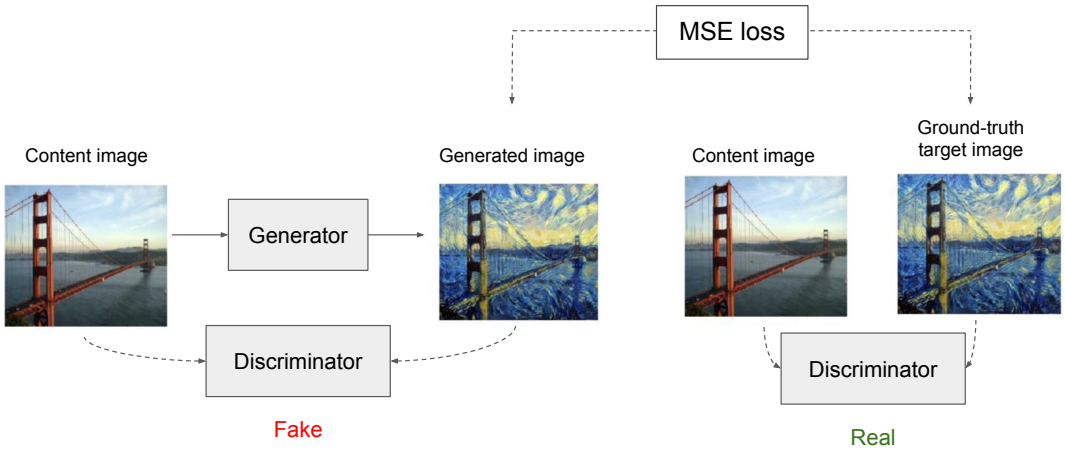

Figure 3.9 – Schematic representation of the learning process of Pix2Pix (Isola et al., 2017). One generator and one discriminator are trained to convert data between two domains, loss functions are adversarial and mean squared error (MSE).

- $G_{B\to A}$ from $\mathcal{B}$ to $\mathcal{A}$
- D_A aims to distinguish between images x_A and translated images $G_{B\to A}(y_B)$
- D_B aims to discriminate between y_B and $G_{A\to B}(x_A)$

The full objective loss consists in two adversarial losses, one between $G_{A\to B}$ and D_B and one between $G_{B\to A}$ and D_A, and one cycle-concistency loss. This cycle-concistency loss is based on the principle that for each image x_A from domain $\mathcal{A}$, the image translation cycle should be able to bring x_A back to the original image: $x_A \to G_{A\to B}(x_A) \to G_{B\to A}(G_{A\to B}(x_A)) \approx x_A$, and similarly for y_B:

$$\begin{aligned}\mathcal{L}_{cyc}(G_{A\to B}, G_{B\to A}) = \mathbb{E}_{x\sim p_{data}(x)}[\| G_{B\to A}(G_{A\to B}(x)) - x \|_1] \\ + \mathbb{E}_{y\sim p_{data}(y)}[\| G_{A\to B}(G_{B\to A}(y)) - y \|_1]\end{aligned} \tag{3.6}$$

StarGAN StarGAN is a generative model architecture designed for multi-domains I2I, its goal is to perform image translation across multiple domains using a single unified model. StarGAN is composed of a single generator and a single discriminator, with some particularities. During training, the generator G takes as input the source image x_A, but also a condition y_B corresponding to a target domain, supposedly leading to image x_{AB}. Then, x_{AB} is set as input to G, this time with a condition y_A to generate x_{ABA}. This

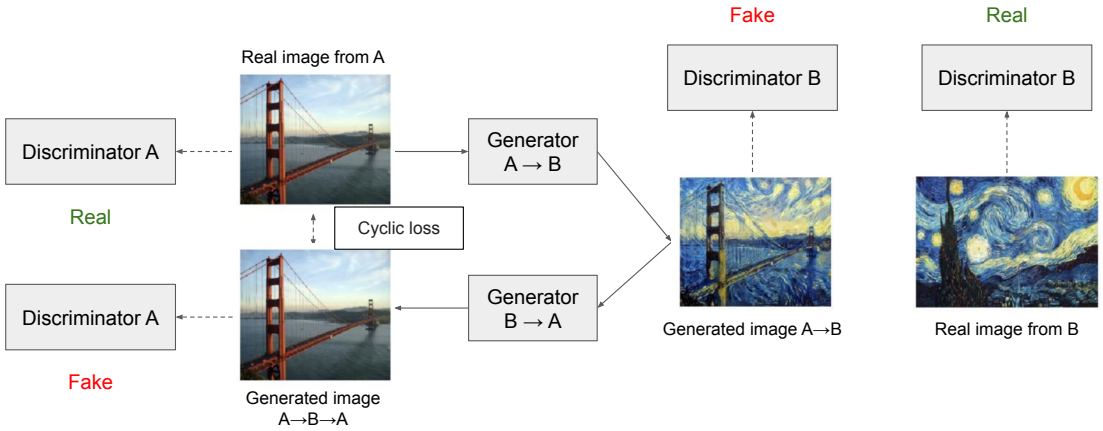

Figure 3.10 – Schematic representation of the learning process of CycleGAN (Zhu et al., 2017). Two generators and two discriminators are trained to convert data between two domains, loss functions are adversarial and cyclic loss.

allows the use of a cyclic-loss that compares x_A and x_{ABA}. The discriminator D tries to distinguish real x_AB images from generated $G(x_A, y_B)$ images, as well as determining the domain of the image x_{AB}.

This model is trained with a loss composed of three components:

- ***Adversarial loss***: to make the generated images indistinguishable from real images. See Equation 3.5.
- ***Cyclic loss***: to guarantee that translated images preserve the content of its input images.

$$\mathcal{L}_{rec} = \mathbb{E}_{x_A, y_A, y_B}[||x_A - G(G(x_A, y_B), y_A)||_1] \tag{3.7}$$

- ***Domain classification loss***: to ensure that generated image are properly classified to the target domain $\mathcal{B}$. To achieve this condition, an auxiliary classifier is added on top of D and an objective loss is decomposed into two terms: a domain classification loss of real images used to optimize D, and a domain classification loss of fake images used to optimize G.

$$\mathcal{L}^r_{cls} = \mathbb{E}_{x_A, y_A}[-log(D_{cls}(y_A|x_A))] \tag{3.8}$$

By minimizing $\mathcal{L}^r_{cls}$, D learns to classify a real image x_A to its corresponding original domain y_A.

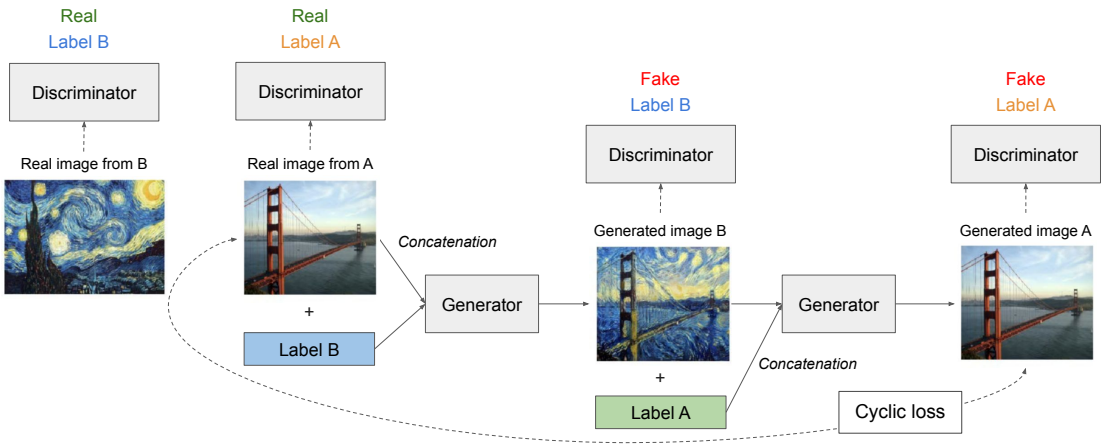

Figure 3.11 – Schematic representation of the learning process of StarGAN (Choi et al., 2018). One generator and one discriminator are trained to convert between multiple domains, loss functions are adversarial, cyclic loss and classification loss.

$$\mathcal{L}_{cls}^{f} = \mathbb{E}_{x_A, y_B}[-log(D_{cls}(y_B|G(x_A, y_B)))] \tag{3.9}$$

By minimizing $\mathcal{L}_{cls}^{f}$, the generator G learns to generate images that are classified in the target domain.

Conditional Denoising Diffusion Probabilistic Models Similarly to GAN, DDPM were rapidly enhanced by adding conditional guidance to the diffusion process. Dhariwal et al., 2021 propose to add conditioning using classifier guidance, *i.e.* use of the gradients of a classifier to guide the diffusion during sampling. The proposed method consists in an unconditional model, and a pretrained a classifier that distinguishes the different labels or domains in the dataset. During sampling, an image from the target domain is passed through the classifier and classifier gradients are injected to the neural network.

In Ho et al., 2021, authors proposed a new framework to dispense with the need for a classifier. In this framework, timestep and conditioning are embedded using 2 MLP and infused with the neural network activations at a certain layer via $a_{L+1} = c_{emb} \cdot a_L + t_{emb}$. An unconditional DDPM is trained along with the conditional one by setting a contrast mask m. This mask changes the conditioning vector to a null token $\emptyset$ with some probability $p_{\text{uncond.}}$, set as an hyper-parameter. During sampling, the framework computes both conditional and unconditional noise prediction and performs a linear combination of the two with a weight w to represent the strength of the conditional guidance using the

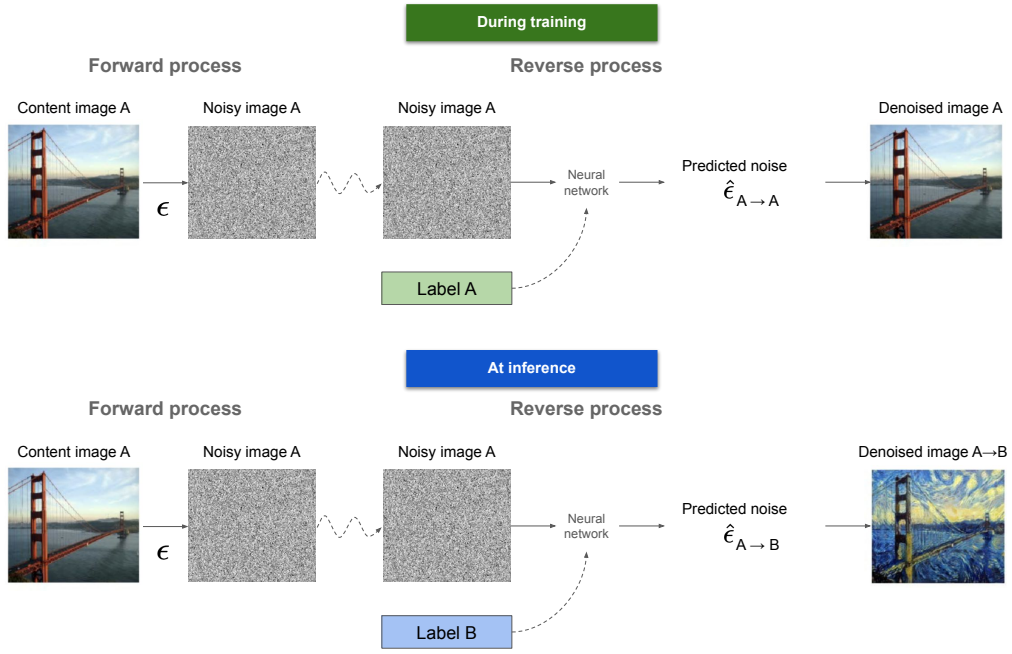

Figure 3.12 – Schematic representation of the learning process of conditional Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2021). During training, a neural network learns to predict the noise added to the image, while knowing its origin domain. At inference, the neural networks predict the noise added to the image, while conditioning on another target domain.

equation from Ho et al., 2021:

$$\widehat{\epsilon_\theta}(X_t, t, C) = (1 + w) \cdot \widehat{\epsilon_\theta}(X_t, t, C) - w \cdot \widehat{\epsilon_\theta}(X_t, t) \tag{3.10}$$

Preechakul et al., 2022 introduce diffusion autoencoders, which consist of a semantic encoder that maps the input image to a latent representation with high-level semantics, and a conditional diffusion model composed of a stochastic encoder to extract a meaningful and decodable representation of an input image and of a decoder for modeling the remaining stochastic variations.

Such frameworks are designed to perform conditional generation, but are not suited for I2I, as the sampling process starts from a random noise. To keep the intrinsic properties of the source image, Saharia et al., 2022 concatenated the source image along with random Gaussian noise to initialize the diffusion. In this paper, a supervised I2I framework is proposed, and conditioning is performed by employing a L_2 regularization between the generated image and the ground truth, similarly to Pix2Pix in GAN (Isola et al., 2017).

For unsupervised settings, *i.e.* with unpaired datasets, Sasaki et al., 2021 developed a framework with two jointly trained DDPM, each one learning the opposite transition $\mathcal{A} \rightarrow \mathcal{B}$ and $\mathcal{B} \rightarrow \mathcal{A}$. During the reverse process, each model is conditioned on the outputs of its counterpart, and a cyclic-concistency loss is added to regularize the training process.

3.4.2.3 Applications in medical imaging

In medical imaging, I2I frameworks are used for multiple tasks (Kaji et al., 2019), including modality transition (Armanious et al., 2020; Denck et al., 2021; Jin et al., 2019; Kong et al., 2021; Lyu et al., 2022; Nie et al., 2018; Ozbey et al., 2023; Qin et al., 2022; Wolterink et al., 2017a; Yang et al., 2020), image denoising (Yang et al., 2018; Wolterink et al., 2017b; Armanious et al., 2020), or data harmonization (Bashyam et al., 2022; Liu et al., 2021). Overall, these frameworks allow researchers to gather more data, and in particular to build multimodal datasets. Observing from multiple modalities offers more comprehensive information, and can reveal more subtle changes in brain tissues, which can be difficult to appreciate with single modality datasets. However, some MR images may become unusable during data acquisition and storage due to various factors such as artifacts or improper scanning parameters. Moreover, rescanning subjects to obtain missing modalities is impractical and costly, as abnormalities in brain structures can change over time, rendering new data incompatible with the original (see 2.1.1). Consequently, cross-modality synthesis of MR images has been explored to address modality absence and inconsistency. Multicentric datasets also offers an opportunity to gather larger datasets, but this can be challenging since different acquisition centers may have different scanning equipment and imaging protocols, leading to unwanted variability in the data (see 2.1.3).

Using a conditional GAN coupled with a perceptual loss and a style transfer loss, MedGAN (Armanious et al., 2020) showed its performance in PET to CT translation, PET denoising and correction of MRI artifacts. In supervised settings, Nie et al., 2018 used a variant of Pix2Pix (Isola et al., 2017) with a gradient-based loss function for MRI to CT translation. Another variant of this model was also used in 3-dimensional for cardiac left ventricle segmentation on echography (Dong et al., 2018). Yang et al., 2018 also used a cGAN for low-dose to high-dose CT translation, with pixelwise loss associated with a minimization of the Wasserstein distance and a perceptual similarity loss. For the same application, Wolterink et al., 2017b proposed to get rid of paired datasets and showed the potential of CycleGAN. This model also showed its potential for stain normalization in histological images (Shaban et al., 2019).

Since their emergence, I2I frameworks focused more and more on the use of diffusion models (Lyu et al., 2022; Ozbey et al., 2023; Pan et al., 2023; Dorjsembe et al., 2024; Jiang et al., 2023). Lyu et al., 2022 showed the superiority of diffusion models compared to GAN in this task for the conversion between MRI and CT using a supervised framework (*i.e.*, with pairs of data from both modalities). In unsupervised settings, Pan et al., 2023 developed a cycle-guided framework composed of two DDPM that condition each other to generate synthetic images from two different MRI pulse sequences. Similarly, Ozbey et al., 2023 proposed SynDiff with a source-conditional adversarial projector that denoises the target image sample with guidance from the source image.

Take-home Message

- Deep representation learning is the process of learning a representation from input data towards a specific task, leading to the identification of meaningful features for the task at hand.
- In computer vision and thus, in medical imaging, the main representatives of deep representation learning models are Convolutional Neural Network (CNN). These are used in many tasks ranging from classification to image registration and denoising.
- The use of medical imaging data comes with challenges, in particular due to the low sample size, low diversity and lack of annotations of datasets.
- To learn better representations of data in such settings, researchers developed several solutions, in particular with transfer learning and generative models.

Chapter 4

Leveraging variability in fMRI results with self-taught learning

---

This chapter was the subject of an article published in *GigaScience*:

- **Title:** On the benefits of self-taught learning for brain decoding
- **Authors:** Elodie Germani, Elisa Fromont*, Camille Maumet*
- **DOI:** 10.1093/gigascience/giad029
- **Code:** swh:1:snp:289ee6f81cd88d26fa3f332eecfb86d3df1f114f
- **Derived data:** Available on Zenodo at 10.5281/zenodo.7566172.
- **Contributions (Credit taxonomy):** Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualisation, Manuscript writing.

* Joint senior authorship.

4.1 Introduction

In the past few years, deep learning approaches have achieved outstanding performance in the field of neuroimaging (Abrol et al., 2021) due to their ability to model complex non-linear relationships in the data. fMRI data are often used as input data to these models for different tasks, such as disease diagnosis (Yin et al., 2022) or brain decoding (*i.e.* identifying stimuli and cognitive states from brain activities) (Firat et al., 2014), with a common goal: linking a target with highly variable patterns in the data and ignoring aspects of the data that are unrelated to the learning task. Researchers took advantage

of the specific properties of fMRI data to build more and more sophisticated models (Vu et al., 2018; Hu et al., 2019; Koyamada et al., 2015; Wang et al., 2020; Huang et al., 2021; Vu et al., 2020; Oh et al., 2019).

As seen in the previous chapter (see Section 3.3), training effective deep learning models using neuroimaging data comes with many challenges due to the particular properties of data (Thomas et al., 2021; Thijs Kooi, 2018). The field also suffers from a large number of sources of variability in the data (see Chapter 2) at the subject level (brain activity patterns differ across participants), the acquisition level (fMRI scanners and protocols often vary between centers and studies) and the analysis level (different analysis pipelines lead to different brain patterns). In our case, brain decoding models should be robust to all these sources of variability, but this remain difficult due to the low sample size and low variability of datasets (Ricci Lara et al., 2022).

To prevent overfitting and allow for generalizable statistical inference, neuroimaging researchers proposed methods to tackle this lack of training data (Bontonou et al., 2021; Yotsutsuji et al., 2021; Zhuang et al., 2019). For instance, Mensch et al., 2014 built a decoding model using data gathered from 35 studies and thousands of individuals that cover various cognitive domains. Despite the good performance of the models, these can only be applied on restricted sets of studies, discriminating between few cognitive concepts. More annotated training data (*e.g.* using large public databases) would be required to map a wider set of cognitive processes. Lots of studies were also made on inductive transfer learning with labeled source data as defined in Pan et al., 2010 (*e.g.* source task and target task are different, as well as source domain and target domain) (Thomas et al., 2023; Y. Gao et al., 2019; Svanera et al., 2019) (see 3.4.1.4).

However, labeled databases are not always available in neuroimaging, despite the growing effort in data sharing to build public databases (Poldrack et al., 2014), such as OpenNeuro for raw data (Markiewicz et al., 2021) and NeuroVault for fMRI statistic maps (Gorgolewski et al., 2015). The unconstrained annotations and the heterogeneity of tasks and studies make them difficult to use to pretrain a supervised deep learning model. To compensate this, weakly supervised learning techniques such as automatic labelling of data has proven its worth. For instance, Menuet et al., 2022 enriched NeuroVault annotations using the Cognitive Atlas ontology (Poldrack et al., 2011b) and used these labeled data to train a multi-task decoding model that successfully decoded more than 50 classes of mental processes on a large test set.

A specific type of inductive transfer learning named *self-taught learning* (Raina et al.,

2007; Wang et al., 2013) showed strong empirical success in the field of machine learning. It does not require any labels as it consists in training models to autonomously learn latent representations of the data and using these to improve learning in a supervised setting. This approach is motivated by the observation that data from similar domains contain patterns that are similar to those of the target domain. By initializing the weights of a supervised classifier with the pretrained weights of an unsupervised model trained on many images, the aim is to improve the model performance by placing the parameters close to a local minimum of the loss function and by acting as a regularizer (Erhan et al., 2010).

In the field of neuroimaging, latent representations have recently been used in a task-relevant autoencoding framework. Orouji et al., 2023 used an autoencoder with a classifier attached to the bottleneck layer on a small fMRI dataset. This model outperformed the classifier trained on raw input data by focusing on cleaner, task-relevant representations. This suggests that a low-level representation of fMRI data, learned for a reconstruction task, can be helpful in a classification task, as in a self-taught learning framework.

In this chapter, we propose to take advantage of NeuroVault – a large public neuroimaging database that was built collaboratively and therefore displays a good level of variability in terms of fMRI acquisition protocols, machines, sites and analysis pipelines – in a self-taught learning framework. We pretrain an unsupervised deep learning model to learn a latent representation of fMRI statistic maps and we fine-tune this model to decode tasks or mental processes involved in several studies. In a first part, we leverage the NeuroVault database to select the most relevant statistic maps and train a Convolutional AutoEncoder (CAE) to reconstruct these maps. In a second part, we use the final weights of the encoder to initialize a supervised Convolutional Neural Network (CNN) to classify the cognitive processes, tasks or constrasts of unseen statistic maps from large collections of the NeuroVault database (an homogeneous collection of more than 18,000 statistic maps and an heterogeneous one with 6,500 maps). Our goal is to investigate how the use of a large and diverse database in a self-taught learning framework can be beneficial in the field of brain imaging for deep learning models.

4.2 Materials and Methods

Figure 4.1 illustrates the overall process used to implement our self-taught learning framework: a CAE was first trained to reconstruct the maps of a large dataset extracted

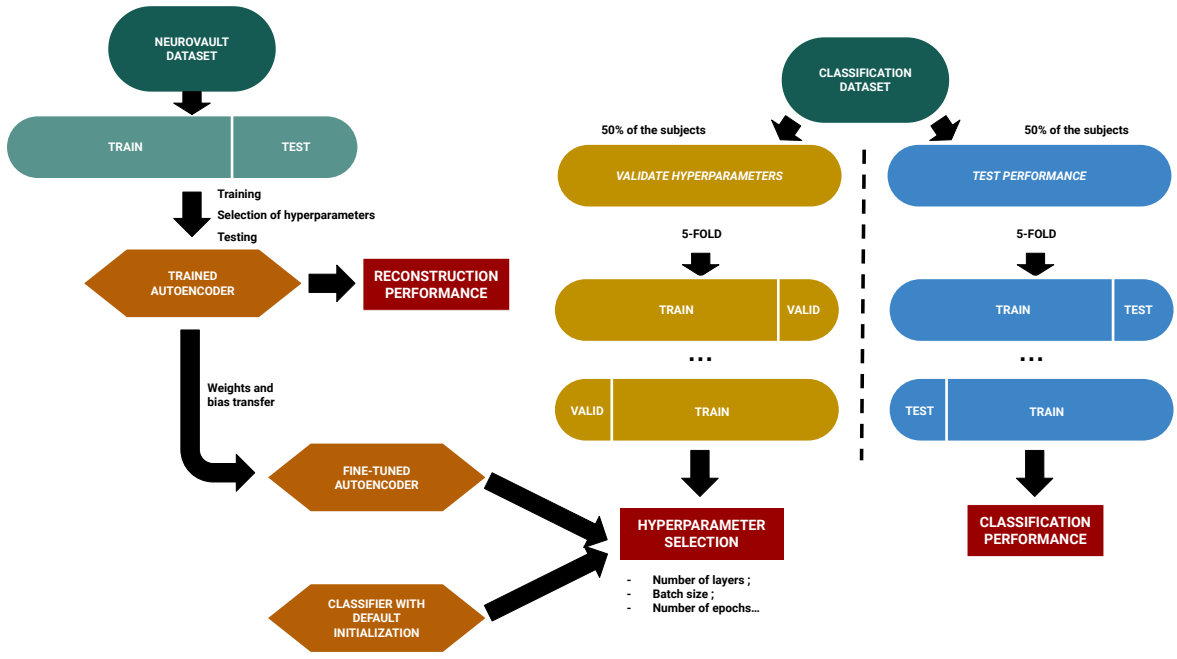

Figure 4.1 – Flow diagram of the self-taught learning methodology. NeuroVault dataset is used to train a Convolutional AutoEncoder (CAE). The encoder of this CAE is used to initialize a Convolutional Neural Network (CNN) and to train it to classify other datasets. These classification datasets are split in two disjoints datasets: a "validation" one used to optimize hyperparameters and a "test" one to evaluate performance. In each one, a 5-fold cross-validation is performed.

from NeuroVault. Then, the encoder part of the CAE was fine-tuned to answer a classification problem on another dataset (with labels). After hyperparameters optimisation, performance of the pretrained classifier was compared to those of a classifier initialized with a default algorithm. Details regarding the datasets (NeuroVault dataset and classification datasets) can be found in the next subsection. The models of the CAE and the CNN are presented in Appendix D. Further explanations on the workflow used to train the CAE and the CNN and to evaluate their performance are available in Sections 4.2.4 and 4.2.5 respectively.

4.2.1 Overview of the datasets

A summary of the different datasets can be found in Table 4.1. Details are given below.

Table 4.1 – Overview of the datasets. For each dataset, number of statistic maps are presented, as well as the number of participants, number of studies and the type of labels (if available).

Dataset	**Maps**	**Participants**	**Studies**	**Labels**
NeuroVault	28,532	-	-	-
HCP	18,070	787	1	Tasks (*7*) Contrasts (*23*)
BrainPedia	6,448	826	29	Cognitive processes (*36*)

4.2.1.1 NeuroVault dataset

NeuroVault (Gorgolewski et al., 2015) (RRID:SCR_003806) is a web-based repository for statistic maps, parcellations and atlases produced by MRI and PET studies. This is currently the largest public database of fMRI statistic maps. NeuroVault has its own public Application Programming Interface (API) that provides a full access to all images (grouped by collections) and enables filtering of images or collections with associated metadata. At the time of experiment (19/01/2022), a total of 461,461 images in 6,782 collections were available. Among the available metadata, some are mandatory and specified for all maps such as the modality (*e.g.* "fMRI-BOLD" for Blood-Oxygen Level Dependent Functional MRI; "dMRI" for Diffusion MRI, etc.), the type of statistic (*e.g.* "T map"

or "Z map") or the cognitive paradigm (*e.g.* "Working memory" or "Motor fMRI task paradigm"), and others are optional and only available if additionally entered at the time of the upload.

From this large database, relevant maps were selected based on multiple criteria. First, we chose maps for which the modality was "fMRI-BOLD" to exclude other modalities such as structural or diffusion MRI. To get comparable maps, we set three additional inclusion criteria and selected maps: 1/ for which all required metadata were provided ("is_valid" to True) 2/ that were registered in MNI space ("not_mni" to False) – to ensure that anatomical structures were located at the same coordinates in each map – and 3/ referenced as "T map" or "Z map" – to exclude maps in which voxel values did not have the same meaning (*e.g.* P value maps, Chi-squared maps, etc.) –. Among these, thresholded statistic maps were excluded.

We found that some maps in our initial dataset, were wrongly referenced as T map or Z map. These misclassified maps were removed by filtering the "filename" column of the dataframe to exclude *SetA_mean SetB_mean* (AFNI contrast maps), *con* (SPM contrast maps), *cope* (FSL contrast maps).

Using these criteria, a total of 28,532 statistic maps were selected from the NeuroVault database and constituted our "NeuroVault dataset". Most of these maps were unlabeled (*i.e.* cognitive processes or tasks performed described as "None / Other") or not labeled in a standardized way (*i.e.* use of terms that are specific for a study instead of generic terms, such as those defined in Poldrack et al., 2011b: *e.g.* some maps were labeled as 'word-picture matching task' for the cognitive paradigm whereas others in which a similar task was performed were referenced as 'working memory fMRI task paradigm' which is a label that includes other specific tasks).

4.2.1.2 HCP dataset (NeuroVault Collection 4337)

NeuroVault collection 4337 (Collection n°4337, n.d.) includes 18,070 z-statistic maps, for base contrasts (task vs baseline), corresponding to 787 participants of the Human Connectome Project (HCP) Young Adult S900 release (Van Essen et al., 2013). This collection was excluded from our pretraining dataset (see section 4.2.1.1) due to missing metadata (*i.e.* 'is_valid' is False).

All maps in this collection were grouped together and referred to as the "HCP dataset" in the following. Multiple labels were entered for each map including: mental concepts ("cognitive_paradigm_cogatlas"), tasks ("task") and contrasts ("contrast_definition")

(as defined in Poldrack et al., 2011b). For each participant, 23 contrasts distributed in 7 tasks were available:

- Working memory: '0-back body', '0-back face', '0-back places', '0-back tools', '2-back body', '2-back face', '2-back places', '2-back tools'
- Motors: 'cue', 'left foot', 'left hand', 'right foot', 'right hand'
- Relational: 'relational', 'match'
- Gambling: 'punish', 'reward'
- Emotion: 'faces', 'shapes'
- Language: 'math', 'story'
- Social: 'tom'

For more details on contrasts, tasks and mental concepts of this study, see Van Essen et al., 2013.

4.2.1.3 BrainPedia dataset (NeuroVault collection 1952)

NeuroVault collection 1952 (Collection n°1952, 2016), known as BrainPedia (Varoquaux et al., 2018), contains fMRI statistic maps of about 30 fMRI studies from OpenNeuro (Markiewicz et al., 2021), the Human Connectome Project (Van Essen et al., 2013) and from data acquired at Neurospin research center, together they were chosen to map a wide set of cognitive functions.

This collection contains 6,573 statistic maps corresponding to 45 unique mental concepts derived from 19 sub-terms (*e.g.* 'visual, right hand, faces' for maps associated with the task of watching an image of a face and responding to a working memory task). These images were previously used to build a multi-class decoding model (Varoquaux et al., 2018) and labels corresponded to the mental concepts associated with the statistic map, *e.g.*, 'visual', 'language' or 'objects'. Here we excluded the nine classes that had less than 30 samples each, leaving 6,448 images corresponding to 36 classes. These 6,448 images were grouped together and referred to as the 'BrainPedia' dataset in the following.

4.2.2 Preprocessing

All statistic maps included in this study were downloaded from different collections of NeuroVault and therefore were processed using different pipelines (see the original studies for more details (Varoquaux et al., 2018; Van Essen et al., 2013)). We resampled all maps

to dimensions (48, 56, 48) using the MNI152 template available in Nilearn (Abraham et al., 2014a) (RRID: SCR_001362) as target image. A min-max normalization was also performed on all resampled maps to get statistical values between -1 and 1. Finally, the brain mask of the MNI152 template in Nilearn was used to exclude statistical values outside the brain in all statistic maps.

4.2.3 Model architectures

Description of model architectures and corresponding Figures are described in Appendix D.

4.2.4 Convolutional AutoEncoder (CAE) training

To train our CAE to reconstruct the statistic maps of the NeuroVault dataset, we used an Adam optimizer (Kingma et al., 2017) with a learning rate of $1e-04$ and all other parameters with default values. The loss function was the Mean Squared Error (MSE: the squared L2 norm) which is the standard reconstruction loss.

4.2.4.1 Dataset split

NeuroVault dataset was randomly split in two subsets: training and test with respectively 80% and 20% of the maps. The training set (N=22,772 maps) was used to train the CAE with the different architectures and the test set (N=5,760 maps) to assess the performance of the different models (with different hyperparameters).

4.2.4.2 Architecture comparison

To limit the computational cost of our experiments, we fixed some of the hyperparameters of the CAE and only compared those who were of interest for the later experiments. Here, we use the term model "hyperparameters", to distinguish with model "parameters", to represent the values that cannot be learned during training, but are set beforehand *e.g.*, the batch size or the number of hidden layers. Thus, a batch size of 32 and a learning rate of $1e-04$ were chosen to train the CAE for a number of 200 epochs (*i.e.* values that are often used in experiments). The only hyperparameter for which different values were compared were the number of hidden layers of the model: 4 layers vs 5 layers for each part (encoder/decoder) of the model.

4.2.4.3 Performance evaluation

To assess the performance of the CAE, we estimated Pearson's correlation coefficient between the reconstructed statistic map and the original statistic map. The correlation coefficient was computed using numpy version 1.21.2 (RRID: SCR_008633) (Harris et al., 2020). The closer to 1 the correlation coefficient was, the stronger the relationship between the maps and the more accurate the reconstruction. Note that we did not use MSE in this context as its individual values (for each data point) were not easily interpreted.

4.2.5 Convolutional Neural Network (CNN) training

We trained two types of classifiers for all the experiments:

— the *classifier with default algorithm* initialized with the original algorithm from He et al., 2015 (*i.e.* Kaiming Uniform algorithm for convolutional and fully-connected layers with a parameter of $\sqrt{5}$) and

— the *classifier with pretrained CAE* initialized using the weights and bias of the convolutional layers of the CAE pretrained on NeuroVault dataset.

The CNN were trained using the Adam optimizer with a learning rate of $1e-04$. We used the cross-entropy loss function for training the classifier. Both were implemented in PyTorch.

4.2.5.1 Dataset split

As described in Fig. 4.1 (on the right), the classification datasets were split in two disjoint subsets: the *'validation dataset'* used to optimize the hyperparameters, and the *'test dataset'* used to test the performance. Each subset contained 50% of the participants of the overall dataset with no overlap to avoid any data leakage (see Varoquaux et al., 2022; Kapoor et al., 2023).

For each experiment, the validation and test datasets were then split into 5 folds for cross-validation. participants were randomly sampled in each fold in order to ensure that there was no overlap of participants across folds. The identifiers of the participants included in the different folds were saved for reproducibility. More details on the methods used to perform the 5-folds split for each dataset are specified in subsection 4.2.6.

4.2.5.2 Evaluation of performance

The performance of each model was measured using several metrics: accuracy (Acc), precision (P), recall (R) and F1-macro score (F1). All metrics were implemented using scikit-learn (Abraham et al., 2014a) with default parameters, except for F1-score for which the "average" parameter was specified with "macro" to deal with multi-class classification.

To evaluate the performance of a model, all metrics were averaged among the 5 folds of cross-validation and standard error of the mean was computed.

To compare the final performance of models with default initialization versus fine-tuned weights, we used paired one-tailed two-sample t-tests between the performance values (accuracy or F1-score) of the 5 models trained during cross-validation. T-statistic and p-value were provided and value of 0.05 was used for the p-value significance threshold.

4.2.5.3 Hyperparameters optimisation

To select the best hyperparameters for each dataset and each type of initialization, we evaluated the performance of each model by performing a 5-fold cross-validation on the validation dataset.

For each type of classifier (*i.e.* initialized with default algorithm versus pretrained), we refined and optimised the hyperparameters using the largest datasets (Large BrainPedia and HCP). However, the large amount of training data made it computationally extremely costly to perform a full grid-search. We therefore limited our research to predefined values of batch sizes (32 or 64), number of epochs (200 or 500) and model architectures (4 layers or 5 layers). All batch sizes, number of epochs and architectures were tested for each type of classifier and each dataset. We did not perform any optimization on the learning rate to limit the computational cost of our experiments. Every model was trained using a learning rate of $1e-04$.

We selected the best set of hyperparameters based on the performance of the corresponding model in terms of accuracy and F1-score, averaged across folds.

4.2.6 Benefits of self-taught learning and impact of different factors

To investigate the benefits of self-taught learning for neuroimaging data, different brain decoding experiments were studied. For all, after optimizing the hyperparameters of the two models (*i.e.* the model with default initialization -or- with pretrained CAE

and fine-tuned weights) we assessed the performance of these optimized models on the test dataset using a 5-fold cross-validation.

4.2.6.1 Homogeneous dataset (single study)

The HCP dataset was used to compare the performance of the models for the task of decoding on a homogeneous dataset (*i.e.* from a single study). We studied the impact of two factors on the classification: sample size and number of target classes. For sample size, subsets of the global HCP dataset were created with different number of participants: N=50, 100 and 200. Each smaller subset being a subset of the immediately larger one. To create these subsets, we first split the global HCP test dataset into 5 folds, with different participants in each fold. In each of these 5 folds, we randomly sampled 200/5 = 40 participants and obtained 5 sub-folds that together composed the smaller subset of 200 participants. This process was repeated for subsamples N=100 and 50 by sampling from their superset. This insured that the 5 models trained on different combinations of the 4 folds of a smaller subset could be tested on the remaining fold of the global test dataset with no overlap between the training and test data. The process is illustrated in Fig. 4.2(a).

In the end, we obtained 4 datasets with respectively N=50, 100 and 200 participants in addition to the global dataset with all participants (N=393). These datasets respectively contained 1150, 2300, 4590 and 9017 statistic maps in the test subset and 1150, 2300, 4591 and 9053 in the validation subset (note: some contrasts were missing for part of participants). Since we use a 5-fold validation scheme, the models were trained on approximately 80% of the statistic maps in the corresponding subset (*i.e.* validation for hyperparameter optimization and test for performance evaluation).

Three types of classification were investigated. First, the 'contrast classification' which consisted in identifying the contrast associated with a statistic map (23 different contrasts). Second, the 'task classification' which consisted in identifying the task associated with a statistic map (7 different tasks, with multiple contrasts per task). Third, the 'one contrast task classification'. This time, we selected a single contrast per task and classified the tasks (7 different tasks, with one contrast per task). The selected contrasts were '2-back places', 'faces', 'punish', 'relational', 'right hand', 'story' and 'tom' respectively for the tasks 'Working Memory', 'Emotion', 'Gambling', 'Relational', 'Motor', 'Language', 'Social'. We selected these contrasts similarly to what was done in Wang et al., 2020 in which the HCP dataset was used in a decoding model. For each task, the contrast

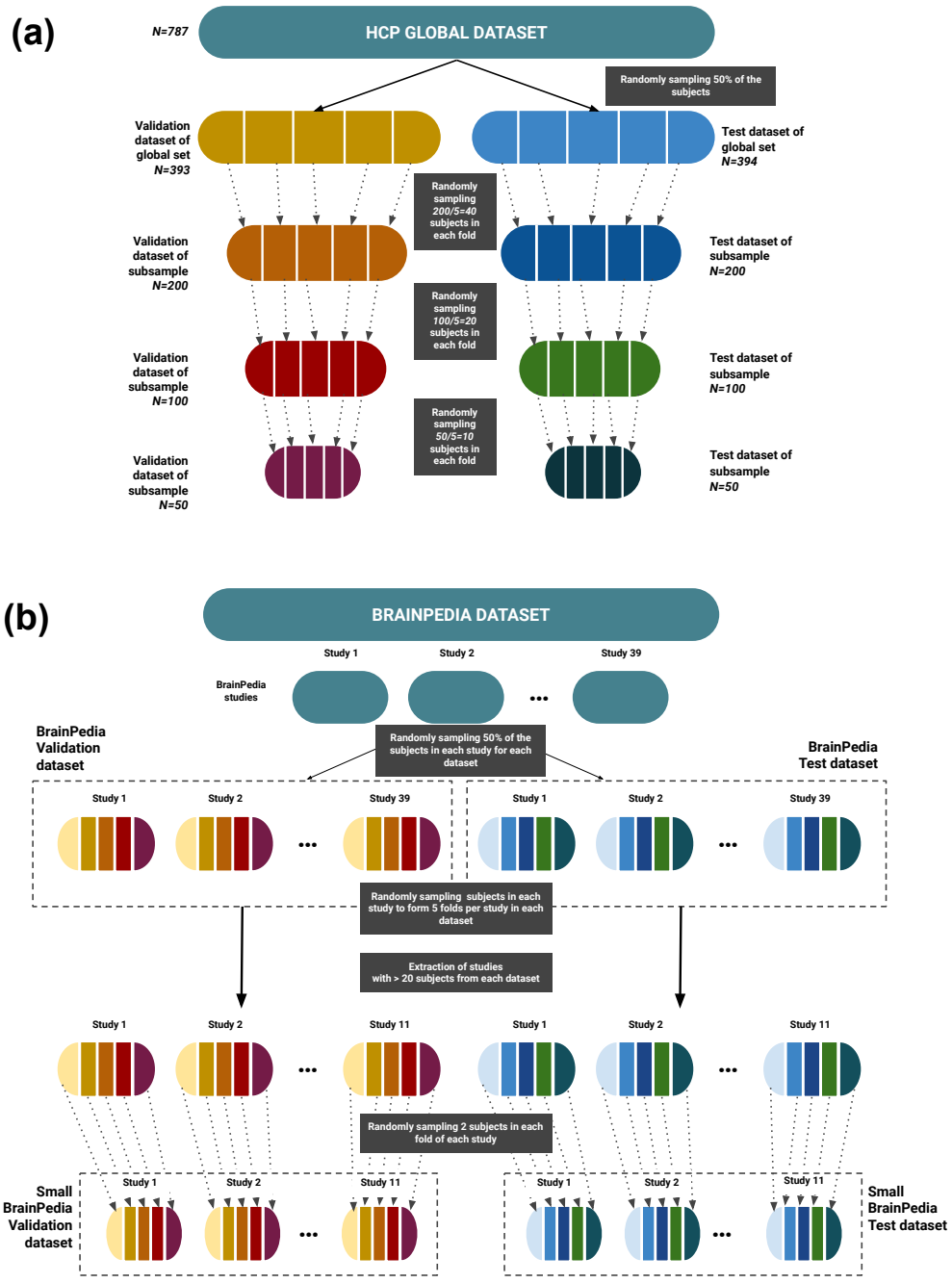

Figure 4.2 – Overview of the process used to split the datasets for cross-validation. (a) shows the method performed for HCP dataset and its subsamples and the one used for BrainPedia and Small BrainPedia datasets is presented in part (b). In both cases, the global dataset is first split into two subdatasets 'validation' and 'test' with respectively 50% of the participants and then each subdataset is divided into 5 folds for cross-validation.

that showed a greater association with the task had priority over the other (for instance, 'punish' for the 'Gambling' task). For 'Working Memory' and 'Motor' tasks, which contained more than one task condition, they randomly chose one ('2-back body' for Working Memory and 'right hand' for Motor). The dataset used for this third type of classification was thus smaller than the others (only one map per task per participant). For this classification task, the number of statistic maps was respectively 300, 598, 1198 and 2355 for N=50, 100, 200 and for the global dataset.

4.2.6.2 Heterogeneous dataset (multiple studies)

To study the benefits of self-taught learning on a heterogeneous dataset (*i.e.* from multiple studies), we used BrainPedia. For these experiments, we focused on the classification of mental concepts (as available in NeuroVault metadata). Fig. 4.2(b) illustrates the process used to split this dataset. To perform the split while maintaining the heterogeneity in each fold, we randomly sampled 50% of the participants of each study to form the 'validation' and 'test' datasets of BrainPedia. Then, each dataset, each study was split into 5-folds and the n-th folds of the different studies were combined to form the n-th fold of the dataset. Validation and test datasets included $N = 428$ participants and were respectively composed of 3179 and 3269 statistic maps.

We also studied the impact of sample size in the presence of heterogeneity by extracting smaller datasets. Among the 29 studies of the BrainPedia dataset, we only kept those which were composed of more than 20 participants. In these remaining studies, already split into 5 folds in BrainPedia validation and test subdatasets, 2 participants were randomly drawn per fold per study per subdataset to obtain 10 participants per study per subdataset. Like above, the n-th folds of the different studies were combined to form the n-th fold of each subdataset of the 'Small BrainPedia' dataset. In the end, this smaller dataset was composed of 1,844 maps, divided in 30 classes, from 11 studies and 220 participants. This dataset was also split into test and validation subsets with 50% of the participants in each (N=110). The test and validation subsets were thus composed respectively of 917 and 927 maps.

4.2.7 Explainability

4.2.7.1 Exploring feature maps to understand the generalizability across participants

To investigate the reasons for the difference in performance between the pretrained and default models, we visualized and analyzed the feature maps of the different convolutional layers of the model. Visualizing these features was useful to better understand how each model made its predictions.

With a generalizable classifier, we hypothesized that features of different participants from the same class should be similar (and therefore not be impacted by individual differences). To study this, we computed for each classifier, each layer and each class, the correlations between the feature maps for all pairs of participants. A high mean correlation highlighted a higher similarity between the feature maps extracted by this layer for a classifier and thus a higher generalizability.

4.2.7.2 Investigating the contribution of each layer to the overall performance

We explored which pretrained layer had the strongest impact on the classification performance. This could be made at two stages: before and during training.

Before training, we only transferred the weights of some parts of the CAE. In particular, we kept the weights of the last convolutional layers with a default initialization and initialized the first layers with the weights of the pretrained CAE. Multiple configurations were explored: transferring only the weights of the first one up to the first four convolutional layers.

During training, we froze some layers of the model initialized with the weights of the pretrained CAE, *i.e.* some layers (the first ones) were not fine-tuned. Multiple types of freezing were tested: freezing of the first two to the first five convolutional layers.

4.3 Results

4.3.1 Convolutional AutoEncoder (CAE) performance

Reconstruction performance of the CAE is presented in Table 4.2. When comparing the two CAE architectures (4-layers vs 5-layers) trained on NeuroVault dataset, the mean correlations between original and reconstructed maps were better for the 4-layers archi-

tecture (86.9% vs 77.8%). These results suggest that the reconstruction capabilities of the CAE are dependant on the model architecture and the size of the latent space. Figure 4.3 shows the reconstruction of a statistic map randomly drawn from the NeuroVault test dataset with the two CAE architectures. With the 4-layers architecture, details of the map were better reconstructed than with the 5-layers architecture (see the green square on the map). This was due to the level of compression of the data that was higher in the 5-layers CAE and that learned only the most useful features with less emphasis in learning specific details. Both models were used as pretrained model for classification to see if the benefits of the CAE were related to their reconstruction performance.

Table 4.2 – Reconstruction performance of the Convolutional AutoEncoder (CAE) depending on model architecture and training set. Values are the mean Pearson's correlation coefficients (standard error of the mean).

Model	**4-layers** *Latent space 18,432*	**5-layers** *Latent space 4,096*
Correlation (*std error*)	86.9 (*0.18*)	77.8 (*0.23*)

4.3.2 Hyperparameters optimisation for Convolutional Neural Network (CNN)

The best hyperparameters and corresponding performance can be found on Table 4.3.

Table 4.3 – Hyperparameters chosen for each dataset and corresponding performance of the classifier on the validation set of the dataset.

Dataset	**Initialization**	**Model**	**Epochs**	**Batch**	**Accuracy (%)** (*std. err.*)	**F1-Score (%)** (*std. err.*)
HCP	Default algorithm	4-layers	500	32	90.8 (*1.5*)	90.8 (*1.6*)
	Pretrained CAE	5-layers	200	64	91.8 (*0.9*)	91.8 (*0.9*)
BrainPedia	Default algorithm	5-layers	500	64	67.1 (*1.7*)	61.0 (*1.6*)
	Pretrained CAE	5-layers	200	64	73.8 (*2.7*)	70.0 (*2.3*)

4.3.2.1 Choice of hyperparameters for HCP dataset

Performance of the different models trained with the different hyperparameters can be found in Supplementary Table S1, available at Germani et al., 2023. For the default

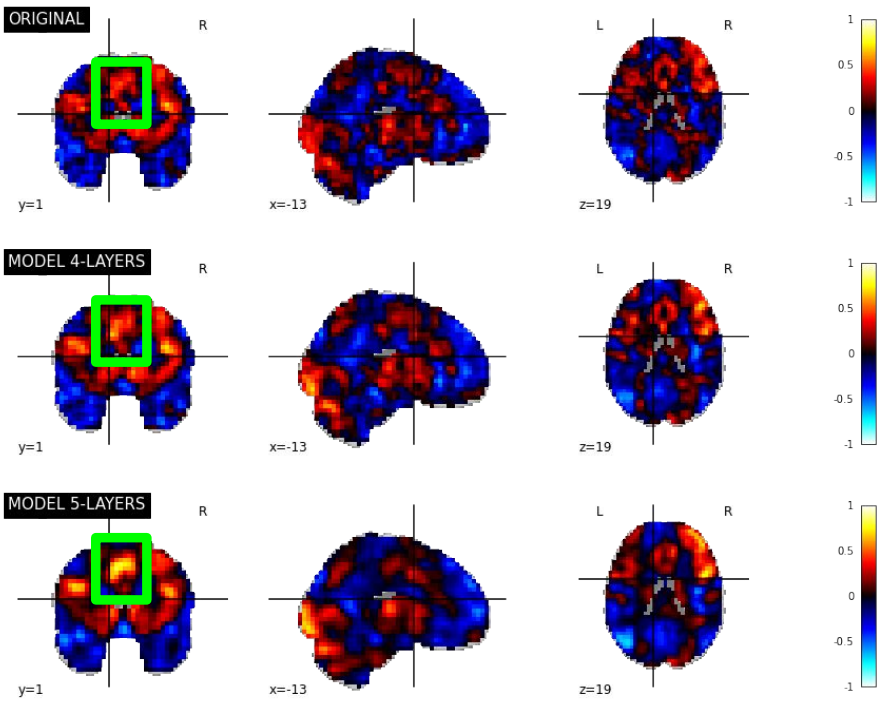

Figure 4.3 – Original version and reconstruction of a randomly drawn statistic map of NeuroVault test dataset (image ID: 109) with the two Convolutional AutoEncoder (CAE) (4-layers and 5-layers). The green square corresponds to a highlighted part of the map for which reconstruction performance are better using the 5-layers architecture.

algorithm initialization, the best model had 4 layers and was trained with a batch size of 32 for 500 epochs. This model achieved an accuracy of 90.8% on average of the 5-folds of cross-validation. For the pretrained CAE initialization, the best model had 5 layers and was trained with a batch size of 64 for 200 epochs (average accuracy of 91.8%). The best hyperparameters for each type of initialization (default and pretrained) were used in all subsequent experiments.

4.3.2.2 Choice of hyperparameters for BrainPedia dataset

Results for all sets of hyperparameters are available in Supplementary Table S2, available at Germani et al., 2023. For the default algorithm initialization, the model who achieved the best performance had 5 layers and a batch size of 64 for 500 epochs. This model classified the BrainPedia dataset with an average accuracy of 67.1% and an average

F1-score of 61%. The performance of the pretrained CAE was the best using a 5-layer architecture, a batch size of 64 and a training time of 200 epochs.

4.3.3 Benefits of self-taught learning on a homogeneous dataset

Table 4.4 summarizes the results for the different classification experiments on the HCP datasets.

Table 4.4 – Classification performance on HCP datasets of models initialized with default algorithm vs with the weights of the pretrained CAE. Mean accuracies and standard errors of the means among the 5-folds of cross-validation are shown. Paired two samples t-tests are performed between the accuracies of the 5 models obtained with cross-validation for each type of initialization. DA: Default Algorithm initialization ; PT: pretraining initialization.

Participants	**50**		**100**		**200**		**Global (393)**	
Maps	1150		2300		4590		9017	
Init.	**DA**	**PT**	**DA**	**PT**	**DA**	**PT**	**DA**	**PT**
Contrast classification (*23 classes*)								
Mean Acc. (%)	83.6	87.0	86.8	89.9	88.6	90.2	90.9	92.4
(std. err.)	(0.61)	(0.51)	(0.69)	(0.34)	(0.84)	(1.46)	(0.38)	(0.44)
Paired T-test (*4 dof*)		**-11.52**		**-4.77**		-1.42		**-4.74**
p-value		***0.0003***		***0.009***		*0.23*		***0.009***
Task classification (*7 classes, multiple contrasts per class*)								
Mean Acc. (%)	96.6	97.3	95.4	98.0	97.9	98.5	98.4	99.0
(std. err.)	(0.47)	(0.43)	(1.49)	(0.25)	(0.44)	(0.16)	(0.17)	(0.13)
Paired T-test (*4 dof*)		***-3.57***		-1.4		-1.5		***-5.65***
p-value		***0.02***		*0.2*		*0.2*		***0.005***
One constrast task classification (*7 classes, one contrast per class*)								
Mean Acc. (%)	97.9	99.1	98.9	99.4	99.3	99.6	99.4	99.6
(std. err.)	(0.3)	(0.3)	(0.17)	(0.25)	(0.2)	(0.2)	(0.2)	(0.14)
Paired T-test (*4 dof*)		**-4.17**		**-3.32**		-2.33		-2.06
p-value		***0.01***		***0.03***		*0.08*		*0.1*

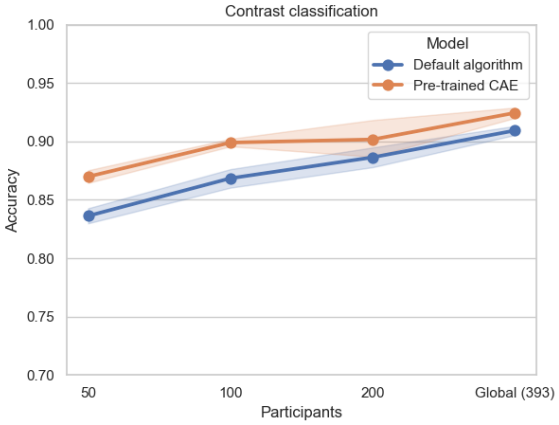

Figure 4.4 – Mean accuracies and standard errors of the mean on contrast classification with the HCP dataset for the models initialized with default algorithm (blue) and pretrained CAE (orange). Pretraining improves contrast classification performance for small sample sizes and at a lower level of improvement, also for large sample sizes.

4.3.3.1 Impact of the sample size

For all classification experiments, the size of the training set (in terms of number of participants) had a strong impact on the benefits of self-taught learning. With 50 participants, the performance of the pretrained CAE outperformed the performance of the classifier initialized with the default algorithm in all our experiments (improvements of 0.7% to 3.4% in mean accuracies). These improvements were always significant ($p < 0.05$). When sample size increased, this improvement reduced and was sometimes not significant. If we focus on contrast classification (Figure 4.4), which was the hardest classification task between the three presented here due to the higher number of classes, the difference between the performance of the two classifiers decreased with sample size (mean accuracies of 88.6% and 90.2% respectively for default initialization and pretrained model respectively for N=200 which corresponded to an improvement of 1.6% compared to almost 3% for N=100). For N=200, the difference of performance was not significant, probably due to the presence of an outlier value in the accuracies of the pretrained CAE. Indeed, accuracies of the pretrained CAE model were superior to the ones of the default model, except for the pretrained model tested on the 3rd fold of cross-validation which was lower. This value was also significantly lower than those of models tested on other folds of cross-validation (see Supplementary Table S3, available at Germani et al., 2023).

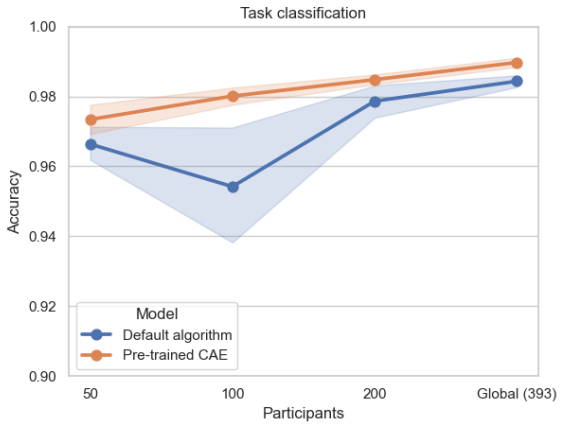

Figure 4.5 – Mean accuracies and standard errors of the mean on task classification with the HCP dataset for the models initialized with default algorithm (blue) and pretrained CAE (orange). Pretraining improves task classification performance for all sample sizes but sample sizes did not have a huge influence on the level of improvement.

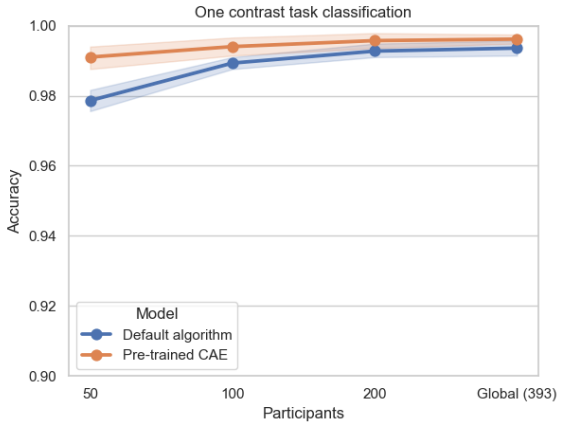

Figure 4.6 – Mean accuracies and standard errors of the mean on one contrast task classification with the HCP dataset for the models initialized with default algorithm (blue) and pretrained CAE (orange). Pretraining does not always improve one-contrast task classification performance: for large sample sizes, pretraining and default initialization give very similar results.

4.3.3.2 Impact of the target classification task

For simpler classification experiments (*i.e.* with less classes to separate), pretraining was not always useful. In these experiments, performance was already nearly perfect (accuracies close to 1) and therefore difficult to improve. For large sample sizes ($N > 100$), performance was close (difference between mean accuracies lower than 0.6%) between models initialized with default algorithm and pretrained models (see Figures 4.5 and 4.6). However, for smaller sample sizes (N=50), pretraining improved classification – similarly to what had been shown for more complex tasks – with accuracies of the pretrained models higher than default models of 0.7% and 1.2% for task classification and one contrast task classification respectively. These results suggest that pretraining can be beneficial when studying difficult classification problems such as those with few training samples or complex classification tasks.

4.3.4 Benefits of self-taught learning on a heterogeneous dataset

Table 4.5 summarizes the results for the classification of mental concepts on the small and the large BrainPedia datasets. These results are illustrated in Figure 4.7.

Table 4.5 – Classification performance on BrainPedia datasets of models initialized with default algorithm vs with the weights of a pretrained CAE. DA: Default Algorithm initialization ; PT: pretraining initialization

Dataset	**Small BrainPedia**		**BrainPedia**	
Init.	**DA**	**PT**	**DA**	**PT**
Mean acc. (%)	56.8	64.5	67.1	74.2
(std. err.)	(1.5)	(2.1)	(0.9)	(2.3)
Paired T-test (*4 dof*)		**-8.72**		**-3.43**
p-value		***0.001***		***0.02***
Mean F1-score (%)	50.5	62.0	64.9	73.6
(std. err.)	(3.5)	(2.1)	(0.8)	(2.2)
Paired T-test (*4 dof*)		**-4.89**		**-2.89**
p-value		***0.008***		***0.04***

On a the small BrainPedia dataset, pretraining improved the performance of the classifier. When looking at the mean accuracies, respectively 56.8% and 64.5% for the classifier initialized with the default algorithm and the pretrained classifier, the difference was high

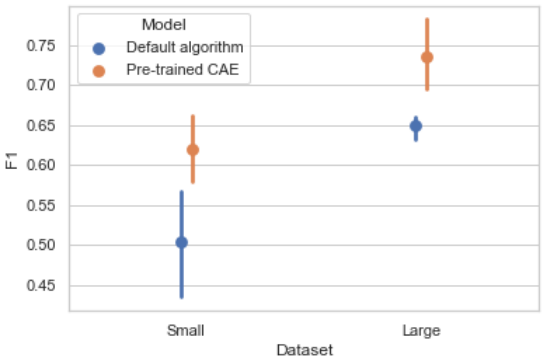

Figure 4.7 – Mean F1-scores and standard errors of the mean of the classification of mental concepts on BrainPedia datasets (Small and Large) for the models initialized with default algorithm (blue) and pretrained CAE (orange). Pretraining improves classification performance, in particular for the small dataset.

(almost 8% of improvement). But in this case, the F1-score was a better metric to assess the performance. Indeed, this metric focuses more on classification errors and is a better indicator of performance when classes are imbalanced, which was the case in this dataset in which some classes were more represented than others (*e.g.* in the small BrainPedia training set, 205 maps corresponded to the class "visual words, language, visual" whereas only 19 are in the class "left foot, visual"). When focusing on this metric, the pretrained classifier performance was markedly higher than the ones of the classifier with default initialization (11.5% of improvement in mean F1-score). Performance (accuracies and F1-scores) was both significantly improved with the pretrained model compared to the default one ($p < 0.05$).

On the global BrainPedia dataset, performance also increased with pretraining. Mean accuracy and F1-score were higher for the the pretrained model (F1-score of 73.6% against 64.9% for the model with default initialization) even if the sample size of the dataset was higher and more classes were represented. Indeed, the classification task was also more complex for this dataset since data were separated into 36 classes instead of 30 for Small BrainPedia due to the presence of maps from other studies in the dataset.

4.3.5 How do we explain these benefits?

4.3.5.1 Features

To better understand the behaviour of each model – in particular on what features

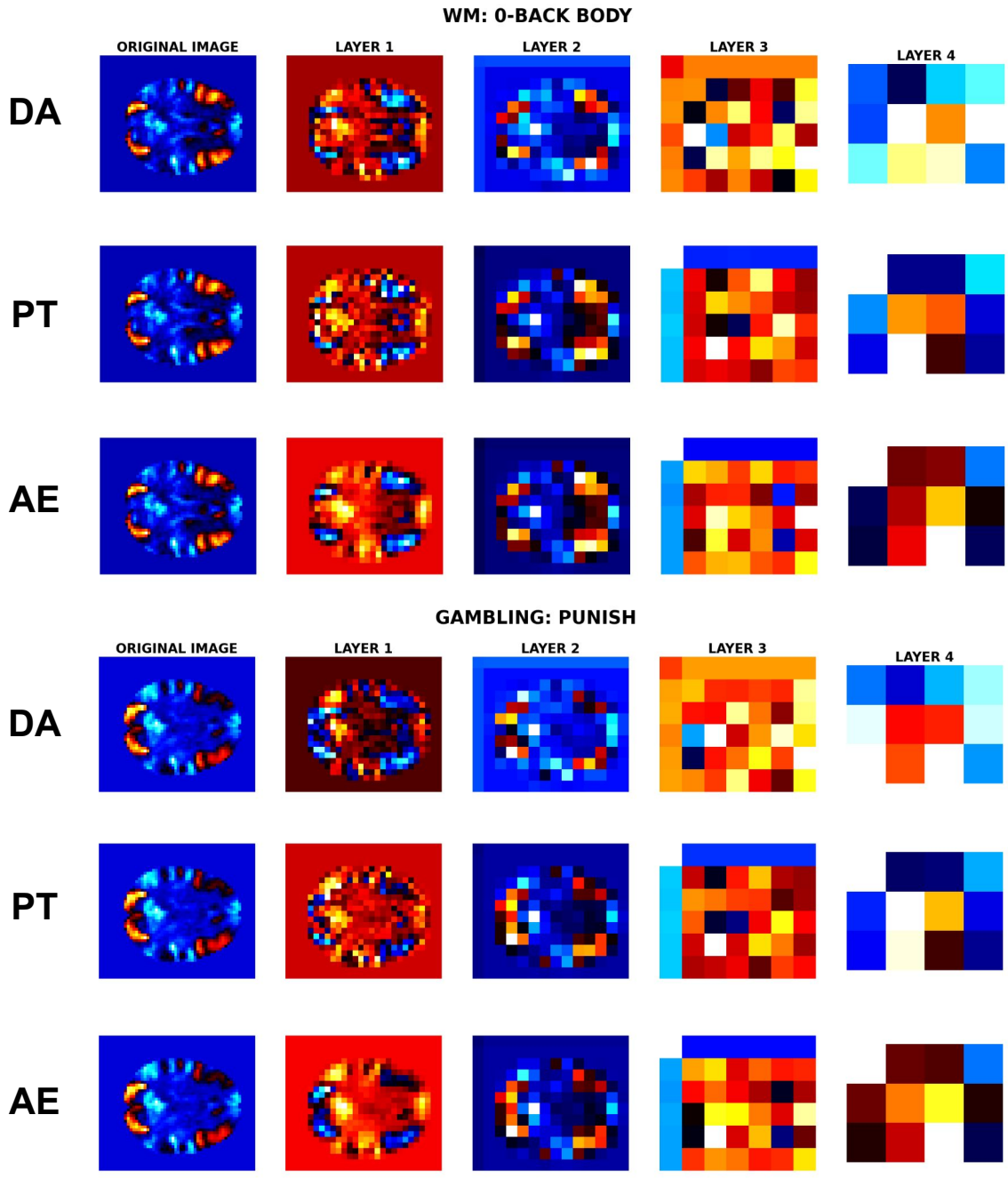

Figure 4.8 – Original mean statistic maps (column 1) and mean feature maps across participants of the fold 1 of the test dataset of HCP 50 for the first four convolutional layers of each model (columns 2-5): CNN with default algorithm initialization (DA), pretrained CNN (PT) and CAE, for two of the eight selected contrasts (WM: 0-back body and Gambling: Punish).

Contrast	**Per-class accuracy**	
	DA	**PT**
WM: 0BK BODY	57.7	60.3
WM: 0BK PLACE	70.5	79.5
WM: 0BK TOOL	57.8	66.7
WM: 2BK BODY	74.3	73.1
WM: 2BK TOOL	47.4	60.3
GAMBLING: PUNISH	55.1	67.9
RELATIONAL	58.9	75.6
GAMBLING: REWARD	57.7	66.7

Table 4.6 – Per-class accuracies for classification of contrasts with HCP dataset sample N=50 for DA (Default Algorithm) and PT (pretrained CAE). Only lowest per-class accuracy ($< 80\%$) are shown in the Figure. For other per-class accuracy, please refer to Supplementary Table S7, available at Germani et al., 2023

they based their predictions on – we visualized the mean features across participants of each layer of the pretrained, default models and baseline CAE for each class label (*i.e.* contrast). Specifically, we studied the mean feature maps obtained across participants in the test set (fold 1) of the N=50 sample of the HCP dataset for different contrasts. This configuration was chosen due to the large difference between performance of default and pretrained models on this classification task. Our main interest was to see if the model would focus on general patterns of activation or more individual features. We focused on the contrasts that led to the most difficult classification tasks (*i.e.* had the lowest per-class accuracy (less than 80%)). Per-class accuracy for selected contrasts are shown in Table 4.6 and for all contrasts in Supplementary Table S7, available at Germani et al., 2023. Eight contrasts were selected: 'Working Memory': '0-back body', '0-back places', '0-back tools', '2-back body', '2-back tools' , 'Gambling: punish', 'Gambling: reward' and 'Relational: relational' and among these 8 contrasts, 7 (all except '2-back body') had a better per-class accuracy with the pretrained CAE, see 4.3.3.

Figure 4.8 shows the mean feature maps for two of the selected contrasts and for the first four convolutional layers of the models: CNN with default initialization, pretrained CNN and CAE. The first convolutional layer features (column two of Figure 4.8) were

similar across models but different between the contrasts: see, for instance, the activation patterns of contrasts WM: '0-back body' and 'Gambling: Punish', which were localised in the same areas, had different shapes. These were high level features: brain shape and main activation patterns. However, the second convolutional layer (third column) seemed to learn more important features for classification. The shape of the brain was still visible but patterns of activation were more blurry, as if they were lower resolution representations of the original statistic maps. However, features started to be different between models at this layer with some modifications of the shape of the main activation patterns between the default model (first row of each contrast) vs. the pretrained model and the CAE (second and third lines). The same observation was made for the third convolutional layer (fourth column), which began to learn deeper representations. Due to the size of the features (6 * 7 * 6), the brain shape and activation patterns were not visible, these features were thus less interpretable and required a quantitative analysis.

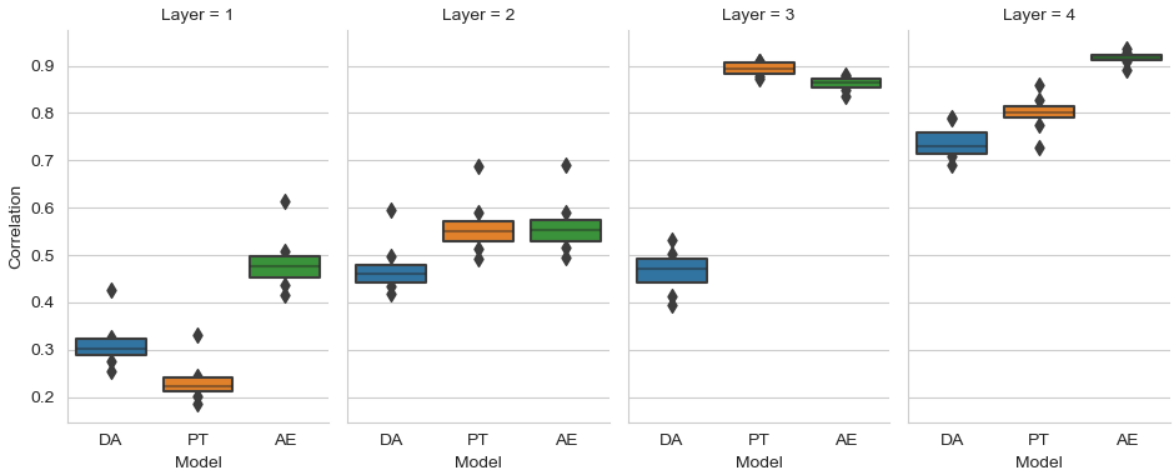

Figure 4.9 – Boxplots of mean correlations between the feature maps of different participants for the eight selected contrasts ('Working Memory': '0-back body', '0-back places', '0-back tools', '2-back body', '2-back tools' , 'Gambling: punish', 'Gambling: reward' and 'Relational: relational') for different models at layer 1, 2, 3 and 4. DA: Default Algorithm initialization ; PT: pretraining initialization ; AE: Baseline AutoEncoder. For Layers 3 and 4, pretrained CNN and baseline CAE show larger correlation between participants than default CNN, meaning a lower attention to individual variabilities.

Mean correlations between the feature maps of the same contrast were computed for each pair of participants. A high mean correlation indicates a higher similarity between the feature maps produced in a given layer of a neural network, and thus potentially, a

higher generalisation power since the feature maps are less different between participants and thus less sensitive to individual variations. Figure 4.9 shows the mean correlations for the 8 selected contrasts and for the first four convolutional layers of the models (different values represent different contrasts). For layers 1 and 2, mean correlations were low (<60%) and not very different between the models even if the pretrained CNN seemed to account more about individual differences than the default model and baseline CAE. The main change was visible at layer 3 where there was an important difference (more than 30% for every contrast) between the mean correlation between the features learned by the default CNN and the pretrained one. The features of this layer seemed more similar between different participants and more generalizable across participants for the pretrained model (mean correlations>80% for all contrasts) than for the default model for which the mean correlations were lower than 50% for every contrast. Correlations started to converge for the fourth layer, but were still lower for the default model.

4.3.5.2 What layers benefit the most from weight transfer from the CAE?

N. of transferred layers	Mean classification accuracy (standard error) (%)
0 (Default initialization)	83.6 (0.61)
1	82.67 (0.45)
2	84.79 (0.52)
3	85.51 (0.8)
4	86.6 (0.4)
Full pretrained model	87.0 (0.51)

Table 4.7 – Classification performance (mean accuracy and standard error, in %) of pretrained models with different numbers of transferred layers on classification of contrasts for HCP dataset sample n=50.

To explore the impact of each layer and the benefits of the baseline weights of the CAE, we tried several experiments with different numbers of frozen layers and several weight transfer configurations: transferring only the weights of the first convolutional layer to transferring the weights of the first four convolutional layers. Performance of the different models with different numbers of transferred layers is shown in Table 4.7. When only the weights of the first layer were transferred, classification performance was lower than

N. of frozen layers	Mean classification accuracy(standard error) (%)
2	86.7 (0.54)
3	86.82 (0.66)
4	86.1 (0.64)
5	80.42 (0.99)

Table 4.8 – Classification performance (mean accuracy and standard error, in %) of pretrained models with different numbers of frozen layers on classification of contrasts for HCP dataset sample n=50.

with other configurations (82.7% of accuracy compared to more than 84% for at least 2 transferred layers). This suggests that features learned by the CAE at this layer were less important for classification. However, when increasing the number of transferred layers, performance started to grow and became closer to the accuracy obtained when transferring all layers (87%). This growth was quite constant and there was no large improvement of performance when transferring the weights of a layer in particular, except when moving from transferring the first layer to the first two layers. Thus, pretraining the deeper layers of the model was beneficial to improve classification performance, probably because of the ability of these layers to extract more general features, less sensitive to individual variations, as we saw above. Transferring the weights of the last convolutional layer (5th) was however not very impactful, performance of model with four transferred layers was very close to the ones of fully pretrained model (86.6% vs 87.0%). We suppose that this layer was important to extract task-related features that were different from the ones learned by the CAE, explaining the limited impacts of transferring the CAE weights.

4.3.5.3 Faster fine-tuning: what happens if we freeze some layers?

Table 4.8 shows the results of the different experiments with different numbers of frozen layers. When we froze the first convolutional layers (from 2 to 4 frozen layers) on the pretrained model, the performance did not decrease. This suggests that the features extracted by the baseline autoencoder for these layers were general enough to perform a classification task with only one fine-tuned convolutional layer in addition to the dense layer. However, when freezing all convolutional layers of the model (5 layers), there was a large drop in terms of performance (86 to 80% of accuracy between freezing 2-4 layers vs 5 layers), this confirmed the observation made before on the difference between the features

extracted by the fifth layer for reconstruction (CAE) and for classification (CNN). In conclusion, the first four convolutional layers of our model extracted more general features whereas the last one extracted deeper and more specific features for classification.

4.4 Discussion

4.4.1 Summary

In this work, we showed the benefits of self-taught learning with a large and variable database on the classification of two large public datasets with different sample sizes and classification tasks. In all cases, pretraining a classifier with an unsupervised task (in our case: reconstruction) was beneficial but the level of improvement varied depending on the classification task and the size of the training dataset.

When sample sizes were small, pretraining always improved the classification performance, regardless of whether the dataset was homogeneous or heterogeneous and of the complexity of the classification task. In medical imaging, where the dimensions of the data are often very large and few samples are typically available due to high financial and human costs, learning a good representation of the data can be very difficult (Thomas et al., 2021). Unsupervised pretraining can thus be helpful by initializing the weights of the CNN to preserve the (brain) structure learned by the autoencoder, and facilitate the learning process. However, when the sample size increases, benefits are less remarkable since the amount of available training data is probably sufficient to learn a good representation.

This observation can also be made for classification tasks. When trying to classify the data in a small number of classes, performance of the pretrained classifier was better but not with a high improvement of performance, even for small sample sizes (*e.g.* 100 participants for task classification). But when trying to separate data into more classes, for a more fine-grained classification, the representation learned during the pretraining was beneficial.

Another benefit of self-taught learning we found was the reduction of the training time. Performance of the pretrained classifier was better even with less training epochs. This was the case for both datasets results which were computed for 500 epochs for the default algorithm and 200 epochs for the pretrained model. This is in line with Neyshabur et al., 2020 in which researchers showed that the pretrained models remain in the same

basin of the loss function when trained on new data and since the weights are already initialized close to a good representation of data, less epochs are necessary to adapt this representation for classification.

Architectures of the models also had an impact on the benefits of self-taught learning. With both datasets, pretrained models performed better using the 5-layers architecture. This effect was studied by Erhan et al., 2010 who showed that, while unsupervised pretraining helps for deep networks with more layers, it appears to hurt for too small networks. The size of the latent space of the CAE with 5-layers being almost 5 times smaller that the 4-layers one, it suggests that only a small subset of features of the input are relevant for predicting the class label.

However, the classification accuracies of the pretrained models were not related to the reconstruction performance of the CAE since the 4-layers CAE reconstructs maps with better precision than the 5-layers CAE. This confirms that the features learned by the 4-layers CAE for reconstruction were not all useful for classification and focusing on a smaller number of features (with 5-layers) facilitates the learning process.

This observation was confirmed by the large drop in performance when freezing the first fifth convolutional layers of the pretrained model and when transferring only part of the layers. Deeper pretrained layers had more impact on classification performance, meaning that the features extracted by these layers were different from those learned by layers initialized with the default algorithm. In particular, the third and fourth convolutional layers showed the best benefits when being transferred, due to the generalizability of the extracted features. This was not the case for the fifth layer, for which features need to be specific to the classification task.

The pretrained model improved the performance in terms of classification due its ability to focus on more generalizable features. By pretraining a model on a large variable dataset such as NeuroVault, we built a model that is less sensitive to the training data and less sensitive to individual differences, thus more generalizable and applicable to new participants.

4.4.2 Limitations

Due to the high computational time required to train a model, we only compared two model architectures (4 and 5-layers). Indeed, training a CAE model can be very time consuming, particularly in our case since we use a large training dataset (N=22,772) and high dimensional data (k=48 * 56 * 48). With the 4-layers model, for 200 epochs it took

approximately 48h to train on 1 GPU. With parallel computing (use of 2 GPUs in parallel), we could hope to shorten this time to 24h, with the cost of using more computing resources. Other types of architectures with different number of fully-connected or convolutional layers could have been tested to see the effect of other latent space sizes as it was done in Erhan et al., 2010.

The main limitation of our work is the classification experiments and datasets we chose. In fMRI, the number of possible labels and thus, classification tasks is very high due to a lack of consensus in the field with respect to standardizing tasks, contrasts and mental concepts (Poldrack et al., 2011b). In our experiments, we used the labels provided by NeuroVault as specified in the original studies (Van Essen et al., 2013; Varoquaux et al., 2018). We chose to compare multiple types of classification on the HCP dataset to illustrate different approaches used in the field or that were used by other studies (Y. Gao et al., 2019; Thomas et al., 2023). For BrainPedia, a multi-label decoding was performed in the original study since multiple concepts are associated with most maps. Labels we had access to were then the list of labels associated with each map. To be able to compare our results with those of the homogeneous dataset (HCP), we chose to classify these as unique labels, which was less complex and less precise in practice. This type of issue is due to the lack of harmonization in the way tasks and cognitive processes are defined. Using ontologies such as Cognitive Atlas (Poldrack et al., 2011b), NeuroVault annotations could be harmonized and enriched, as it was done by Menuet et al., 2022 by mapping the original labels to target ones from Cognitive Atlas or Walters et al., 2022 in which cognitive conditions were annotated by a group of expert using the same atlas.

In neuroimaging, many sources of variability can impact the results of an experiment and the generalizability of the results. Here, we investigated the generalizability of our model by assessing the benefits of pretraining on a heterogeneous dataset (BrainPedia). While this dataset was heterogeneous in terms of the studies that were included, all maps were obtained using the same processing pipeline. Multiple studies have shown that the exact pipeline used to obtain an fMRI result can have a non-negligeable impact on fMRI statistic maps (Carp, 2012a; Botvinik-Nezer et al., 2020). In the future, investigating performance of classification on a more variable target dataset with statistic maps from different studies but also processed using different pipelines would be of great interest. In a recent study (Vu et al., 2020), the authors tried to compare the performance of different classifiers trained on fMRI 3D volumes series obtained with various scenarios of minimal preprocessing pipelines. A similar experiment was recently made by Li et al., 2023 who

found that preprocessing pipeline selection can impact the performance of a supervised classifier. Comparing the adaptation capacities of models on volumes preprocessed with different pipelines could be also interesting to evaluate the impact of analytical variability on deep learning with fMRI and to see if the generalizability of our pretrained models also works for inter-pipeline differences.

Note that self-supervised (instead of self-taught) learning could have also been used to pretrain our model, as it was done by Thomas et al., 2022 who designed self-supervised learning frameworks, inspired by the field of natural language processing, to pretrain mental state decoding models. Self-supervised learning is a supervised machine learning setting where the supervision is generated directly from the data and the model is pretrained using a supervised surrogate task. Self-supervised is particularly relevant if the surrogate task is close to the final one targeted by the user, *e.g.* if they can share the same feature representation. It is possible that, by designing a relevant supervised surrogate task that could be relevant for all very diverse usage of our model, the pretrained model would have performed better than the one presented in this article. Designing and experimenting with such a surrogate supervised task could be interesting for future work.

In our self-taught context, using unsupervised models could allow us to build a space capturing the similarities and differences of statistic maps, *i.e.* to learn a robust latent representation of the important features of statistic maps in a specific context. By adding other constraints to this latent space and/or choosing an adapted pretraining dataset, we could use this for other purposes than brain decoding. For instance, building a space that captures the analytical variability in statistic maps could help us understand the difference between the pipelines but also identify the more robust pipelines. Future works will focus on building such a space with specific constraints to evaluate distance between different pipelines.

Take-home Message

- Transfer learning, and in particular self-taught learning, is a solution to make use of the unlabeled statistic maps shared on public databases.
- By re-using these data in such framework, we obtained better performance than standard supervised models in classification experiments.
- Representations learned by the pretrained model were more generalizable and less sensitive to the sources of variations in the target data (here, different participants, different studies and potentially acquisition parameters).
- This framework could be adapted to other target tasks (*e.g.* disease classification, other decoding tasks, etc.). We shared the pretrained CAE with the community on Zenodo (see Germani et al., 2023) to facilitate re-use.

Chapter 5

Mitigating analytical variability in fMRI results with style transfer

This chapter was the subject of a paper that will soon be submitted to *Human Brain Mapping*.

- **Title:** Mitigating analytical variability in fMRI results with style transfer
- **Authors:** Elodie Germani, Camille Maumet*, Elisa Fromont*
- **HAL:** inserm-04531405
- **Code:** swh:1:dir:75ffda70e008d7efe57b21db93e61007d77330f5
- **Contributions (Credit taxonomy):** Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualisation, Manuscript writing.

* Joint senior authorship.

5.1 Introduction

In the previous chapter, we showed that large databases can be leveraged to build more generalizable representations of fMRI statistic maps and to improve performance of brain decoding models. This study was an example of data re-use for deep learning tasks, in which the presence of variability in the training data is beneficial as it allows the model to learn more generic features, and thus prevent over-fitting and increase generalizability. In other data re-use settings, for instance meta- or mega-analyses (Costafreda, 2009), the goal is to perform a larger statistical analysis by re-using data from multiple previous studies. These analyses would provide more flexibility as to which research question can

be investigated, and increase the sample sizes, leading to higher statistical power and more robust results.

Usually, mega-analyses are performed using raw data coming from different sources, which are then processed and analysed using the same pipeline. As data sharing becomes more prevalent, in particular for derived data, *i.e.* after preprocessing and statistical analysis, these analyses could also be built by combining subject-level or group-level statistic maps shared from different studies. In fMRI, due to the high flexibility of the analytical pipelines (Carp, 2012a), derived data shared on public databases often come from different pipelines. However, different pipelines lead to different results (see Chapter 2) and combining results from different pipelines in mega-analyses can lead to a higher risk of false positive findings (Rolland et al., 2022). To benefit from these large amount of derived data available, it is necessary to find a way to mitigate the effect of analytical variability.

In Chapter 3, we showed that Image-to-image transition (I2I) frameworks, based on neural style transfer, were giving promising results in many conversion tasks in medical imaging, *e.g.* converting data between imaging modalities, image denoising or data harmonization. Considering the achievements of these models in modality transition, which involves transitioning between distinct acquisition modalities, there is reason to anticipate their success in transitioning between other image types, such as statistic maps coming from different analysis pipelines. In this work, we propose to use I2I frameworks to convert statistic maps between pipelines and build more valid mega-analyses.

To be useful in real practice, the proposed method should rely on unpaired data (*i.e.* could be trained without access to the ground-truth target images) and perform multi-domain transitions (*i.e.* learn multiple transfers using a single model). However, to the best of our knowledge, this application of I2I to conversion of data between different analysis pipelines is new and off the shelf I2I methods do not directly apply as these were not designed on the same type of data and were not evaluated with the same metrics. Thus, we test and compare other frameworks than multi-domain unsupervised ones, for instance using supervised datasets or one-to-one transitions. In particular, we study frameworks based on GAN (Goodfellow et al., 2014), namely Pix2Pix (Isola et al., 2017), CycleGAN (Zhu et al., 2017), StarGAN (Choi et al., 2018), and also design a DDPM (Ho et al., 2020) framework to tackle our task.

DDPM models have achieved state-of-the-art performance in synthesizing natural images, overpassing GAN by producing complex and diverse images (Nichol et al., 2021),

while reducing the risk of modality collapse (Li et al., 2022). But, these models are challenging to control when the objective is to generate images that maintain the intrinsic properties of the source images while transferring the extrinsic properties to the target domain, *i.e.* in I2I frameworks. DDPM are iterative generative models, *i.e.* they learn to model the transition from a Gaussian distribution to a target data distribution. Thus, data generated by the DDPM depend on the initial samples drawn from the Gaussian distribution, usually done at random. In this context, we adapt an existing conditional DDPM, initially built for conditional generation, to perform I2I. We compare the performance of such model with GAN and explore the impact of several modifications of the framework on the conversion performance.

In the following section, we describe the dataset used for our experiments and the different frameworks implemented. We also detail the different variations of DDPM-based I2I frameworks that we explored and the evaluation metrics that we used. In section 5.3, we compare the results of GAN-based frameworks and those of DDPM-based frameworks. Finally, in section 5.4, we discuss these results and conclude on the success of style transfer in the context of pipeline transition.

5.2 Materials and Methods

5.2.1 Dataset

In this work, we used group-level statistic maps from the HCP multi-pipeline dataset, that we will present in greater details in Chapter 6. We explored in particular the data from four different pipelines that differed in terms of software package (SPM (Penny et al., 2011) or FSL (Jenkinson et al., 2012)) and presence or absence of the derivatives of the Haemodynamic Response Function (HRF) for the first-level analysis. We used all the available group-level statistic maps ($N = 1,000$) for each pipeline for the task "right-hand". In the following, these pipelines will be denoted as "software-derivatives", for instance "fsl-1" means use of FSL software package and HRF derivatives.

The selected group-level statistic maps were resampled to a size of 48 x 56 x 48 and masked using the intersection mask of all groups. The voxel values were normalized between -1 and 1 for each statistic maps using a min-max operation. The 1,000 groups were split into train, valid and test with a 90/8/2 ratio and all models were trained and evaluated on the same sets. Further investigation about possible data leakage across

groups is provided in Appendix E (Figure E.1).

5.2.2 Generative Adversarial Network (GAN) frameworks

First, we assessed the potential of GAN-based frameworks to convert statistic maps between pipelines. In particular, we evaluated the performance of Pix2Pix (Isola et al., 2017), CycleGAN (Zhu et al., 2017) and StarGAN (Choi et al., 2018). These frameworks are described in larger details in Chapter 3, Section 3.4.2.2. We provide a quick description of the main properties of these models in Table 5.1.

Framework	Learning	Transition	Loss
Pix2Pix (Isola et al., 2017)	Supervised	One-to-one	Adversarial Reconstruction
CycleGAN (Zhu et al., 2017)	Unsupervised	One-to-one	Adversarial Cyclic
StarGAN (Choi et al., 2018)	Unsupervised	Multi-domain	Adversarial Cyclic Classification

Table 5.1 – Description of Generative Adversarial Network (GAN)-based frameworks

We used the default architecture of these models, as described in their respective papers, and we only modified the 2-dimensional convolutions and batch normalization layers to 3-dimensional, to cope with our 3-dimensional statistic maps.

5.2.3 Denoising Diffusion Probabilistic Model (DDPM) frameworks

Due to the promising performance of DDPM in natural images and medical imaging (see Chapter 3, Section 3.4.2.3), we also assessed the potential of DDPM-based frameworks. However, there is only few frameworks developed for this application, and most of them rely on paired datasets (Saharia et al., 2022) or one-to-one transitions (Pan et al., 2023). Thus, to perform multi-domain transitions, we used traditional conditional DDPM that we adapted to answer I2I tasks. In particular, we used the conditional DDPM from Ho et al., 2021, which generates images conditioned using a one-hot encoding of the class. We also extended this model to a conditioning based on the latent space of the classifier, inspired from Preechakul et al., 2022. Both are unsupervised frameworks, learning multi-domains transitions. A more detailed description of the original frameworks from

Ho et al., 2021 and Preechakul et al., 2022 is available in Chapter 3, Section 3.4.2.2, and we summarize their main properties in Table 5.2.

Framework	Conditioning	Target images
Ho et al., 2021	One-hot	None
Preechakul et al., 2022	Classifier-conditional	N=1

Table 5.2 – Description of Denoising Diffusion Probabilistic Model (DDPM)-based frameworks

In Figure 5.1, we illustrate the design of DDPM-based frameworks, with the main modifications applied to the basis of Ho et al., 2021. Figure 5.1 (A), (C) and (D) represent the conditional diffusion used in Ho et al., 2021, that we enhanced using source content preservation and classifier conditioning (Figure 5.1 (B)).

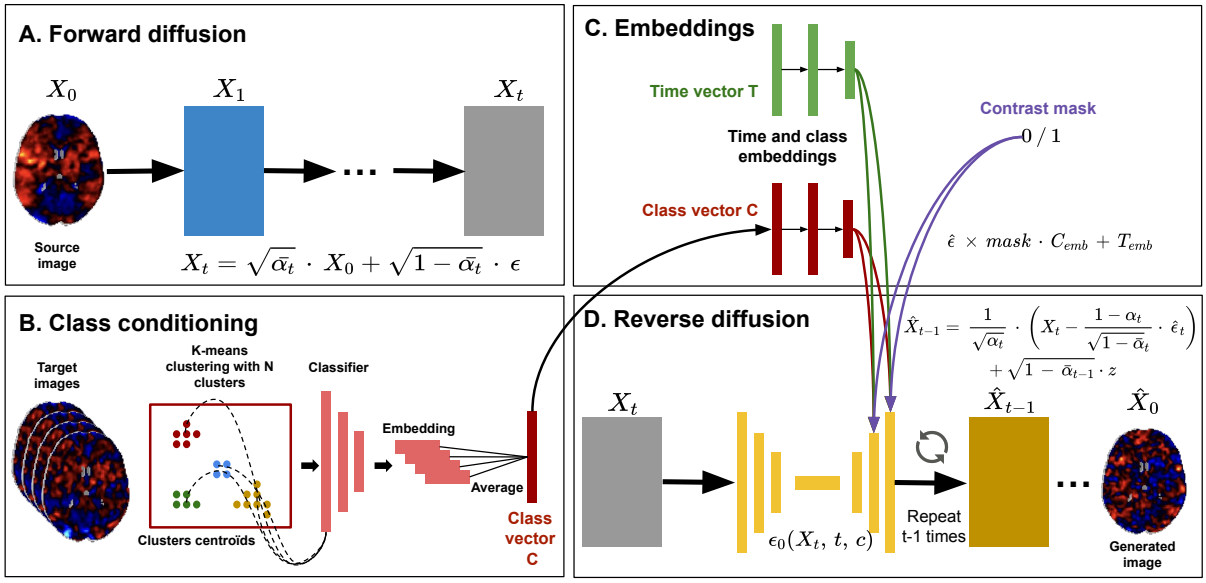

Figure 5.1 – Diagram of the workflow. During the forward diffusion (A), original maps X_0 are turned into X_t after t steps of noise addition ϵ. (B) Class conditioning uses latent vectors extracted from a classifier. These are averaged across N images, which are the centroids of N clusters identified using a K-Means algorithm. (C) Time and class are embedded using two Multi-Layers Perceptrons (MLP). A mask is applied to the class conditioning vector to jointly train an unconditional model with a pre-defined probability. (D) During the reverse diffusion, the neural network $\epsilon_\theta(X_t, t, c)$ learns to predict the noise added to the image and reconstructs X_{t-1} iteratively until $t = 0$.

Source content preservation. To adapt these models to I2I, we made several modifications. First, our main objective was to find a solution to generate images that still contained the intrinsic properties of the source image. In Saharia et al., 2022, authors concatenated the source image along with random Gaussian noise to initialize the diffusion. Here, we fixed the initial state of the DDPM by directly using the forward diffusion process to generate a noisy version of the source image X_t (Equation 3.3). Then, the noisy source image is iteratively denoised using the predicted noise and the reverse diffusion process (Equation 3.4) with an additional conditioning on the target domain.

Classifier conditioning. We also developed an extension of the model from Ho et al., 2021 to condition the generation based on the latent space of a classifier (see Figure 5.1 (B)). Indeed, in Ho et al., 2021, the diffusion is conditioned using a one-hot encoding of the domain, which decreases the diversity of samples. In Preechakul et al., 2022, a semantic encoder is used to guide sampling. Thus, we extended this idea by using a pretrained CNN that identifies the pipeline used to obtain the statistic maps (*i.e.* their domain) to condition the model. The features are extracted just before the fully connected layer, to get a latent vector with the most important features that distinguish images across pipelines.

Multi-target images. To condition on the latent space of this classifier during sampling, target images must be selected. In Choi et al., 2021, authors showed that conditioning on multiple images generates images that share coarse or fine features with the target ones depending on the number of selected images. Selecting multiple target images to convert images between domains can help to generate images that represent the diversity of the target domain. In practice, the whole set of images available in the target domain could be used. This is impractical for large datasets and might lead the model to focus on specific patterns of the target domain if these are over-represented in the dataset. Here, we implemented several variations to explore the impact of the choice of target images.

- **Number of target images**: N=5, 10 or 20.
- **Target images selection**: random (∞), using a K-means algorithm, or using a K-Nearest Neighbors algorithm.

For the target image selection, we proposed several algorithms. We used K-Means algorithm (MacQueen, 1967) to identify N clusters of images in the target domain (see

Figure 5.1 (B)). Then, we extract the centroïd of these clusters and average their latent vector for conditioning. We also compared the selection process with a random sampling of target images and with a sampling based on the identification of images that are close to the source image using a K-Nearest Neighbors algorithm (Mucherino et al., 2009).

Details regarding architecture and training of the models are available in Appendix E.

5.2.4 Evaluation of performance

We evaluated the performance of the frameworks using different metrics. In the following equations, we use X_A, X_B and X_{AB} to respectively define the source image, target image and translated image.

- Pearson's correlation (Corr.) in percent

$$r = \frac{\sum_{i=1}^{n}(X_{AB_i} - \overline{X_{AB}})(X_{B_i} - \overline{X_B})}{\sqrt{\sum_{i=1}^{n}(X_{AB_i} - \overline{X_{AB}})^2}\sqrt{\sum_{i=1}^{n}(X_{B_i} - \overline{X_B})^2}} \tag{5.1}$$

- Mean Squared Error (MSE)

$$MSE = \frac{1}{n} * \sum_{i=1}^{n}(X_{AB_i} - X_{B_i})^2 \tag{5.2}$$

- Inception Score (IS) (Salimans et al., 2016) computed using the pipeline classifier. In the following equation, X refer to any generated image, and Y the corresponding target label.

$$IS(G) = exp(\mathbb{E}_{x \sim p_g} D_{KL}(p(Y|X) \parallel p(Y))) \tag{5.3}$$

The first two metrics were used to study the adequacy of generated images to the ground truth target, whereas IS was used to explore the confidence of the conditional class predictions (quality) and the integral of the marginal probability of the predicted classes (diversity).

		fsl-1 → spm-0		spm-0 → fsl-1		fsl-1 → spm-1		fsl-1 → fsl-0	
	IS	Corr.	MSE	Corr.	MSE	Corr.	MSE	Corr.	MSE
Initial	*3.69*	*76.2*	*0.008*	*76.2*	*0.008*	*82.6*	*0.004*	*91.0*	*0.0022*
Pix2Pix	-	**91.4**	**0.0029**	**89.2**	**0.0015**	**90.3**	**0.0026**	**97.4**	**0.0006**
CycleGAN	-	86.0	0.0046	66.6	0.0052	71.0	0.0069	71.8	0.0047
StarGAN	3.63	90.6	0.0034	87.1	0.0021	87.7	0.0036	91.8	0.0016

Table 5.3 – Performance associated with four transfers for Generative Adversarial Network (GAN)-based frameworks. IS means "Inception Score" across all transfers. Pearson's correlation (%) and Mean Squared Error (MSE) computed between generated and ground-truth target image for 20 images per transfer. *Initial* represents the metrics between the source image (before transfer) and the ground-truth target image. **Boldface marks the top model**. Note: Inception score was not computed for Pix2Pix and CycleGAN as different transfers are learnt by different models.

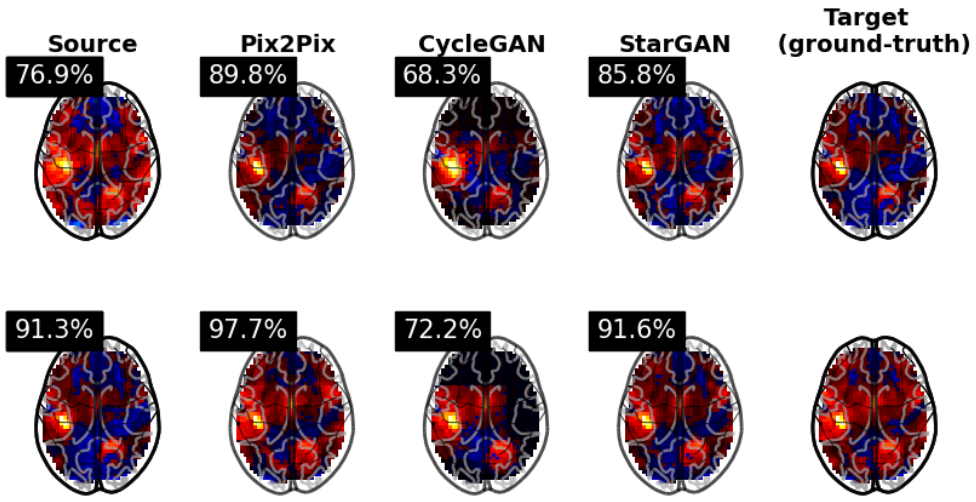

Figure 5.2 – Generated images for two transfer and different competitors: Pix2Pix (Isola et al., 2017), CycleGAN (Zhu et al., 2017) and starGAN (Choi et al., 2018). Correlation with target ground-truth are indicated below generated and source images.

5.3 Results

5.3.1 Generative Adversarial Network (GAN) frameworks

In Table 5.3, we show the performance of GAN-based frameworks for four transfers, between pipelines with: different HRF and different software (columns 1-4), same HRF

and different software (columns 4-6) and, different HRF and same software (columns 6-8). Overall, using Pix2Pix (Isola et al., 2017) and StarGAN (Choi et al., 2018), the conversion of statistic maps between pipelines seem to be sucessful, with increased correlations between target and generated maps compared to correlations between source and target (similar observations can be done with decreased MSE), *e.g.* 91.4% for target-generated compared to 76.2% for source-target with Pix2Pix for conversion "fsl-1 to spm-0".

We can point out the large superiority of the supervised framework (Pix2Pix (Isola et al., 2017)) compared to the other, which are all unsupervised. By benefiting from paired data, this model outpass the performance of all the other frameworks, and even the initial metrics obtained when comparing with the source image. Correlations between target and generated images are close to 0.9, which is nearly perfect. On the other hand, the CycleGAN (Zhu et al., 2017) framework gives surprising results, relatively low compared to the other GAN-based frameworks. While it makes use of a cyclic-loss in unsupervised settings, similarly to StarGAN, this framework only learn transfers between two domains. We can suppose that StarGAN benefit from learning from other transfers and from the additional classification loss, leading to higher performance in similar settings.

In Figure 5.2, we illustrate two transfers: (first row) between pipelines with different HRF and different software packages (spm-0 to fsl-1) and (second row) between pipelines with different HRF (fsl-1 to fsl-0). Maps generated using Pix2Pix (Isola et al., 2017) remain closer to the target ground-truth, with more similar patterns, as stated by the similarity metrics.

5.3.2 Denoising Diffusion Probabilistic Model (DDPM) frameworks

In Table 5.4, we show the performance of DDPM-based frameworks for the same four transfers as in Table 5.3. Performance of different frameworks are compared: one-hot encoding conditioning from Ho et al., 2021, classifier-conditioning with $N = 1$ target image selected randomly, inspired from Preechakul et al., 2022, and classifier-conditioning with $N = 10$ target images selected randomly (named $N = 10, \infty$ in the Table).

Using such frameworks, the conversion between pipelines seems more difficult. While all models succeed in changing the class identified by a pipeline classifier to the target domain, the success of the conversion in terms of similarity to the target ground-truth image is variable across transfers. For instance, all DDPM-based frameworks succeed in

		fsl-1 → spm-0		spm-0 → fsl-1		fsl-1 → spm-1		fsl-1 → fsl-0	
	IS	Corr.	MSE	Corr.	MSE	Corr.	MSE	Corr.	MSE
Initial	*3.69*	*76.2*	*0.008*	*76.2*	*0.008*	*82.6*	*0.004*	*91.0*	*0.0022*
One-hot	3.66	83.8	0.0096	75.0	0.0048	78.7	0.0087	81.1	0.0044
N=1	3.70	85.5	0.0053	77.8	0.0035	79.9	0.0072	82.8	0.0033
N=10, ∞	**3.86**	**86.5**	**0.0047**	**79.0**	**0.0032**	**81.8**	**0.0049**	**84.3**	**0.0028**

Table 5.4 – Performance associated with four transfers for Denoising Diffusion Probabilistic Model (DDPM)-based frameworks. IS means "Inception Score" across all transfers. Pearson's correlation (%) and Mean Squared Error (MSE) computed between generated and ground-truth target image for 20 images per transfer. *Initial* represents the metrics between the source image (before transfer) and the ground-truth target image. **Boldface marks the top model**.

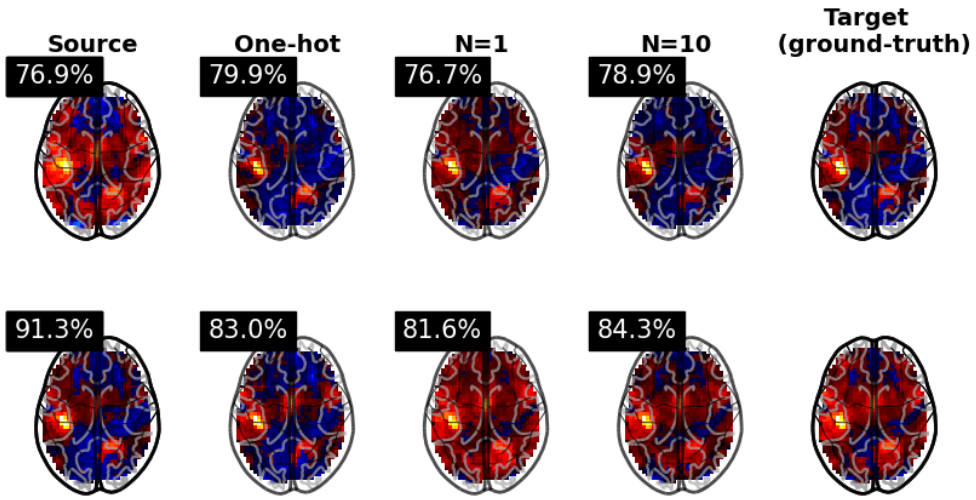

Figure 5.3 – Generated images for two transfer and different competitors: conditioning with one-hot encoding (Ho et al., 2021), with a classifier and N=1 (Preechakul et al., 2022) and N=20 with random selection. Correlation with target ground-truth are indicated below generated and source images.

converting statistic maps for the transfer "fsl-1 to spm-0", while none is successful for the transfer "fsl-1 to fsl-0". These low performance could be explained by the difficulty of the models to learn differences between close pipelines. In Table 5.5 and Figure 5.4, we show the performance of the pipeline classifier and compare the similarity of features, as done in Chapter 4, Figure 4.9. In particular, we observe that features learned at Layer 4

(*i.e.* the features used for conditioning) are close for pipelines sharing the same software, which might explain the difficulty to rely on these features to perform transfer.

Pipelines	Layer 1	Layer 2	Layer 3	Layer 4
Same software, different parameters				
fsl-5-0-0 / fsl-5-0-1	86.5	91.4	95.4	99.2
spm-5-0-0 / spm-5-0-1	86.5	90.9	94.2	98.4
Same parameters, different software				
fsl-5-0-0 spm-5-0-0	88.8	88.2	93.6	98.2
fsl-5-0-1 spm-5-0-1	84.8	85.8	92.4	98.0
Different software, different parameters				
fsl-5-0-0 spm-5-0-1	74.5	81.0	88.7	97.1
fsl-5-0-1 spm-5-0-0	74.8	77.7	88.2	97.3

Table 5.5 – Mean correlations between features maps learned at each layers for each pair of pipelines

The use of a DDPM with classifier-conditioning and multiple target images ($N = 10, \infty$) seems to improve performance compared to other DDPM models. Both quality and diversity of images is increased ($IS = 3.86$), and in terms of similarity to the ground-truth target image, this frameworks outperforms the other DDPM models by up to 4% in correlations between target ground-truth and generated image compared to Ho et al., 2021 for transfer "spm-0 to fsl-1" and up to 3% for "fsl-1 to spm-0".

The first row of Figure 5.3 illustrates a transfer between pipelines with different HRF and different software packages ("spm-0 to fsl-1"). The second row shows a transfer between pipelines with different HRF ("fsl-1 to fsl-0"). The DDPM with multiple target images generates statistic maps close to the ground-truth for both transfer, representing the intrinsic properties of the map while modifying its extrinsic properties to the target domain. Using the one-hot encoding conditioning, the generated statistic maps seem far from the target image, failing to represent the whole characteristics of the target domain. When using only one target image, statistic maps are more similar to the target in terms of activation area.

The performance of such frameworks remain highly inferior to the ones obtained with Pix2Pix (Isola et al., 2017) or StarGAN (Choi et al., 2018). This superiority can be explained by the differences between frameworks: GAN-based methods use adversarial training and StarGAN improves this by using a classifier loss and a cyclic-reconstruction loss. Moreover, GAN sampling rely on the source image directly and do not require to set an initial state, which might facilitate the source content preservation.

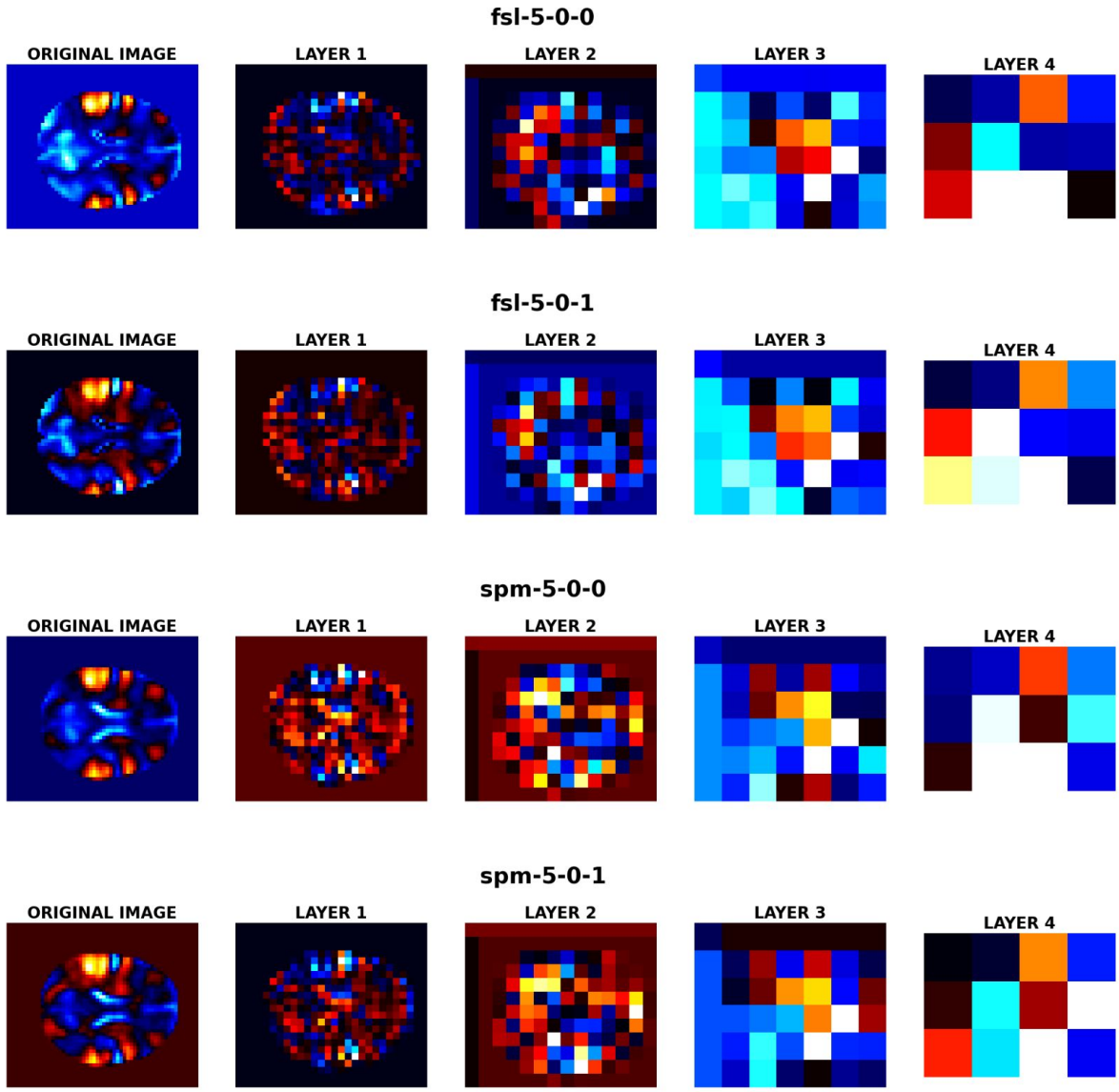

Figure 5.4 – Original mean statistic maps (column 1) and mean feature maps across groups learned by the pipeline classifier for the first 4 convolutional layers for the different classes. Pipelines with the same software show similar feature maps at Layer 2 and 3.

5.3.3 Impact of multi-target images

		fsl-1 → spm-0		spm-0 → fsl-1		fsl-1 → spm-1		fsl-1 → fsl-0	
	IS	Corr.	MSE	Corr.	MSE	Corr.	MSE	Corr.	MSE
N=5, ∞	3.89	86.5	0.0046	79.1	0.003	82.0	0.0051	84.2	0.0031
N=10, ∞	3.86	86.5	0.0047	79.0	0.0032	81.8	0.0049	84.3	0.0028
N=20, ∞	3.85	86.7	0.0048	79.3	0.003	81.5	0.0051	84.4	0.0028
N=5, Kmeans	3.86	86.4	0.0046	78.7	0.003	81.2	0.0051	84.5	0.0031
N=10, Kmeans	3.86	86.1	0.0047	79.0	0.0032	81.2	0.0049	84.1	0.0028
N=20, Kmeans	3.87	86.1	0.0048	79.2	0.003	81.3	0.0051	83.9	0.0028
N=10, KNN	3.75	84.9	0.0047	78.7	0.0032	81.6	0.0049	83.6	0.0028

Table 5.6 – Performance associated with four transfers with DDPM-based frameworks with different implementation. IS means "Inception Score" across all transfers. Pearson's correlation (%) and Mean Squared Error (MSE) computed between generated and ground-truth target image for 20 images per transfer. *Initial* represents the metrics between the source image (before transfer) and the ground-truth target image. ∞ means random sampling.

In Table 5.6, we show the influence of the number of target images and of the selection methods. The number of images does not seem to impact the performance, correlations are very similar between $N = 5, N = 10$ and $N = 20$. Performing selection using K-Means algorithm does not seem to improve performance compared to a random selection, for any N values, probably due to the low diversity in our dataset. However, selection using a K-Nearest Neighbors (KNN) algorithm decreases the performance from 1.6%, meaning that the diversity of target images is beneficial for a good transfer.

5.4 Discussion

In this work, we made the assumption that statistic maps could be converted between pipelines to facilitate re-use of derived data in mega-analyses (Costafreda, 2009). We explored different frameworks based on GAN and DDPM with the aim to develop an unsupervised multi-domain framework that researchers could use to convert the derived data available in public databases such as NeuroVault (Gorgolewski et al., 2015). Our results are promising, with satisfying performance in transferring statistic maps between pipelines with distant results (*e.g.* from different software packages). In these cases, generated statistic maps were closer to the target image than the original ones, and generated statistic maps were all classified in the target domain by the pipeline classifier.

In Chapter 8, in a follow-up work of Rolland et al., 2022, we will see that combining data from different pipelines in mega-analyses leads to invalid results with different levels of false positive rates, and that studies combining data from different software packages are the ones that led to the largest false positive rates, and thus largest invalidity. This possibility to transfer statistic maps between software packages using I2I frameworks is thus highly hopeful for the future of data re-use.

We compared several frameworks and found that, in our case, GAN-based frameworks always overpass DDPM-based ones in terms of adequation with target ground-truth image. While the largest performance of DDPM was demonstrated in many papers (Dhariwal et al., 2021; Müller-Franzes et al., 2023), we believe that our particular results are related to the specific properties of the task and data. These two studies showed the superiority of DDPM compared to GAN for the task of image synthesis, in both natural and medical images, but not for I2I. The traditional sampling strategy of DDPM is not suited for such task, as it relies on random noise, which makes it difficult to maintain intrinsic properties of the source images while changing the style. On the contrary, GAN sampling relies on the source images directly and do not require to set an initial state, which might facilitate the source content preservation. In addition, DDPM are trained to minimize a MSE loss between the predicted noise and the actual noise added to the image, without any component related to style transfer, whereas in the GAN frameworks, and in particular StarGAN (Choi et al., 2018), the classifier loss seems to greatly improve performance. Another issue related to DDPM is the high dimensionality of images, here 3-dimensional images with hundreds of thousands of values, which, associated with the large number of trainable parameters of the model, makes it difficult to train performing models. Recently, the potential of latent diffusion models was shown, these frameworks act in the latent space of a Variational AutoEncoder to reduce the size of data and facilitate training (Rombach et al., 2022).

Across GAN-based frameworks, we obtained better performance with the supervised framework compared to the unsupervised ones, in particular for conversion between pipelines giving already close results (*e.g.* same software package, different parameters). However, gathering paired data is impractical and far from real life practice. In large databases, for instance NeuroVault, we have no information about the pipeline used to obtain statistic maps and potentially no access to raw data to build paired datasets. The goal is to build a model that could be applied on two or more datasets with different statistic maps of the same task, but obtained with different pipelines. In such unsupervised

settings, performance of StarGAN remain statisfying, the model succeeds in generating data that are close to the target image for all transfers, and closer than the source image in long-distance transfer. We believe that this model could be a good candidate for further development in real-life practice.

In this study, we showed the ability of the frameworks to convert unseen statistic maps of the same task (here, *right-hand*). We started to test the generalizability of the frameworks to other tasks, for instance *right-foot*. Ideally, researchers would be able to re-use a framework trained to convert statistic maps of a task for another one. For now, our results show that the StarGAN framework could not be applied to another task, as it leads to generated statistic maps with low correlation with their corresponding target maps (see Supplementary Table E.1). These results makes us suppose that the mapping from a pipeline to another is different between tasks. Future works would be needed to explore these relationships between pipelines, in order to develop a more generalizable framework. This exploration of the stability of relationships between pipelines will be treated in Chapter 7.

Take-home Message

- We explore the ability to convert fMRI maps between pipelines using generative models (GAN-based and DDPM-based frameworks, in supervised and unsupervised settings).
- To enhance DDPM conversion performance, we explore several modifications of traditional DDPM frameworks by conditioning on multiple target images in the latent space of a classifier.
- Our results show that images can be converted successfully using DDPMs, but with lower similarity with the ground-truth target compared to GANs, in particular in supervised settings.

Part III

How to explore the fMRI analytical space?

Chapter 6

The HCP multi-pipeline dataset: an opportunity to investigate analytical variability in fMRI data analysis

---

This chapter was the subject of a paper that will soon be submitted to *Scientific Data*:

- **Title:** The HCP multi-pipeline dataset: an opportunity to investigate analytical variability in fMRI data analysis
- **Authors:** Elodie Germani, Elisa Fromont, Pierre Maurel*, Camille Maumet*
- **HAL:** inserm-04356768.
- **Code:** swh:1:snp:17870c3d782aa25a7ffdd6165fe27ce6eac6c90b
- **Data:** currently working with our DPO for sharing on Public nEUro
- **Contributions (Credit taxonomy):** Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualisation, Manuscript writing.

* Joint senior authorship.

6.1 Introduction

As we saw in the previous chapters (see Chapters 1 and 2), neuroimaging data, such as functional Magnetic Resonance Imaging (fMRI) data, can be used for a wide range

of application, including diagnosis (Yin et al., 2022) or brain decoding (*i.e.* identifying stimuli and cognitive states from brain activities) (Firat et al., 2014). But the workflows used to analyze these data are highly complex and flexible. Different tools and algorithms were developed over the years, leaving researchers with many possible choices at each step of an analysis (Carp, 2012a) (see Chapter 1). This flexibility of analyses pipelines induces a phenomenon called "analytical variability", which describe the variations of the results obtained when varying the pipeline used to process and analyze data (see Chapter 2). As there is usually no ground-truth that can be used to benchmark pipeline results, this phenomenon calls for a better understanding of the pipeline-space to try to identify the cause of the observed differences amongst the final results.

The pipeline-space is especially large (Carp, 2012b) and challenging to explore due to its interaction with other properties of a dataset: for instance, with sample size and sampling uncertainty (Klau et al., 2020) or even with the research question (Botvinik-Nezer et al., 2020). However, due to the high computational cost of storing and analyzing task-fMRI data, recent studies investigating analytical variability in neuroimaging focused on a restricted number of participants (N=108, N=30, N=15, and N=10 respectively for Botvinik-Nezer et al., 2020; Li et al., 2021; Carp, 2012a; Xu et al., 2023) and cognitive tasks (one paradigm for Botvinik-Nezer et al., 2020; Carp, 2012a with respectively k=9 and k=1 contrasts and use of resting-state fMRI for Li et al., 2021; Xu et al., 2023).

Multiple efforts for collecting datasets with larger number of participants have arisen in the field of neuroimaging in the past 10 years with for instance the Human Connectome Project (HCP) (Van Essen et al., 2013) or the UK Biobank (Sudlow et al., 2015; Miller et al., 2016). In particular, the HCP Young Adult most recent releases provide task-fMRI data for more than 1,000 participants and for different tasks and cognitive processes. These data are also available as minimally processed versions, *i.e.* preprocessed using a common pipeline chosen by the HCP collaborators (Glasser et al., 2016). In brief, this pipeline consists in the following steps: removal of spatial distortions, volumes realignment to correct for participant motion, registration of the functional volumes to the structural one, bias field reduction, normalization to a global mean and masking using a structural brain mask computed in parallel.

A set of group-level statistic maps of the HCP-Young Adult have also been made publicly available (see NeuroVault Collection 457 (Collection n°457, 2015) and corresponding publication (Van Essen et al., 2013)). These were obtained using data from a subset of the participants (68 subjects scanned during the first quarter (Q1) of Phase II data collection.

Z-scored statistic maps are available for all base contrasts (23 different contrasts) using a single analysis pipeline. This is beneficial for studying individual differences and contrasts but it does not allow for analytical variability studies for which multiple pipelines are needed, or to perform other analyses such as group-level analyses that could be used to explore interaction with sampling uncertainty or sample size.

Statistic maps published during the Neuroimaging Analysis Replication and Prediction Study (NARPS) study (Botvinik-Nezer et al., 2020) are also publicly available on NeuroVault (Gorgolewski et al., 2015) with one collection per team. For each of the 70 teams, 9 group-level statistic maps are shared (one per research hypothesis) based on two groups of N=54 participants. Additionally, for a limited number of teams (K=4), subject-level contrast maps are also available. The pipeline space studied in this dataset is unconstrained since teams were instructed to use their usual pipelines to analyze the data.

In this Chapter, we describe the *HCP multi-pipeline dataset*, composed of a large number of subject and group-level statistic maps and representing a non-exhaustive but controlled part of the pipeline space. Contrast and statistic maps were obtained for the 5 contrasts of the motor task of the HCP for the 1,080 participants of the S1200 release, with 24 analysis pipelines that differ on a predefined set of parameters as typically used in the literature. We also computed group-level contrast and statistic maps for 1,000 randomly sampled groups of 50 participants for each pipeline and contrast.

While solutions have been proposed to standardize fMRI preprocessing (*e.g.* fMRIprep (Esteban et al., 2019)), practitioners still face multiple choices regarding first-level statistical analyses. Here, we focus on a set of parameters that often varies across pipelines and this even when standardized preprocessing are used: smoothing kernels, HRF modelling and the inclusion/exclusion of motion regressors as nuisance covariates. Group-level statistical analyses were performed uniformly for all pipelines.

6.2 Methods

6.2.1 Raw Data: the Human Connectome Project

This work was performed using data from the Human Connectome Project Young Adult (Van Essen et al., 2013). Written informed consent was obtained from participants and the original study was approved by the Washington University Institutional Review

Board. We agreed to the Open Access Data Use Terms available at *Human Connectome Project: Data Usage Agreement* 2013.

The HCP Young Adult aimed to study and share data from young adults (ages 22-35) from families with twins and non-twin siblings, using a protocol that included structural and functional magnetic resonance imaging (MRI, fMRI), diffusion tensor imaging at 3 Tesla (3T) and behavioral and genetic testing. The S1200 release includes behavioral and 3T MR imaging data from 1206 healthy young adult participants (1113 with structural MR scans) collected in 2012-2015.

Unprocessed anatomical T1-weighted (T1w) and task-fMRI data (Moeller et al., 2010; Feinberg et al., 2010; Setsompop et al., 2012; Xu et al., 2012) were used in this work. The task-fMRI data includes seven tasks, each performed in two separate runs. Among these tasks, we selected data from the motor task in which participants were presented with visual cues asking them to tap their fingers (left or right), squeeze their toes (left or right) or move their tongue. This task is the simplest one of the tasks performed in the study, and the protocol associated with this task is very standard and robust. We used unprocessed data for the $N = 1080$ participants who completed this task.

6.2.2 Analyses pipelines

Multiple preprocessing and first-level analyses were performed on the task-fMRI data, giving rise to 24 different analysis pipelines. These pipelines differ in 4 parameters:

— Software package: SPM (Statistical Parametric Mapping, RRID: SCR_007037) (Penny et al., 2011) or FSL (FMRIB Software Library, RRID: SCR_002823) (Jenkinson et al., 2012).

— Smoothing kernel: FWHM was equal to either 5mm or 8mm.

— Number of motion regressors included in the GLM for the first-level analysis: 0, 6 (3 rotations, 3 translations) or 24 (the 6 previous regressors + 6 derivatives and the 12 corresponding squares of regressors).

— Presence (1) or absence (0) of the derivatives of the HRF in the GLM for the first-level analysis. Only the temporal derivatives were added in FSL pipelines and both the temporal and dispersion derivatives in SPM.

For more details on the meaning of such parameters, the reader may refer to Chapter 1.2. In the following, we will denote the pipelines by 'software-FWHM-number of motion regressors-presence of HRF derivatives'. For instance, pipeline with FSL software,

smoothing with a kernel FWHM of 8mm, no motion regressors and no HRF derivatives will be denoted by 'fsl-8-0-0'.

All pipelines were implemented using Nipype version 1.6.0 (RRID: SCR_002502) (Gorgolewski, 2017), a Python project that provides a uniform interface to neuroimaging software packages and facilitates interaction between these packages within a single workflow.

6.2.2.1 Computing environment

To limit the variability induced by different computer environments and versions of the software packages, we used NeuroDocker (RRID: SCR_017426) (Kaczmarzyk et al., 2018) to generate a custom Dockerfile. To build this image, we chose NeuroDebian (Halchenko et al., 2012) and installed the following software packages: FSL version 6.0.3 and SPM12 release r7771. To install Python and Nipype, commands were added to the Dockerfile to create a Miniconda3 environment with Python version 3.8 and multiple packages, such as Nilearn (Abraham et al., 2014a) (RRID: SCR_001362), Nipype and NiBabel (RRID: SCR_002498) (Brett et al., 2020). This docker image is available on DockerHub (Germani, 2021) and the command to generate the DockerFile can be found in the README of the software heritage archive (see 6.2.2).

6.2.2.2 Preprocessing

Preprocessing consisted of the following steps for all pipelines: spatial realignment of the functional data to correct for motion, coregistration of realigned data towards the structural data, segmentation of the structural data, non-linear registration of the structural and functional data towards a common space and smoothing of the functional data. Depending on the software package used, these steps were performed in a different order, following the default behavior of each software package.

In SPM, for each participant, functional data were first spatially realigned to the mean volume using the 'Realign: Estimate and Reslice' function with default parameters (quality of 0.9, sampling distance of 4 and a smoothing kernel, 2nd degree B-spline interpolation and no wrapping). Realigned functional data were then coregistered, with the 'Coregister: Estimate' function, to the anatomical T1w volume acquired for the participant using Normalized Mutual Information. In parallel, we segmented the different tissue classes of the same anatomical T1w volume using the 'Segment' function. The forward deformation field provided by the segmentation step was used to normalize the functional data to a standard space (Montreal Neurological Institute (MNI))('Normalize: Write' function)

with a voxel size of 2mm and a 4th degree B-spline interpolation. Normalized functional data were then smoothed with different FWHM values depending on the pipeline (5 or 8mm).

In FSL, we reproduced the preprocessing steps used in FEAT (Woolrich et al., 2001) within Nipype. Functional data were realigned to the middle functional volume using MCFLIRT. Brain extraction was applied with BET and we masked the functional data using the extracted mask. We smoothed each run using SUSAN with the brightness threshold set to 75% of the median value (default value in FSL) for each run and a mask constituting the mean functional. Different values were used for the FWHM of the smoothing kernel depending on the pipeline. We also performed temporal highpass filtering on the functional data with a value of 100s. In parallel, we computed the transformation matrix to register functional data to anatomical and standard space (MNI) using linear (FLIRT function) and non-linear registration (FNIRT function). Contrary to SPM, the first-level statistical analysis is performed on the smoothed data in subject-space. Only the transformation matrix was computed at this stage, using boundary-based registration and applied on the contrast maps output after the statistical analysis.

6.2.2.3 First level statistical analyses

To obtain the contrast maps of the different participants and contrasts, we modeled the data using a GLM. Each event was modelled using the onsets and durations provided in the event files of the HCP dataset. Six events, corresponding to the six contrasts studied, were modeled: cue (which represent any visual cue), right hand, right foot, left hand, left foot and tongue. Each condition was convolved with the canonical HRF. For both SPM and FSL pipelines, we used the Double Gamma HRF (default in SPM).

Different numbers of motion regressors (0, 6 or 24) were included in the design matrix to regress out motion-related fluctuations in the BOLD signal. The modelling of the HRF also varied: Double Gamma HRF with or without derivatives (time+dispersion for SPM and time for FSL).

In SPM, temporal autocorrelations in the BOLD signal timeseries were accounted for by highpass filtering with a 128s filter cutoff and modelling of serial correlation using an autoregressive model of the first order (AR(1)). In FSL, highpass filtering was already performed during preprocessing with a 100s filter cutoff, modelling of serial correlation was also performed using an AR(1) model. Model parameters were estimated using a Restricted Maximum Likelihood approach for both SPM and FSL software packages.

Subject-level contrast maps were computed and saved for 5 contrasts (right hand, right foot, left hand, left foot and tongue) and each participant. In the end, for each of the 24 pipelines, we had 5,400 contrast and statistic maps (5 contrasts for each of the 1,080 participants). These maps constituted the subject-level dataset. Figure 6.1(A) presents the statistic maps for the contrast *right-hand* obtained with the different pipelines for a representative subject.

6.2.2.4 Second-level statistical analyses

Group-level statistical analyses were performed using the contrast maps obtained with the different analyses pipelines. 1,000 groups of 50 participants were randomly sampled among the 1,080 participants.

For each analysis pipeline, we performed one sample t-tests for each group and each contrast in SPM (default parameters). We purposely used the same second-level analysis method and software for all pipelines in order to focus on first-level analysis differences.

For each of the 24 pipelines, the group-level dataset was thus composed of 5,000 contrast maps and statistic maps (5 contrasts for each on the 1,000 groups). Figure 6.1(B) presents the statistic maps obtained with the different pipelines for one group for the contrast *right-hand*.

6.3 Data Records

The contrast and statistic maps will be accessible on Public nEUro (*Public nEUro* 2020), the preprint will be updated to include the link as soon as possible. We now have the agreement of our Data Protection Officers to share data, the contract with Public nEUro has been validated by both sides, and is currently in signature phase.

The dataset will be organized in BIDS format (Gorgolewski et al., 2016). Discussions are underway with BIDS maintainers to find the best way to rename and reorganize our data.

6.4 Technical Validation

To assess the quality of the statistic maps, we checked that all contrasts led to an activation of the primary motor area.

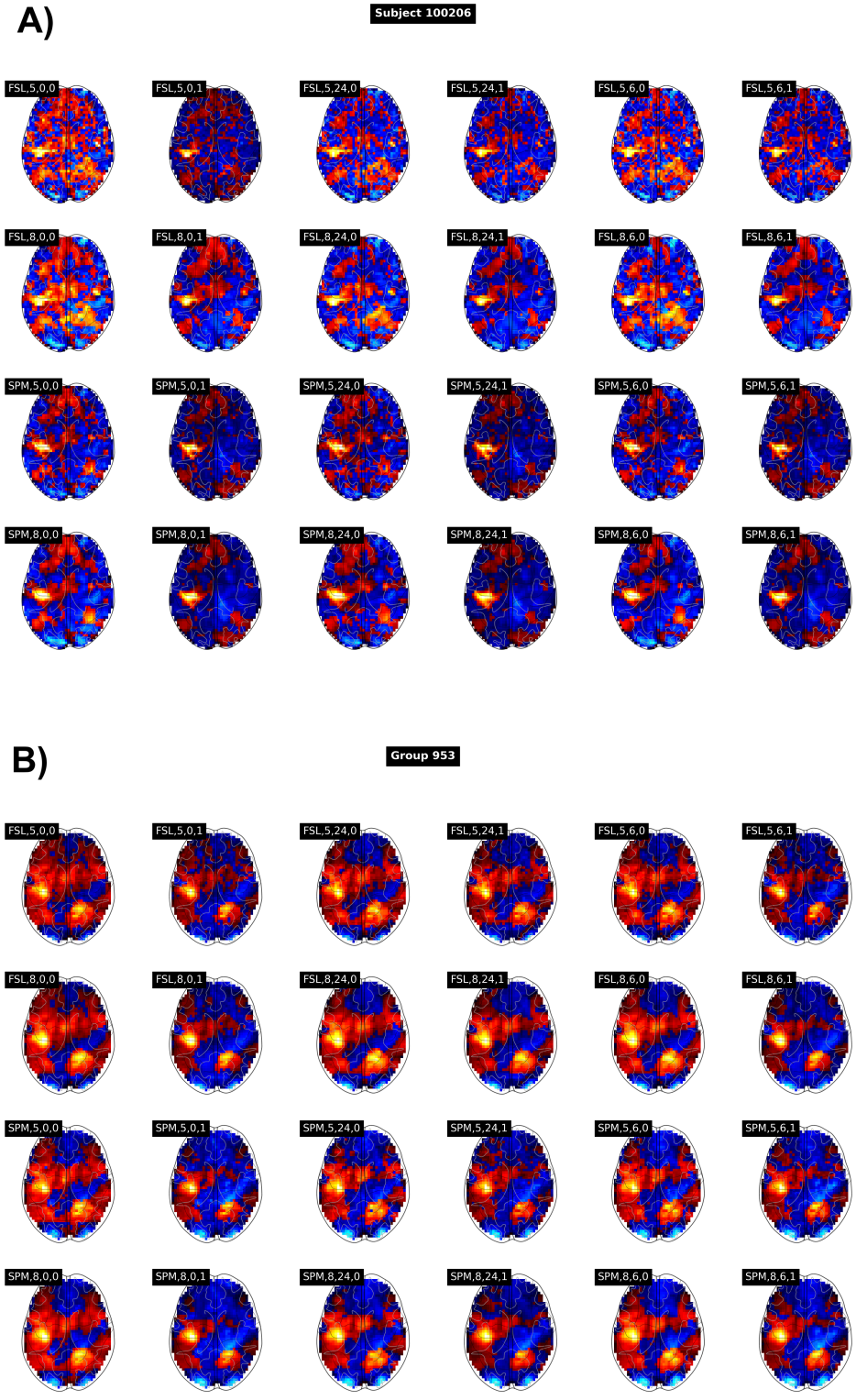

Figure 6.1 – Example of subject (A) and group-level (B) statistic maps obtained for subject 100206 and group 953 for each pipeline for the contrast *right-hand*. Pipelines are denoted by 'software-FWHM-motion regressors-HRF derivatives'.

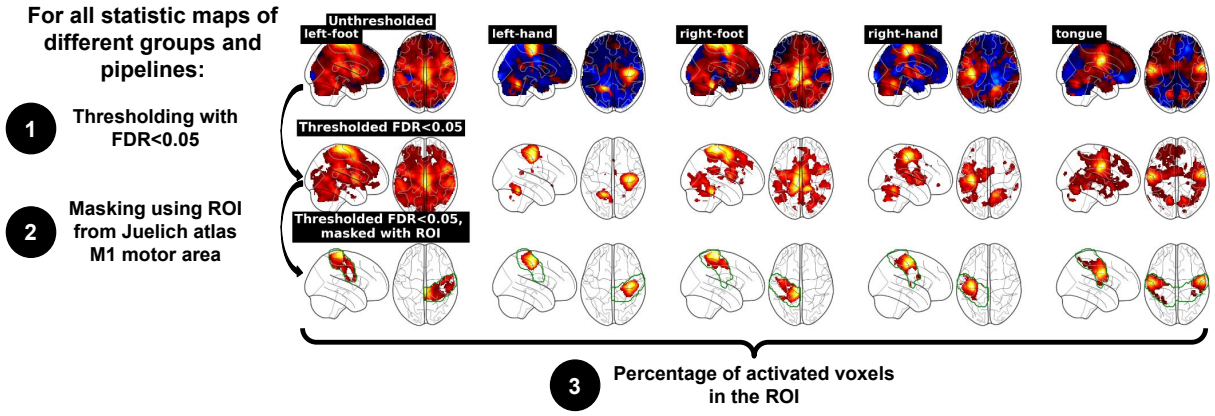

Figure 6.2 – Workflow of technical validation of statistic maps. We thresholded each statistic map of each group, each pipeline and each contrast using a FDR-corrected voxelwise p<0.05 and masked the thresholded map using the ROI of Juelich atlas of the Primary Motor Cortex. We then computed the percentage of activated voxels in the ROI of the Primary Motor Cortex.

As described in Figure 6.2, we looked at the significant activations inside the Primary Motor Cortex (M1) of the brain for each statistic map of each group, each contrast and each pipeline. Our group-level statistic maps were thresholded using an FDR-corrected voxelwise p-value of $p < 0.05$ and masked using the probabilistic Juelich Atlas (Amunts et al., 2020) available from Nilearn. We selected the Region of Interest (ROI) corresponding to the Primary Motor Cortex, Brodmann Area 4. Depending on the contrast, both left and right hemisfer's ROI ('tongue'), only the left hemisfer ('right hand' or 'right foot') or only the right hemisfer ('left hand' or 'left foot') ROI were selected, to focus on controlateral activations in the motor cortex.

For each map, we computed the percentage of activation inside the Primary Motor Cortex, which is the percentage of voxels of the ROI that are activated, *i.e.*:

$$PercentageOfActivation = \frac{N_{activated\ voxels}}{N_{total\ voxels}} \times 100 \tag{6.1}$$

where $N_{activated\ voxels}$ is the number of activated voxels in the ROI and $N_{total\ voxels}$ is the total number of voxels in the ROI.

Figure 6.3 represents the distribution of mean percentage of activation inside the Primary Motor Cortex per contrast for all studied pipelines. Results were different depending on the contrast: for all contrasts, mean percentages of activation were between 20% and

40% but those of contrasts *left foot* and *right foot* were lower than for *right hand*, *left hand* and *tongue*. When looking at the activations of different contrasts in the ROI for one of our group-level statistic maps (see Figure 6.4), we could see that the activations of the foot contrast seemed widespread with a small area of high activation. For a hand contrast, the high activation area was larger and covered nearly the entire ROI. This observation was consistent with the literature (Schott, 1993) and with statistic maps obtained from NeuroSynth (Yarkoni et al., 2011) (RRID:SCR_006798) in which the identified area of activation inside the motor cortex for the foot was smaller than the hand one. In the Primary Motor Area, the statistic maps of the foot contrasts thus have less activated voxels. Overall, the technical validation was successful. The goal of this quality check was to have a low-level estimation of the accuracy of the statistic maps to represent the task performed, thus we chose to define a single ROI covering the entire motor area. The definition of a specific ROI of the foot activation area could help having better metrics.

We observed consistent metrics across pipelines, with high percentages of activation for hand and tongue contrasts and lower ones for foot contrasts. An example of the distribution of percentage of activations for all group maps of each contrast is shown in Figure 6.5 for the pipeline spm-5-0-0.

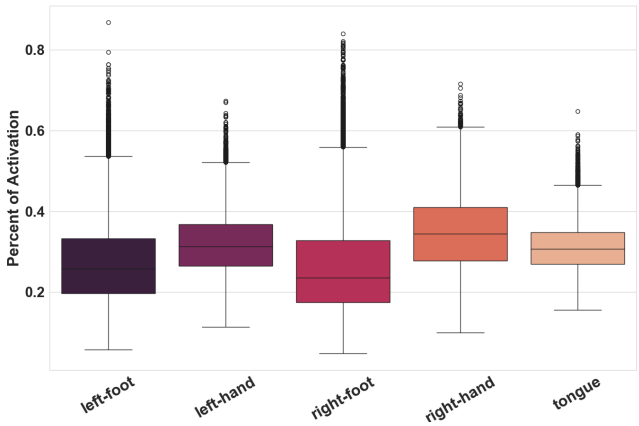

Figure 6.3 – Distribution of mean Percentage of Activation inside the Primary Motor Cortex for all groups and pipelines in the different contrast maps.

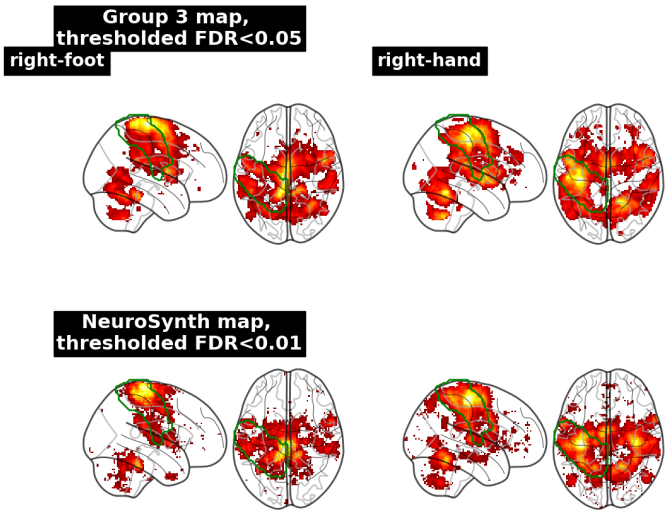

Figure 6.4 – Thresholded statistic maps for contrasts *right foot* (right) and *right hand* (left) for group-level analysis of group 3 with pipeline spm-5-0-0 (upper). Percentage of Activation inside the Primary Motor Cortex were respectively 0.34 and 0.41 for the contrasts *right foot* and *right hand*. NeuroSynth activation maps corresponding to the forward inference of the "hand" and "foot" paradigms (lower). Green borders correspond to the motor area ROI.

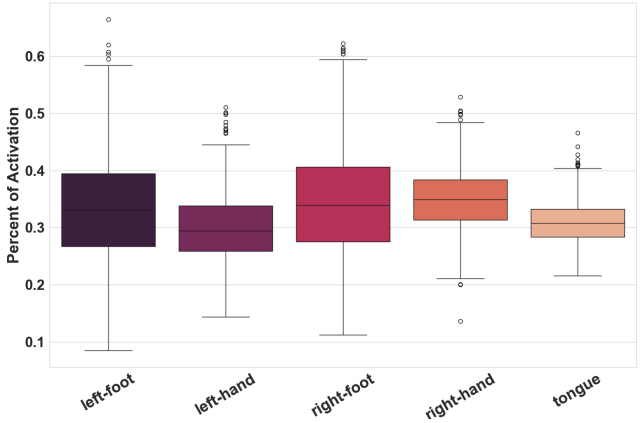

Figure 6.5 – Distribution of Percentage of Activation inside the Primary Motor Cortex for all group-level statistic maps for pipeline spm-5-0-0 in the different contrast maps.

6.5 Discussion

The *HCP multi-pipeline dataset* provides researchers with a re-usable dataset of fMRI contrast maps. The data will be accessible on Public nEUro (*Public nEUro* 2020), the preprint will be updated to include the link as soon as possible. We now have the agreement of our Data Protection Officers to share data, the contract with Public nEUro has been validated by both sides, and is currently in signature phase.

This dataset brings together a wide range of analysis conditions, covering many aspects of inter-subject, inter-groups, inter-contrasts and inter-pipelines variability. Data from 1,080 participants were used to form 1,000 different groups of 50 participants, 5 contrasts were analyzed with 24 different pipelines.

While many aspects of variability have been studied in the field of neuroimaging, changes in analytical choices are still hardly understood. Due to the computational cost in time and storage capacity of analysing fMRI data, datasets dedicated to the exploration of analytical variability (*i.e.* in which multiple pipelines are applied to the same data) are rare. Recently, the results of the NARPS study (Botvinik-Nezer et al., 2020) were made publicly available on NeuroVault, but even if 70 different analytic conditions were described, it only gives access to one group level statistic maps for 9 different contrasts.

Analytical variability is not limited to neuroimaging and has been studied in many other disciplines (Hoffmann et al., 2021), such as psychology (Simmons et al., 2011) or software engineering (Alférez et al., 2019). These different fields have brought solutions to explore and handle analytical variability. These techniques have begun to be used in neuroimaging, with, for instance, the implementation of continuous integration, a software engineering technique, to facilitate the reproducibility of neuroimaging computational experiments (Sanz-Robinson et al., 2022) or multiverse analyses that help to find the most efficient pipelines depending on the data and the goal of the study (Dafflon et al., 2022).

By sharing directly the results obtained from different analysis strategies, we hope to facilitate the use of these data by researchers from other fields, that could apply their own methods to help explore the neuroimaging analytical space. Using the code provided to create the pipelines, other researchers could be able to enhance this dataset with other combinations of parameters, giving rise to other pipelines, or apply these pipelines to other participants, groups or contrasts.

Take-home Message

- We developed and (soon) publicly shared a multi-pipeline dataset, with 24 different pipelines varying in terms of 4 criteria including software package, smoothing kernel FWHM, number of motion regressors and use of HRF derivatives.
- This dataset contain statistic maps for a wide range of context (5 cognitive paradigm, 1,080 participants, 1,000 groups). This is to our knowledge the largest dataset available to explore analytical variability.
- The goal is to provide other researchers with a set of 24 pipelines to transpose methods to explore analytical variability from other fields to neuroimaging.

Chapter 7

Uncovering communities of pipelines in the task-fMRI analytical space

---

This chapter is the subject of a paper accepted at the 2024 IEEE International Conference on Image Processing (ICIP).

- **Title:** Uncovering communities of pipelines in the task-fMRI analytical space
- **Authors:** Elodie Germani, Elisa Fromont*, Camille Maumet*
- **HAL:** hal-04331232.
- **Code:** swh:1:snp:8286215df8022543630bbbb20c5b0bd78eced45e.
- **Contributions (Credit taxonomy):** Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualisation, Manuscript writing.

* Joint senior authorship.

7.1 Introduction

In the previous chapters (1 and 2), we saw that a large number of software packages and methods are available to analyze fMRI data, making the choice of pipeline a challenging process for practitioners. These choices can have a large impact on the results, and a single change can lead to variations in the final statistic maps. Yet, there is no ground truth that can be used to measure and compare the performance of competing fMRI

pipelines. Also, there are only limited best practices to guide the pipeline choice. In an effort to guide practitioners into the pipeline space, Dafflon et al., 2022 proposed a new method to identify the pipeline that are best suited to answer a problem for which ground truths are available, such as predicting the age of participants.

In our case, a potential solution to take into account analytical variability in results is the use of multiverse analyses (Steegen et al., 2016). In these analyses, a set of pipelines is selected and used to provide a consensus results across different analytical conditions. But, the pipeline space is very large, and there is still little understanding as to which factor in the fMRI pipelines are the main drivers of analytical variability. Thus, choosing a subset of pipelines can be challenging. Here, we propose to investigate the relationships between pipeline results to help in understanding the homogeneity (*i.e.*, pipelines that give similar results) but also the heterogeneity (*i.e.* pipelines that have a different behavior) of the pipelines.

In Chapter 5, we discussed the work of Rolland et al., 2022 who recently showed the invalidity of studies combining subject-level results obtained from different pipelines for group-level analyses. We proposed a method to combine such data by converting data between pipelines using style transfer. An open question is whether patterns observed across pipelines remain stable in different contexts (*e.g.* for different groups of participants, cognitive paradigms, acquisition parameters, etc.). Style transfer frameworks aim at learning a mapping between two domains (see 3.4.2) and apply this mapping to data. If the mapping is different between contexts (*e.g.* different cognitive paradigm), a framework trained to transfer statistic maps of a particular paradigm would not be applicable to other statistic maps. To verify the potential of generalizability of our method, we also propose to explore the stability of the relationships between pipeline results. This is of particular importance to assess the potential of our solution and beyond of any solution that aims at being generalizable across different set of participants or fMRI cognitive tasks.

To measure distances between pipelines, clustering algorithms can be applied to statistic maps. However, because the data are high-dimensional and suffer from large number of sources of variability at different level (at the subject and group-level as brain activity patterns differ across participants, at the acquisition level since fMRI scanners and protocols often vary between centers and studies, etc.), distance measures between statistic maps are often meaningless and unrelated dimensions might mask existing clusters. In such case, subspace clustering algorithms (Parsons et al., 2004) are typically used to find

clusters in different subspaces within a dataset.

Here, we used community detection algorithms (*i.e.* clustering on graphs) to explore the pipeline space and assess the stability of relationships between pipeline results across different groups and cognitive paradigm. Using a clustering in two steps, we first look for clusters of pipelines and then, we explore how these clusters are similar across different groups. We explore the factors that impact the relationships between pipeline results, *i.e.* which parameters lead to more distant pipeline results and how do these parameters impact the statistic maps of the pipeline. We also aim at identifying groups of pipelines that give similar results whatever the contexts (*i.e.* different contrasts or group of participants). If two pipelines are located in the same community (*i.e.* the two pipelines present similar results) in different contexts, we can consider that their relationship is relatively stable.

7.2 Materials and Methods

To study the relationships between pipeline results and the stability of these relationships across different contexts, we computed graphs of similarity between the statistic maps of different pipelines for each group and used the *Louvain* community detection algorithm (Blondel et al., 2008) to partition each graph. *Stability* was measured for each pair of pipelines as the number of groups (out of 1,000) for which the two pipelines were located in the same community. Graphs and communities were computed using NumPy (Harris et al., 2020) (RRID:SCR_008633) and Networkx (Hagberg et al., 2008) (RRID:SCR_016864).

7.2.1 Dataset

Data used in this work are part of the HCP multi-pipeline dataset presented in the previous chapter (see Chapter 6). We used the 1,000 group-level statistic maps available for each cognitive paradigm.

7.2.2 Data processing

Group-level statistic maps obtained with different software packages did not have the same dimension, as default MNI templates used for spatial normalization are different across software packages. To be able to compute correlation between maps obtained with

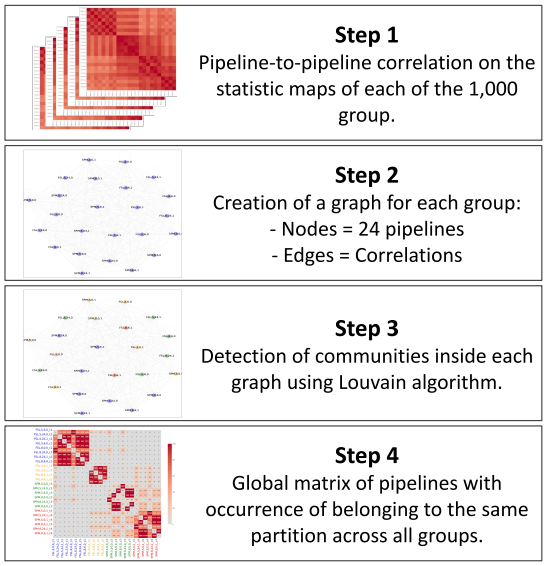

Figure 7.1 – Workflow of community detection in the pipeline space across different groups of participants and contrasts

the two software packages, we had to resample group-level statistic maps maps onto a common grid. We used Nilearn (Abraham et al., 2014a) (RRID: SCR_001362) to resample all statistic maps from all pipelines to the MNI152Asym2009 brain template with a 2mm resolution using continuous interpolation. We computed a brain mask as the intersection of all group-level brain masks from all pipelines. This mask was also resampled to the MNI brain template using nearest-neighbors interpolation and applied to the resampled group-level data. In the end, group-level statistic maps from all pipelines were resampled to the same dimensions and masked using the same brain mask.

7.2.3 Graph computation and community detection

We computed the similarity for each pair of pipelines in terms of Pearson's correlation coefficient between their statistic maps (Figure 7.1 - Step 1). This correlation matrix was used as an adjacency matrix to build an undirected weighted multi-graph for each group, with nodes representing the statistic maps of the different pipelines (V = 'fsl,0,0,0', 'fsl,0,0,1', etc.) and edges weighted by the correlation coefficient between each pipeline and labeled E={('fsl,0,0,0','fsl,0,0,1'), etc.} (Figure 7.1 - Step 2). After computation, each graph was partitioned using the Louvain algorithm (Blondel et al., 2008) to detect the best partitions based on *modularity* optimization (Figure 7.1 - Step 3), which represents the

density of links inside communities as compared to links between communities. Therefore, the communities detected in each graph represent the pipelines that give similar results for the corresponding group.

To explore the stability of the communities across different groups of participants, we counted, for each pair of pipelines, the number of groups for which the two pipelines were located in the same community (Figure 7.1 - Step 4). The higher the value the higher the similarity and stability across groups. This matrix was used to build a second graph, global across groups, in which nodes represent the different pipelines and edges represent the stability measure mentioned above. Louvain community detection algorithm was again applied to this second graph to detect communities in which pipeline provided similar statistic maps across different groups. These global graphs were computed for each contrast.

7.2.4 Communities statistic maps

Within each pipeline, the statistic map was obtained by averaging statistic maps across groups. For display purposes, we selected one pipeline in each community (see Figure 7.3 and 7.6). All other pipeline average statistic maps are available in supplementary (see Supplementary Figures F.4 and F.5).

These were thresholded assuming a Standard Normal distribution (and effectively leading to conservative estimates since we did not take into account the dependency of the different groups of participants on which these maps were averaged) and using a voxelwise False Discovery Rate (FDR) of $p < 0.05$. For each community map, we computed the number of activated voxels in the thresholded maps, but also within the ROI of the Primary Motor Cortex (M1), extracted from the probabilistic *Jülich Atlas*, available in *Nilearn* (Abraham et al., 2014a) (RRID: SCR_001362). This ROI is usually used to extract regional statistic values inside a whole-brain statistic brain of the motor task. The goal was to identify the specific patterns of each community, to understand why a pipeline was located inside a community, and to explore the potential impact on the results of the pipelines.

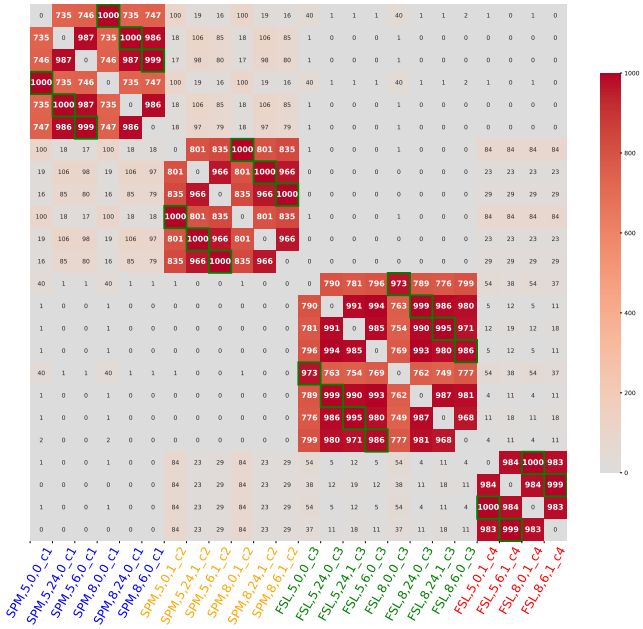

Figure 7.2 – Adjacency matrix representing the number of times each pair of pipelines belong to the same community across different group-level statistic maps of the contrast *right-hand*

7.3 Results

7.3.1 Communities for the contrast *right-hand*

The adjacency matrix representing the number of times each pair of pipelines belonged to the same community across different group-level statistic maps of the contrast *right-hand* is shown in Figure 7.2. The graph corresponding to this adjacency matrix was partitioned using the Louvain community algorithm and 4 communities were identified. These communities correspond to groups of pipelines that are frequently located in the same community across groups of participants (*i.e.*, that give similar results for a high number of groups). The partitioning of this graph achieves a modularity of 0.64 (modularity (Blondel et al., 2008) takes values between -0.5 and 1 and considered high above 0.3).

We can see that pipelines inside each partition share specific parameters, these parameters are the main factors that distinguish pipelines between communities, *i.e.* that drives the variability of the pipeline space. Here, in each community, we can find pipelines with the same software package and the same use of HRF derivatives. This means that

for this contrasts, pipelines sharing these parameters provide closer results.

Inside communities, pairs of pipelines show a large number of co-occurrence in the same community across groups (more than 700 for all pairs of pipelines in each community). This means that the relationships observed between pipelines results are stable across different groups of participants. In particular, pairs of pipelines sharing all parameters except smoothing kernel FWHM are more than 99% of the time identified in the same community (see green highlight in Figure 7.2). For instance, pipelines 'spm-5-0-0' and 'spm-8-0-0' are located in community 1 for all groups of participants.

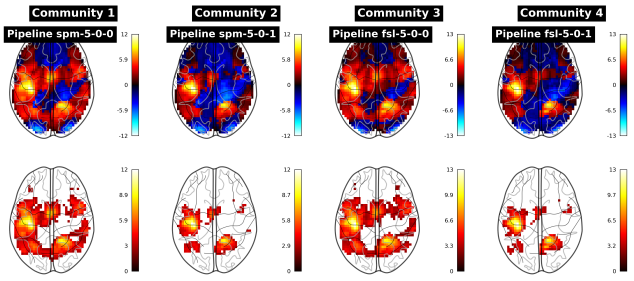

Figure 7.3 – Mean statistic map for the contrast *right-hand* across groups (of participants) for a representative pipeline in each community. Unthresholded maps (upper) and thresholded maps (lower) with voxelwise FDR-corrected $p < 0.05$.

Mean unthresholded (upper) and thresholded (lower) statistic maps of a representative pipeline in each community identified for the contrast *right-hand* are displayed in Fig.7.3. Mean maps of other pipelines per communities are available in Appendix F. This representative pipeline was arbitrarily selected, by construction all pipelines in each community show similar activation patterns. The global activation patterns are similar across communities, but the activation area is larger for the pipeline of communities 1 and 3. These communities are composed of pipelines that do not include HRF derivatives. We can suppose that this parameter has an impact on the number of significant voxels detected in the analysis. This observation is confirmed by the number of activated voxels in the thresholded maps of the pipelines inside each community (Table 7.1). Statistic maps of the representative pipeline of communities 1 and 3 show a high number of activated voxels ($N = 2,786$ and $2,539$) compared to communities 2 and 4 ($N = 796$ and 727). The numbers of activated voxels inside the ROI of the Primary Motor Cortex are similar between communities but remain more elevated in communities 1 and 3.

These maps also show that pairs of communities can have similar activation area. We can suppose that the pipelines of these pairs of communities are closer to each other than

to the ones from other communities. This suppose that there are distant & close pipelines (inside vs outside a community), but also distant & close communities. In this case, pipelines sharing the same use of HRF derivatives (community 1 and 3) seem closer than those having different use of HRF derivatives but the same software package (community 1 and 2).

Table 7.1 – Mean number of activated voxels in the thresholded mean statistic maps of the representative pipeline of each community (1st row) and inside the ROI of the Primary Motor Cortex (2nd row) for the contrast right-hand.

Community	**1**	**2**	**3**	**4**
Whole maps	2,786	796	2,539	727
ROI	382	252	337	215

7.3.2 Communities for the contrast *right-foot*

Figure 7.4 shows the adjacency matrix for the contrast *right-foot*. For this contrast, only 3 communities are identified and the distribution of pipelines inside the communities differ compared to the one observed for the contrast *right-hand*. In Figure 7.2, for contrast *right-hand*, communities are composed of pipelines with different software packages (communities 1 vs 3) and different use of HRF derivatives (communities 2 vs 4). For the contrast *right-foot*, the main factors that drive the clustering of pipelines into communities do not seem to be related to the software package: communities 1 and 2 contain pipelines from different software packages, but community 3 is composed of both SPM and FSL pipelines. In this case, the use of different numbers of motion regressors seems to have a larger impact on community identification (pipelines with 6 or 24 motion regressors are located in communities 1 and 2 vs. 0 or 6 motion regressors in community 3).

This demonstrates that the relationships between pipeline results can vary across different contexts, here cognitive paradigm. In Appendix F, we also show the adjacency matrices obtained for the contrasts *left-hand* (Figure F.1) and *left-foot* (Figure F.2), *i.e.* same cognitive paradigm as those presented in Figure 7.2 and 7.3 but located in the controlateral brain hemisphere. Communities identified for these left paradigms are similar to those observed for the counterpart right contrasts. This shows that pipeline behaviors, and thus relationships between different pipelines, are related to the effect under study (here, activation of the brain when performing a motor action with the hand or the foot).

We can also observe that the detected communities are slightly less stable across

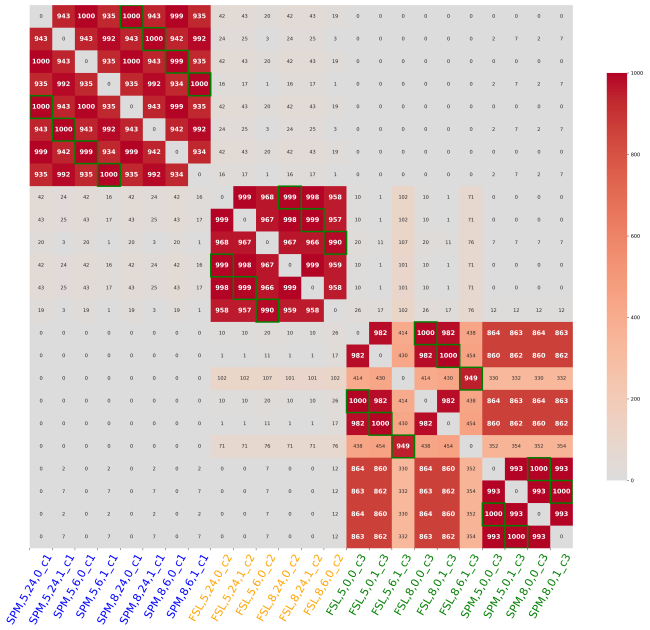

Figure 7.4 – Adjacency matrix representing the number of times each pair of pipelines belong to the same community across different group-level statistic maps of the contrast *right-foot*.

groups of participants for contrast *right-foot*, in particular for community 3 some pairs of pipelines show a number of co-occurence in the same community of less than 500 out of 1,000.

To explore this findings, we looked at the matrix of pipeline-to-pipeline correlations (averaged across groups) (see Figure 7.5). Pipelines of community 3 for which the number of co-occurrence in the communities with other pipelines is low are highlighted in blue. We can see that correlations between these pipelines are lower than other correlations inside the community, for instance: pipelines 'fsl,5,6,1' and 'spm,8,0,1' are both located in community 3 for only 55 groups out of 1,000 and the mean correlation between their statistic maps is of 0.75. In comparison, pipelines 'spm,8,0,0' and 'spm,8,0,1' are co-located in community 3 for 972 groups and the correlation between their maps is of 0.93. These observations might explains the low stability observed in this community.

This matrix also shows that results of pipelines inside a community can be close to those of a community but distant from those of another. Here, statistic maps of pipelines in community 1 seem closer to the ones of community 2 than to those of community 3. However, this does not impact the stability of relationships since between-communities correlations (around 0.8) are still lower than intra-communities correlations (0.9).

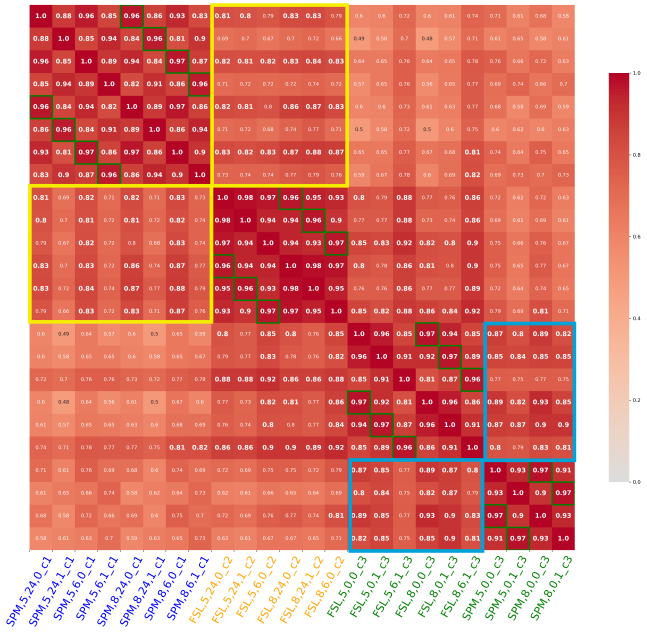

Figure 7.5 – Mean correlations (across groups) between statistic maps for each pair of pipelines with the contrast *right-foot*. Correlations between statistic maps of pipelines located in community 1 and community 2 are shown in a **yellow** box. Correlations between statistic maps of pairs pipelines located in community 3 that have a low number of co-occurence in the same community are shown in a **blue** box.

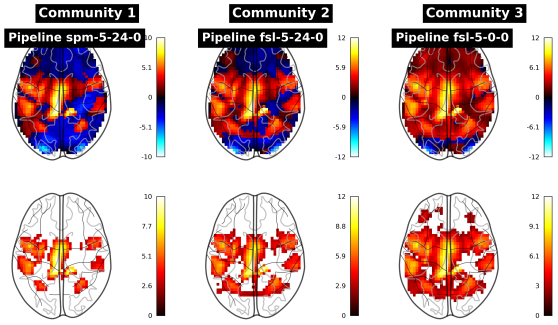

Figure 7.6 – Mean statistic map for the contrast *right-foot* across groups (of participants) for a representative pipeline in each community. Unthresholded maps (upper) and thresholded maps (lower) with voxelwise FDR-corrected $p < 0.05$.

Mean unthresholded (upper) and thresholded (lower) statistic maps of a representative pipeline of each community identified for the contrast *right-foot* are displayed in Figure 7.6. Observations are similar as those made for maps of the contrast *right-hand* (Figure 7.3), but the differences between statistic maps of communities in terms of size of activation

is larger than for the right-hand contrast. The size of activation areas seems to be an important criteria to group pipelines in communities and seems to be related to different pipeline parameters depending on the context, here on the cognitive paradigm.

7.4 Discussion

In this work, we used community detection algorithms to explore the relationships between statistic maps obtained with different pipelines in task-fMRI. Our goal was to gain a better understanding of the relationships between the results of different pipelines and of the stability of these relationships in different contexts (*i.e.* across groups of participants and cognitive paradigms). We were able to identify communities of pipelines that were giving close results across different groups of participants, but not across cognitive paradigm. Pipelines inside each community shared specific parameters values: for instance, same software package and use of HRF derivative for the communities identified in contrast *right-hand* and same use of motion regressors for contrast *right-foot*. Identification of these parameters that drives the relationships between pipeline results, is crucial to select the pipelines to explore in multiverse analyses.

Pipelines statistic maps in communities shared similar activation patterns. In particular, we found that the main distinguishing factor between communities seemed to be related to the size of the activation area, in particular for communities 1-3 and 2-4 for the contrast *right-hand*. In this context, regarding the composition of communities, we could suppose that the use of HRF derivatives in the pipelines led to a more restricted activation area in the resultant statistic maps. However, statistic maps of representative pipelines of communities with different software packages (communities 1-3 and 2-4) showed very similar activation patterns but differed in terms of the scale of statistic values. This can be explained by differences in terms of method implementation between software packages, *e.g.* pre-whitening methods which have been shown to impact the number of significant voxels (Olszowy et al., 2019). FSL analyses tend to lead to higher statistical values, which might explain the lower correlations between the maps of pipelines coming from different software packages. Thus, we could conclude that these pipelines parameters had an impact on the size of the activation area and on the scale of statistical values, which were sufficiently different in pipelines results to group them into different communities.

One of the main findings of this work is that the relationships between pipelines is not stable and depend on the contrast and on the group of participants studied. In particular,

for specific pairs of pipelines, the relationship between their results can vary across groups of participants with a number of co-occurence in the same community less than 500/1,000. This means that two pipelines can give very close results for a group of participants and more distant ones for another group. Relationships between pipeline results are even more variable when comparing different cognitive paradigm, as two pipelines can be identified in different communities (*i.e.* giving distant results) for a paradigm and located in the same one (*i.e.* giving close results) for another paradigm. Here, the communities of pipelines identified for the contrast *right-hand* were different from those identified for *right-foot*, but these were similar to those identified for the contrast *left-hand*. As we may suppose that controlateral paradigms are similar, this suggests that pipeline behaviors, and thus relationships between pipelines results, might depend on specific characteristics of the paradigm, for instance of the size of the activated area, as previous findings in the literature showed a larger number of activated voxels detected by functional MRI during the execution of finger movements than toe or tongue movements (Ehrsson et al., 2003).

This relative instability of the relationships between pipeline results puts into question the ability to learn a mapping between pipelines. Indeed, to facilitate data re-use with maps coming from different pipelines, a possible solution would be to learn a mapping between pairs of pipelines to convert statistic maps. This can be done through style transfer, as we did in Chapter 5. However, this requires a stable relationship between the data from the two pipelines to train the style transfer model and for inference, for instance if we want to apply a model trained on data from a cognitive paradigm on another one. Our findings suggest that such models might be hard to generalize to different datasets, in particular if the data explore other cognitive paradigms than the ones seen during training. This is also in line with our findings on the poor generalizability of StarGAN (Choi et al., 2018) observed in Chapter 5 and Appendix E.

The main limitation of our work is the use of a constrained set of pipelines. The HCP multi-pipeline dataset contains statistic maps output from 24 different pipelines that varies across 4 parameters (software packages, smoothing kernel FWHM, number of motion regressors, and use of HRF derivatives). These pipelines were chosen to represent typical pipelines found in the literature (Carp, 2012a). We also selected these pipelines to represent parameters that have been shown to impact the results when using a different value (Cignetti et al., 2016; Carp, 2012a). As we wanted to explore the stability of the pipeline space across different contexts, the 1,000 groups and the 5 contrasts of the motor task present in this dataset were a major advantage. In future work, it would be

interesting to explore other datasets, such as the statistic maps resulting from the NARPS many-analyst study (Botvinik-Nezer et al., 2020) (one group-level statistic maps for 70 pipelines and 9 research hypotheses), which would allow to explore a less constrained set of pipelines and a different cognitive paradigm for which the effect size is lower (mixed gamble task).

Take-home Message

- We used community-detection algorithm to explore relationships between pipelines and the stability of these relationships in different contexts (*e.g.* groups of participants, cognitive paradigms).
- We showed that relationships are relatively stable across groups of participants, but not across different paradigms.
- The different communities of pipelines identified vary in terms of extent of activation area, showing that some pipelines parameters influence the sensitivity to the signal.
- This work is also in line with our findings in Chapter 5 where we found poor generalizability of style transfer frameworks to unseen tasks.

Chapter 8

Validity of fMRI mega-analyses with data processed with different pipelines

---

This chapter was the subject of a paper under review at *Imaging Neuroscience*.

- **Title:** On the validity of fMRI mega-analyses using data processed with different pipelines.
- **Authors:** Elodie Germani*, Xavier Rolland*, Pierre Maurel, Camille Maumet
- **HAL:** inserm-04466478.
- **Code:** swh:1:snp:585d3a0a3388a928ab3c6211c1826702aa618190.
- **Contributions (Credit taxonomy):** Formal analysis (reproduction of the analyses performed by X. Rolland), Investigation, Methodology (Between-software analysis), Software, Visualisation, Manuscript writing.

* Co-first authors.

8.1 Introduction

As discussed in Chapter 2.2.3, small sample sizes undermine the reliability of neuroimaging research findings. With the increased adoption of open science practices (Poline et al., 2012; Poldrack et al., 2014; Niso et al., 2022) and the development of dedicated research infrastructures (Gorgolewski et al., 2015; Markiewicz et al., 2021; Barillot et al., 2016), such as NeuroVault (Gorgolewski et al., 2015), OpenNeuro (Markiewicz et al.,

2021), more and more neuroimaging data from various studies have been made available to the scientific community. In this context, re-using data from previous studies seems a promising solution to increase sample sizes using meta- (Salimi-Khorshidi et al., 2009) or mega-analyses.

In fMRI, shared data often include raw data at the subject-level, that can be re-analyzed using the same processing steps and combined in a mega-analysis, but also derived data (*i.e.* already processed) at the subject or group-level. At the group-level, derived data can be used in meta-analyses to build consensus results across multiple studies (Salimi-Khorshidi et al., 2009). At the subject-level, individual contrast maps (after the subject-level processing) from different studies can be combined using mega-analyses. But, there are several limitations to these method due to publication bias (Ioannidis et al., 2014).

The re-use of subject-level data is more optimal compared to raw data, not only because sharing of statistic maps is easier due to reduced privacy constraints, but also because it avoids having to perform costly re-computations. As explained in Chapter 1, fMRI studies require multiple processing steps on the data, called a "pipeline", both at the subject-level (preprocessing of the raw fMRI data to prepare them for statistical analysis, and first-level analysis for each participant) and at the group-level (second-level statistical analysis using the subject-level contrast maps resulting from first-level analysis).

However, we showed in Chapter 2 that researchers have multiple choices to make to build their pipeline due to their high flexibility. Thus, it is likely that derived data shared on public databases come from different pipelines. In Rolland et al., 2022, the validity of mega-analyses combining data processed differently at the subject-level was explored within the SPM software package, and results showed that these studies were invalid for almost all combinations.

Here, in a follow-up work, we further explore the validity of these mega-analyses that combine data processed differently at the subject-level. We extend the work from Rolland et al., 2022 by adding subject-level data processed using FSL, and explore the validity of mega-analyses combining data processed within SPM, within FSL and between software packages. Similarly to Rolland et al., 2022, we carry out a series of between-groups analyses, with each group corresponding to subject-level contrast maps processed with different pipelines and randomly sampled from the Human Connectome Project (HCP) Young Adult dataset (Van Essen et al., 2013). Since participants in each groups are sampled from the same population, all differences detected are therefore false positives

offering an empirical estimation of the false positive rate.

8.2 Materials and Methods

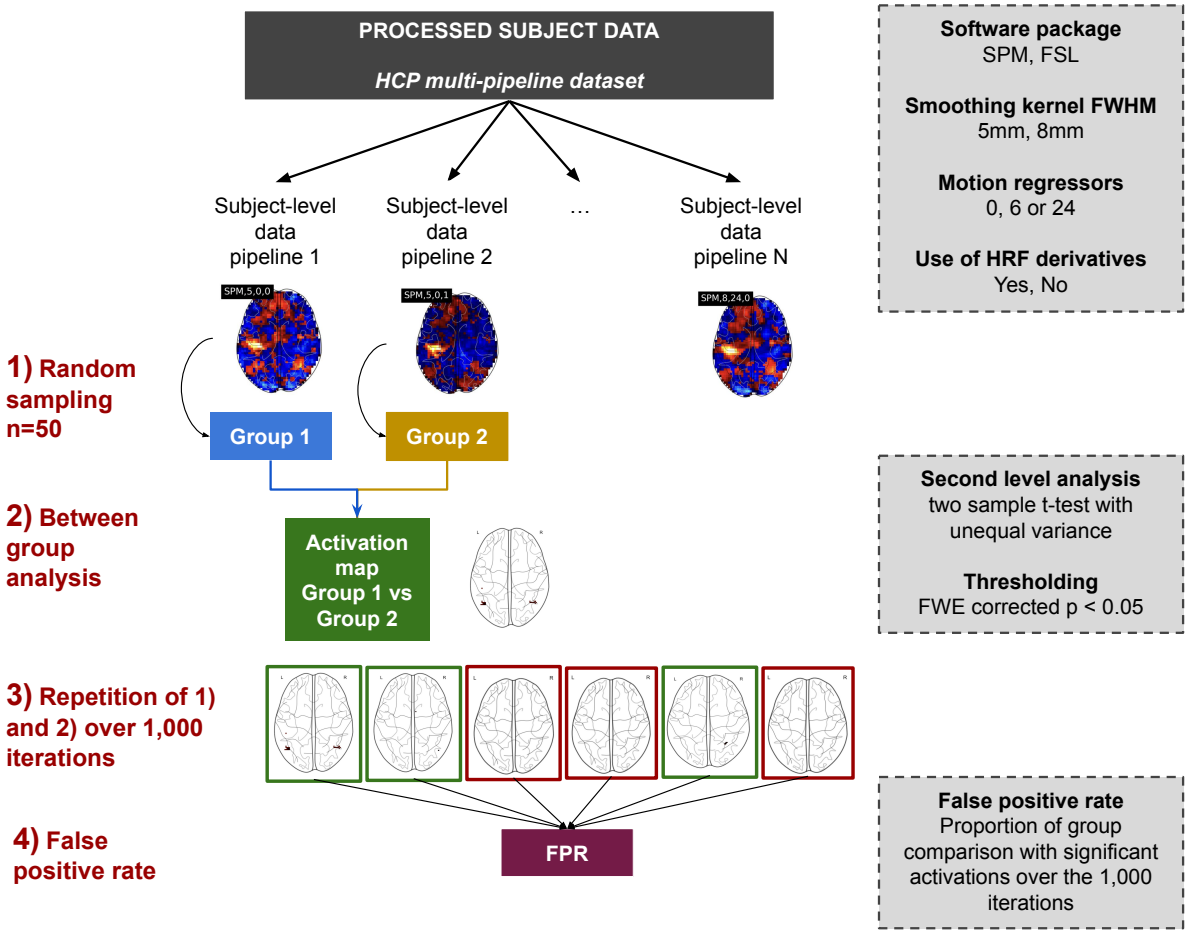

Figure 8.1 – Overview of the method: 1) sampling of n=50 subject-level contrast maps for each group (i.e. one group = one pipeline) from the HCP multi-pipeline dataset (Chapitre 6), 2) between-group analysis "Group 1 ≠ Group 2", 3) running 1,000 iterations of 1) and 2), and 4) estimation of the false positive rate.

The goal of this study is to test the validity of between-group analyses using subject-level contrast maps processed with different pipelines. In the following sections, the term "pipeline" is used to refer to the subject-level pipelines only.

The steps performed in order to estimate this validity are presented in Figure 8.1. First, we randomly sampled subject-level contrast maps processed through different pipelines

from the HCP multi-pipeline dataset (see section 8.2.1 and Chapter 6). Then, for each pair of pipelines, we performed a between-group analysis (see section 8.2.2). This group comparison was repeated 1,000 times in order to estimate the empirical false positive rate (see section 8.2.3).

8.2.1 Dataset

In this work, we used subject-level contrast maps from the HCP multi-pipeline dataset, that we presented in greater details in Chapter 6. We used contrast maps for the paradigm *right-hand* for the 24 pipelines implemented for this dataset.

In brief, the pipelines implemented in the dataset varied on the following set of parameters:

— Software package: SPM (Statistical Parametric Mapping, RRID: SCR_007037) (Penny et al., 2011) or FSL (FMRIB Software Library, RRID: SCR_002823) (Jenkinson et al., 2012).

— Smoothing kernel: FWHM of 5 mm or 8 mm.

— Number of motion regressors included in the GLM for the first-level analysis: 0, 6 (3 rotations, 3 translations) or 24 (3 rotations, 3 translations + 6 derivatives and the 12 corresponding squares).

— Presence (1) or absence (0) of the derivatives of the HRF in the first-level GLM. The temporal derivative was added in FSL and both the temporal and dispersion derivatives in SPM.

In total, this led to 24 different subject-level pipelines (2 software packages $\times$ 2 smoothing kernels $\times$ 3 numbers of motion regressors $\times$ 2 HRF).

8.2.2 Between-group analyses

In this work, we explored the validity of between-group studies with subject-level contrast maps from different pipelines in three settings: within-pipeline (baseline), within-software (*i.e.* pipeline implemented in the same software package with different parameters) and between-software (*i.e.* pipeline implemented in different software packages with similar parameters).

8.2.2.1 Contrast post-processing

As FSL and SPM use different MNI templates (Evans et al., 2012), subject-level contrast maps from different software packages had different dimensions. To compute between-software comparisons, we therefore had to post-process the contrast maps to put them in the same target space and on the same grid. We used Nilearn (Abraham et al., 2014a) (RRID: SCR_001362) to resample all subject-level contrast maps to the MNI152Asym2009 brain template with a 2 mm resolution using continuous interpolation. We masked the contrast maps using the intersection of all subject-level brain masks (all pipelines).

FSL and SPM contrast maps are also scaled differently (see Nichols, 2012). In both software packages, contrast maps are theoretically expressed in percent BOLD change but there are important differences in how this percent BOLD change is computed that effectively lead to scaling differences. Hence, in SPM, contrast maps units are closer to 2.5 times percent BOLD change due to the mask used to compute the global in-brain mean intensity. On the other hand, FSL contrast maps are scaled to 10,000 (*i.e.* 100 times percent BOLD change). We applied a factor to each contrast map to make them closer to percent BOLD change. Contrast maps in SPM and FSL were therefore rescaled by multiplying by $100/250 = 0.4$ and $100/10,000 = 0.01$ respectively.

All between-group analyses were performed on resampled, masked and re-scaled subject-level contrast maps. As a sanity check, we also ran the between-group same-pipeline analyses on the original (*i.e.* not resampled, masked nor unit-re-scaled) subject-level contrast maps. As expected, no differences were identified in the estimated false positive rate (see Supplementary Table G.1).

8.2.2.2 Analysis setup

For each between-group analysis, we randomly sampled 100 participants without replacement among the full set of 1,080 participants and splitted them into two groups ($N = 50$ in each group). In each group, subject-level contrast maps were obtained with a different pipeline. This process was repeated for different groups and pairs of pipelines. We performed a one-tailed two-sample t-test with unequal variance and computed the statistic maps associated to H_0: "no mean difference of activation between groups". We used a voxelwise $p < 0.05$ FWE-corrected threshold. All between-group analyses were performed in SPM in order to keep consistent second-level analysis conditions.

8.2.3 False Positive Rates Estimation

For a given pair of pipelines, the between-group analysis was repeated 1,000 times with different sets of participants. Since participants in each group were sampled from the same population, H_0 is true by construction. All differences detected were therefore false positives and the empirical false positive rate was estimated as the proportion of between-group analyses, across the repetitions, with at least one significant detection (see Figure 8.1).

Since the null hypothesis is true, we expect the p values associated with the between-group statistic maps to be uniformly distributed, and in particular, the empirical false positive rate is expected to be equal to the α-level (here 0.05). A higher rate highlights invalidity (*i.e.* an inflated rate of false positive) and a lower rate conservativeness (*i.e.* reduced sensitivity).

8.2.4 Statistical distributions and P-P plots

P-P plots are usually used to observe how a given set of statistical values diverge from an expected distribution by plotting, for each k^{th} ordered statistical value, the expected associated p value on the x-axis and the obtained p value on the y-axis. Here, under the null hypothesis, p values were expected to follow a uniform distribution $U(0, 1)$. Thus, for a set of N statistical values, the k^{th} ordered p value was expected to be equal to $k/(N+1)$.

We used a Bland-Altman (Giavarina, 2015) variant of P-P plots by replacing the p values by the following:

— on the x-axis: the expected p value in $-log_{10}$

— on the y-axis: the difference between the $-log_{10}$ obtained and the $-log_{10}$ expected p values.

This update made it easier to observe the behaviour in the tails of the p value distribution (which is of interest here). High statistical values (right tail of our sample) are associated to low p values, *i.e.* to high $-log_{10}$ p values. We also looked at the distributions of the statistical values for multiple between-group analyses, and compared with a Student distribution $\mathcal{T}_{98}$.

8.3 Results

8.3.1 Analyses using the same pipeline (baseline)

Table 8.1 shows the false positive rates obtained for all analyses with the same pipeline in both groups, separately for SPM and FSL. For all combinations, the false positive rates were below the expected value of 0.05, ranging between 0.012 and 0.028 for SPM and between 0.013 and 0.024 for FSL.

SPM

	Smooth 5mm		Smooth 8mm	
	No derivatives	Derivatives	No derivatives	Derivatives
0 motion regressors	0.012	0.013	0.016	0.023
6 motion regressors	0.015	0.006	0.024	0.013
24 motion regressors	0.023	0.016	0.025	0.028

FSL

	Smooth 5mm		Smooth 8mm	
	No derivatives	Derivatives	No derivatives	Derivatives
0 motion regressors	0.014	0.013	0.015	0.023
6 motion regressors	0.018	0.014	0.018	0.018
24 motion regressors	0.015	0.013	0.016	0.024

Table 8.1 – False positive rates for between-groups analyses with the same pipeline in both groups, with SPM and FSL and for all possible sets of parameters (number of motion regressors, smoothing kernel FWHM and presence or absence of HRF temporal derivatives). The false positive rates were always under 0.05.

These results, obtained with the same pipeline in both groups, are used as a baseline in the following. False positive rates obtained with original contrast maps (non resampled, masked and corrected) were similar (see Appendix G - Supplementary Table G.1).

8.3.2 Analyses using pipelines with different parameters

The following subsections present the results obtained with pipelines using different set of parameters (within software). In each case, we looked at the false positive rates

(Figure 8.2), the statistical distributions (Figures 8.7 and 8.8) and the associated P-P plots (Figure 8.3, 8.4, 8.5 and 8.6). To present the results, we chose a default value for each studied parameter – smoothing 5 mm FWHM, HRF with derivatives and 24 motion regressors – and compared our results to those obtained with the default.

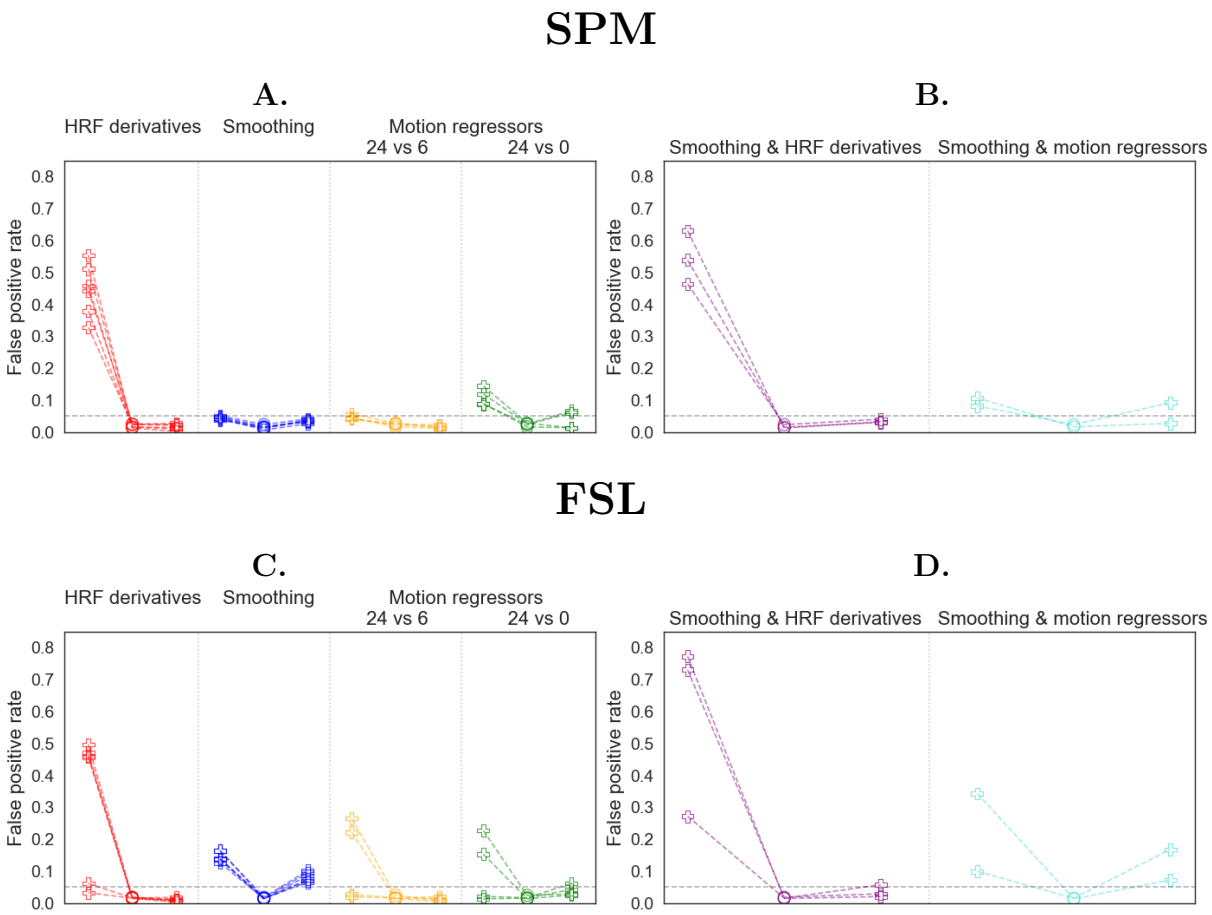

Figure 8.2 – False positive rates for pipelines with different parameters within SPM (A, B) and FSL (C, D). For each studied parameter (HRF derivatives, smoothing and motion regressors), we provide the false positive rates obtained for: 1/ both tails, *i.e.* pipeline 1 > pipeline 2 and reverse (crosses) and 2/ for the corresponding analysis in which pipeline 1 and 2 are identical, *i.e.* the baseline (circles). Panels B and D, provides the false positive rates when two parameters vary. The grey dashed line corresponds to the expected theoretical value (0.05).

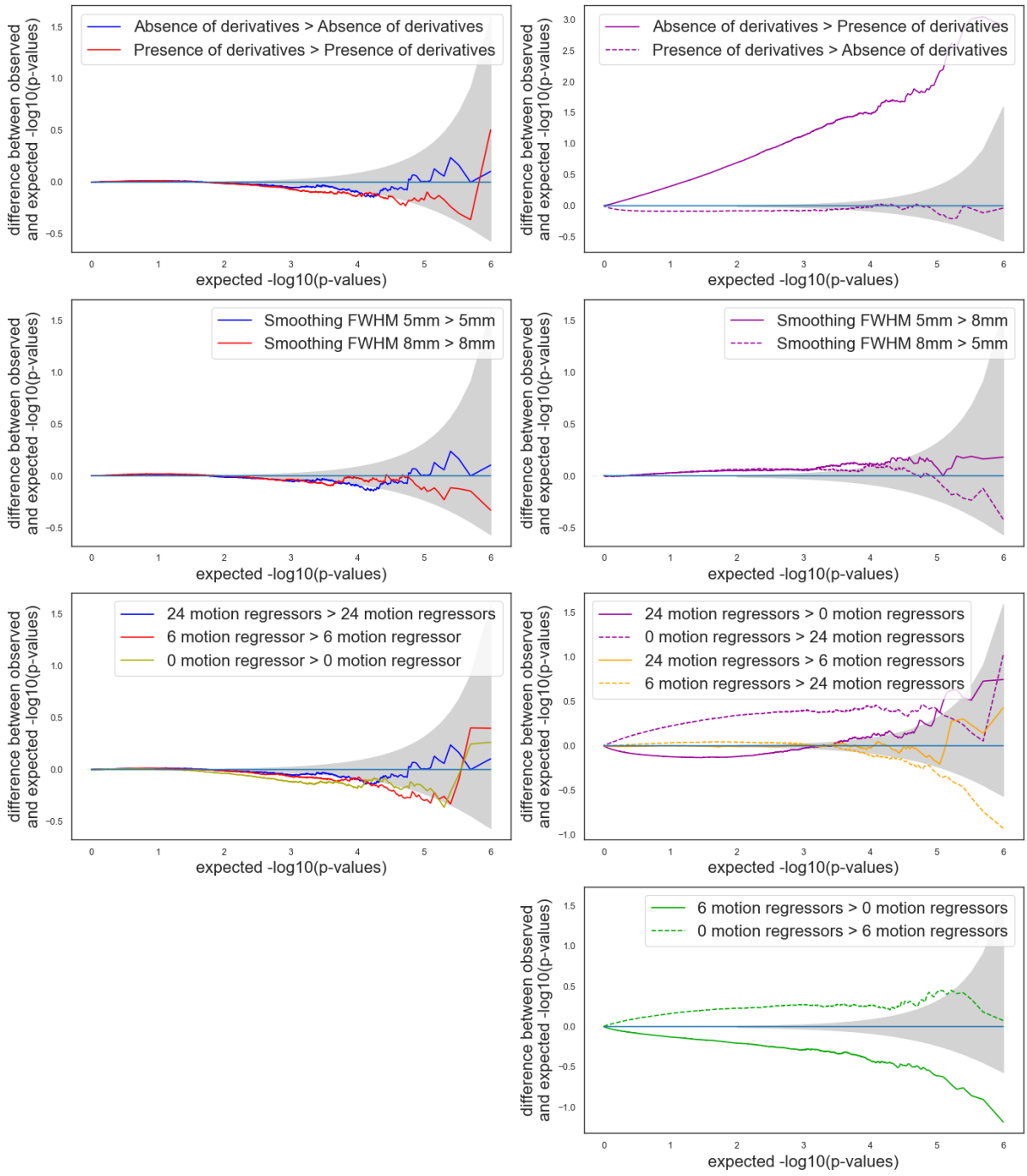

Figure 8.3 – Bland-Altman P-P plots for pipelines with different (right column) parameters and with the same (left column) parameters within SPM. The grey shade corresponds to the 0.95 confidence interval. A curve above (respectively below) the confidence interval indicates invalidity (respectively conservativeness). Default parameters: 5 mm smoothing, 24 motion regressors and no HRF derivatives.

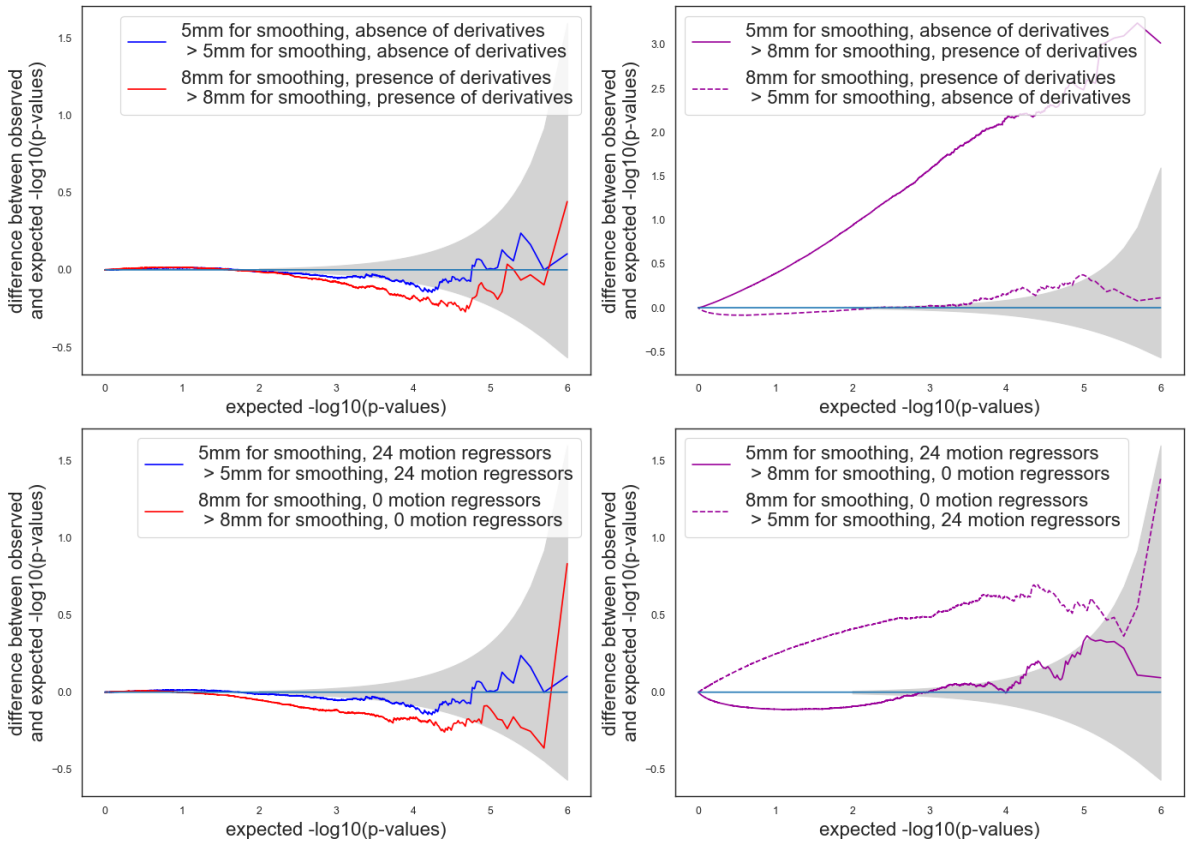

Figure 8.4 – Bland-Altman P-P plots for pipelines with two different (right column) parameters and with the same (left column) parameters within SPM. The grey shade corresponds to the 0.95 confidence interval. A curve above (respectively below) the confidence interval indicates invalidity (respectively conservativeness). Default parameters: 5 mm smoothing, 24 motion regressors and no HRF derivatives.

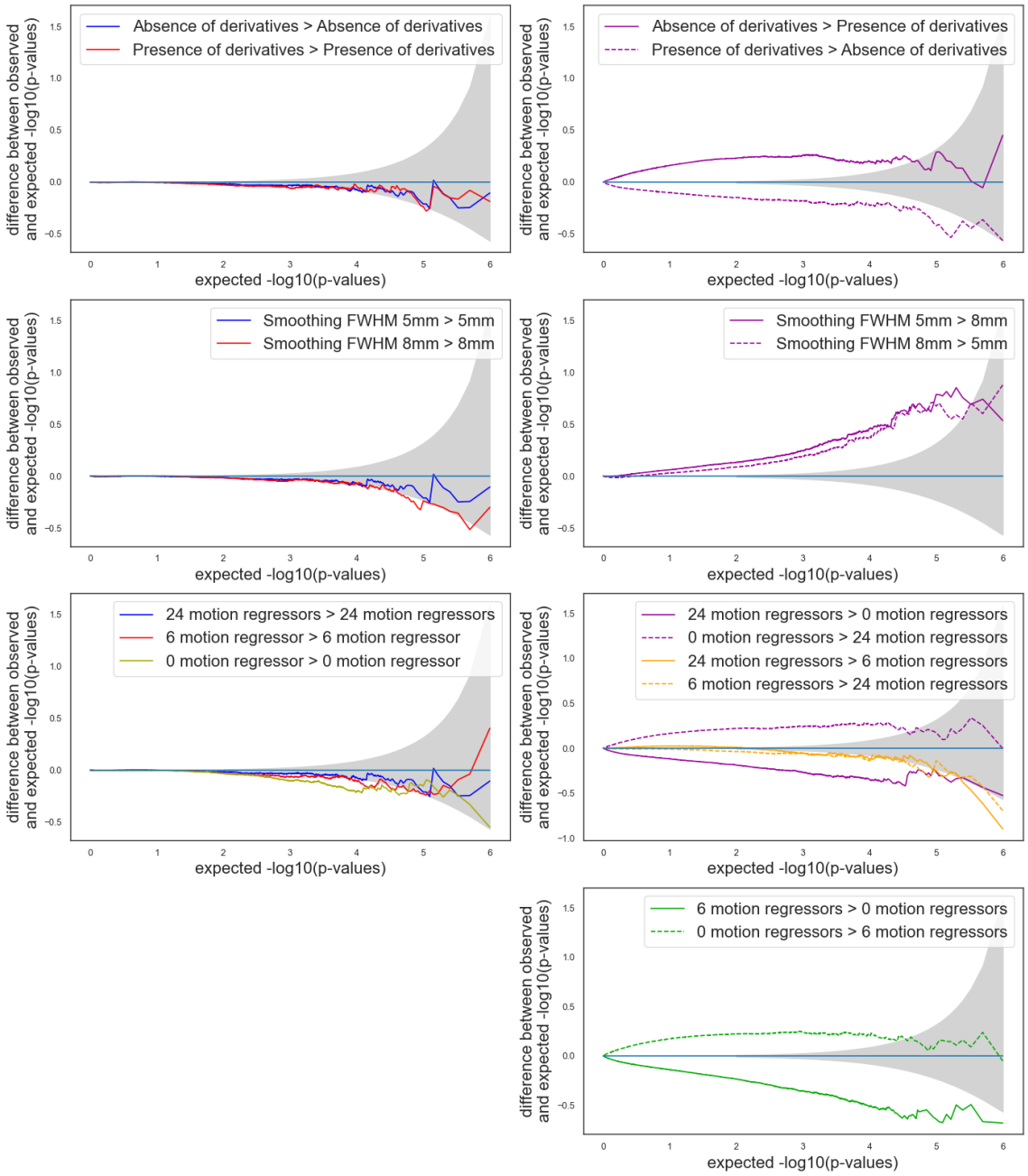

Figure 8.5 – Bland-Altman P-P plots for pipelines with different (right column) parameters and with the same (left column) parameters within FSL. The grey shade corresponds to the 0.95 confidence interval. A curve above (respectively below) the confidence interval indicates invalidity (respectively conservativeness). Default parameters: 5 mm smoothing, 24 motion regressors and no HRF derivatives.

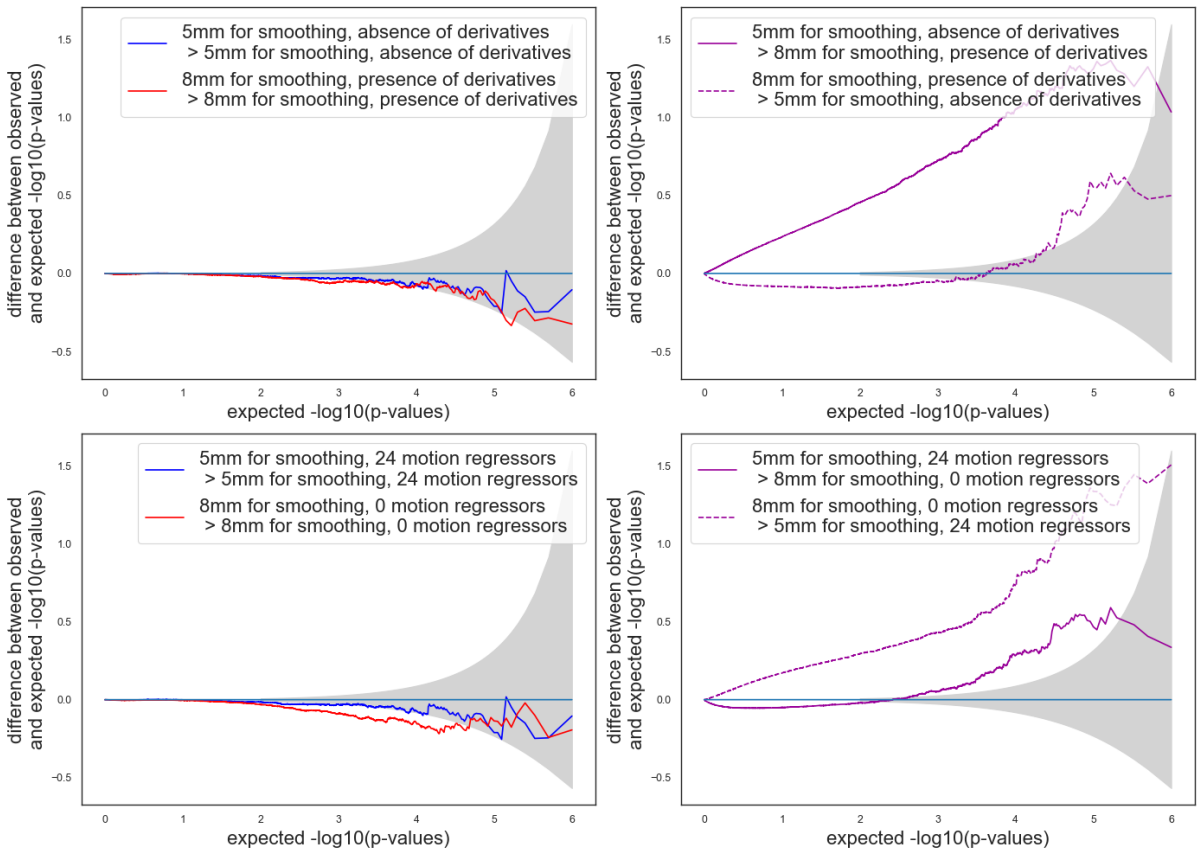

Figure 8.6 – Bland-Altman P-P plots for pipelines with two different (right column) parameters and with the same (left column) parameters within FSL. The grey shade corresponds to the 0.95 confidence interval. A curve above (respectively below) the confidence interval indicates invalidity (respectively conservativeness). Default parameters: 5 mm smoothing, 24 motion regressors and no HRF derivatives.

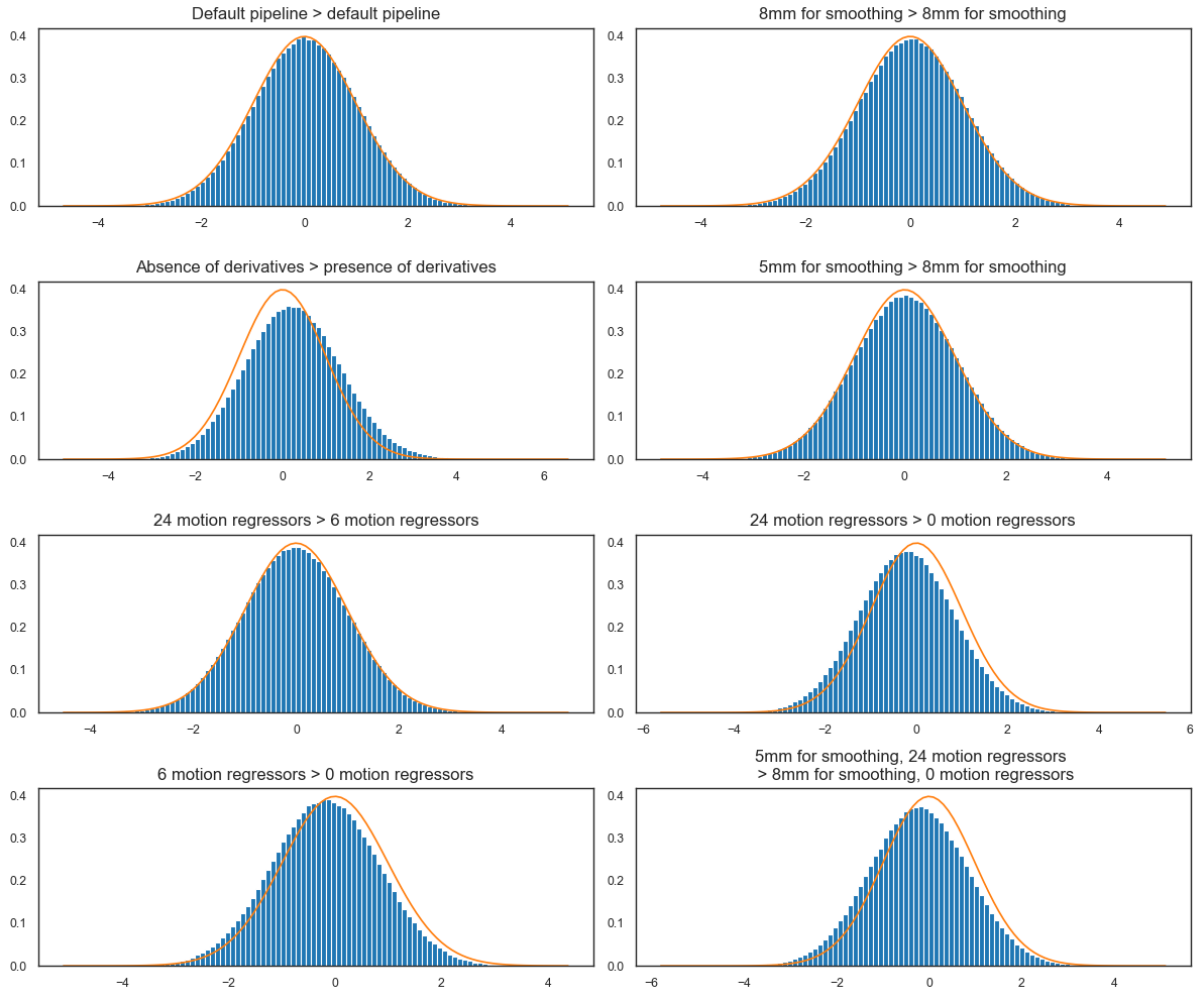

Figure 8.7 – Distribution of statistical values for multiple between-group analyses under SPM, compared to the expected distribution. Default parameters: 5 mm smoothing, 24 motion regressors and no HRF derivatives. Pipelines which differ from the default pipeline are put in bold. The orange curve represents the Student distribution with 98 degrees of freedom, which is the expected distribution in our case under null hypothesis.

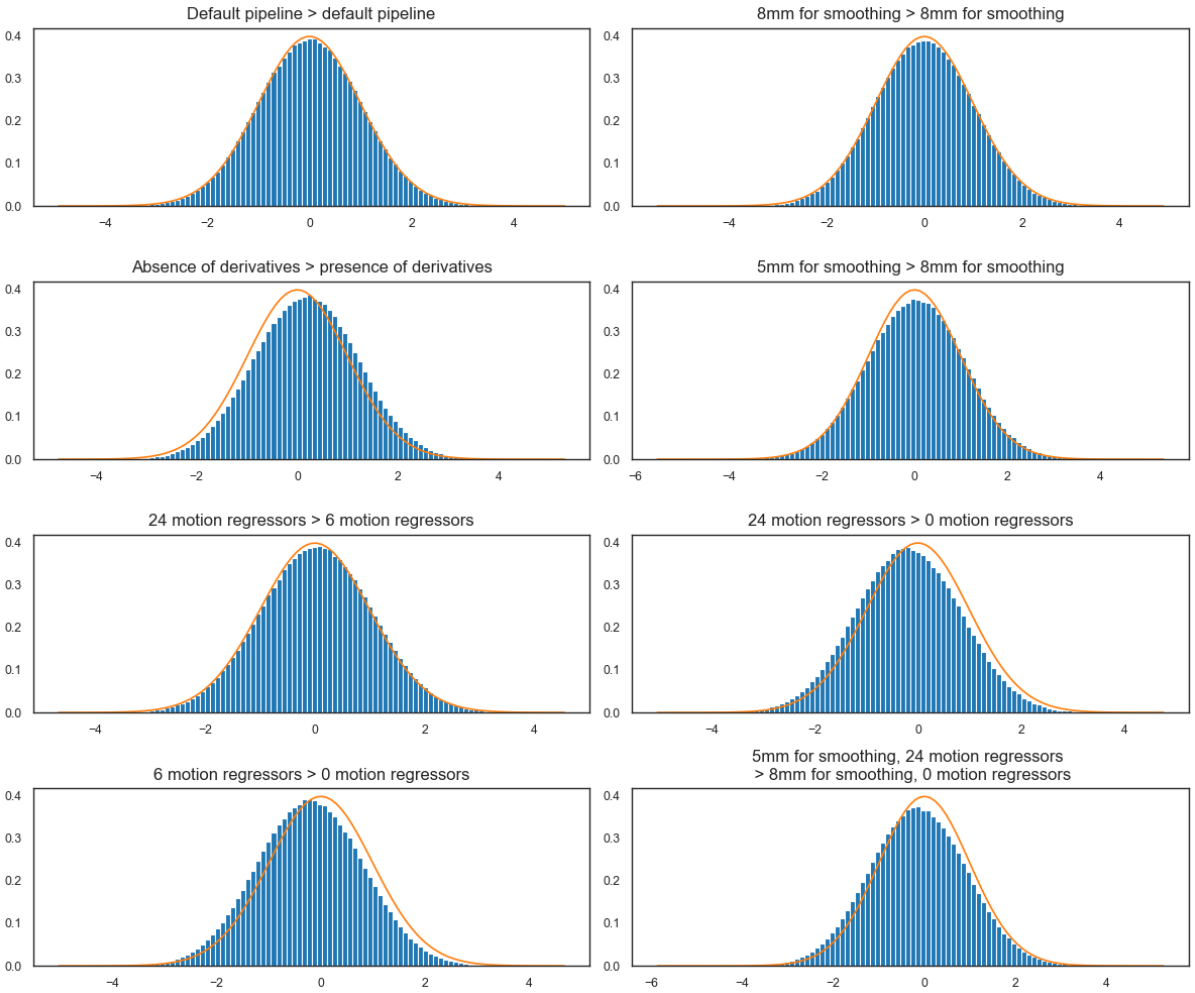

Figure 8.8 – Distribution of statistical values for multiple between-group analyses under FSL, compared to the expected distribution. Default parameters: 5 mm smoothing, 24 motion regressors and no HRF derivatives. Pipelines which differ from the default pipeline are put in bold. The orange curve represents the Student distribution with 98 degrees of freedom, which is the expected distribution in our case under null hypothesis.

8.3.2.1 Different HRF

Adding derivatives to model the HRF was the most impacting of all three varying factors in both software packages. The false positive rates obtained with different HRF (*canonical HRF* versus *HRF with derivatives*) in the pipelines are presented in Figure 8.2 A and C (red curves) for the six analyses performed (*i.e.* with varying levels of smoothing and number of motion regressors – with the same setting in both pipelines).

In SPM, the comparison *canonical HRF > HRF with derivatives* (Figure 8.2 A, red curve on the left) showed invalid false positive rates (above the 0.05 threshold) for all pipeline combinations. Similarly, in FSL, all combinations gave invalid results for this same comparison except two combinations: *5 mm* or *8 mm smoothing FWHM* and *24 motion regressors*. These two analyses led to values close to the 0.05 threshold (0.032 and 0.061 respectively). For the opposite comparison (*i.e. canonical HRF < HRF with derivatives*) all combinations resulted in valid results with false positive rates under 0.05.

Figures 8.3 and Figure 8.5 show the corresponding Bland-Altman P-P plots for comparisons with different HRF and otherwise default parameters. In both software packages, consistently with what we observed for the false positive rates, the comparison *canonical HRF > HRF with derivatives* led to values that were outside of the 95% confidence interval (grey area). In SPM, values were further away from the 95% confidence interval than in FSL.

The same observations could be made on the statistical distributions for both SPM and FSL (Figures 8.7 and 8.8): both showed a shift in mean and variance, but this was smaller for FSL. The combination of pipelines parameters used in this Figure (*i.e.* pipelines with *5 mm* FWHM and *24 motion regressors*, with different HRF derivatives) showed nearly valid false positive rates, as stated in the previous paragraph (see Figure 8.2), which could explain why the shift seemed smaller in FSL compared to SPM. We also observed the P-P plots for a different combination of FSL pipelines with other parameters (5 mm, 0 motion regressors) in Supplementary Figure G.1 and found a similar shift as the one observed for SPM.

8.3.2.2 Different smoothing

The false positive rates obtained with different levels of smoothing (*5 mm* or *8 mm*) in the pipelines are presented in Figure 8.2 A and C (blue curves) for the six analyses performed (*i.e.* with varying HRF models and number of motion regressors – with the

same setting in both pipelines).

The false positive rates obtained with different levels of smoothing (*5 mm* versus *8 mm*) in the pipelines were above the 0.05 theoretical false positive rate in FSL (ranging from 0.07 to 0.16) and below or close to the theoretical rate in SPM (ranging from 0.03 to 0.05). Compared to the baseline analyses using the same pipelines, the false positive rates were always inflated and were slightly higher for the tail *5 mm > 8 mm*.

The Bland-Altman P-P plots (Figure 8.3 and 8.5) are consistent with the observations made on the false positive rates. Between-group analyses using pipelines with different smoothing gave invalid results in FSL and values within the 95% confidence interval in SPM, with only a small positive difference in the direction *5 mm > 8 mm*.

The behaviors observed on the P-P plots can be explained by the positive shift in mean values and standard deviations observed on the statistical distribution for *5 mm > 8 mm* for FSL (Figure 8.8), which is less pronounced for SPM (Figure 8.7).

8.3.2.3 Different number of motion regressors

The false positive rates obtained with different number of motion regressors (0, 6 and 24) in the pipelines are presented in Figure 8.2 A and C (yellow and green curves) for the six analyses performed (*i.e.* with varying levels of smoothing and different HRF – with the same setting in both pipelines). We studied the combinations *24 motion regressors* versus *6 motion regressors* (yellow curves) and *24 motion regressors* versus *0 motion regressors* (green curves).

In SPM, false positive rates were below the 0.05 theoretical rate for all comparisons of *24 motion regressors* versus *6 motion regressors*. For the comparison with no motion regressors, the false positive rates were higher and above 0.05 for *24 motion regressors > 0 motion regressors* and slightly below for the opposite. In FSL, the validity of the results was dependant on the other pipeline parameters. All combinations led to invalid results (*i.e.* above the theoretical 0.05 threshold) except for *24 motion regressors > 0/6 motion regressors* when using the canonical HRF (*i.e.* no HRF derivatives) in both pipelines.

In Section 8.3.2.1, we showed that all combinations of pipelines with varying HRF models led to invalid results except those with *5 mm* or *8 mm smoothing* and *24 motion regressors*. Here, we also observe invalid results for all combinations of pipelines with *24 motion regressors* versus *0/6 motion regressors*, except those with *5 mm* or *8 mm smoothing* and *no HRF derivatives*. We can suppose that in FSL, when using *24 motion regressors*, the use of HRF derivatives in the GLM has a low impact on the results and

similarly, when using the *canonical HRF*, using *0, 6 or 24 motion regressors* does not change the results much, and thus has a low impact on the validity of the mega-analyses combining subject-level data obtained from pipelines with different parameters.

In the Bland-Altman P-P plot for SPM (Figure 8.3), we observed more extreme values in the P-P plots for the comparisons "*24 motions regressors* versus *0 motion regressors*" than for those of "*24 motions regressors* versus *6 motion regressors*", which is consistent with our observations on false positive rates. The Bland-Altman P-P plot (Figure 8.5) for FSL with 5 mm smoothing and an HRF with derivatives, the comparison *24 motion regressors* versus *0 motion regressors* were consistent with the invalid false positive rates found with such parameters: we found conservative results for the comparison *24 motion regressors* > *0 motion regressors* (plain line) and invalid results in the opposite direction (dashed line).

Statistical distributions (Figures 8.7 and 8.8) also show a shift in mean and variance for the comparison "*24 motion regressors* versus *0 motion regressors*", for both SPM and FSL. This shift is not as important for the comparison "*24 motion regressors* versus *6 motion regressors*". The comparison "*6 motion regressors* versus *0 motion regressors*" was also showed for comparison, and showed similar results as the "*24 motion regressors* versus *0 motion regressors*" comparison.

8.3.2.4 Combined effects of parameters

We observed the combined effects of:

— differences in smoothing and in HRF model

— differences in smoothing and in motion regressors

The false positive rates obtained with different smoothing and different HRF model or different motion regressors in the pipelines are presented in Figure 8.2 (B and D) for the different analyses performed.

In both SPM and FSL, the first set of between-group analyses (*5 mm smoothing*, *canonical HRF*) > (*8 mm smoothing*, *HRF with derivatives*) led to invalid results, with false positive rates largely above the 0.05 theoretical threshold (around 0.60). The opposite test provided valid results.

In SPM, the results for (*5 mm smoothing*, *canonical HRF*) > (*8 mm smoothing*, *HRF with derivatives*) were close to those obtained for the analyses with a single varying parameter *canonical HRF* > *HRF with derivatives* (from 0.46 to 0.63 in the combined effect

analysis and from 0.32 to 0.52 in the exploration of HRF derivatives effect only, see Figure 8.2). In the isolated analyses, the effect of changing the smoothing kernel FWHM was not very important in SPM ("5mm vs 8mm smoothing kernel FWHM"), which might explain why the false positive rates did not increase much in the combined effect analyses.

Under FSL, the previous analyses on the effect of each of these parameters separately (changing smoothing kernel FWHM and changing HRF model separately) both gave inflated false positive rates, and their combined effect largely increased the false positive rates (up to 0.77) compared to the effect of changing the use HRF derivatives alone (up to 0.49).

Similar observations can be made on the P-P plots on Figure 8.4 and 8.6.

In both SPM and FSL, the second set of analyses (*5 mm smoothing*, *24 motion regressors*) versus (*8 mm smoothing*, *0 motion regressors*), we found invalid results for nearly all combinations. In SPM, false positive rates were only slightly above the theoretical threshold of 0.05 (0.081 and 0.11), which is consistent with our previous observation: initially, changing smoothing kernel FWHM and number of motion regressors separately led to false positive rates close to 0.05, consistently, their combination led to rates that were only slightly invalid.

For both SPM and FSL, we observed shifts in the distributions of statistical values (Figures 8.7 and 8.8). These shifts were similar to those obtained for changes in motion regressors only.

8.3.3 Analyses using pipelines with different software packages

We also explored the ability to use in a same between-group analysis subject-level data obtained with different software packages (here FSL and SPM). We performed the analyses for all possible combinations SPM versus FSL: 2 smoothing kernels $\times$ 3 numbers of motion regressors $\times$ 2 HRF models, corresponding to 12 between-software comparisons – with the same setting for both SPM and FSL pipelines. The false positive rates are displayed in Figure 8.9. For all between-software analyses, the false positive rates were above 0.05. We obtained lower values for $SPM > FSL$ (between 0.10 to 0.32), than for the opposite test (between 0.56 to 0.95). In all cases, false positive rates were largely increased compared to the reference analyses (i.e. using the same software in both groups). This observation was consistent with the P-P plot, which showed a large deviation from the 95% confidence interval for the direction $SPM < FSL$ (Figure 8.10). Figure 8.11 shows the distribution of statistical values for the between-software comparison with all other

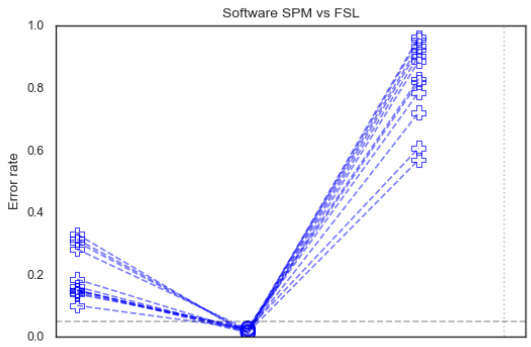

Figure 8.9 – False positive rates for pipelines with different software packages. For each pipeline combination, we provide the false positive rates obtained for: 1/ both tails, *i.e.* pipeline 1 > pipeline 2 and reverse (crosses) and 2/ for the corresponding analysis in which pipeline 1 and 2 are identical, *i.e.* the baseline (circles). The grey dashed line corresponds to the expected theoretical value (0.05).

parameters set with default values (i.e. 5mm smoothing kernel, 24 motion regressors and no HRF derivatives). We can see a shift in terms of mean and standard deviation of values. This shift was larger than those observed, for instance, for the effect of HRF derivatives, which was the most impacting factor on within-software comparisons.

8.4 Discussions

In this work, we showed that between-group analyses that use data generated by different pipelines can lead to invalidity (*i.e.* inflated false positive rates). In almost all cases, combining data processed with different pipelines led to false positive rates above the theoretical 0.05 threshold. These invalid results, obtained when combining subject-level contrast maps processed differently, suggest that it is necessary to consider how analytical variability may affect the results when combining data.

When performing analyses using the same pipeline on all participant data (as traditionally done in the literature), results were valid for all analyses. Although the false positive rates obtained in this situation were lower than the expected 5% rate, the results were similar to those obtained for a similar framework in Eklund et al., 2016.

Our results for different pipeline analyses suggest that some factors have a larger impact than others. We saw that for differences regarding the size of the smoothing kernel and number of motion regressors (6 versus 24 motion regressors) within SPM software

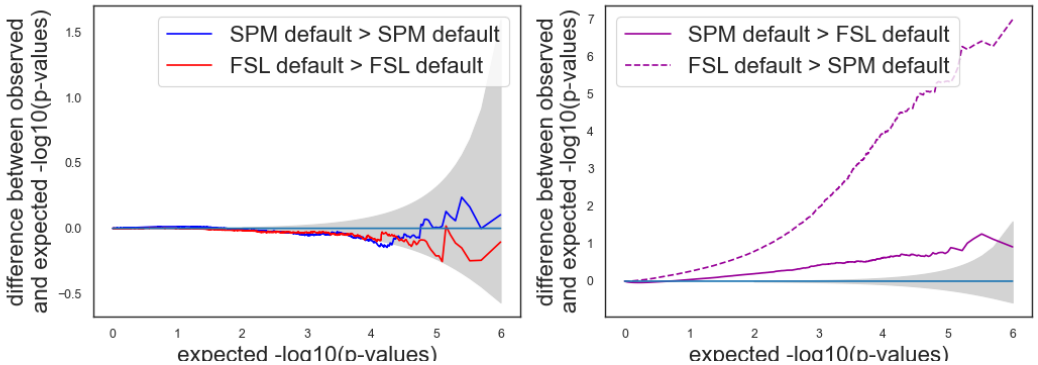

Figure 8.10 – Bland-Altman P-P plots for pipelines with different software packages. The grey shade corresponds to the 0.95 confidence interval. A curve above (respectively below) the confidence interval indicates invalidity (respectively conservativeness). Default parameters: 5 mm smoothing, 24 motion regressors and no HRF derivatives.

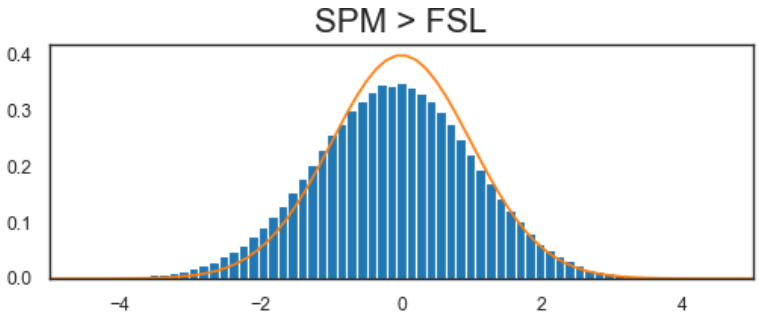

Figure 8.11 – Distribution of statistical values for between-software analyses, compared to the expected distribution.

package, results were similar to those obtained with identical pipeline analyses, suggesting that participant data can be combined without having to consider the differences in pipelines, if this is the only difference. This is not the case for differences in the use of HRF derivatives and use of motion regressors (0 motion regressors versus 6 or 24 motion regressors), which gave invalid results.

We also saw that combining multiple differences in parameters could result in bigger effects, depending on the effect of each parameter alone. The combination of two parameters that both have a high effect on compatibility led in our case to inflated false positive rates, while the combination of parameters that had a limited effect on validity did not lead to higher false positive rates (*e.g.* smoothing and motion regressors in SPM). This suggests that it may be possible to model the effect caused by specific variations in the

subject-level pipelines. To enable this in the future, it is essential that the pipelines used is shared with enough details to allow a reproduction of the exact processing applied on the data.

However, the ability to model the effect of parameters is limited to specific variations. For example, for each variation of parameter, we saw different effects across the two software packages under study (SPM and FSL). Overall, observations were similar, but false positive rates were often increased in FSL compared to SPM for the same comparison. This suggests that some parameters values are more robust to changes when combined together, here, in FSL, when using 24 motion regressors, combining data with different use of HRF derivatives led to false positive rates closed to the baseline analysis (*i.e.* same pipeline in both groups).

The most important source of invalidity was found when studying the effect of differences in software packages. SPM and FSL both implement similar pipeline steps with different settings. While we tried to align parameters between the two software packages by changing the software package default values (*e.g.* smoothing kernel, type of HRF, etc.), some steps are specific to each software and cannot be changed by the user, causing potential differences between the results. We tried to correct some of these differences, in particular for the unit scale of subject-level contrast maps. But, even with these corrections, we still found highly inflated false positive rates when comparing pipelines with the same values for the parameters under study and different software packages. We suppose that differences in how software packages scale the data were not compensated by our simple rescaling approach and that more work will be needed to be able to combine subject-level data from two different software packages in the same analysis.

In this work, we focused on between-group analyses in which each group of participants was processed with a different pipeline. In practice, other combinations may be observed, for example with multiple pipelines used within a group. The setup that we used here – in which processing pipelines varied depending on the group – was justified by the use-case in which data from various public datasets are used in the same analysis. For example, specific datasets have been created to study various neurological disorders, usually associated with a minimal processing pipeline dedicated to the study, and the corresponding minimally processed data (Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack Jr et al., 2008) for Alzheimer's disease, Autism Brain Imaging Data Exchange (ABIDE) (Di Martino et al., 2014) for autism, etc). Researchers may want to use these minimally processed data and compare groups of participants where each group corresponds to a

specific disease.

We chose to study variations induced by 4 types of parameters (software package, HRF, smoothing and number of motion regressors), within each software package based on their widespread use in the neuroimaging community (Carp, 2012a). Yet, in practice, there are many more variations: researchers might use different software versions, perform or not specific sub-steps in the analysis (for example, the use or not of slice-timing correction), use different HRF models etc. Therefore, in real conditions, the differences observed between pipelines will likely be more important. In future works, other analyses may be done for other varying parameters using the same framework.

Here, we showed that the effects of analytical variability often prevent doing a direct analysis without considering the differences in processing pipelines. For other sources of variability, methods have been proposed to remove unwanted variance: for example, correcting the variability resulting from imaging site and scanner effect (technical variability) in neuroimaging (Beer et al., 2020; Fortin et al., 2016). Recently, deep learning frameworks, and in particular generative models used for style transfer (Gatys et al., 2016), showed their potential for such task in converting data between different domains (*e.g.* acquisition site) (Liu et al., 2021). Considering the achievements of these models, there is reason to anticipate their success in transitioning between other domains, such different analysis pipelines. We explored the ability to convert data between pipelines using style transfer frameworks in Chapter 5.

Take-home Message

- We explored the validity of mega-analyses using subject-level data processed with different pipelines by performing between-group analyses under the null hypothesis.
- We extend the work from Rolland et al., 2022 by integrating within-software combinations with FSL and between-software combinations.
- We showed that the combination of data processed with different software packages lead to the higher level of inflation of the false positive rates.
- Our results suggest that it is impossible to combine processed fMRI data without taking into account differences in subject-level processing.

Conclusion and perspectives

Conclusion

Summary

Building fMRI studies is challenging, in particular with the high flexibility of analyses methods and the limited generalizability of findings, to different populations and/or to different pipelines. The aim of our work was to embrace two main challenges related to analytical variability in fMRI results: *(i)* facilitating the re-use of derived data shared on public databases, and *(ii)* exploring and understanding relationships between pipelines to guide researchers.

In our first set of contributions, we derived practical solutions to facilitate data re-use for researchers and thus, increase sample sizes to help solve the issue of low statistical power identified as one of the main obstacle to the robustness of brain imaging studies. To do so, we learned meaningful lower-dimensional representations of fMRI statistic maps with deep representation learning, and applied two techniques to transfer and manipulate these representations.

First, in Chapter 4, we used self-taught learning to improve the performance of deep learning models on fMRI statistic maps for supervised tasks such as brain decoding. We showed that pretraining with an unsupervised task on a large and diverse database (NeuroVault (Gorgolewski et al., 2015)) was beneficial, in particular for small sample sizes and complex classification tasks. This benefit was associated with a better generalizability of pretrained models, as these learned more general features and less individual characteristics. By sharing our pretrained unsupervised model to the community, we give the opportunity for researchers to re-use it for other tasks and thus, improve the robustness of future studies on classification tasks with fMRI.

We also proposed to apply specific representation learning frameworks to facilitate the re-use of derived data shared on public databases for statistical studies. In Chapter 5, we made the assumption that the pipeline used to compute statistic maps could be seen as a style component of images. We adapted several state-of-the-art I2I frameworks, as well as a newly developed framework based on DDPM. Our results are promising, in particular with GAN frameworks and complex transfer (*i.e.* from pipelines with distant results),

and give hope for the future of data re-use.

In the second set of contributions, we focused on the exploration of the fMRI analytical space. Our goal was to better understand the contexts in which the solutions that we proposed in the first part could be or must be used. We developed and shared a multi-pipeline dataset that can be re-used by other researchers (potentially outside neuroimaging) to explore analytical variability in fMRI results. We described this dataset in Chapter 6. Using this dataset, we derived two studies in which we aimed to gain a better understanding of the relationships between pipelines results. By exploring the stability of these relationships (Chapter 7), our goal was to better understand the lack of generalizability of the style transfer frameworks that we developed in Chapter 5. While these relationships seemed stable across different groups of participants, the communities of similar pipelines identified were different between paradigms. This confirmed our observations, but also provided some knowledge for future works on building more generalizable style transfer frameworks that could be applied to statistic maps from any paradigm.

Our study on the validity of mega-analyses with data processed from different pipelines in Chapter 8 allowed us to identify the critical cases were combining such data would lead to high false positive rates, and thus invalid studies. The most critical case was identified as the studies combining data from different software packages. We tested the ability of our style transfer frameworks on this context and found satisfying results in terms of similarity to the ground-truth statistic maps of the target pipeline. While we found substantially lower performance for other transfers, the identification of these critical cases, combined with the performance of the frameworks, give hope for the future of data re-use.

In Appendix, we describe a third set of contributions in which we explored the impact of analytical variability on the reproducibility of fMRI studies. In collaboration with Prof. Tristan Glatard and Prof. Jean-Baptiste Poline (see Appendix B), we studied the replicability of resting-state fMRI derived biomarkers of Parkinson's disease. We showed that variations in cohort selection, image processing pipelines and machine learning frameworks could have a large impact on the performance of prediction models, making challenging their application in clinical practice. In Appendix C, we present the project NARPS Open Pipelines led by our lab. The goal is to reproduce the pipelines used in a many-analyst study and share these as a resource for the community. We describe the main challenges of pipeline reproduction using textual description and provide an example of application of the codebase to explore the impact of sample size on analytical variability.

Perspectives

Short-term perspectives

Future works would be needed to strengthen our contributions, in particular to validate the solutions that we developed for researchers to more easily re-use shared data.

Exploring other benefits of self-taught learning. In Chapter 4, we showed the benefits of using data shared on NeuroVault (Gorgolewski et al., 2015), a public database composed of a large number of statistic maps coming from different studies, to pretrain an unsupervised deep learning models that can be fine-tuned for other purposes. Here, we limited our experiments to supervised tasks with the labels provided by the database (*e.g.* paradigm, task) and we applied the frameworks to two datasets (*e.g.* homogeneous, with data from a single study and heterogeneous, with data from several studies but analyzed with the same pipeline). Further work would be needed to investigate the performance of classification on other target datasets with other sources of variability, for instance with statistic maps from different studies but also processed using different pipelines. Several studies (Vu et al., 2020; Li et al., 2023) showed that deep learning models might fail to generalize to new data analyzed with a different pipeline than the one used on the training set. Comparing the adaptation capacities of models on volumes preprocessed with different pipelines could be interesting to evaluate the impact of analytical variability on deep learning with fMRI statistic maps and to see if the generalizability of our pretrained models also works for inter-pipeline differences.

Optimizing DDPM-based frameworks for style transfer In another attempt to facilitate data re-use for statistical studies (Chapter 5), we explored the potential of image-to-image transition frameworks to convert statistic maps between pipelines. Our results were promising and showed that GAN-based frameworks, even in unsupervised settings with StarGAN (Choi et al., 2018), were able to transfer statistic maps in a target domain with high similarity to the ground-truth target. We obtained lower performance using DDPM-based models, probably due to the specific functioning of such models, which is not initially suited for image-to-image transition (Ho et al., 2020). Training DDPM-based frameworks was relatively time consuming, and we did not perform any hyper-parameter optimization, on the number of denoising steps for instance. Moreover, recent studies showed the benefits of using latent diffusion models, that act in the latent space

of a Variational AutoEncoder (VAE) to reduce the size of data and facilitate training of DDPM-based frameworks (Rombach et al., 2022). In future works, we would like to focus on these frameworks to better understand their lower performance compared to GAN-based frameworks. This was the topic of a Master's degree intern that we co-supervised since October 2024 with Pr. Elisa Fromont.

Re-using data converted with style transfer The practical usability of the proposed frameworks remain questionable. Further work would be needed to assess the potential of these newly transferred statistic maps for statistical studies. In Chapter 8, we computed false positive rates of mega-analyses combining subject-level data obtained from different pipelines. This method could be applied with between-group analyses composed of 1) a group with data from the target pipeline and 2) a group with data originally obtained with another pipeline and that have been converted to the target pipeline.

However, for now, due to the standardization of the data used as input to deep learning models and the architecture of the models, voxel values in generated maps are constrained between -1 and 1. First attempts have been made to de-normalize data using a scale factor derived from the source map. Further work would be needed to deal with the case of maps coming from different software packages. For instance for a transfer from FSL to SPM, differences in percent BOLD change (*i.e.* unit of fMRI contrast maps (see 8.2.2.1) would have to be taken into account.

Towards a better understanding of the relationships between pipeline results In Chapter 7 and Appendix E, we showed that relationships between pipeline results were stable across groups of participants, but not across paradigms, which explained the lack of generalizability of style transfer frameworks to data from unseen paradigm.

In the context of mega-analyses, researchers would have access to data from two (or more) pipelines A and B in the same paradigm. Thus, a single framework could be trained to transfer data from A to B, or the inverse. However, researchers could also have access to data from a pipeline A with the paradigms 1 and 2 and to data from a pipeline B with only the paradigm 2. In this context, they might want to apply our frameworks to convert data from A in paradigm 1 to pipeline B to visualize the differences between the results obtained from the two pipelines without having to recompute the whole pipeline (or if they do not have access to raw data, or to information about the pipeline B). Transductive transfer learning (Arnold et al., 2007) aims to improve the learning of a

target task in a target domain using knowledge from a similar task in a source domain. In particular, in unsupervised transductive transfer learning, there is no labeled data from the target domain. Further experiments would be needed to investigate whether transfer learning frameworks could be helpful to obtain more generalizable results in our context.

Long-term perspectives

In this work, we proposed several solutions to leverage publicly available data (*e.g.* NeuroVault (Gorgolewski et al., 2015), Human Connectome Project (Van Essen et al., 2013), etc.). In practice, even if these data provide an easy and accessible solution for researchers to build larger and more generalizable studies, re-using more sensitive data from hospitals or clinical trials comes with more challenges. First, hospitals may be reluctant to share sensitive patient data, in particular due to data privacy regulations like GDPR. Moreover, real-life data are usually noisy: they often contain inconsistencies and missing values, they are imbalanced and have inherent biases that can lead to unfair model prediction (Ricci Lara et al., 2022). Finally, even if hospitals were sharing re-usable models and not the data themselves, as done with our *self-taught learning* framework, the models could be sensitive to privacy attacks such as membership inference or data retrieval (Aguiar et al., 2023).

Federated learning: benefiting from data without sharing data Federated learning (Rieke et al., 2020) has shown promises in the field of machine learning for healthcare, by enabling the training of several models across multiple clients on their local data, without exchanging the data itself. As stated by data protection regulations such as GDPR, since individual characteristics can be used to retrieve the identity of a participant, data are usually not considered as anonymized. With federated learning, each client defines its own data governance system, allowing the model to benefit from numerous and diverse data without directly hosting the data. This also makes the process of data storage easier, as healthcare data can be high-dimensional and thus, very costly in terms of storage space.

By leveraging diverse, decentralized datasets across different clients, this technique protects sensitive information while improving the fairness of models by using larger and more diverse data. This is crucial to capture subtle relationships between patterns and outcomes. However, for federated learning frameworks to perform accurately, data from different clients still need to be independent and identically distributed, which is

sometimes not the case. Research addressing this issue includes, for instance, works on federated learning coupled with domain adaptation (Li et al., 2020). Moreover, this technique does not prevent privacy attacks (Lyu et al., 2024), as described in Chapter 3.

The development of federated learning in healthcare centers is thus a promising solution to re-use sensitive data in accordance to data protection regulations, but also to build more generalizable models with diverse sets of data.

Synthetic datasets: a solution for unbiased and privacy-preserving datasets Another solution to avoid sharing sensitive data while building large and diverse datasets is to make use of synthetic data. With generative models such as GAN (Goodfellow et al., 2014) (described in Chapter 3), it is possible to generate synthetic data that mimic the distribution of an original dataset while not being associated to a particular participant. This allows to encompass the data protection regulation rules and to share healthcare data with the community to build more robust models. In machine learning for natural images, the release of ImageNet (Deng et al., 2009) was the major breakthrough and led to a large increase in terms of performance. The goal of using large datasets of synthetic data would be to reach the same level of performance.

However, synthetic data are not real data, and we might wonder if the use of synthetic data would not lead to a decrease in terms of performance compared to real-life data. Indeed, generative models might fail to represent some characteristics of the data, or overfit to some others. For now, there is no standard method to evaluate the quality of synthetic data in medical imaging and metrics used for natural images might fail in this context (Thijs Kooi, 2018). Finally, similarly to federated learning, the question of the privacy of synthetic data remains an issue, as generative models can also suffer from privacy attacks. Techniques based on differential privacy started to emerge to build more privacy-preserving generated models (Yoon et al., 2020).

Such large synthetic datasets of healthcare data would facilitate the development of machine learning model since they would allow researchers to work with data that mimic the real-world, with minimized privacy issues.

Differential privacy: towards more robust models As stated in the previous paragraphs, the lack of privacy of machine learning models gives rise to concerns, in particular in healthcare settings with sensitive data. Differential privacy can be defined as the absence of changes in the outcome of any computation done on the model when including

or removing an individual record from the dataset (Abadi et al., 2016). This is usually done by adding noise to the model to mask the contribution of any individual while preserving accuracy (Dwork et al., 2014). However, there are no guidelines to tell researchers what differential privacy entails and which guarantees to aim for. Moreover, it is usually difficult to achieve a good benefit-risk tradeoff (*i.e.* increasing privacy without decreasing performance). Recent works focused on the implementation of privacy losses in different machine learning frameworks, by focusing on iterative models (such as DDPM (Ho et al., 2020)) and better privacy-utility trade-offs (Das et al., 2024).

Incorporating differential privacy constraints would be helpful to facilitate the sharing of machine learning models while taking into account privacy concerns. This would lead to an increase confidence when sharing models in the context of federated learning, but also to build synthetic data that are different enough from the original data so that these could not be re-identified.

Overall, these are promising techniques for leveraging healthcare data while adhering to privacy regulations. Federated learning enables to use decentralized data without direct sharing, and synthetic data generation creates privacy-preserving datasets that mimic real-world data. Additionally, incorporating differential privacy enhances model robustness by protecting individual data contributions. These approaches are poised to drive future advancements in building generalizable and privacy-compliant machine learning models in healthcare.

Appendix - Contributions on reproducibility

Appendix A

State-of-the-art on reproducibility

In the contributions of this thesis exposed in the previous chapters, we explored the fMRI analytical space and proposed solutions to facilitate the re-use of data from public databases. Here, we extend this work and describe the challenges related to the methodological flexibility of studies in the reproducibility of results. In this chapter, we give an overview of the state-of-the-art on the reproducibility crisis that arised in the last decade. In the next chapters, we propose two contributions that explore the reproducibility of fMRI studies: in clinical settings in Appendix B, and in the context of a many-analyst study in Appendix C.

In experimental research, researchers follow the scientific method to make new contributions. This typically start by the formulation of a research question and the exploration of the state of the art of this topic. Then, researchers establish an hypothesis and design one or more experiments to test it. This process usually begin by the reproduction of a published claim, to investigate the method in more details and to build new advances. After analyzing the results, researchers draw a conclusion and determine if we can trust this finding. At this stage, researchers usually try to reproduce their results with the same settings and evaluate the impact of variations to improve their confidence in their results and to identify any bug or error. At each stage of a project, the notion of reproducibility is present. Robust and reproducible research is thus the foundation on which new findings are developed (Begley et al., 2015). During the last ten years, experimental research faced a "reproducibility crisis", in which the validity of published results was put into question (Ioannidis, 2005). Researchers attempted to reproduce or replicate several research findings, with low rates of success (see for example (Open Science Collaboration, 2015)). This crisis encouraged researchers to question their practices and to develop solutions to build a more trustworthy research.

In this chapter, we first give an overview of the issues and concerns that led to the

"reproducibility crisis" (A.1). In (A.2), we explore the different aspects of reproducibility. We then describe the causes that have been identified for this lack of reproducibility in experimental research (A.3) and expose the solutions proposed by researchers to adress these issues (A.4). For each cause and solution, we first describe the general case and then we detail the specificities related to neuroimaging research.

A.1 The reproducibility crisis

The first concerns regarding the validity of research findings appeared in 2005 with a theoretical analysis made by Ioannidis, 2005. In this paper, Ioannidis, 2005 stated that due to publishing and analysis practices, in particular the low statistical power of studies, it was likely that more than half of the published results were false, and therefore irreproducible.

Numerous studies emphasized this issue by attempting to reproduce the results of published research findings. In the "Reproducibility Project", led by Open Science Collaboration, 2015, several researchers tried to reproduce the methods and results of 100 psychological studies published in three renowned journals. They used different criteria to assess reproducibility and obtained relatively low agreement between original and reproduced results. In the end, only 39 reproductions were concluded as successful. In another study on drug development, an industrial laboratory reported having successfully replicated the original results of landmark findings in only 25% of the attempted cases (Prinz et al., 2011). Such studies led researchers to consider more seriously the concept of reproducibility and to question their research practices, as it might undermine the reliability of research results.

In Baker, 2016, the *Nature* journal took a survey on 1,500 researchers about their opinion on the crisis. Most scientists (more than 70% of the respondents) have experienced failure in trying to reproduce the results of an experiments and 52% agreed that there was a significant "reproducibility crisis". However, 31% were still trusting the published literature and believed that the lack of success of reproduction studies was not related to the validity of the original results, but more likely to errors in the reproduction process.

Such errors can arise due to the complexity of reproduction experiments, and these are mostly due to missing information about the protocol and data used in the original study. The question of the ability to reproduce a study with sufficient materials also takes part of the crisis. Multiple reproduction studies concluded that reproduction was impossible

due to missing informations (Begley et al., 2015). Without transparency, it is impossible to assess the validity of a protocol, or to detect any error or fraud, which emphasizes the lack of reliability of published results and the importance of reproducibility.

A.2 Evaluating reproducibility

The term "reproducible research" can be understood differently across fields (Barba, 2018). The Turing way – a collaborative open science handbook and community (Community et al., 2019) – divides reproducibility into four different aspects.

— *Reproducible*: same data, same analysis

— *Replicable*: different data, same analysis

— *Robust*: same data, different analysis

— *Generalisable*: different data, different analysis

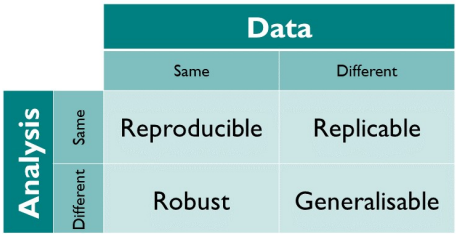

Figure A.1 – Reproducible research defined by The Turing way.
Credits: CC-BY reproducible matrix ©The Turing Way Community.

The first one is also known as *computational reproducibility* (Claerbout et al., 1992) and requires the exact same code and data as in the original study. In practice, the difficulties encountered to ensure long-term availability of such materials makes the evaluation of reproducibility a challenging process (Perkel, 2020). Moreover, as stated in the *Nature* survey (Baker, 2016), researchers who tries to reproduce an experiment rarely contact the authors of the original publication when they fail or if they need additional materials. In the following, we will use *reproducibility* as a general term englobing the four aspects.

Researchers can also evaluate the *replicability*, the *robustness* and the *generalisability* of a study. Experimental protocols are subject to several sources of variability including those related to the choice of dataset (intra-subject variability, *i.e.* differences in data acquired at different moments for the same participant, inter-subject variability, *i.e.* differences

between samples of the population or between different populations) or to the choice of analysis (different data acquisition or analysis protocol). Variations that might appear in the results under different conditions informs the researchers on the finding: Is the finding specific to the data and analysis setup of the study? Or is it generalizable under different conditions?

Reproducibility can be explored under different conditions but the success of the reproduction can also be assessed using different criteria. In the Reproducibility Project (Open Science Collaboration, 2015), conclusions of the studies were presented with the results of statistical inference tests and effect sizes. The first criteria was the comparison of the final conclusions of the original study with the reproduced results, *i.e.* the finding is significant (or not) in both studies (success), or the finding is significant in one and not significant in the other (failure). They also compared the effect sizes detected in the original study and the one of the reproduction: Is the original effect size in the 95% confidence interval of the reproduced effect? Are the original and reproduced effected in the same order of magnitude? Original and reproduced studies were also combined in meta-analyses to obtain a mean estimates of the effect size, which can help to evaluate the reliability of the results. Finally, researchers also gave a subjective opinion on the success of the replication. All these criteria evaluate reproducibility at different levels and can lead to different conclusions on the success of the reproduction. For instance, in Open Science Collaboration, 2015, 36% of the reproductions were found successful with the significance criteria, whereas 47% were successful when comparing effect sizes with the 95% confidence interval.

A.3 Causes of irreproducibility

This "reproducibility crisis" encouraged researchers to re-think the way they were doing research, and in particular to identify the practices that were leading to the low reproducibility of findings.

A.3.1 Lack of information on datasets

As mentioned in (A.1), missing information about the analysis protocol and the data used can make it challenging to reproduce a study. Data used in published studies are particularly useful for reproducibility. Although the importance of data sharing is ac-

knowledged today in many fields of research, it was not always the case and their availability differs across scientific domains. Tedersoo et al., 2021 evaluated data availability in different disciplines and found percentages of 54% and 72% respectively for full and at least partial data availability, with a lower availability rate for fields like ecology or psychology. After contacting the authors, these percentages increased but some authors remained reluctant to share their data. The perceived barriers to data sharing are numerous (Gomes et al., 2022), and can be related to lacks of knowledge of the process, concerns about data re-use and lack of dedicated infrastructures. Sharing data requires to use appropriate tools, or databases, to make them available to the public. Such process can be time consuming for researchers who need to learn to master the tools but also think about the best ways to share their data in a way that is easily accessible for the public while respecting legal constraints.

Specificities in neuroimaging. Neuroimaging data are specific and data sharing requires a dedicated methodology (Poline et al., 2012). Indeed, datasets are usually composed of high dimensional data that take up a large amount of storage space. The platforms used to store the data must be adapted to facilitate the access for future reuse. Moreover, sharing neuroimaging data can be difficult due to specific regulations, such as the General Data Protection Regulation (GDPR) in Europe which puts protection on shared data that could be individualized. In the US, anonymisation solely means that any primary identifiable information (*i.e.* name, date of birth, address and face) should be removed, while in Europe this includes any kind of correlation, individualization and inference. Despite this issue, even when reusing data from public databases in the US, researchers usually signs a Data Usage Agreement (DUA) that sometimes stipulate that any derived data must be shared under the same DUA. Moreover, many levels of derived data can be shared by researchers, including preprocessed data, statistical maps at the subject-level and at the group-level. All these data requires large storage space, but also a specific organization with formatting and addition of metadata for example. Such formats would make data easily reusable, notably for the reproduction of experiments, but requires specific tools for data conversion for instance and to re-organize the data (Li et al., 2016), and researchers might lack knowledge for such tools, preventing them from sharing data. In neuroimaging, the BIDS (Gorgolewski et al., 2017) format is now widely used in the community to standardize the organisation of datasets.

A.3.2 Lack of information on the analysis protocol

With advances in research, workflows became more and more complex with many steps and parameters. Detailed informations are necessary to reproduce the experiments but also to evaluate the quality and validity of the results by detecting misconducts or errors. In many fields, researchers rely on software packages that implement an analysis workflow with default parameters and thus, they do not share the details of the workflow when default values are used. However, these values can vary from a software package to another and between versions of a same software package, which can prevent from reproducing the same results. Even when there is enough information, the way information is delivered can impact the success of the reproduction and the difficulties encountered to reproduce. For instance, a recent study (Laurinavichyute et al., 2022) showed that reproducibility increased by almost 40% when the analysis code was provided. When manual steps are performed in the workflow, sharing the process in a fully reproducible way can be difficult. While the best practice is to use reproducible code, some workflows might require manual inputs and some steps might be hardly translated to code. In such cases, researchers should provide detailed explanations on their workflows and the potential manual steps involved.

Specificities in neuroimaging. Neuroimaging studies requires multiple processing steps for which multiple options are available. Reporting the full detailed analysis in a paper is impractical, even with the development of standard ways to describe the neuroimaging workflows (*e.g.* COBIDAS (Nichols et al., 2017)). In Carp, 2012b, 241 fMRI studies were analyzed and authors showed that detailed information regarding certain crucial steps such as data acquisition, processing and statistical analysis was missing in a large number of studies. Moreover, the neuroimaging fields is composed of researchers from different domains, which might not be familiar with programming and code sharing platforms. Also, even if the code is shared, software packages used in neuroimaging can evolve to different versions, which can lead to changes in algorithm used to perform certain operations (Gronenschild et al., 2012). Other studies showed the impact of Operating System (OS) and versions in the way calculations are done (Vila et al., 2024; Glatard et al., 2015), which may lead to important changes in the results when applied on successive operations.

A.3.3 Lack of generalizability and fairness

The sucessful reproduction of a finding with same data and analysis protocol does not guarantee similar results in another experiment addressing the same research question with different conditions, in particular with other data. Research experiments are ideally performed by randomly sampling from the target population. This process can lead to sampling variability, which refers to the fact that the statistical information of a sample varies as the random sampling is repeated. In machine learning for instance, models are trained on a set of data and may "overfit" on these data by fitting the random noise rather than a true signal likely to generalize across samples. This leads to good predictions in the training data that do not generalize to the testing data. In a recent study on treatment outcome prediction for schizophrenia, Chekroud et al., 2024 showed that models that performed well in terms of accuracy in their tranining sample or in a test set sampled from the same population routinely failed to generalize to unseen patients, in particular when those are sampled from a different context (*e.g.* age, disease type, etc.).

Indeed, samples can be biased against some specific criteria depending on how they were recruited, *e.g.* same neighborhood or place of work. Such samples are not representative of the whole population, and thus, an experiment might give different results when applied to a different sample. A recent study in medical imaging (Larrazabal et al., 2020) showed the importance of representative samples in machine learning. They trained a classifier to detect lung opacity using data from males participants only and found a large drop in performance when applying it to women (and vice-versa). This issue is present in multiple fields, in particular when using artificial intelligence related tools that are usually trained on biased datasets (Buolamwini, Joy, 2019).

Specificities in neuroimaging. Inter-individual differences are particularly important in neuroimaging, where the inter-scan variability within an imaging session is very small in comparison to the variability of responses from subject to subject (Holmes et al., 1998). Thus, sampling variability can lead to poor generalizability of studies, in particular when using small sample sizes. Moreover, due to the complexity and the cost of acquiring data and the difficulty to gather participants for specific studies, participants often come from the same location (*e.g.* in a close neighborhood around the MRI facility), leading to unrepresentative samples.

A.3.4 Publication bias

Publication bias is often identified as one of the main cause of the "reproducibility crisis". This term refers to the fact that negative results (non significant) are usually withholded from publication, leading to the presence in scientific litterature of a large proportion of papers with significant results. At first, the beliefs were that non-significant results were associated with poorly formulated and tested experiments, which would be rejected by the competition for publishing and fundings. However, due to the low statistical power of studies, numerous published significant results are actually false positives (Ioannidis, 2005). Studies showed that this competition for fundings and the pressure of publication that led researchers to submit mostly significant results for publication (Munafò et al., 2008; Munafò et al., 2009). Such results are more likely to be considered favorably by editors, receive more favorable peer reviews, and subsequently, to be cited more often once published.

Such publication practices can have devastating consequences on research. For instance, more than a thousand published papers showing the effectiveness of antidepressant treatment on depression were put into question by meta-analyses of FDA (Food and Drugs Administration) data showing potential biases such as selective publication of positive results (Ioannidis, 2008b). This overestimation of the effect of antidepressant had a large impact on treatment decisions for patients suffering from depression, but also on pharmaceutical laboratories and researchers that put some time, money and efforts on research built on nonexistent foundations.

A.3.5 Low statistical power

In experimental research, findings are often tested using statistical tests in which the research hypothesis - named H_A for alternative hypothesis - is compared to a null hypothesis - H_0 - set as the opposite of the alternative hypothesis, *e.g.* absence of effect. The likelihood of rejecting H_0 when it is actually false (and thus, H_A is actually true) is called the *statistical power* of the test (see Table A.1). The power of a study depends on the sample size used and of the true effect size. Studies with small sample sizes and low effect sizes will have a low statistical power, which thus reduces the chance of detecting a true effect in these studies. In many research fields, such studies are quite common as data acquisition can be complex and costly, and modern studies increasingly target small sample sizes (Vesterinen et al., 2011). However, this also impacts the probability of a

detected effect to be effectively true, *i.e.* the Predictive Positive Value, which depends on the statistical power of the study and the level of statistical significance α (Ioannidis, 2008a). Button et al., 2013 demonstrated that both small sample sizes and low effect sizes in neuroscience led to reduced statistical power and thus, small positive predictive values, leading to high proportion of false positive findings in the literature. The α-level controls for false positives in situations where no effect exists, but does not prevent from the overestimation of the effect. Moreover, positive results are favored for publication, leading to a biased scientific literature.

	State of the world	
	H0 True	H0 False
Accept H0	True accept $(1-\alpha)$	False accept Type II error (β)
Reject H0	False reject Type I error (α)	True reject Power $(1-\beta)$

Table A.1 – Principles of statistical testing.

Specificities in neuroimaging. As stated in A.3.3, gathering participants for neuroimaging studies can be difficult, leading to small datasets. A study evaluated the evolution of sample sizes in neuroimaging studies until 2015, and pointed out that statistical power was too low to find reasonable effect sizes, in particular in a field with low effect sizes. In 2015, the median sample size of fMRI studies was estimated at approximately 30, corresponding to a median effect size associated to a standard 80% statistical power equal to 0.75, which is very high for such studies.

A.3.6 Analytical flexibility

With the advances in analysis methods, researchers often have the possibility to select the method to use amongst a collection of possible analytical choices (Simmons et al., 2011). This phenomenon is also known as *analytical flexibility* or as the *researcher's degree of freedom*. While it is common to explore different analytic conditions, results can be biased if these conditions are not set beforehand and if researchers only report the results of one analysis. In theory, when multiple conditions are explored, a correction for multiple comparisons should be applied in order to keep the guarantee provided by the α-level across the family of tests.

Moreover, a study can obtain different estimates of the effect depending on the analytical options it implements, defined as *vibration of effects* (Ioannidis, 2008a). This phenomenon is exacerbated in small studies in which the results are more prone to uncertainties and variations (Loken et al., 2017). In addition to the multiple testing, this can lead to an increase of false positive findings in the literature. In particular, when the different analytical conditions are tested in order to obtain the desired (usually more significant) results, this selective reporting can be seen as a case of malpractice (referred to as p-hacking or data dredging). Such practice usually arise due to the complexity of analysis workflows, a lack of statistical knowledge and the absence of consensus for analysis methodologies.

Specificities in neuroimaging. As stated in Chapter 2, *vibration of effects* is particularly present in neuroimaging studies due to the large number of steps required to analyze data. Minimal processing workflows were developed to reduce the number of decision to make for authors (Esteban et al., 2019), but these are still not fully deployed in the community and researchers still have to make decision for other steps of the analysis. These choices can greatly impact the results. In a many-analyst study (Botvinik-Nezer et al., 2020), 70 research teams analyzed the same fMRI dataset and overall, there were no identical pipeline. Variations were present in the results in terms of final statistical maps but also answer to binary research hypothesis, showing the large impact of analytical flexibility on the possibility of finding false positive results in neuroimaging studies due to the *vibration of effects*.

A.4 Proposed solutions

A.4.1 Sharing data

Making data available to researchers allows them to reproduce the exact results of a study. This might also help them to detect any potential fraud or error in the analysis and thus, increase the reliability of the results. To address this, several journals have implemented 'open data' policies to encourage or require data sharing. Some editorial boards or reviewers also aknowledge the efforts that are made by authors to make their studies reproducible. Initial findings suggest that these policies have led to a significant rise in articles reporting publicly available data (Hardwicke et al., 2018b; Laurinavichyute

et al., 2022). For instance, Hardwicke et al., 2018b examined the impact of this policy on the journal *Cognition* and found an increase of shared data. However, a considerable portion of supposedly available data still remain missing, incomplete, incorrect, or poorly documented, making it unusefull for reproduction.

To face these issues, data sharing platforms were also developed and enhanced: some are non-specific to a field such as Zenodo European Organization For Nuclear Research et al., 2013, while others are specifically used by researchers from a domain. For instance, in neuroimaging, data can be shared on OpenNeuro (Markiewicz et al., 2021) (for raw data) or NeuroVault (Gorgolewski et al., 2015) (for derived data) following the US reglementations and on Public nEUro (*Public nEUro* 2020) following European regulations. To facilitate data sharing and reuse, the Brain Imaging Data Structure (BIDS) (Gorgolewski et al., 2016) - a standard organization of files and directories for neuroimaging datasets - is often applied to the shared datasets.

These solutions are valid for future publications, but other initiatives try to retrieve the data used in passed publications. Hardwicke et al., 2018a report the outcome of our efforts to retrieve, preserve, and liberate data from 111 of the most highly-cited articles published in psychology and psychiatry between 2006 and 2016. Even if some authors were reluctant to share their data, such initiative can help surface barriers to data sharing, and advance community discussions on data management.

A.4.2 Sharing analysis protocol and code

The first step to make the protocol easily available for the community is to share the analysis code on version control platforms like GitHub or Gitlab. Only sharing the repositories URLs on these platforms may not be sufficient to ensure long-term accessibility of the code. The Software Heritage initiative (Cosmo et al., 2017), which aims to collect, preserve, and share the entire corpus of publicly accessible software source code, provides a good opportunity to tackle this issue.

The second step is to provide researchers a good documentation and a high-quality code to facilitate the reproduction. In most neuroimaging software packages, researchers usually have the possibility to save pipelines in batch or script formats. This can facilitate reproducibility of processing steps, but does not take into account for pipelines implemented on several software packages and scripts are often very specific to the software package, which can make their comprehension difficult for new users. Nipype is a Python project that provides a uniform interface to existing neuroimaging software

packages and facilitates interaction between these packages within a single workflow. Combining such frameworks with "literate programming" tools like Jupyter Notebook or reproducible preprints like NeuroLibre (DuPre et al., 2022) enables researchers to present detailed analysis alongside executable code, enhancing comprehension of each step in the workflow.

Finally, the exact reproduction of experiments through execution of code requires access to the environment in which the code was executed when it is possible to limit the impact of variations in code or OS. Containerization of the workflow using Docker (Merkel, 2014) or Singularity (Kurtzer et al., 2017) may be used to allow the reproduction of results with the exact software versions, and Virtual Machines can be used to adapt the computing environment.

A.4.3 Increased results validation

An important issue with the generalizability and replicability of findings is the over-estimated effect size detected in the original study. To face this issue, guidelines and best practices were developed to improve validation procedures. For instance, for machine learning studies (Varoquaux et al., 2023; Ricci Lara et al., 2022), solutions have been derived to deal with imbalanced datasets, to mitigate model biases through data augmentation or adversarial training and to evaluate performance objectively with no leakage (*e.g.* using cross-validation).

A.4.4 Sample sizes and data re-use

The low statistical power of studies can be counterbalanced by increasing sample sizes. To this end, large scale studies emerged with the Human Connectome Project ($N = 1,000$ participants) or the UK Biobank (Miller et al. 2016), ($N = 100,000$ participants. However, these studies provide data for a limited number of cognitive tasks and thus, cannot be used to answer all research questions.

Another advantage of data sharing is the possibility to reuse shared data in new studies. Such combined datasets would increase the sample sizes of neuroimaging studies, but also would provide a larger variability in datasets, making the studies theoretically more generalizable. In practice, re-using data is not easy: one must first search for the different data to combine and process them while taking into account the differences in data acquisition or pre-processing if using derived data. For the first issue, tools

like NeuroBagel (Jahanpour et al., 2023) were created to query databases depending on specific criteria to build a new cohort composed of data from different datasets.

A.4.5 Practices in data analysis

To facilitate the choice of analysis workflow for researchers and limit the impact of analytical flexibility, metrics were developed to optimize pipelines based on various factors including reproducibility (Strother et al., 2004). Standard workflows were also developed, to minimize the number of input to make by researchers (Esteban et al., 2019). These solutions limit the number of workflows tested by researchers during their experiment and thus, limit the potential false positive findings.

Another solution to avoid these is to pre-register the analysis (Chambers et al., 2015). Researchers first design their experiment, send a first version of the paper, without any results. The review process is thus made on based on the methodology and on the research question, but the results are not taken into account. If the paper is accepted after the review process, the researcher can begin the experiments and add the results to the paper. The paper can thus be published, whatever the significance of the results.

Appendix B

Reproduction and replication of a study: Predicting Parkinson's disease trajectory using clinical and functional MRI features

---

This chapter was the subject of an article in revision at *PLOS ONE*:

- **Title:** Predicting Parkinson's disease trajectory using clinical and functional MRI features: a reproduction and replication study
- **Authors:** Elodie Germani, Nikhil Baghwat, Mathieu Dugré, Rémi Gau, Albert A. Montillo, Kevin P. Nguyen, Andrzej Sokołowski, Madeleine Sharp, Jean-Baptiste Poline, Tristan Glatard
- **HAL:** inserm-04465765
- **Code:** swh:1:snp:ac39cd7495afa754e5d0d298a502cda8684c7eca
- **Contributions (Credit taxonomy):** Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualisation, Manuscript writing.

B.1 Introduction

Parkinson's Disease (PD) is the second most common neurodegenerative disorder with more than 10 million people affected in the world. Disease manifestations are heterogeneous and their evolution varies between patients, dividing them in different subtypes and

stages (Bloem et al., 2021). Identification of these stages or subtypes is essential for clinical trials as well as for clinical practice to track the disease progression. However, there is currently no established biomarker of disease severity or progression (Gwinn et al., 2017; Mitchell et al., 2021).

Neuroimaging techniques are able to capture rich and descriptive information about brain structure and functional architecture non-invasively. In conjunction with computational algorithms based on pattern recognition and machine learning, neuroimaging measures began to emerge as candidate PD biomarkers in the past few years. Among other imaging modalities, Functional Magnetic Resonance Imaging (fMRI), which estimates the Bold Oxygen Level Dependent (BOLD) effect to represent neural activity, showed a high potential in identifying specific biomarkers related to PD and its progression (Hou et al., 2022). While disease phenotypes are heterogeneous, neuronal dysfunction patterns were shown to be highly replicable between patients (Warren et al., 2013).

resting-state functional Magnetic Resonance Imaging (rs-fMRI) features are particularly promising. Region-wise measurements such as Regional Homogeneity (ReHo) and Amplitude at Low Frequency Fluctuation (ALFF) were used in multiple studies to predict PD trajectory or motor subtypes (Hou et al., 2016; Hu et al., 2015; Pang et al., 2021; Nguyen et al., 2021; Wu et al., 2009; Yue et al., 2020). ReHo quantifies the connectivity between a voxel and its nearest neighboring voxels and was shown to be affected by neurodegenerative diseases (Zang et al., 2004). ALFF and its normalized form, Fractional Amplitude at Low Frequency Fluctuation (fALFF), measure the power of the low frequency signals at rest, which mostly consists in spontaneous neuronal activity (Zou et al., 2008).

However, despite their potential, neuroimaging measures are sensitive to multiple sources of variability that impact their replicability and may explain why the derived biomarkers are not well established in clinical and research practice. In particular, neuroimaging analyses require specific methodological choices at various computational steps, related to the software tools, the method, and the parameters to use. These choices, also known as "researchers degrees of freedom" (Simmons et al., 2011), might have a large impact on the results of an experiment as they can impact the predictiveness of the signal extracted and can lead to a lack of agreement when analyzing the same neuroimaging dataset with different analysis pipelines (Bowring et al., 2019; Botvinik-Nezer et al., 2020). For instance, in task-based fMRI, 70 research teams were asked to analyze the same fMRI dataset using their usual analysis pipeline and results were substantially variable across

teams (Botvinik-Nezer et al., 2020).

Furthermore, neuroimaging results have been shown to be impacted by differences in hardware architectures or software package versions (Glatard et al., 2015; Gronenschild et al., 2012), questioning the robustness of the results. This suggests that a single pipeline evaluation is not sufficient to obtain robust results, though the reliability of results may be increased when studying their distributions across perturbations.

There are also concerns about the reproducibility of machine learning studies. Indeed, in a recent study, Kapoor et al., 2023 attempted to reproduce several machine learning experiments, revealing multiple issues which could lead to the non-reproducibility of findings. These issues can be split in three categories (Varoquaux et al., 2023): data leakage, computational reproducibility, and choice of evaluation metrics. In particular, Wen et al., 2020 performed a review of CNN-based classification of Alzheimer's subtypes and found a potential data leakage in half of the 32 surveyed studies due to a wrong data split at the subject-level, a data split after data augmentation or dimension reduction, transfer learning with models pre-trained on parts of the test set or the absence of an independent test set. Such a data leakage, which we did not notice in our study, might cause an over-optimistic performance assessment of models and thus, a lack of reproducibility and replicability of the findings. Evaluation procedures can also cause the non-reproducibility of findings, due to unsuitable metric choices when using unbalanced datasets for instance or questionable cross-validation procedures, in particular with low sample sizes. Random choices in a training procedure, for instance initial weights or hyper-parameters random selection, which all impact computational reproducibility, might also lead to uncontrolled fluctuations in results when using different random initialization states.

Conflicting terminologies exist for the terms reproducibility and replicability (Barba, 2018). Here, we define reproducibility as attempts made with the same methods and materials. Replicability, on the other hand, is tested with different but comparable materials or methods, assuming that the tested pipelines are all suitable to extract signal from the data. Note that comparable is ambiguous, but defined further in this case in Method Section.

Replicability experiments have shown different degrees of variability between findings obtained with different analytic conditions. These studies are usually done using healthy populations and in general research practice (as opposed to clinical research), as in Botvinik-Nezer et al., 2020. For clinically-oriented research, however, the topic remains understudied. Such studies requires a specific attention as they are useful to

develop new biomarkers that can influence treatment development and clinical trial applications. These studies also often target specific populations of patients with unique characteristics, in particular for PD for which inter-individual variability is high (Wüllner et al., 2023). Such studies often use small sample sizes, which has been shown to lead to a lower reproducibility of findings (Klau et al., 2020; Poldrack et al., 2017). Reproducibility and replicability of studies in clinical settings is of higher importance to improve the trustworthiness of new biomarkers and to facilitate their development.

In this paper, we evaluate the reproducibility and replicability of the study in Nguyen et al., 2021, a clinically-oriented research on a PD population. The study in Nguyen et al., 2021 is of particular interest as it uses the Parkinson's Progression Markers Initiative (PPMI) dataset (Marek et al., 2018), a large open access dataset to study Parkinson's disease. Moreover, it investigates the clinically relevant problem of trying to predict an individual's current and future disease severity over up to 4 years and it uses two different rs-fMRI-derived biomarkers: ReHo and fALFF. In Nguyen et al., 2021, the authors, including current co-authors KPN and AAM, trained several machine learning models using regional measurements of ReHo or fALFF along with clinical and demographic features to predict Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) total score at acquisition time and up to 4 years after. They selected n=82 PD patients by searching for all patients available at that time with rs-fMRI and MDS-UPDRS score at the same visit from the PPMI database and preprocessed the functional images to extract whole-brain maps of fALFF and ReHo. They compared three atlases, splitting the brains in different numbers of regions to extract mean region-wise features which are fed to the machine learning models. They achieved better than chance performance for prediction at each time point with both fALFF and ReHo, *e.g.* r-squared of 0.304 and 0.242 for prediction of current severity with ReHo and fALFF respectively. Finally, the authors discussed the most important brain regions for prediction. Although most studies do not perform external validation, authors of Nguyen et al., 2021 confirmed the predictiveness of their models on an external dataset, the next largest dataset available at the time: the Parkinson's Disease Biomarkers Program (PDBP) from NIH. On this dataset, they found reproducible model performance.

Different criteria could be used to conclude on success of the reproduction and replication of this study: 1) if the models trained on fALFF and ReHo at each time points showed better than chance performance in terms of R-Squared (R2) (R-squared - coefficient of determination (R2) >0 and R2 >chance-model R2) when tested on the PPMI

dataset using the evaluation procedure proposed in Nguyen et al., 2021 and 2) if these models showed similar performance (R2 greater than 0 and absolute difference between original and reproduction R2 less than 0.2) to those proposed in the original study. Our main interests were to assess the difficulties and challenges of reproducing fMRI research experiments, but also to further evaluate the impact of different analytical choices (*e.g.* processing pipeline, choice of feature set, etc.) on the results of these experiments. In this paper, we explore how these choices affect different parts of the analysis:

— Cohort selection and sample size,

— fMRI pre-processing pipeline,

— fMRI feature quantification,

— Choice of input features for machine learning models,

— Machine learning models choice and results reporting.

A primary purpose of this investigation is also to learn about the difficulties encountered to reproduce neuroimaging studies, in particular in clinical research settings, and to provide some recommendations on best practices to facilitate the reproducibility of such studies in the future.

B.2 Materials and Methods

Our study consisted of two steps: a first replication attempt without contacting the authors, using only publicly-shared resources available with the original paper, and a second replication attempt after contacting the authors, to obtain more accurate information on the original study. This two-step reproduction was meant to assess the challenges of reproducing a study using only publicly available materials and to evaluate the contribution of data and code sharing platforms to results reproducibility.

B.2.1 Dataset

As in the original study, we used data available from the PPMI dataset (Marek et al., 2018), a robust open-access database providing a large variety of clinical, imaging data and biologic samples to identify biomarkers of PD progression. The PPMI study was conducted in accordance with the Declaration of Helsinki and the Good Clinical Practice (GCP) guidelines after approval of the local ethics committees of the participating

sites. We signed the Data User Agreement and submitted an online application to access the data. More information about study design, participant recruitment and assessment methods can be found in Marek et al., 2018. We note that access to such data does not permit us to share such data on our own. Moreover, unlike code repositories with version control numbering, most data repositories are not version controlled, making re-retrieval of data years later thorny.

B.2.2 Summary of experiments

Reproducing an analysis can be challenging due to (1) the lack of specific information on analysis pipelines, software versions, or specific parameter values, (2) the presence of confusing terms in the available information, (3) the evolution of the software and data materials used in the original study. Our reproduction study consisted of 5 global steps: cohort selection, image pre-processing, imaging features computation, choice of input features and model choice and reporting. We used the information available in the original paper and for some parts of the analysis, we also had access to the code shared by the authors on GitHub (*e.g.* for feature computation and machine learning models). Though the authors also made their contact information plainly available, in our first experiment we wished to work independently of any author contact. Under this scenario, we had to make informed guesses due to the 3 types of challenges stated above, which resulted in a high number of possible workflows. To evaluate the effect of each variation at each step, we defined a *default workflow* to which each variation was compared to. At each step, if a variation of the workflow was tested, the other steps were implemented as in the default one. This default workflow was the most likely according to the code shared along with the paper. Figure B.1 summarizes the different variations tested and the *default workflow*.

B.2.3 Cohort selection

The cohort reported in Nguyen et al., 2021 was composed of the largest set of PPMI available at the time, and consisted in 82 PD participants with rs-fMRI and MDS-UPDRS scores obtained during the same visit. MDS-UPDRS Part III (motor examination) was conducted when patients were under the effect of PD medication. Of these 82 participants, 53 participants also had MDS-UPDRS scores available at Year 1 after imaging, 45 at Year 2, and 33 at Year 4.

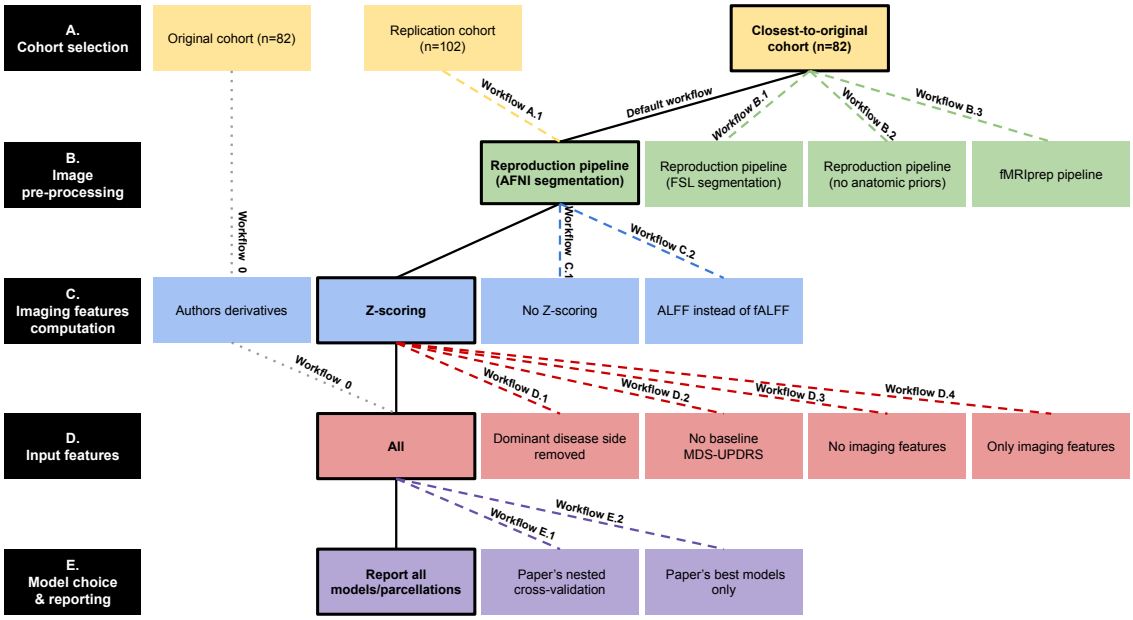

Figure B.1 – Summary of the different workflows implemented to reproduce the results of Nguyen et al., 2021 and explore their robustness to different analytic conditions. Bold and bordered cells represent the implementation of the default workflow at each step, this whole workflow is labeled *Default workflow* and is represented using a plain bold line. The different variation workflows are represented in dashed lines: all steps different from the variation follow the default workflow and each workflow corresponds to one variation from the default one.

- *Workflow 0* - reproduction using authors derivatives.

Variations of cohort selection (A):
- *Workflow A.1* - default workflow with replication cohort.

Variations of pre-processing pipeline (B):
- *Workflow B.1* - default workflow with FSL segmentation,
- *Workflow B.2* - default workflow without structural priors,
- *Workflow B.3* - fMRIprep pipeline.

Variations of feature computation (C):
- *Workflow C.1* - default workflow with no Z-scoring,
- *Workflow C.2* - default workflow with ALFF.

Variations of input features (D):
- *Workflow D.1* - default workflow with no dominant disease side,
- *Workflow D.2* - default workflow with no Baseline MDS-UPDRS,
- *Workflow D.3* - default workflow with no imaging features,
- *Workflow D.4* - default workflow with only imaging features.

Variations in model choice and reporting (E):
- *Workflow E.1* - default workflow with paper's nested cross-validation,
- *Workflow E.2* - default workflow with only paper's best model reporting.

B.2.3.1 Replication cohort

We first attempted to reproduce the cohort of Nguyen et al., 2021 using only the information available in the code shared on GitHub and the paper. Based on this information, we filtered the PPMI database using 4 criteria:

— Participants belong to the "Parkinson's disease" cohort, as defined in PPMI.

— Participants have an fMRI acquisition and a MDS-UPDRS score, with MDS-UPDRS Part III conducted ON-medication ("PAG_NAME" different from "NUPDRS3" in the PPMI score file) computed at the same visit (same visit code in PPMI database). Thus, only participants with valid values for MDS-UPDRS Part III score were included in the cohort.

— Participants and visits were also filtered depending on the type of fMRI acquisition. We queried the database with the exact same information as in the S1 Table of the original paper (field strength = 3T, scanner manufacturer = Siemens, pulse sequence = 2D EPI, TR = 2400ms, TE = 25ms).

— We also filtered the database to keep only participants for which the visit date and archive date of the image was set before January 1st, 2020 (more than a year before the original study publication) since without contacting the authors we had somewhat imprecise information about the date the authors accessed the database.

This query involved both fMRI metadata obtained using a utility functions from the Python packages livingpark-utils v0.9.3 and ppmi_downloader v0.7.4 and the MDS-UPDRS-III file from the PPMI database.

Since the PPMI database does not permit querying the database at any prior time point, we queried the database at the then current time. Specifically, we queried the PPMI database on August 21st, 2023 and we included the participants selected using these filters in the Baseline time point of our replication cohort. To find the participants who also had a score available at Year 1, Year 2, or Year 4 follow-up, we looked for the visit date associated with the MDS-UPDRS score at Baseline and searched for participants that also had a score at 365 days (1 year) +/- 60 days (2 months), 2 × 365 days (2 years) +/- 60 days (2 months) and 4 × 365 days (4 years) +/- 60 days (2 months). This method was also used by the original authors to search for their cohort at Year 1, Year 2, and Year 4 follow-up.

B.2.3.2 Closest-to-original cohort

After contacting the authors (KPN and AAM), the exact participant and visit list used at Baseline was provided to us. We queried the PPMI database using this list and compared with our replication cohort.

The 82 participants of the original Baseline cohort were all included in our replication cohort. For 4 of them, the visit used in our replication cohort was different from the one used in the original cohort. For two participants, we used an earlier visit than the authors: V06 (2 years) instead of V10 (4 years) and BL (baseline) instead of V04 (1 year). For the last two participants that had different visits selected in the replication cohort, images of the visits used by the original authors were not available in the PPMI database when we queried it. We assumed that this issue resulted from the update of the PPMI database in September 2021, and that there is no way to query prior versions of the database, and that the original authors are not allowed to share the original images they obtained when they accessed the database.

The 82 participants of the original cohort that were also included in our replication cohort were used to build a “closest-to-original" cohort to compare with our original cohort. The authors also provided the participant identifiers included at Year 1, Year 2 and Year 4, but we did not have the exact visit used at these time points. Thus, for each time point, we searched for the participants involved in our replication cohort for this time point that were in the list provided by the authors. Several participants from the list provided by the authors were not found in our cohorts. When checking the MDS-UPDRS-III files for these missing participants, we found the potential visit used by the authors, but these did not meet the criteria set to select the valid MDS-UPDRS-III scores (*i.e.* “PAG_NAME" was equal to “NUPDRS3" for these visits, but these were discarded when selecting only ON medication scores). For one participant missing in the Year 2 time point, we have not found any visit 2 years +/- 2 months after the Baseline visit. The visit selected for this participant was different in our cohort compared to the original authors cohort due to missing images, which could explain the reason for not finding back this participant for the Year 2 time point. Table B.1 summarizes the cohort selection process.

B.2.4 Image pre-processing

We downloaded functional images from the PPMI database manually for all participants selected in the replication cohort by using the image identifiers corresponding to the participants and visits selected. We also downloaded T1w images corresponding to the participants and visits selected in the replication cohort. If multiple T1w images were available for a participant at a given visit, we selected the one with the smallest identifier number (1st one in the meta-data table). Imaging data from the PPMI online database were available in DICOM format. We converted them into the NIfTI format and we reorganized the dataset to follow the Brain Imaging Data Structure (BIDS) (Gorgolewski et al., 2016) (RRID:SCR_016124) using HeuDiConv v0.13.1 (Halchenko et al., 2024) (RRID:SCR_017427) on Docker v20.10.16.

B.2.4.1 Default reproduction pipeline

To pre-process the data, we built a pipeline reproducing the one described by the authors in Nguyen et al., 2021 without contacting them for any additional information or code (which has since been provided). The paper mentions that fMRI images were first realigned to the mean volume with affine transformations to correct for inter-volume head motion, using the MCFLIRT tool in the FSL toolbox (Jenkinson et al., 2012) (RRID:SCR_002823). Then, images were brain-masked using AFNI 3dAutomask (Cox, 1996) (RRID:SCR_005927). Non-linear registration was performed directly to a common EPI template in MNI space using the Symmetric Normalization algorithm in ANTS (Avants et al., 2011) (RRID:SCR_004757). For denoising, motion-related regressors computed using ICA-AROMA (Pruim et al., 2015) were concatenated with the nuisance regressors from affine head motion parameters computed with MCFLIRT and mean timeseries of white matter and cerebrospinal fluid. These nuisance signals were regressed out of the fMRI data in one step (*i.e.* all confounds concatenated in a single matrix and regressed from voxels timeseries).

Using this information, we reproduced the closest-possible pipeline to this description. We implemented this pipeline — referred to as the *default workflow* — using Nipype v1.8.6 (RRID:SCR_002502) (Gorgolewski, 2017), FSL v6.0.6.1, AFNI v23.3.01 and ANTs v2.3.4. We executed the pipeline with a custom-built Docker image available on Dockerhub `https://hub.docker.com/repository/docker/elodiegermani/nguyen-etal-2021/general` and built using NeuroDocker (Kaczmarzyk et al., 2018) with base image fedora:36 and

a miniconda v23.5.2-0 (*Anaconda Software Distribution* 2020) environment with Python v3.10. All pre-processing, feature computation and model training were run using home-made Boutiques descriptors using Docker v20.10.16 and Boutiques v0.5.25 (Glatard et al., 2017). Boutiques descriptors for image processing and model training are available in Zenodo (Germani, 2023a; Germani, 2023b).

In this *default reproduction workflow*, functional images were first realigned to the middle volume using FSL MCFLIRT, using affine registration (6 degrees of freedom), b-spline interpolation and mutual information cost function. The motion-corrected images were then skull-stripped using AFNI 3dAutomask with default parameters (clip level fraction of 0.5). Following this, ANTs symmetric normalization algorithm was used to normalize images to the MNI template. First, rigid, affine, and symmetric normalization transformations from native to MNI space were computed using the first volume of the brain-extracted functional images as source image and the MNI152NLin6Asym template, with a 2mm resolution as reference. The exact MNI template used for registration was not mentioned in the original paper. The choice of this particular template for our reproduction was due to the use of ICA-AROMA after registration. Indeed, to run ICA-AROMA in the MNI space or without FSL registration transform matrices, images must be in FSL's default MNI space, which is the MNI152NLin6Asym (*ICA-AROMA & fmriprep using child template - fmriprep - Neurostars* 2019). We downloaded this EPI template from C-PAC: `https://github.com/FCP-INDI/C-PAC/blob/main/CPAC/resources/templates`. We applied the computed transformations to functional images using ANTs also with B-Spline non linear registration.

For denoising, we regressed out several nuisance signals from the fMRI data, as in the original study. The 6 affine motion parameters computed using MCFLIRT were used as regressors. In addition, we ran ICA-AROMA v0.4.3-beta on data already registered in MNI space to extract motion-related components. All the components classified as motion-related were added as regressors to each participants.

For White Matter (WM) and CerebroSpinal Fluid (CSF) signals, there was no information about the method used by the authors to compute these signals in the original paper. Thus, we implemented three different methods to reproduce the original workflow but also to compare the impact of pre-processing pipelines on the results of the study. In the *default workflow*, we chose to use AFNI to compute these regressors. We used the structural T1w images downloaded from PPMI and ran several analysis steps: brain extraction using 3dSkullstrip, segmentation using 3dSeg with defaults parameters, 3dCalc

to extract the mask for WM and CSF, 3dResample to resample the masks to the functional image using nearest-neighbors interpolation and 3dMaskave to extract timeseries of voxels inside the WM and CSF masks. Then, we computed the mean timeseries across these voxels for WM and CSF and added these signals as nuisance regressors.

B.2.4.2 Variations of the reproduction pipeline

We also compared this workflow with two other methods to extract WM and CSF signals. The first method (pipeline *B.1 - default workflow with FSL segmentation*) used tools from FSL instead of AFNI to extract structural-derived masks. In this pipeline, BET was used to remove non-brain tissues from structural images, then the images were segmented using FAST to extract WM and CSF masks. The masks were resampled to functional images using affine registration implemented in FLIRT, and mean timeseries inside each mask were extracted using FSL's ImageMeants function in Nipype.

The second method (pipeline *B.2 - default workflow without structural priors*) did not involve image segmentation. We used mask templates available in FSL and Nilearn: MNI152_T1_2mm_VentricleMask from FSL for CSF, and WM brain-mask in MNI152 template resolution 2mm in Nilearn v0.10.2 (Abraham et al., 2014b) (RRID:SCR_001362) for WM. The masks were resampled to the functional images using a nearest neighbors interpolation in Nilearn, and mean timeseries inside each mask were also computed using Nilearn.

In all reproduction pipelines, the nuisance signals were regressed from the functional images in MNI space using FSL RegFilt. The denoised images were then used to compute the imaging features passed as input to the machine learning models.

B.2.4.3 Other pipelines variations

To explore the robustness of the original results to variations in the workflow, we also analyzed the functional and structural images using fMRIprep v23.0.2 (Esteban et al., 2019) (RRID:SCR_016216), a robust pre-processing pipeline that requires minimal user input. We used default parameters for fMRIprep, except for the reference template that we set to MNI152NLin6Asym with a resolution of 2mm to be able to run ICA-AROMA afterwards (*ICA-AROMA & fmriprep using child template - fmriprep - Neurostars* 2019).

Final preprocessed functional images in MNI space were then passed as input to ICA-AROMA to obtain motion-related components. The 6 motion regressors, WM and CSF

mean timeseries extracted by fMRIprep were concatenated to the timeseries of the motion-related components identified by ICA-AROMA and regressed out from the pre-processed images using FSL RegFilt, as in the reproduction pipeline. This pipeline is referred to as *B.3 - fmriprep pipeline*.

B.2.4.4 Quality control

We implemented quality control checks at different steps of the pipelines. The purpose of these controls was to explore quality of data, but we did not exclude any participant due to data low quality, as this step was not performed in the original paper.

For each participant, we controlled the quality of functional pre-processing (motion correction, brain masking, and registration to MNI space) by superposing the pre-processed functional volume at each time point to an MNI-space brain mask, and visually inspecting a pre-defined image slice for incorrect registration or masking. We also visually inspected the 6 motion parameters identified during motion correction (rotation and translation in the x, y and z directions). We also computed the frame-wise displacement (FD) of head position as done in Power et al., 2014, calculated as the sum of the absolute volume-to-volume values of the 6 translational and rotational motion parameters converted to displacements on a 50 mm sphere (multiplied by $2 \times \pi \times 50$). We explored these values using the threshold used in Parkes et al., 2018 for the lenient strategy: identification of participants with mean FD > 0.55mm. Segmentations masks for WM and CSF obtained with the 2 different workflow variations were also visually inspected for failed segmentations. For the fMRIprep pipeline, we validated the quality of the processing using the log files produced by the pipeline, since these produce the same outputs as the quality control steps mentioned above.

B.2.5 Imaging features computation

In the original study, mean regional values of z-scored fALFF and ReHo maps were used as input features to the machine learning models, in addition to several clinical and demographic features. fALFF and ReHo were computed on the denoised fMRI data using C-PAC (Cameron et al., 2013) (RRID:SCR_000862). Voxel-wise ReHo was computed using Kendall's coefficient of concordance between each voxel and its 27-voxel neighborhood. For ALFF and fALFF, linear de-trending and band-pass filtering were first applied to each voxel at 0.01–0.1 Hz, then the standard deviation of the signal was computed to

obtain ALFF whole-brain maps. These maps were divided by the standard deviation of the unfiltered signal to obtain whole-brain fALFF maps. Z-scores maps for ReHo and fALFF were calculated at the participant-level.

For our reproduction, we used the original code used by the authors (see authors code). We followed the exact same steps as in the original paper to compute the raw ReHo and fALFF maps. However, a mask file was needed in the authors' code to compute the features. We thus applied AFNI 3dAutomask on the denoised fMRI data to obtain a brain mask for each participant.

The initial code shared by the authors did not include any z-scoring of the whole-brain maps for fALFF and ReHo, thus we used FSL's ImageMaths function to compute the z-score maps. Non z-scored maps (*C.1 - default workflow with no Z-scoring*) were also saved and set as input to the models for comparison. We also considered ALFF instead of fALFF as input measure (*C.2 - default workflow with ALFF*) as the authors also mentioned having tested this feature. We note that for the second step of the reproduction experiment, the authors of Nguyen et al., 2021 have supplied us with all derived maps.

In the original paper, regional features were extracted from the ReHo and fALFF whole-brain maps using three different parcellations. These included the 100-ROI Schaefer (Schaefer et al., 2018) functional brain parcellation, modified with an additional 35 striatal and cerebellar ROI, and the 197-ROI and 444-ROI versions of the Bootstrap Analysis of Stable Clusters (BASC) atlas (Bellec et al., 2010). These parcellations were used to compute the mean regional ReHo or fALFF values for each participant and performance of the machine learning models were compared between the parcellations. For the first step of the reproduction, we did not have access to the modified version of the Schaefer atlas used by the original authors. Thus, we derived a similar custom atlas by using the 100-ROI Schaefer atlas available in Nilearn, the probabilistic cerebellar atlas available in FSL, from Diedrichsen et al., 2009, and the Oxford-GSK-Imanova connectivity striatal atlas from Tziortzi et al., 2014, also available in FSL. The cerebellar and striatal atlases were respectively composed of 28 and 7 ROI, which was consistent with the 35 ROI mentioned in the original paper. We merged the ROI from the Schaefer, cerebellar and striatal atlas in this order to build a custom 135-ROI atlas which we used to extract regional features.

The three atlases were resampled to the whole-brain ReHo and fALFF maps using Nilearn and a nearest-neighbor interpolation, as done by the authors. Mean regional values for each imaging feature and parcellation were also extracted using Nilearn.

We obtained from the authors the custom atlas used in the original analyses. We found some slight differences between the cerebellar and striatal regions in the two atlases, *e.g.* in terms of size of the regions or division in subregions. We compared the mean regional values for the corresponding regions in the two atlases using paired two-sample t-tests. Among the 82 participants at baseline, 19 had significantly different values at $p < 0.05$ for fALFF and none at $p < 0.01$. Considering these small differences, we decided to report the results only using our reproduction atlas.

B.2.6 Input features

B.2.6.1 Clinical and demographic features

In addition to imaging features, to better mirror clinical practices, the authors endeavored to integrated several clinical and demographic features as additional inputs to the machine-learning models. Clinical features included disease duration, symptom duration, dominant symptom side, Geriatric Depression Scale (GDS), Montreal Cognitive Assessment (MoCA), and presence of tremor, rigidity, or postural instability at Baseline. Baseline MDS-UPDRS score was also included as a feature when training models to predict outcomes at Year 1, Year 2, and Year 4. Demographic features included age, sex, ethnicity, race, handedness, and years of education.

We searched for the mentioned input features using the study files in the PPMI database, as done by the authors (see authors code). For each feature, we searched for the corresponding columns in the study files and used the same character encoding method as the authors.

To evaluate the robustness of the findings to different analytical conditions, we also compared the results obtained with different sets of features. In pipeline *D.4 - default workflow with only imaging features*, we trained models using only imaging features (regional measures of fALFF and ReHo), *i.e.*, without clinical or demographic features. In pipeline *D.3 - default workflow with no imaging features*, we removed imaging features and trained models only on clinical and demographic features. Following an update of the PPMI database, the feature for dominant disease side was deprecated and only available as an archive file in the version of the database we had access to. We included the feature in the *default workflow* and removed it in another variation workflow, to assess the impact of this feature (*D.1 - default workflow with no dominant disease side*). We did not contact the authors for the values of these features that they had downloaded, through they did

factor prominently into their results, in order to understand better the relevance of the database update.

For models trained to predict MDS-UPDRS scores at Year 1, Year 2, and Year 4, Baseline MDS-UPDRS score was included as feature. However, due to the potential large effect of including this variable on the results, we trained a model with all features except this one and compared the performance of prediction models with and without the feature (*D.2 - default workflow with no Baseline MDS-UPDRS*).

B.2.6.2 Outcome measurement

In Nguyen et al., 2021, the authors used the above-mentioned imaging, clinical, and demographic features to predict MDS-UPDRS total scores. The MDS-UPDRS score consists of 4 parts with 51 items, each item values from 0 to 5. To compute the total scores, we summed the values of the 4 different parts available in PPMI study files. We used: MDS-UPDRS part Ia entered by a rater (PPMI column "NP1RTOT"), part Ib for the patient questionnaire (column "NP1PTOT"), part II ("NP2TOT"), part III ("NP3TOT") and part IV ("NP4TOT"). Missing values in "NP4TOT" columns were replaced with zeros, as done by the authors. There were no participants with missing values for the other parts of the score.

B.2.7 Model selection and performance evaluation

We trained and optimized separate machine learning models to predict MDS-UPDRS scores from either ReHo or fALFF features, along with clinical and demographic features. Four machine learning models architectures were implemented using the latest version of scikit-learn at the time of this experiment, v1.3.0 (Abraham et al., 2014b), and were tested for each target-imaging feature (fALFF or ReHo) combination: ElasticNet regression, Support Vector Machine (SVM) with a linear kernel, Random Forest with a decision tree kernel, and Gradient Boosting with a decision tree kernel. We recognize that this version of scikit-learn is likely newer than that used by the authors in 2022 and that we could download a prior version of scikit-learn, but did not because we wish to evaluate the relevancy of ML source code update. Each parcellation was also implemented, which resulted in 12 different combinations of model and parcellation per imaging feature and time point. All models were trained using our newer version of scikit-learn, we used the set of hyperparameters available in the authors code to train and optimize the models.

For hyperparameter optimization (1) and performance estimation (2), the authors used a nested cross-validation scheme, *i.e.*, each model architecture × hyperparameter × parcellation combination was evaluated using (1) a 10-fold cross-validation inner-loop applied to the n-1 participants in the cohort and from which the combination with the lowest Root Mean Squared Error (RMSE) was selected, (2) a leave-one-out (LOO) cross-validation outer-loop where each iteration trained the selected model on all the participants in the cohort except one, and tested the model on the remaining held-out participant. To evaluate the impact of the evaluation pipeline on the results, we implemented a different nested cross-validation loop for model selection and evaluation for the *default workflow*. Fig B.2 illustrates the different methods implemented. We evaluated the performance of each combination of model × parcellation separately: the 10-fold cross-validation inner-loop was used to select the set of hyperparameters (*e.g.* maximum tree depth for Random Forests) with the lowest RMSE, this set was used to train a model on all except one participants in the outer-loop and we tested the model on the held-out participant. Thus, we obtained performance estimates for each model × parcellation combination.

We also reported results obtained using the exact nested cross-validation scheme explained in the paper (*E.1 - Workflow with paper's nested cross-validation*), *i.e.*, the performance on each outer-fold is assessed with the best model × hyperparameter × parcellation combination found on the 10-fold cross-validation of the inner-loop and averaged across outer-folds. Finally, as authors reported only the best performing model and parcellation for each imaging feature type and time point, we also reported the results we would have obtained had we only used the best model and parcellation reported in the paper (*E.2 - Workflow with only paper's best model reporting*).

B.2.7.1 Evaluation metrics

As in the original paper, performance metrics included the R2, which represents the percentage of variance explained by the model, and the RMSE, as implemented in scikit-learn.

We defined a null performance to compare our R2 values to using permutation test. We fixed the model and parcellation scheme with ElasticNet and Schaefer atlas. We ran 1000 permutations on the target labels and obtained performance for each feature and timepoint. At each permutation, we performed a nested cross-validation with 5-folds cross-validation as inner-loop and outer-loop. We optimized the hyper-parameter set of the model as done with the “real" models in the inner-loop and evaluated performance on

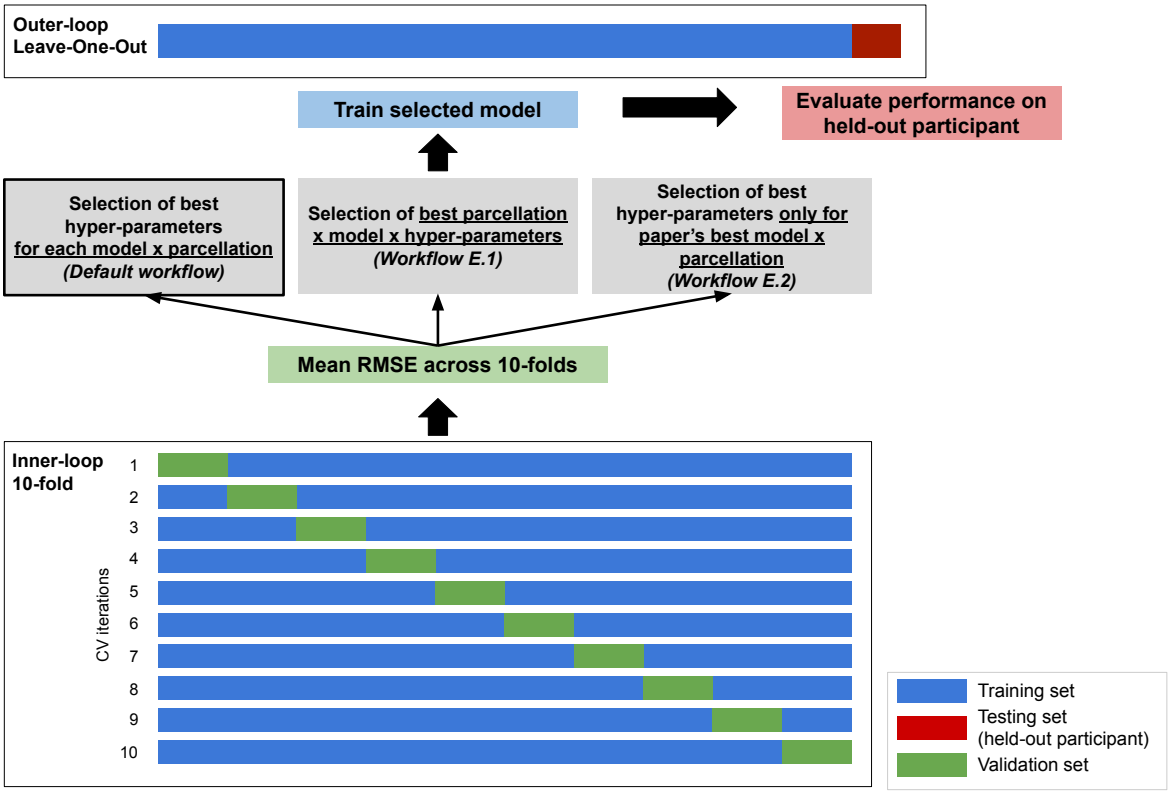

Figure B.2 – Workflow of model selection and performance evaluation. This workflow represents one iteration of the outer-loop with Leave-One-Out cross-validation and is iterated over all the dataset to estimate mean performance.

the outer-loop. R2 values obtained using the different workflows were compared to this null performance to check if the models did not learn to predict only the average value.

To evaluate the models' ability to classify high versus low severity participants, a threshold was set to separate the participants and each model's predictions were thresholded post-hoc. This threshold was computed by using the average of the median MDS-UPDRS score at each of the four time points. In Nguyen et al., 2021, the threshold was 35. We computed this threshold the same way for the replication cohort and for the closest-to-original cohort. We obtained a value of 36 for the replication cohort and 35 for the closest-to-original one. Authors also mentioned having found no significant difference ($p > 0.05$) between the high and low-severity groups in motor predominance (Part III score as a percentage of total score) at each time point. With our thresholds, we ran two sample t-tests between high and low severity groups in the two cohort and did not find any significant difference with $\alpha = 0.05$ either in any cohort or time point. Performance

metrics for this secondary classification outcome included area under the receiver operating characteristic curve (AUC), positive predictive value (PPV), negative predictive value (NPV), specificity, and sensitivity.

B.2.7.2 Authors derivatives

Authors shared with us the derived data used in the original study (*i.e.* whole-brain fALFF and ReHo maps for the original cohort). We applied our input features selection (clinical and demographics) and machine learning model training and selection to these data and computed the results for the *Workflow 0*. This allowed us to verify the reproduction of these steps and to get more information on the potential factors of variations in the results (*e.g.*, suppressing differences in the imaging processing, while retaining some potential differences in the version of scikit-learn).

B.2.8 Feature importance

As in Nguyen et al., 2021, we measured feature importance in the models trained for each time point and imaging feature (fALFF or ReHo). For the ElasticNet and SVM models, we used the coefficients of the trained models to determine feature importance, since coefficients of higher magnitude indicate more important features in these two models. The sign of the coefficient was indicative of whether the feature was positively or negatively associated with the prediction target. For Random Forest and Gradient Boosting models, we used impurity-based feature importance coupled with univariate linear correlation to determine the direction of the association. Feature importance was computed on each iteration of the outer-loop and the median importance was reported for each feature.

To name the imaging features, we used the same method as the authors of Nguyen et al., 2021: the centroid of each feature's ROI was computed, if the feature was located in a ROI of the Automated Anatomical Labeling (AAL) atlas (Rolls et al., 2020), this label was allocated to the ROI. If not, we searched for the nearest ROI of the AAL atlas. Authors also sent us their ROI labels. However, since we decided to use the reproduced Schaefer atlas, we used the reproduced labels in the figures for consistency.

B.3 Results

B.3.1 Cohort selection

Using the method described above, we built two cohorts from the PPMI database: the replication cohort and the closest-to-original cohort.

Table B.2 shows the demographics and Baseline clinical characteristics of the replication and closest-to-original cohorts compared to the original cohort reported in Nguyen et al., 2021. The replication cohort was composed of respectively 102, 67, 61 and 46 participants for time points Baseline, Year 1, Year 2, and Year 4. The closest-to-original cohorts at the same time points were composed of respectively 82, 51, 41 and 30 participants.

Compared to the original cohort, our replication cohort showed similar demographics characteristics at each time point, except at Year 4 where our replication cohort showed a significantly higher age on average than in the original cohort ($p < 0.01$). Regarding clinical variables, mean MoCA score, GDS total score and Hoehn-Yahr stage were similar between the two cohorts at all time points. However, we found higher mean disease durations in the replication cohort than in the original one at all time points, for instance at Baseline with (866.9 days $\pm$ 598.7 days) in replication vs (770 days $\pm$ 565 days) in original. This difference was not significant at threshold $p < 0.05$. We also observed lower mean MDS-UPDRS scores at Baseline in the replication cohort for all time points except Baseline, with significant difference at Year 2 ($p < 0.05$) only. For these two time points, even if mean Baseline scores in the replication cohort significantly differed from the original ones, mean MDS-UPDRS scores at prediction time point were more similar to the original one. At Year 4, however, we also found a higher mean MDS-UPDRS score at prediction time point than in the original cohort, but this difference was not significant at $p < 0.05$.

The closest-to-original cohort exhibited almost the same characteristics as the original one at Baseline. For subsequent time points, we found some differences, in particular at Year 2 and at Year 4: participants were older in the closest-to-original cohort than in the original study at Year 4 ($p < 0.05$), Baseline mean MDS-UPDRS score was lower (significant for Year 2 and Year 4 at $p < 0.05$ and $p < 0.01$ respectively) and mean MDS-UPDRS score at prediction time point was similar to the original cohort except at Year 4.

These differences for the Year 1, Year 2, and Year 4 cohorts could be related to the evolution of the PPMI database in which sessions were added and removed since the

authors queried it for the original study. For these time points, we were not able to find all the participants that were included in the original cohort: the patients included in our closest-to-original cohorts represented respectively 96% (Year 1), 91% (Year 2) and 91% (Year 4) of the patients included in the original cohort. However, only represented 76% (Year 1), 67% (Year 2), and 65% (Year 4) of the replication cohort was composed of patients of the original cohort.

Criterions	N
PPMI global query - Baseline	**102**
Participants belonging to the list provided by the authors at Baseline	**82**
Participants not belonging to the corresponding session list	4
Original session after the one obtained with PPMI query	2
Image of original session not available anymore in PPMI	2
PPMI global query - Year 1	**67**
Participants belonging to the list provided by the authors at Year 1	**51**
Participants not belonging to original list	2
PAG_NAME was NUPDRS3	2
PPMI global query - Year 2	**61**
Participants belonging to the list provided by the authors at Year 2	**41**
Participants not belonging to original list	4
PAG_NAME was NUPDRS3	3
Absence of corresponding score at follow-up time point	1
PPMI global query - Year 4	**46**
Participants belonging to the list provided by the authors at Year 4	**30**
Participants not belonging to original list	3
PAG_NAME was NUPDRS3	3

Table B.1 – Summary of cohort selection procedure. PPMI global query corresponds to the replication cohort, highlighted in **blue**. Participants belonging to the list provided by the authors composed the closest-to-original cohort, highlighted in **green**.

	Baseline			Year 1			Year 2			Year 4		
	Orig.	Repro.	Closest	Orig.	Repro.	Closest	Orig.	Repro.	Closest	Orig.	Repro.	Closest
% Caucasian	95.1	95.1	93.9	94.4	94.0	94.1	97.8	95.1	95.1	97.0	97.8	96.7
% African-American	2.4	2.0	2.4	1.9	1.5	0.0	0	1.6	0.0	0	0.0	0.0
% Asian	3.7	2.9	3.7	5.6	4.5	5.9	4.4	3.3	4.9	3.0	2.2	3.3
% Hispanic	1.2	1.0	0.0	0	1.5	0.0	0	1.6	0.0	0	0.0	0.0
% Male	67.0	66.7	67.1	68.5	65.7	68.6	82.2	80.3	85.4	75.8	67.4	73.3
% right-handed	89.0	89.2	89.0	85.2	85.1	84.3	88.9	90.2	90.2	87.9	84.8	86.7
Mean age, years	62.1 ± 9.8	62.0 ± 9.5	62.1 ± 9.7	61.9 ± 10.3	62.2 ± 9.9	63.0 ± 10.4	63.6 ± 9.2	64.7 ± 9.1	65.9 ± 9.4	59.5 ± 11.0	**66.2 ± 10.1****	**63.8 ± 11.0***
Mean years of education	15.6 ± 3.0	15.6 ± 2.8	15.7 ± 2.9	15.1 ± 3.2	15.5 ± 2.9	15.4 ± 2.9	15.1 ± 3.3	15.4 ± 2.8	15.5 ± 3.0	15.0 ± 3.4	15.3 ± 3.0	15.2 ± 3.4
Mean disease duration at Baseline, days	770 ± 565	866.9 ± 598.7	760.3 ± 559.2	808 ± 576	904.1 ± 614.5	808.5 ± 580.0	771 ± 506	867.5 ± 516.3	732.0 ± 462.8	532 ± 346	746.6 ± 624.6	464.6 ± 294.9
Mean MDS-UPDRS at Baseline	33.9 ± 15.8	34.5 ± 15.6	33.9 ± 16.1	38.0 ± 20.9	33.4 ± 15.1	34.1 ± 15.4	40.2 ± 18.2	**35.0 ± 15.1***	**35.2 ± 16.1***	34.9 ± 15.7	30.7 ± 13.9	**26.1 ± 11.4****
Mean MDS-UPDRS at timepoint	-	-	-	39.2 ± 21.6	40.7 ± 24.5	39.9 ± 22.0	40.9 ± 18.5	40.0 ± 18.7	40.7 ± 18.7	35.9 ± 16.5	41.5 ± 19.8	34.2 ± 16.2
Mean MoCA at Baseline	26.7 ± 2.8	26.5 ± 3.0	26.4 ± 2.8	26.9 ± 3.2	27.0 ± 2.9	26.7 ± 3.1	26.7 ± 3.5	27.0 ± 2.5	26.5 ± 2.4	27.5 ± 2.3	26.8 ± 3.2	27.4 ± 2.6
Mean GDS at Baseline	5.4 ± 1.4	5.4 ± 1.4	5.4 ± 1.5	5.4 ± 1.6	5.5 ± 1.8	5.5 ± 1.9	5.4 ± 1.2	5.5 ± 1.3	5.6 ± 1.3	5.4 ± 1.7	5.8 ± 1.8	5.6 ± 1.7
Mean Hoehn-Yahr stage	1.8 ± 0.5	1.7 ± 0.5	1.7 ± 0.5	1.8 ± 0.5	1.8 ± 0.6	1.7 ± 0.5	1.8 ± 0.5	1.9 ± 0.5	1.9 ± 0.5	1.7 ± 0.5	1.9 ± 0.5*	1.8 ± 0.5
Number of subject	82	102	82	53	67	51	45	61	41	33	46	30

Table B.2 – Demographic and clinical variables for the different cohorts. Orig. = original paper cohort. Repli. = replication cohort. Closest = closest-to-original cohort. Values are reported in percentages of the cohort or in mean values ± standard deviation. Significance testing was performed using two sample t-test between the original cohort and the replication and closest cohort respectively. Bold text represent features showing a significant difference, * represent significance at $p < 0.05$ and ** at $p < 0.01$.

Fig B.3 compares the distribution of MDS-UPDRS scores in our cohorts with the one in the original cohort reported in Fig S1 in Nguyen et al., 2021. Distributions of MDS-UPDRS scores at Baseline were similar between our two cohorts but seemed different from the original cohort one. The observed difference between the original and closest-to-original distributions might result from differences in MDS-UPDRS score calculations, or from the fact that different sessions were used for 4 of the participants in the closest-to-original cohort compared to the original one. At Year 1, however, the closest-to-original cohort presented a MDS-UPDRS score distribution more similar to the original one than the replication one, suggesting that the differences at Baseline did not originate in differences in MDS-UPDRS score calculations. We found no significant difference between the distribution of MDS-UPDRS scores in the replication and closest-to-original cohort neither at Baseline nor at Year 1 using Kolmogorov-Smirnov distribution testing.

B.3.2 Image quality control

After running the pre-processing pipelines, we checked the resulting images and looked for potential pipeline failures. Regarding registration, all participants brains were correctly registered to the MNI space after visual inspection. Brain masking was also successful for most of the participants, except for 2 in which we found a small artifact in the inter-hemispheric area. Given the low magnitude of this artefact and its location, we decided to keep these two participants in the study.

Most participants of the study showed high movement parameters. Indeed, out of 102, 80 showed at least one time point with a frame-wise displacement superior to 0.5mm. However, since the authors in Nguyen et al., 2021 did not remove high-motion volumes within participants, that removing volumes entirely can disrupt some derived values, and that completely removing participants with high-motion volumes would highly decrease our cohort's sample size, we chose to keep all participants and all volumes.

Regarding segmentation masks, after visual inspection no significant artifact was found for any participants using AFNI segmentation in default workflow. For some participants, small distortions were found in particular close to brain extremities (inter-hemispheric area or close to the skull in occipital and parietal regions). Using FSL segmentation however, we found brain masking issues that had impacts on segmentation quality. We used BET using default parameters to skullstrip images before segmentation and since we chose to explore the impact of different default implementations of pipelines, we did not exclude the segmentations for any participant nor segmentation workflow.

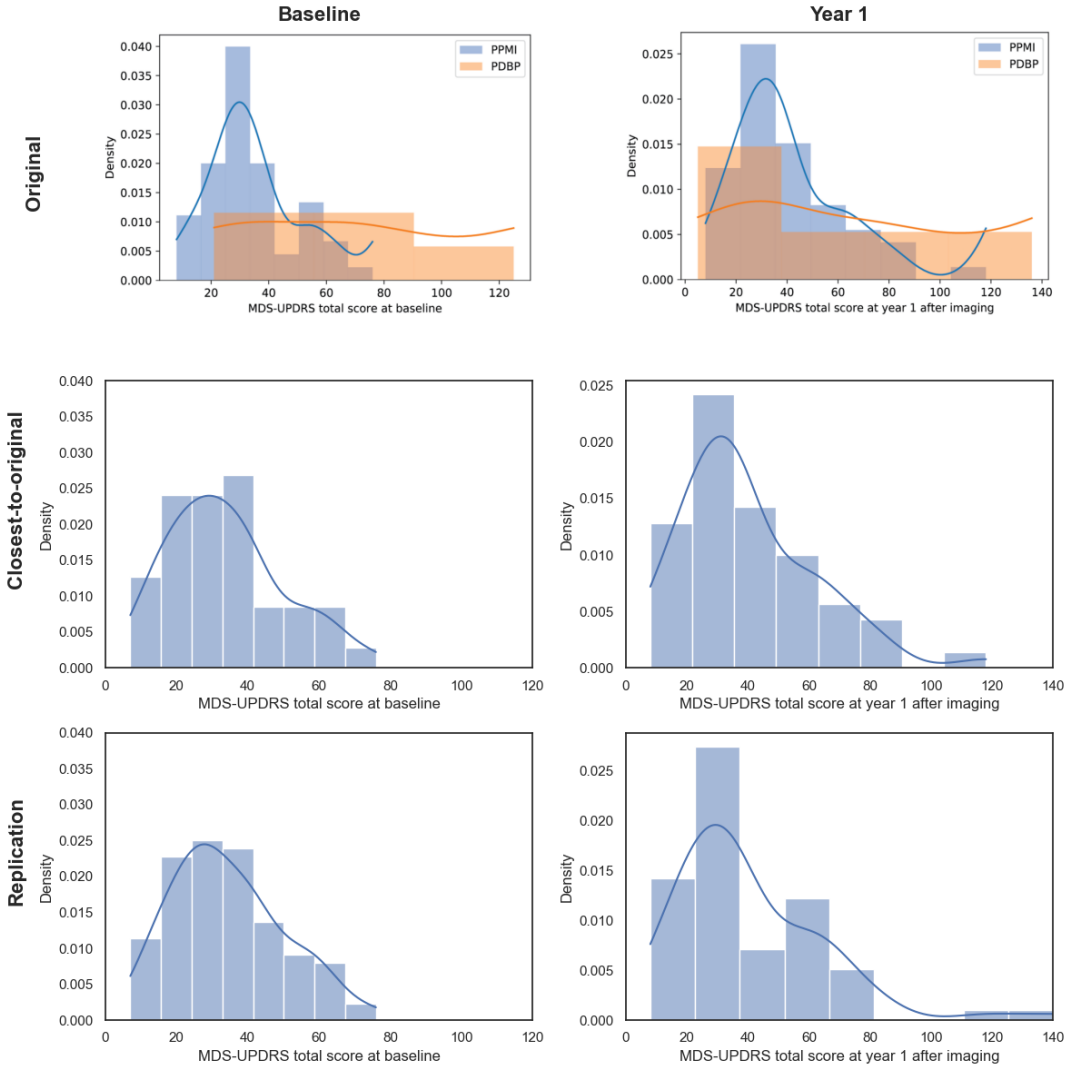

Figure B.3 – Distribution of MDS-UPDRS scores reported in the original paper's cohort (*top*: Fig S1 extracted from Nguyen et al., 2021), the replication cohort (*middle*) and the closest-to-original cohort (*bottom*).

With the fMRIprep pipeline, observations were similar regarding movement parameters and registration. There was no large artefact in the segmentation masks.

B.3.3 Performance of the *default workflow*

The first objective of this study was to reproduce the models described in Nguyen et al., 2021 and to compare their performance with the one in the original study. In our default workflow, we implemented the default choices described in Fig B.1: closest-

to-original cohort, image pre-processing pipeline with AFNI segmentation, z-scoring of whole-brain fALFF and ReHo maps, use of all demographic, clinical and imaging features described in the original paper, and the model selection method derived from the authors' code.

We trained 12 models per time point (Baseline, Year 1, Year 2, Year 4) and imaging feature (fALFF or ReHo), corresponding to 4 machine learning models × 3 brain parcellations. We reported for each imaging feature and time point the performance of the 12 models in Table B.3.

Chance levels were computed using permutation tests as described in the Evaluation metrics section. We obtained R2 values that represented the chance prediction performance at different time point for fALFF and ReHo. These values are also presented in Table B.3.

Using the default workflow, we obtained prediction scores different but relatively consistent with the results of Nguyen et al., 2021, for all models × parcellation combination. At Baseline, our best model performed better than chance and we obtained a R2 value close to the one reported in the original paper with the best model. However, the best-performing models were different from those reported in the original study: instead of Schaefer atlas and Gradient Boosting for both fALFF and ReHo features, we found for fALFF the Gradient Boosting Regressor with BASC197 atlas, with R2=0.205 (original R2=0.242) and ElasticNet and Schaefer for ReHo with R2=0.124 (original R2=0.304).

At Year 1, the performance of our models was better than reported in the original study, with an increase of the R2 of 28% and 18% for fALFF and ReHo respectively. For other time points (Year 2 and Year 4), results were slightly different from those reported in Nguyen et al., 2021 but overall consistent. These differences were not constant between ReHo and fALFF at Year 2, but were similar at Year 4: for fALFF, we obtained higher R2 scores than in the original study at Year 2 and at Year 4 (0.529 and 0.397 compared to 0.463 and 0.152 in the original paper); for ReHo, we obtained lower R2 scores than in the original ones at Year 2 (0.344 instead of 0.471) and higher R2 scores at Year 4 (0.312 compared to 0.255 in the original study). For these two time points, the mean MDS-UPDRS scores at Baseline were significantly different between the original cohort and our closest-to-original cohort, which might explain these differences in performance. In this context, the results observed remained similar in terms of effect size and reproduction remained satisfactory.

At each time point, the best model x parcellation combination performed better than

chance-level. Some of the combinations led to very low performance, for instance SVM with Schaefer atlas at Year 2. At every time point and with every feature (except at Year 1 with fALFF), at least one combination gave a performance lower than chance. This highlight the importance of model selection and performance reporting, which were also featured prominently in Nguyen et al., 2021. Some models may have not been optimally tuned, and all models do not have equal capability due to their different functioning, leading to lower performance. These low performance obtained with some models do not put into question the other results, as these have been validated on an external dataset by Nguyen et al., 2021.

Time	Feature	Type	ElasticNet			SVM			GradientBoosting			RandomForest		
			schaefer	basc197	basc444	schaefer	basc197	basc444	schaefer	basc197	basc444	schaefer	basc197	basc444
Baseline	fALFF	Orig.							**0.242**					
		Repli.	0.04	-0.035	-0.045	-0.718	-0.241	-0.182	-0.039	**0.205**	0.061	-0.024	0.068	0.02
		Null	**-0.041**											
	ReHo	Orig.							**0.304**					
		Repli.	**0.124**	0.057	0.117	-0.3	-0.4	-0.152	-0.102	0.028	0.027	0.024	0.022	0.099
		Null	**-0.036**											
Year 1	fALFF	Orig.	**0.558**											
		Repli.	0.453	**0.717**	0.5	0.519	0.216	0.185	0.622	0.575	0.506	0.369	0.499	0.444
		Null	**-0.079**											
	ReHo	Orig.	**0.453**											
		Repli.	**0.535**	0.434	0.512	0.04	-0.094	-0.01	0.36	0.261	0.289	0.442	0.392	0.393
		Null	**-0.077**											
Year 2	fALFF	Orig.	**0.463**											
		Repli.	**0.529**	0.277	0.285	-0.031	0.108	-0.413	-0.19	0.08	0.01	0.138	0.206	0.09
		Null	**-0.101**											
	ReHo	Orig.	**0.471**											
		Repli.	**0.344**	0.191	0.287	-0.915	-0.741	-0.051	-0.03	0.001	-0.033	0.267	0.121	0.251
		Null	**-0.094**											
Year 4	fALFF	Orig.					**0.152**							
		Repli.	**0.397**	0.115	0.351	0.196	-0.134	-0.296	0.08	0.411	-0.355	0.079	0.338	0.01
		Null	**-0.129**											
	ReHo	Orig.					**0.255**							
		Repli.	0.072	0.09	-0.175	-0.12	-0.23	-0.139	-0.017	**0.312**	0.041	-0.007	0.02	0.0
		Null	**-0.141**											

Table B.3 – Predictive performance achieved for each MDS-UPDRS time point and each imaging feature type, computed through leave-one-out cross-validation. Metric: R2, coefficient of determination. Green text corresponds to original performance reported in Nguyen et al., 2021; Blue text corresponds to best performance achieved during replication; Red text corresponds to chance level computed using permutation test.

B.3.4 Authors derivatives

In Fig B.4, we can see that using authors derivatives and thus, the original cohort, we achieve performance that are very close to the original ones, except at Year 4. This informs us on the quality of the reproduction of the clinical and demographic features selection, but also on the machine learning models training and selection. We can also suppose that the variations observed between the performance of the default reproduction workflow and the original results are related to imaging features (pre-processing or feature computation) or differences between cohorts.

B.3.5 Robustness to workflow variations

We assessed the performance of the different models for each time point and feature for different variations of the analysis workflow (Fig B.4).

Workflow A.2, in which we trained the different models on the replication cohort instead of the closest-to-original one, showed only small differences in R2 values with the *default workflow*, except for fALFF at Year 1 and ReHo at Year 4. Indeed, performance was slightly lower at Year 1 for fALFF and higher at Year 4 for ReHo, with raw effect size above 0.15. At Year 1, the replication cohort was composed of 16 more participants than the closest-to-original cohort and exhibited a lower mean MDS-UPDRS score at Baseline compared to the original cohort. At Year 4, we also found differences in term of sample size, age of participants and Baseline MDS-UPDRS score between the replication cohort, the original one and the closest-to-original one. These differences might explain the variations between models performance, even if R2 values remained better-than-chance for Year 1 and close to other performance obtained with different variations. Best model performance of *workflow A.2* remained better than chance-level.

Performance of models trained with variations in pre-processing pipeline (*workflows B.1, B.2 and B.3*) was similar to those of the default workflow, with R2 absolute difference with the *default workflow* below 0.15 except at Year 4 with fALFF in which the *B.2 workflow* (no structural segmentation) led to lower R2 values and at baseline with fMRIprep pipeline (*B.3 workflow*). For these, the best performance achieved was better than chance.

Regarding the impact of feature computation variations (*workflows C.1 and C.2*), we found better performance at Baseline for workflows *C.2 - default workflow with ALFF* in which the best model $\times$ parcellation combination led to a better R2 value than the one

reported in the original study (0.325 vs 0.242 in the original paper). We also observed this phenomenon with the *C.1 workflow* in which we used non z-scored ReHo maps: we found a higher performance than the one obtained with the default workflow and reported in the original study ($R2 = 0.374$). For these two variations, R2 differences with default remained lower than 0.1. At Year 1 and Year 4 with fALFF however, the use of ALFF instead of fALFF (*workflow C.2*) led to lower performance (R2 mean absolute difference above 0.15). This observation was not found at Year 2.

For Year 1 and Year 2 predictions, the set of input features (*workflows D.*) had a large impact on the performance of these models. In particular, models trained without Baseline MDS-UPDRS score (D.2) and with only imaging features (D.4) showed lower R2 values with for fALFF and for ReHo at Year 1 and Year 2 (R2 absolute difference above 0.2), which suggests that Baseline MDS-UPDRS played a central role in the prediction of MDS-UPDRS at follow-up visits compared to imaging features. It also explains why variations in the extraction of imaging features (pre-processing or computation) only had a lower impact on the performance for these two time points.

Overall, at Year 1 and Year 2, performance seemed to be driven mostly by clinical and demographic features, in particular by MDS-UPDRS Baseline scores. At Baseline and Year 4, other variations related to image features (pre-processing and feature computation) were associated with larger changes in performance. For all workflows, time points and feature, best performing model x parcellation combination always exhibited better than chance performance.

B.3.6 Model choice and performance reporting

Table B.4 compares the results obtained using different model selection and evaluation methods. Using the nested cross-validation described in the paper (*Workflow E.1*), we obtained lower results than the original ones and than the ones obtained with our best models for all time points (for instance, $R2 = 0.049 vs 0.205$ with our best model for prediction with fALFF at Baseline). Using this method, the models at Year 1 and Year 2 were still well performing compared to other time point, for both ReHo and fALFF, with particularly high R2 values (between around 0.4 and 0.6) obtained using any reporting method.

Results computed using the same model and parcellation as the best performing combinations in the original paper (Table 2 from Nguyen et al., 2021) (*Workflow E.2*) also had lower performance than in the original study, for all time points (*e.g.* $R = -0.102$ for

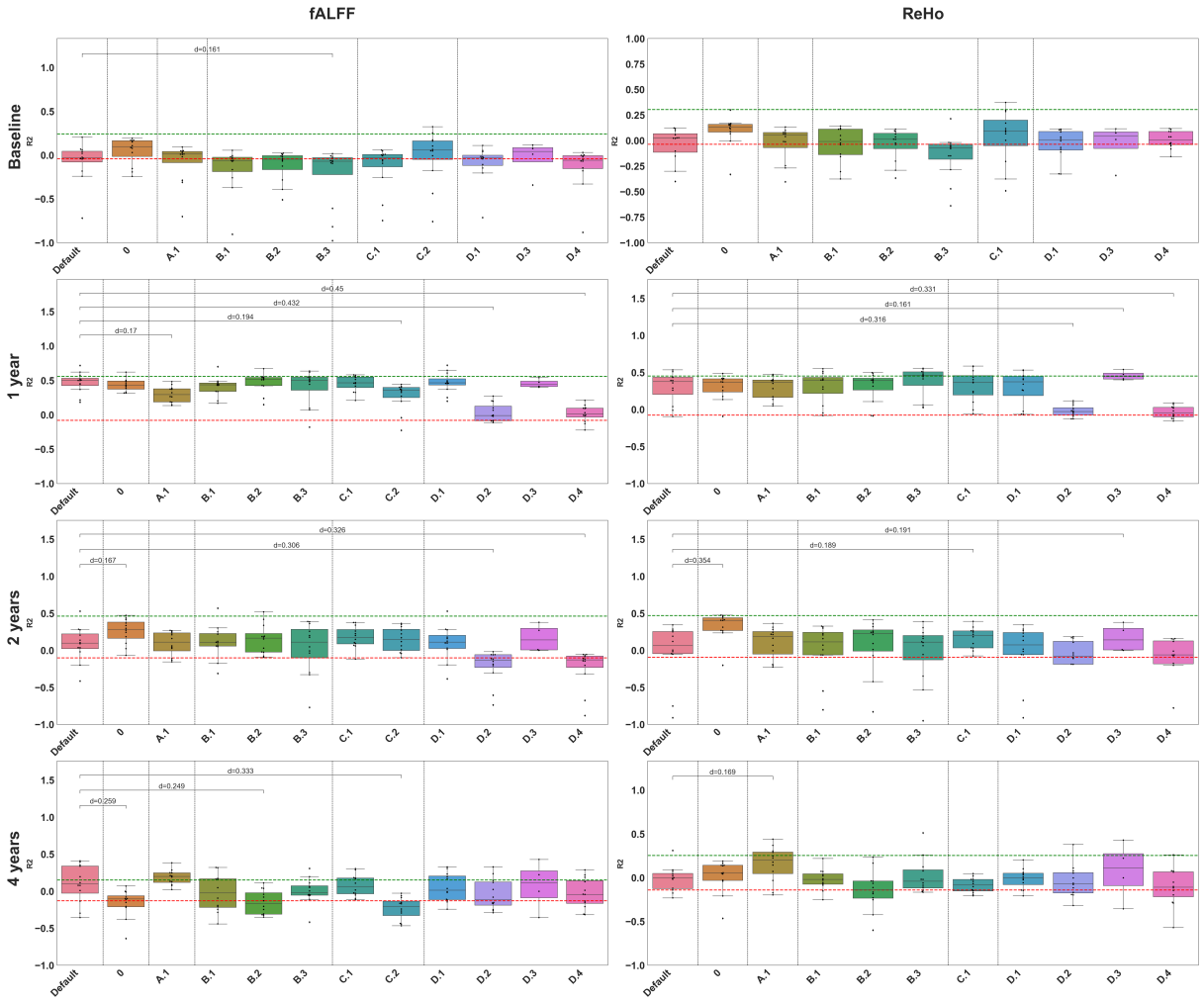

Figure B.4 – Performance of models trained for prediction at each time point, using fALFF or ReHo, with variations in the workflow. Boxes represent the performance (R2 values) of the 12 models (4 models × 3 parcellations). Green horizontal dashed lines show the R2 value reported in the original study for the corresponding time point and feature. Red horizontal dashed lines show the chance-level computed using permutation test. Raw effect sizes (d) are computed as absolute difference between the mean R2 performance with *default workflow* and mean R2 performance with other variations. Only large differences (above threshold $d = 0.15$) are reported.

- *Workflow 0* - reproduction using authors derivatives.
- *Workflow A.1* - default workflow with replication cohort.
- *Workflow B.1* - default workflow with FSL segmentation,
- *Workflow B.2* - default workflow without structural priors,
- *Workflow B.3* - fMRIprep pipeline.
- *Workflow C.1* - default workflow with no Z-scoring,
- *Workflow C.2* - default workflow with ALFF.
- *Workflow D.1* - default workflow with no dominant disease side,
- *Workflow D.2* - default workflow with no Baseline MDS-UPDRS,
- *Workflow D.3* - default workflow with no imaging features,
- *Workflow D.4* - default workflow with only imaging features.

prediction with ReHo at Baseline). However, as observed for nested cross-validation, the performance obtained with these models at Year 1 and Year 2 was still high and close to the ones obtained with our best models. We speculate that the effect size detected with models at these time points was large and thus, tended to be more reproducible across optimization schemes.

In Nguyen et al., 2021, authors also report the model's ability to classify high- versus low-future severity subjects. The performance obtained for this task was consistent with the observation made on R2 values: models with high performance in terms of R2 were usually good at distinguishing high and low severity patients (*e.g.*, AUC of 0.805 and 0.767 for prediction at Year 1 with respectively fALFF and ReHo using the *default workflow*).

B.3.7 Feature importance

To further explore the reproducibility and replicability of findings in Nguyen et al., 2021, we measured feature importance for the ReHo and fALFF imaging features and the default reproduction workflow, across all time points. Fig B.5 and B.6 compare the feature importances obtained with the *default workflow* to the ones reported in the original study.

Feature importance showed relatively few overlap between the ones obtained using our models and those reported in the original study, especially for imaging features, at all time points. Note that the same mask Schaefer atlas that was used by Nguyen et al., 2021 was not used here. For instance, for fALFF at Baseline, the left postcentral region was identified as the most important feature for prediction in our study and was not identified in the original study. For ReHo, we found no important imaging feature that was similar to the ones detected in the original study. However, for some brain regions for which an imaging feature was identified as an important feature, hemispheric opposites or sub-parts of the same global regions were identified in our models compared to the original detected features. For instance, the middle cingulum was identified in our Baseline model with ReHo but in the left hemisphere instead of the right one in the original paper. For this model, regions of the frontal cortex were also detected as important in the original paper, but those we found were very close or were part of the same lobe/region (*e.g.* frontal supero-orbital and middle in original, frontal inferior in ours). Regions identified for fALFF and ReHo were also different at Baseline, consistently with the findings of Nguyen et al., 2021.

For other time points, the main feature of importance was the Baseline MDS-UPDRS score for both fALFF and ReHo and other features had a lower importance value, in par-

Time point	Feature	Type	R2	RMSE	AUC	PPV	NPV	Spec.	Sens.
Baseline	fALFF	Original	0.242	14.006	0.668	60.0%	74.0%	75.5%	58.1%
		Default	0.205	14.26	0.584	51.7%	66.0%	71.4%	45.5%
		Workflow E.1	0.049	15.6	0.514	42.3%	60.7%	69.4%	33.3%
		Workflow E.2	-0.039	16.31	0.493	39.4%	59.2%	59.2%	39.4%
	ReHo	Original	0.304	13.415	0.674	59.4%	75.0%	73.5%	61.3%
		Default	0.124	14.98	0.716	63.9%	78.3%	73.5%	69.7%
		Workflow E.1	-0.164	17.26	0.528	43.8%	62.0%	63.3%	42.4%
		Workflow E.2	-0.102	16.8	0.493	39.3%	59.3%	65.3%	33.3%
Year 1	fALFF	Original	0.558	14.256	0.753	70.4%	80.0%	71.4%	79.2%
		Default	0.717	11.6	0.805	75.9%	86.4%	73.1%	88.0%
		Workflow E.1	0.569	14.3	0.786	73.3%	85.7%	69.2%	88.0%
		Workflow E.2	0.453	16.11	0.69	62.9%	81.2%	50.0%	88.0%
	ReHo	Original	0.453	15.861	0.753	70.4%	80.0%	71.4%	79.2%
		Default	0.535	14.85	0.767	71.0%	85.0%	65.4%	88.0%
		Workflow E.1	0.483	15.67	0.726	70.4%	75.0%	69.2%	76.0%
		Workflow E.2	0.535	14.85	0.767	71.0%	85.0%	65.4%	88.0%
Year 2	fALFF	Original	0.463	13.426	0.765	78.6%	76.5%	68.4%	84.6%
		Default	0.529	12.68	0.669	69.2%	66.7%	55.6%	78.3%
		Workflow E.1	0.478	13.35	0.669	69.2%	66.7%	55.6%	78.3%
		Workflow E.2	0.529	12.68	0.669	69.2%	66.7%	55.6%	78.3%
	ReHo	Original	0.471	13.322	0.739	75.9%	75.0%	63.2%	84.6%
		Default	0.344	14.95	0.635	65.5%	66.7%	44.4%	82.6%
		Workflow E.1	0.272	15.76	0.607	63.3%	63.6%	38.9%	82.6%
		Workflow E.2	0.344	14.95	0.635	65.5%	66.7%	44.4%	82.6%
Year 4	fALFF	Original	0.152	14.957	0.636	64.7%	62.5%	62.5%	64.7%
		Default	0.411	12.19	0.833	91.7%	77.8%	93.3%	73.3%
		Workflow E.1	0.242	13.83	0.733	73.3%	73.3%	73.3%	73.3%
		Workflow E.2	-0.134	16.92	0.633	66.7%	61.1%	73.3%	53.3%
	ReHo	Original	0.255	14.015	0.699	73.3%	66.7%	75.0%	64.7%
		Default	0.312	13.18	0.667	72.7%	63.2%	80.0%	53.3%
		Workflow E.1	-0.044	16.23	0.567	60.0%	55.0%	73.3%	40.0%
		Workflow E.2	-0.23	17.62	0.6	63.6%	57.9%	73.3%	46.7%

Table B.4 – Performance reported using different model selection and evaluation methods. “Original" is the performance reported in the Original study (Nguyen et al., 2021). “Default" is the performance obtained with the model × parcellation that obtained the best performance during reproduction. “Workflow E.1" is the performance obtained when using the nested cross-validation scheme described in the paper (*i.e.* optimizing model × parcellation in the inner fold). “Workflow E.2" is the performance obtained with the model and parcellation reported in the paper.

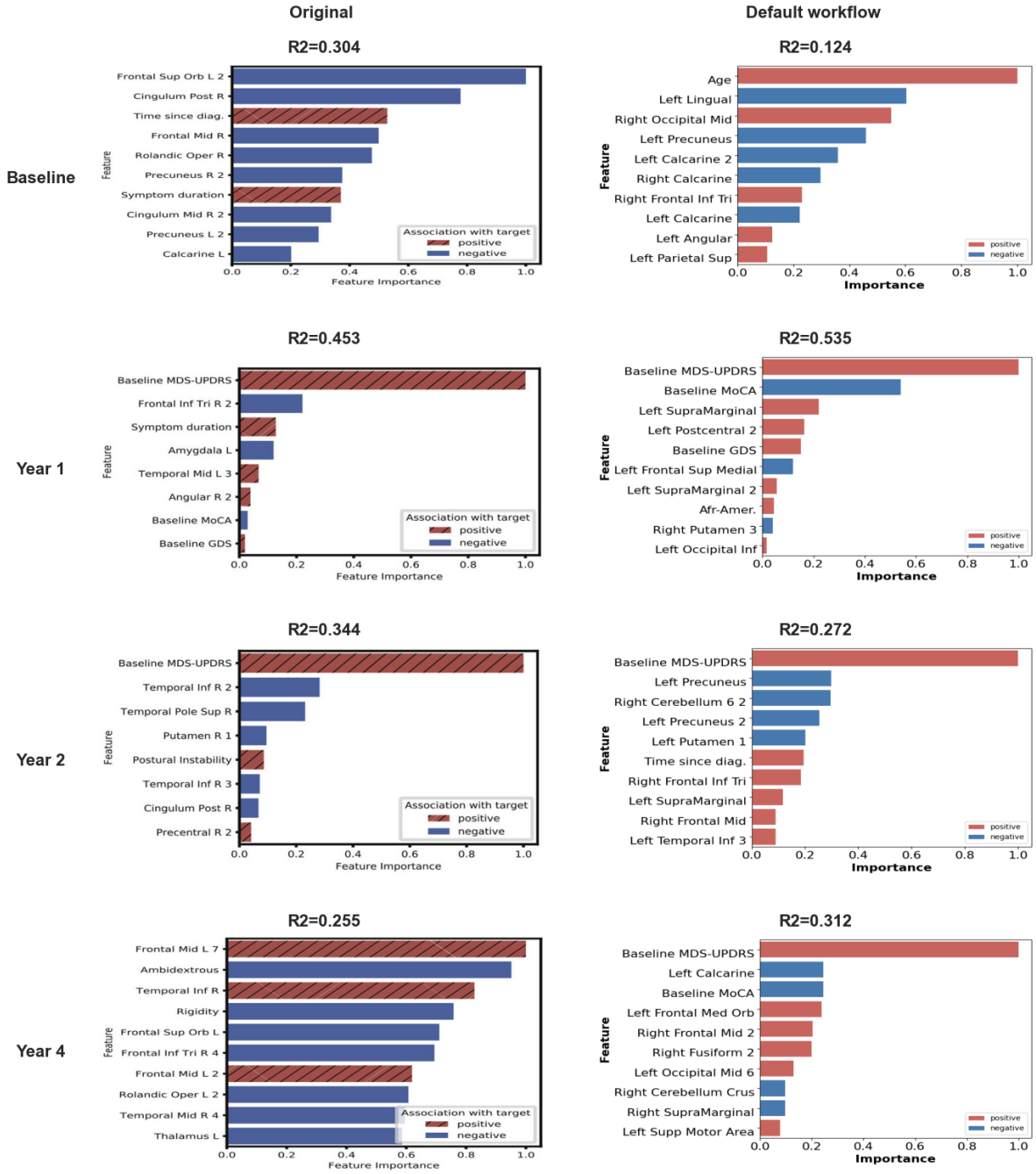

Figure B.5 – Predictive features learned by the best performing models to predict MDS-UPDRS score at each time point for the original study (left - extracted from Nguyen et al., 2021) and the *default workflow* (right) using ReHo. Features with low importance were not shown. Red bars indicate a positive association and blue bars indicate a negative association. Stars (*) represent the presence of this feature in the original study and the reproduction.

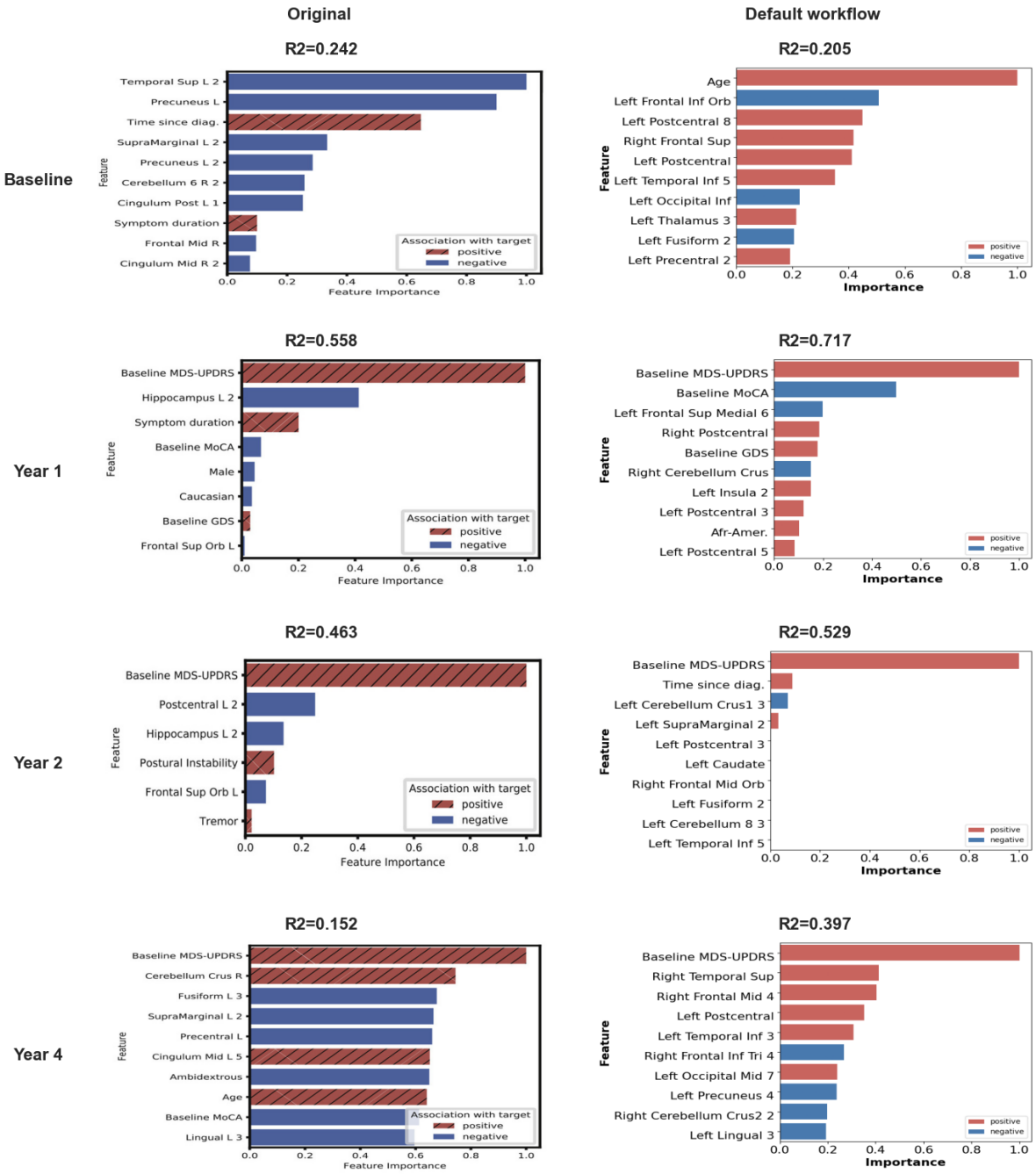

Figure B.6 – Predictive features learned by the best performing models to predict MDS-UPDRS score at each time point for the original study (left - extracted from Nguyen et al., 2021) and the *default workflow* (right) using fALFF. Features with low importance were not shown. Red bars indicate a positive association and blue bars indicate a negative association. Stars (*) represent the presence of this feature in the original study and the reproduction.

ticular at Year 1 and at Year 2. This observation was also supported by the performance of models that did not include the Baseline MDS-UPDRS score in their feature set: these models showed lower performance at these two time points compared to the default models ($p < 0.01$). Note that, as shown in Fig B.6 and B.5, similar R2 is attained, though through different sets of features. This is entirely plausible for multivariate machine learning models, and does not preclude the other set of features from not also being useful (*e.g.* if default gets 0.717, it could be that the features from Original are still informative of outcome).

B.4 Discussion

B.4.1 Summary

We investigated the reproducibility and replicability of the predictive models of PD progression described in Nguyen et al., 2021. Using the *default reproduction workflow*, *i.e.*, with methods and cohorts closest to the ones described in Nguyen et al., 2021, the performance of our best models was better than chance ($R2 > 0$). For both ReHo and fALFF, we found slightly lower performance than the one reported in the original study at Baseline with our *default workflow*. The performance were higher than in the original study at Year 1, Year 2 and Year 4. These values remained close to those reported in the original study and performance were better than chance, supporting the predicting capability of the model reported in the original paper. Thus, using a cohort and methods adapted from Nguyen et al., 2021, we were able to train several machine learning models that predicted Parkinson's disease progression (MDS-UPDRS scores at Baseline, Year 1, Year 2, and Year 4) with a performance higher than chance and with values comparable to those reported in the original study for most models. On these criteria, we could conclude that the replication experiment was successful.

When training the models using the derived data computed by the authors at the time of the original study (fALFF and ReHo whole-brain maps), we found very close performance to the original ones, except at Year 4 with fALFF, where the default worklow found even higher predictability. This confirms the quality of the reproduction for the clinical and demographics feature selection and for the machine learning part. Thus, differences in performance with our reproduction workflow could be explained by the preprocessing pipelines and imaging features computation but also by differences between

cohorts since our reproduction cohort contains, at baseline, 4 participants with different sessions than the original ones. This also impacts follow-up time points cohorts, and potentially the performance of the models. In addition, we found feature importance values that differed —for some predictions— from the ones found by the authors. This step was complex to reproduce since our best performing model x parcellation combination did not match the ones reported in the original paper at several time points, which questions the comparability of the features. When fitting a machine learning model, similar performance can be achieved by different sets of features, which explains why feature importance values might be inconsistent across models.

When introducing specific variations in the workflow, we managed to obtain results that were more similar to the original ones than our reproduction ones, in particular when changing the feature computation method at Baseline. Some changes in the *default workflow* also led to lower performance, for instance at Year 1 and at Year 2 when removing Baseline MDS-UPDRS score or when using only imaging features. For these time points in particular, variations of the pre-processing pipeline (workflows B.), feature computation (workflows C.) and model choice and reporting (workflows E.) had little impact on the performance of the models compared to other time points. We speculate that imaging features were of low importance in the models prediction for these time points compared to other time points (Baseline and Year 4) for which variations on image computation (pre-processing or feature) had a larger impact. Without variations (*i.e.* with the *default workflow*), performance of models at Baseline and Year 4 time points was already low, which also suggests that effect sizes detected by models were small and that these models were underpowered (Button et al., 2013; Ioannidis, 2008a), making them more sensitive to variations.

In the original study, authors also reported performance of the models evaluated on an external dataset (Table 2 of Nguyen et al., 2021) and with Leave-One-Site-Out cross-validation (LOSO CV) in the outer-loop compared to Leave-One-Out (LOO CV) in the main study. They found similar performance at Year 1 (R2 over 0.5) with these variations, comparable to the main results in Nguyen et al., 2021 which reported R2 up to 0.558. Performance at other time points was not available for the external validation, but for LOSO CV, models trained for prediction at Year 2 also performed very well and those of time point Baseline and Year 4 exhibited lower prediction ability compared to the ones tuned using the LOO CV scheme (main original workflow). These two comparisons are consistent with our observations on the robustness to image features variations and of

model selection at Year 1 and Year 2 and the higher sensitivity of models at Baseline and Year 4.

When using a different cohort with distinctions in the distribution of the most important feature (MDS-UPDRS score at Baseline) of the Year 1 model, a lower performance was found using fALFF ($p < 0.05$) and ReHo. This performance remained high and close to the one reported in the original study. Moreover, when removing specific clinical features such as MDS-UPDRS Baseline scores, the performance models at Year 1 and Year 2 significantly dropped. This suggests that the robustness mentioned above was probably dependant on the distribution of these measures. It would be interesting to assess the interaction of variations in both cohorts, imaging features and input features sets to see if the robustness to analytical variations was also present using the replication cohorts and when increasing the importance of image features in the prediction.

B.4.2 Challenges of reproducibility studies

In our reproducibility study, several challenges were encountered, in particular related to cohort selection, fMRI feature pre-processing, and results reporting. To extract the same Baseline cohort as used in Nguyen et al., 2021, we first attempted to query the PPMI database using the information available in the paper and the code shared by the authors. This step was unsuccessful since we could not get the same sample size at Baseline (102 instead of 82 in Nguyen et al., 2021), and we decided to contact the authors who provided us the exact subject and visit list used in the original study. With this list, we were able to build a cohort with the same participants at Baseline. A potential solution to avoid similar difficulties in future reproducibility studies would be to register cohorts obtained from public databases under the same data usage agreements as the original data. In the case of PPMI, a specific section of the online portal could be created to store cohort definitions and associate them with published manuscripts.

Even with the original participant identifiers and visit list at Baseline, we could not retrieve the same Baseline cohort in the PPMI database. Our closest-to-original cohort included the 82 original participants, but for 5 of them, a different visit than the original one was used. For 3 of these visits, we intentionally chose to keep the visits selected by our first query to better fit with the description of the cohort in the paper. For the 2 other visits, the functional images corresponding to these participants and visits were not available anymore in the PPMI database. Since the PPMI database continuously adds new participant visits, we chose to keep only the visits that were added more than a year

before the original study publication, since the original authors did not report the date at which they queried the database. With this filter, the Baseline participants list and the exact same code used to search for follow-up visits, the cohorts obtained for follow-up visits were still dissimilar to the original ones, with more participants and several noteworthy differences in clinical and demographic variables. A first step to solve this particular issue would be to systematically report the date when databases are queried. However, the issues faced when attempting to reproduce the original cohort in fact highlight the need for version control in public databases, using tools such as DataLad (Halchenko et al., 2021) that is for instance adopted in the OpenNeuro database (Markiewicz et al., 2021). With version control, we would be able to retrieve the data from the database as it existed on the date of the original query. In addition, authors would be able to cite the exact version of the database used, which would importantly facilitate cohort reproductions.

Reproducing the fMRI pre-processing and feature computation pipelines described in Nguyen et al., 2021 also raised challenges. First, although authors provided a description of the different pre-processing steps performed and tools used, exact reproductions of neuroimaging pipelines require more detailed information — including specific parameters values, name and version of the standard template used, software versions — given the overall complexity and flexibility of image analysis methods (Carp, 2012a). To reproduce the pipeline used in Nguyen et al., 2021 without contacting the authors, we had to make informed guesses about important parameters of the analysis. Some of these choices were conditioned by the nature of the neuroimaging pipelines (*e.g.*, the choice of standard template to register functional images was constrained by the use of ICA-AROMA) while other decisions were more arbitrary and led to multiple valid variations (*e.g.*, the computation of WM and CSF mean time-series for which we applied three different variations with different software packages and methods). Reporting guidelines, such as COBIDAS (Nichols et al., 2017), were developed to help document analyses and facilitate reproduction studies. However, to reproduce complete analyses, sharing the entirety of the code used in the original experiment remains the most valuable information, as it contains a both human and machine-readable description of the exact method employed. In our case the authors did provide all code and their custom atlas when asked. Code-sharing platforms such as GitHub and GitLab are now widely available for this purpose and long-term preservation of these code is supported by archive systems such as Software Heritage (Cosmo et al., 2017; Abramatic et al., 2018) or Zenodo. We also note that different journals have different requirements regarding what is to be submitted

beyond the manuscript. The original paper (Nguyen et al., 2021) was published in P&RD which at the time of publication of Nguyen et al., 2021 had minimal expectations beyond the manuscript. The authors met these requirements and beyond, providing a public code repository. Harmonization of such practice across journal would be highly beneficial to help reproduction of studies.

The use of a custom-based atlas to parcellate the brain in the original study also created challenges. Future reproducibility studies would benefit from comprehensive descriptions of the methods used to create such custom data, access to the code to create the data, and sharing of the data itself through platforms such as Zenodo, the Open-Science Framework, Figshare, or NeuroVault (Gorgolewski et al., 2015). Such platforms could also be used for sharing derived data, for instance whole-brain fALFF and ReHo maps. However, Data Usage Agreements often requires that derived data have to be shared under the same conditions. We emphasize again the need for specific platforms in public databases to host data associated with a published manuscript, including cohort descriptions and derived imaging data.

The authors of Nguyen et al., 2021 shared code used in the original study, in particular for feature computation (fALFF and ReHo after pre-processing and clinical/demographic features search in PPMI study files) and machine-learning models training. The availability of this code was extremely useful for our reproducibility study, and we warmly acknowledge the authors for taking the time to share reusable code with their analysis. Despite the availability of the code, we still faced some difficulties to reproduce the results presented in the original study, due to discrepancies between the methods reported in the paper and the code shared, especially for the imaging feature computation, the cross-validation procedure and the results reports. For instance, we were not able to retrieve the Z-scoring of whole-brain fALFF and ReHo maps mentioned in the paper. This discrepancy was likely due to the update of the C-PAC pipeline used by the authors for pre-processing, in which the documentation still mentioned the possibility to output Z-scored maps even if this option was not implemented anymore in the pipeline. This reiterate the importance of code versioning and reporting software versions. The use of software container engines such as Docker and Singularity in combination with frameworks such as Boutiques (Glatard et al., 2017) or BIDS-Apps (Gorgolewski et al., 2017) facilitates reproduction and reduces the technical work required to find and install the software versions used in the original study. The authors in Nguyen et al., 2021report that they have begun using both Singularity/Apptainer and Podman for this exact purpose.

Regarding model selection and optimization, we highlight the complexity of nested cross-validation schemes and the on-going debate on the choice of rigorous cross-validation procedures (Wainer et al., 2018; Varoquaux et al., 2023). Here again, code sharing is required to describe the exact evaluation method used in the original study. At this level in the analysis, Jupyter notebooks (Kluyver et al., 2016) are an interesting option to document code and mix it with data, natural text and figures. Initiatives were recently launched to share reproducible Jupyter notebooks, such as NeuroLibre (DuPre et al., 2022), a platform for sharing re-executable preprints. We created a Jupyter notebook for our study, that we made publicly available at `https://github.com/elodiegermani/nguyen-etal-2021`.

To conclude, we highlighted the challenges associated with the reproduction of neuroimaging studies. We discussed some of the specific difficulties encountered in our study, as well as numerous success in reproduction, and provided some potential solutions to further facilitate this process in the future, in terms of time cost and adequacy of the reproduction. Nevertheless, given the complexity of the data, software and analyses required in current neuroimaging studies, reproducing existing papers remains extremely challenging.

Appendix C

Reproduction of analysis pipelines: the NARPS Open Pipelines project

---

In this supplementary chapter, we present the *NARPS Open Pipelines Project*, that was presented at several hackathons (see C.1.4) and a use case of the codebase was the subject of an abstract and a poster presentation at the 28th Annual Meeting of the Organization for Human Brain Mapping (OHBM) in 2022:

- **Title:** fMRI data analysis: how does analytical variability vary with sample size?
- **Authors:** Elodie Germani, Camille Maumet
- **HAL:** inserm-03642535.
- **Code:** swh:1:snp:2e1634838081d6fd46177b674d9d891720f60752.
- **Contributions (Credit taxonomy):** Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualisation, Writing.

C.1 The NARPS Open Pipelines project

A description of the project was published in the Proceedings of the OHBM BrainHack 2022 (Moia et al., 2024).

The goal of the *NARPS Open Pipelines Project* is to provide a public codebase that reproduces the 70 pipelines chosen by the 70 teams of the Neuroimaging Analysis Replication and Prediction Study (NARPS) study (Botvinik-Nezer et al., 2020). The project is

public and the code hosted on GitHub at `https://github.com/Inria-Empenn/narps_open_pipelines`.

C.1.1 Description of the project

This project initially emerged from the idea of creating an open repository of fMRI data analysis pipelines (as used by researchers in the field) with the broader goal to study and better understand the impact of analytical variability. NARPS (Botvinik-Nezer et al., 2020) – a many-analyst study in which 70 research teams were asked to analyze the same fMRI dataset with their favorite pipeline – was identified as an ideal usecase as it provides a large array of pipelines created by different labs. In addition, all teams in NARPS provided extensive (textual) description of their pipelines using the COBIDAS (Nichols et al., 2017) guidelines. All resulting statistic maps were shared on NeuroVault (Gorgolewski et al., 2015) and can be used to assess the success of the reproductions.

C.1.2 Reproduction of the pipelines

C.1.2.1 NARPS dataset

The dataset given to all teams participating in NARPS was designed to study the neural basis of decision-making under risk (Botvinik-Nezer et al., 2019). During fMRI acquisition, participants had to make a choice between options that yield different known outcomes with known probabilities (*e.g.* 50% chance of either gaining 40 or to lose 20); this type of trial is called a "mixed gamble task". On each trial, participants were asked to accept or reject a proposal in which they had an equal 50% chance of either gaining or losing money. Based on previous literature, two groups of participants were defined in which the only difference was the amount of money they could win or lose:

— in the *equal indifference* group, potential losses were half of the potential gains ;

— in the *equal range* group, the range of gains was equal to the range of losses.

The dataset was composed of raw data and preprocessed data. The raw data included anatomical and functional images for each of the 108 subjects, organized using the BIDS standard. The dataset also contained files concerning the tasks: repartition of participants in the two groups and event files. For the preprocessed data, raw data included in this dataset were preprocessed using fMRIprep (Esteban et al., 2019) (RRID: SCR_016216)

version 1.1.4, an fMRI data preprocessing workflow that is robust to variations in acquisition protocols and that does not require a lot of user input. Each participating team could choose to use raw data or preprocessed ones.

C.1.2.2 Pipelines description

In the NARPS study, each participating team was requested to use their favorite analysis pipeline and were asked to describe the pipeline they choose according to the COBIDAS guidelines (Nichols et al., 2017). We used these descriptions to reproduce their pipelines. Among the information they had to provide, there was for instance a description of each preprocessing step, details about the statistical model used and parameters used for the inference.

C.1.2.3 Implementation

To reproduce the pipelines, we used Nipype version 1.6.0 (RRID: SCR_002502) (Gorgolewski, 2017), a Python project that provides a uniform interface to existing neuroimaging software packages and facilitates interaction between these packages within a single workflow. One of the main advantages of Nipype is that it allows efficient and optimized computation through parallel execution plugins. Another one is that users can create workflows using functions from different software packages without the need to switch between scripts in different programming languages with a lot of manual intervention.

This choice of implementation cames with challenges. Indeed, even if the interface with most neuroimaging software packages is easy to understand, it is sometimes not straightforward to convert a pipeline described in the original software package in a Nipype workflow (Chen et al., 2022). Moreover, some functions used in the pipelines to reproduce might not be implemented in Nipype and in this case we have to choose between excluding this pipeline or trying to find a similar function in another neuroimaging software.

C.1.2.4 Evalution of the reproduction

To assess the quality of a reproduction, each pipeline is tested by running on increasing numbers of participants (from 20 to 108) and computing the Pearson's correlation coefficient between the reproduced statistic maps and the original ones shared by the teams on NeuroVault (Gorgolewski et al., 2015).

N. of participants	20	40	60	80	108
Min. correlation for validation	0.30	0.70	0.80	0.85	0.93

Table C.1 – Criteria for validating the reproduction of pipelines in the NARPS Open Pipelines project

C.1.3 Current status of the project

This project was started during my Master 2 internship, and continued during the first year of my thesis. This project obtained fundings from Région Bretagne (Boost MIND) and by Inria (Exploratory action GRASP), which allowed the recruitment of a research engineer (Boris Clenet) for 18 months and of a postdoctoral researcher for 5 months to improve the project.

The work of Boris Clenet has greatly improved the accessibility and the quality of the database. A good and exhaustive documentation was written for potential users and for contributors. Today, thanks to his work, all the pipelines using SPM and FSL using fMRIprep preprocessed data are now implemented in the database, most of them have been validated and only a few remain to be tested (≈ 10). GitHub Actions workflows have also been implemented enabling continuous integration, *i.e.* testing existing pipelines everytime there are changes on them, and to detect typos errors in code comments and documentations.

C.1.4 Outreach

This project was presented and received contributions during the following events:

— Empenn team hackathon (February 2024)

— Brainhack Marseille 2023 (December 2023)

— ORIGAMI lab hackathon (September 2023)

— OHBM Brainhack 2023 (July 2023)

— e-ReproNim FENS NENS Cluster Brainhack (June 2023)

— OHBM Brainhack 2022 (June 2022)

The project will soon be presented at the OHBM BrainHack 2024 in June 2024 at Seoul, South Korea.

C.2 Use case: Evolution of analytical variability with sample size

C.2.1 Context

In Appendix Chapter A, we showed that low sample sizes were one of the cause for the lack of reproducibility of research findings. First, low sample sizes are associated with low statistical power, which decrease the probability of rejecting H0 when it is actually false. Button et al., 2013 estimated the median statistical power of studies in neurosciences between 8% and 31%, which is particularly low. Poldrack et al., 2017 computed the median sample size used in fMRI studies as 28.5 in 2015 and estimated that the standardized effect sizes that would have been required to detect an effect with 80% power was 0.75 with this size of sample. Knowing that this size of effect is considered as large - and given that most modern neuroimaging studies are built to probe effects that are likely to be small such as variations that occur early in the development of a pathology - this suggests that most studies use insufficient sample sizes.

If the use of small sample sizes could lead to incorrect findings due to a lack of statistical power, it also increases the impact of instabilities on results. This has been studied by Loken et al., 2017 who showed the effect of sample sizes on results obtained with different levels of measurement errors. In a recent study, Klau et al., 2020 explored the impact of multiple types of uncertainty for varying sample sizes for two associations in personality psychology. An augmentation of sample size was shown to decrease the vibration of effect caused by sampling uncertainty and other sources of uncertainties, and, even if they remained non negligible, these vibrations stabilized above a certain sample size.

Recently, the *many analyst* approach was used in many fields (Silberzahn et al., 2018) to assess the impact of the flexibility of analytical approaches on the results. In NARPS(Botvinik-Nezer et al., 2020), 70 teams used the same fMRI dataset to answer the same research questions. The 9 hypotheses to answer were about the activation or not of a specific area of the brain during a specific task. Each team was requested to use their favourite pipeline. The 70 teams used 70 different pipelines and contradictory results were found. For some hypotheses, there was a mutual agreement but for others, there was no consensus with a percentage of teams giving a positive answer between 20 to 30%. Statistic maps corresponding to the result of a statistical test used to determine

the significance of an activation on each point of the brain (or voxel of the image) were also compared and few overlap was measured.

However, in NARPS, this instability on the results (which is an illustration of vibration of effects) was observed for analyses made with a dataset containing 108 participants (Botvinik-Nezer et al., 2019), which is nearly 4 times the median sample size of fMRI studies in 2015 (Poldrack et al., 2017). Regarding the observations made by Klau et al., 2020 on the evolution of vibration of effects with sample size, we can then ask how it behaves when analyses are made with a smaller number of subjects in the context of fMRI.

The goal of this project is to investigate the evolution of analytical variability with sample size in the context of fMRI studies. We use the NARPS Open Pipelines codebase to reproduce the statistic maps obtained by several teams of NARPS (Botvinik-Nezer et al., 2020) and modify the number of subjects used to obtain these results to replicate these statistic maps with smaller sample sizes. We use several metrics to measure the evolution of vibration of effects with different sample sizes and try to find a sample size for which vibrations stabilize.

C.2.2 Methods

C.2.2.1 Comparison with NARPS results

In NARPS, the rates of reported significant findings varied across hypotheses and for some, the majority of the teams agreed (see Figure 1 in Botvinik-Nezer et al., 2020). We considered that when the proportion of teams giving a positive answer to a hypothesis was higher than 0.9 or lower than 0.1, it was a converging hypothesis. Using this threshold, a total of 3 hypotheses were considered as converging: H7, H8 and H9.

However, since the proportion of teams giving a positive answer to these three hypothesis was close to 0, we also wanted to study the evolution of vibration of effects for a hypothesis for which a high proportion of teams reported an activation in the studied area. Thus, we also added the hypothesis 5 for which 84% of teams gave a positive answer. The 4 converging hypothesis were then: H5, H7, H8 and H9. To investigate how the vibration of effects would behave with a smaller sample size, we used these converging hypotheses and studied the evolution of this convergence with sample size.

We selected 3 sample sizes: N=20, 40, 60 participants as well as the complete dataset comprising data from 108 participants to verify that the convergence of the hypotheses

was replicated. An illustration of the workflow is presented in Figure C.1. For each sample size, participants were randomly drawn to constitute a group of the wanted sample size. Each sub-dataset of each sample size was analyzed with each pipeline yielding to a result that consisted of 9 sub-results (*i.e.* 1 per hypothesis) each with one unthresholded statistic map, one thresholded statistic map and an answer Yes/No to answer the corresponding hypothesis.

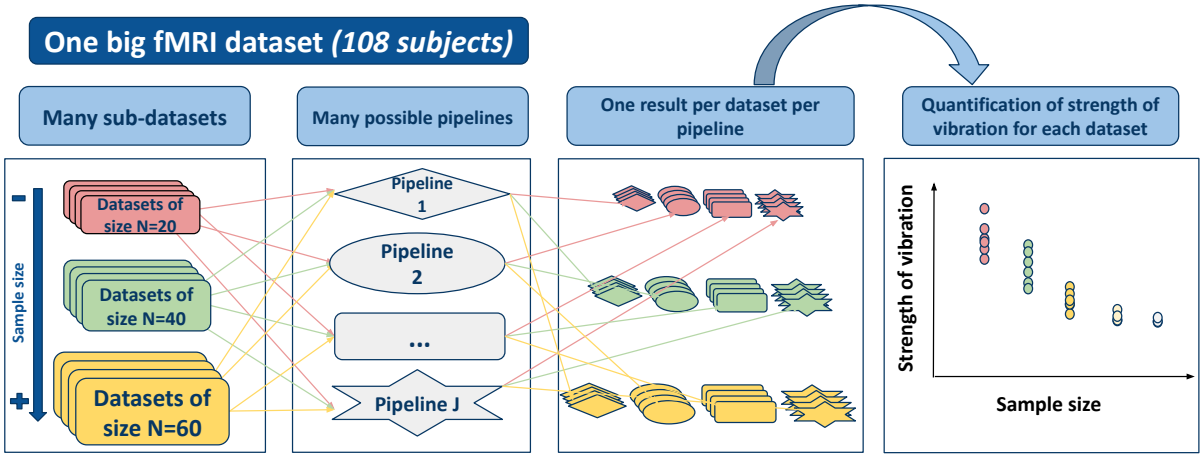

Figure C.1 – Workflow used to study the impact of sample size on vibration of effects

C.2.2.2 Measuring analytical variability

To explore the evolution of analytical variability with sample size, we compared the thresholded and unthresholded statistic maps obtained with different sample sizes qualitatively and quantitatively.

For the quantitative measurements, we first extracted the statistic values of voxels included in the ROI associated with each hypothesis in NARPS using Harvard/Oxford atlas (Desikan et al., 2006). We explored the evolution with sample size of the maximum statistic values inside the ROI for each hypothesis and computed the ratio largest/smallest statistic value inside the ROI.

C.2.3 Results

At the time of the study, 8 pipelines were fully reproduced in the NARPS Open Pipelines Project: 6 using SPM (4 using fMRIprep preprocessed data, 2 using raw data) and 2 using FSL (both using fMRIprep preprocessed data). For each hypothesis, we made a visual between-pipeline comparison of thresholded and unthresholded statistic maps obtained with the different sample sizes. Figure C.2 shows the comparison of statistic maps obtained for H5 with the different pipelines and sample sizes. Looking at these maps, we can see that for $N = 20$, few voxels were found activated for all pipelines. A possible cause might be that, using a small sample size, the power of the statistic test used in the group-level of the analysis was not sufficient to detect an activation, even if there possibly was one. For higher sample sizes, maps answering the same hypothesis showed differences in the number and location of activated voxels. For instance, with $N = 40$, Q6O0's map presented no activation whereas C88N and 2T6S's ones showed some areas of activation within the brain. For $N = 108$, Q6O0's thresholded map also contained less activated voxels than the two others. This observation was probably due to the implementation of Q6O0's pipeline that used age and sex as covariates for the group-level analysis, whereas 2T6S and C88N did not, leading to a reduction of detected effects that were considered as due to one of the covariates.

Figure C.3 shows the evolution of the maximum statistical value inside the ROI depending on sample size for H5. With growing sample sizes, maximum statistical values seems to converge, which is consistent with previous observations from Klau et al., 2020 that the vibration of effects reduces and stabilizes with larger sample sizes. In Figure C.4, we can see the evolution of the ratio largest/smallest statistical value across all teams inside the ROI for H5. We can observe a stabilization and convergence to 1.3 of the ratio between the largest and the smallest maximum statistical values among all teams with sample size, meaning that results across teams are getting closer.

C.2.4 Conclusion

With this work, our objective was to show the impact of sample size on analytical variability. Our findings show that, in fMRI data analysis, the vibration of effects decreases with sample size. Our results also suggest that some variability remains even for large sample sizes. Further work will be needed in order to include more pipelines and investigate which part of the pipelines are the most impactful. The NARPS Open

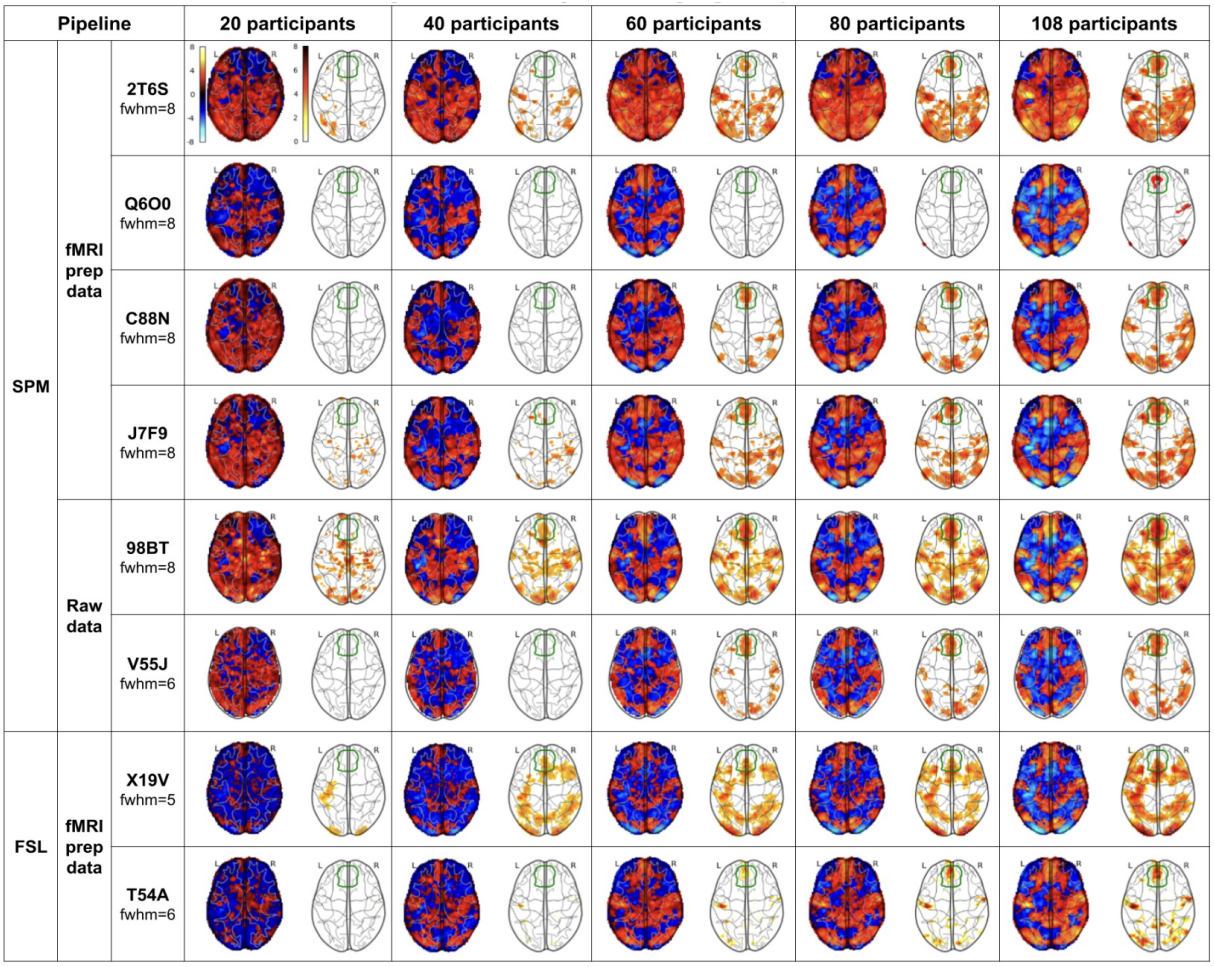

Figure C.2 – Thresholded and unthresholded statistic maps obtained with the different pipelines and sample sizes for H5 "Negative effect in the Ventromedial Prefrontal Cortex - for the equal indifference group".

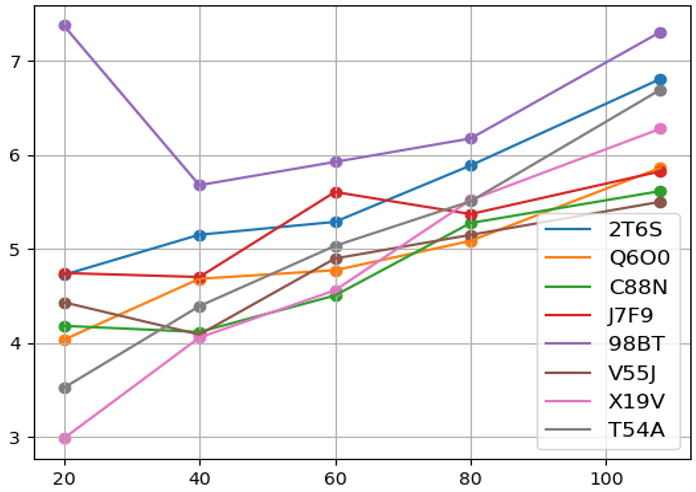

Figure C.3 – Maximum statistical value inside the ROI of the ventromedial prefrontal cortex depending on sample size for the 8 reproduced pipelines from the NARPS Open Pipelines Project

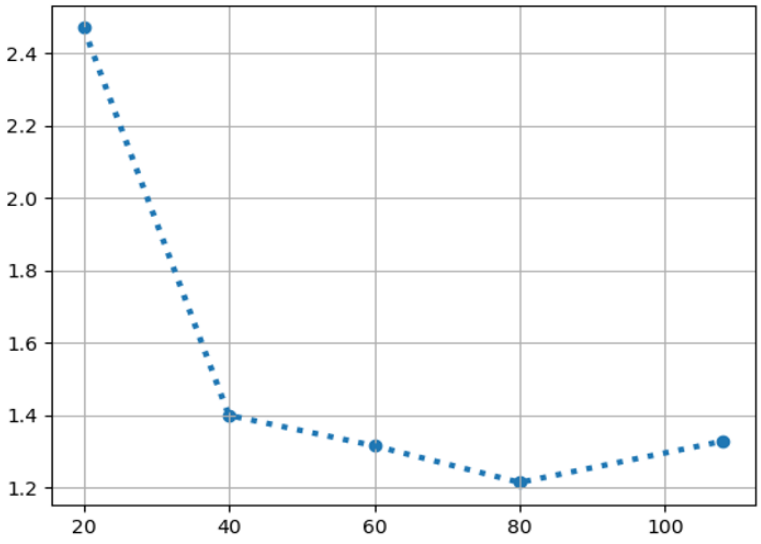

Figure C.4 – Ratio largest/smallest maximum statistical value inside ROI of the ventromedial prefrontal cortex across the 8 reproduced pipelines from the NARPS Open Pipelines Project depending on sample size.

Pipeline project is still ongoing, with more and more pipelines reproduced everyday. We plan to submit a paper to describe the codebase as soon as possible.

Appendix - Supplementary materials

Appendix D

Supplementary materials for Chapter 4

D.1 Model architecture

All models were implemented using PyTorch (Paszke et al., 2019) v1.12.0 (RRID:SCR_018536) with CUDA (Cook, 2012) v10.2. For our model architectures, we chose to use 3-dimensional convolutional feature extractors that take into account the three spatial dimensions of fMRI statistic maps. Schematic representations of the architectures are available in Figure D.1 and Figure D.2.

D.1.0.1 Convolutional AutoEncoder

The base architecture of our CAE was inspired from Zhuang et al., 2019. Two architectures were derived from this base: a 4 layers and a 5 layers architecture, respectively corresponding to the number of convolutional layers in each part of the CAE (encoder and decoder). In the 4-layer model, the encoder part consisted in four 3-dimensional convolutional layers with respectively 64, 128, 256 and 512 channels. Each layer had a kernel size of 3 x 3 x 3, a stride of 2 x 2 x 2 and a padding of 1 x 1 x 1. 3-dimensional batch normalization layers (Ioffe et al., 2015) followed each convolutional layers with respectively 64, 128, 256 and 512 channels and a leaky rectified linear unit (ReLU) activation function was used for all layers. The decoding part of the CAE was symmetric to the encoder, except that 3-dimensional transposed convolutional layers were used instead of classic convolutional layers. Transposing convolutions is a method to upsample an output using learnable parameters. It can be seen as an opposite process to classical convolutions. To keep the number of features symmetric at each layers output, the kernel size of the first layer was set to 4 x 3 x 4 and to 4 x 4 x 4 for all other transposed convolutional layers. Leaky ReLU activation function was also used for all layers except for the last one, *i.e.*

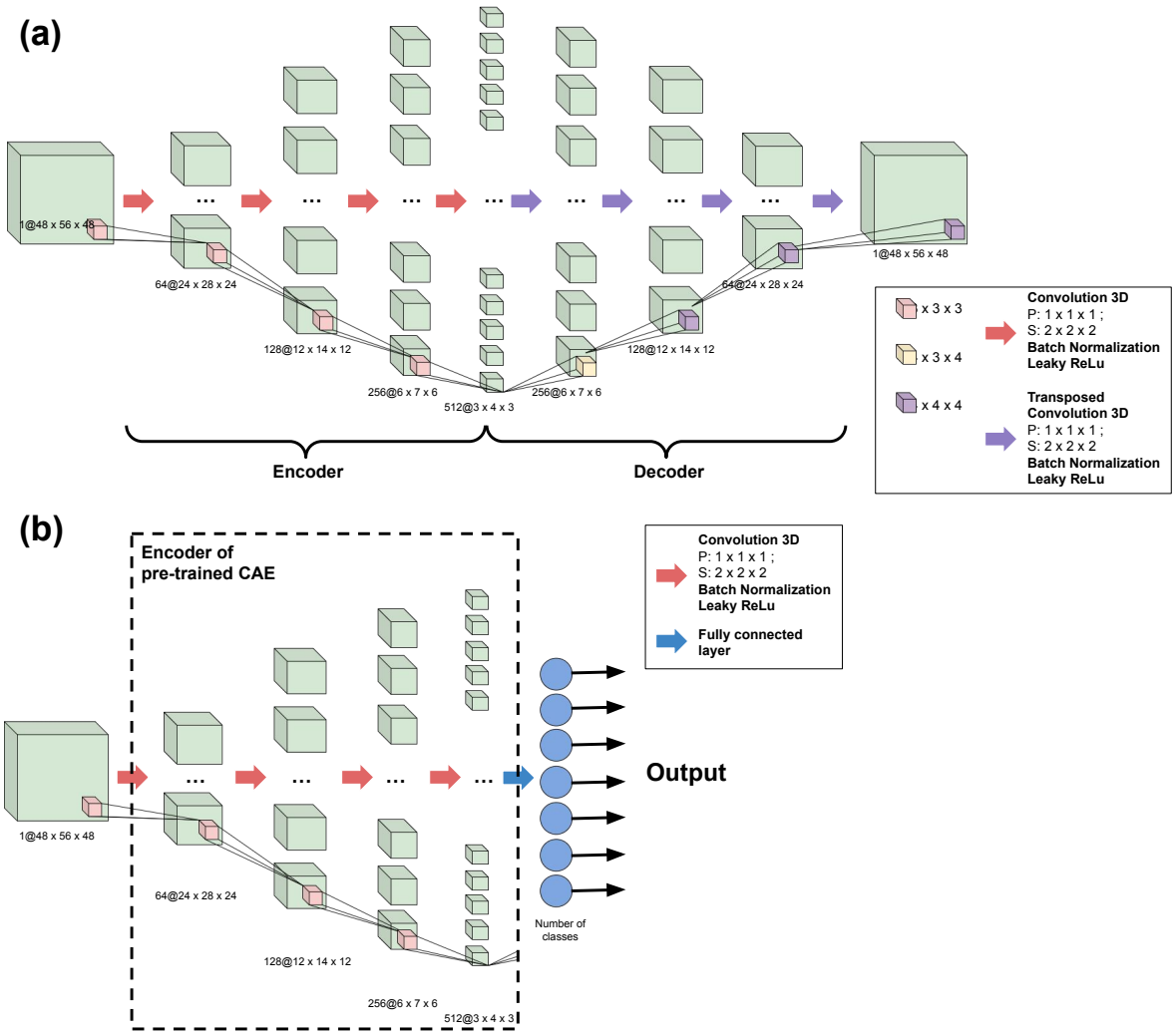

Figure D.1 – Schematic visualisation of the architectures of the CAE (a) and CNN (b) with 4 layers. The CAE is composed of an encoder and a decoder with respectively 4 convolutional and transposed convolutional layers. The size of the latent space is 512 * 3 * 4 * 3. The CNN has the same architecture as the encoder of the CAE with a fully-connected layer added at the end of the network with different numbers of output node depending on the dataset and the classification performed.

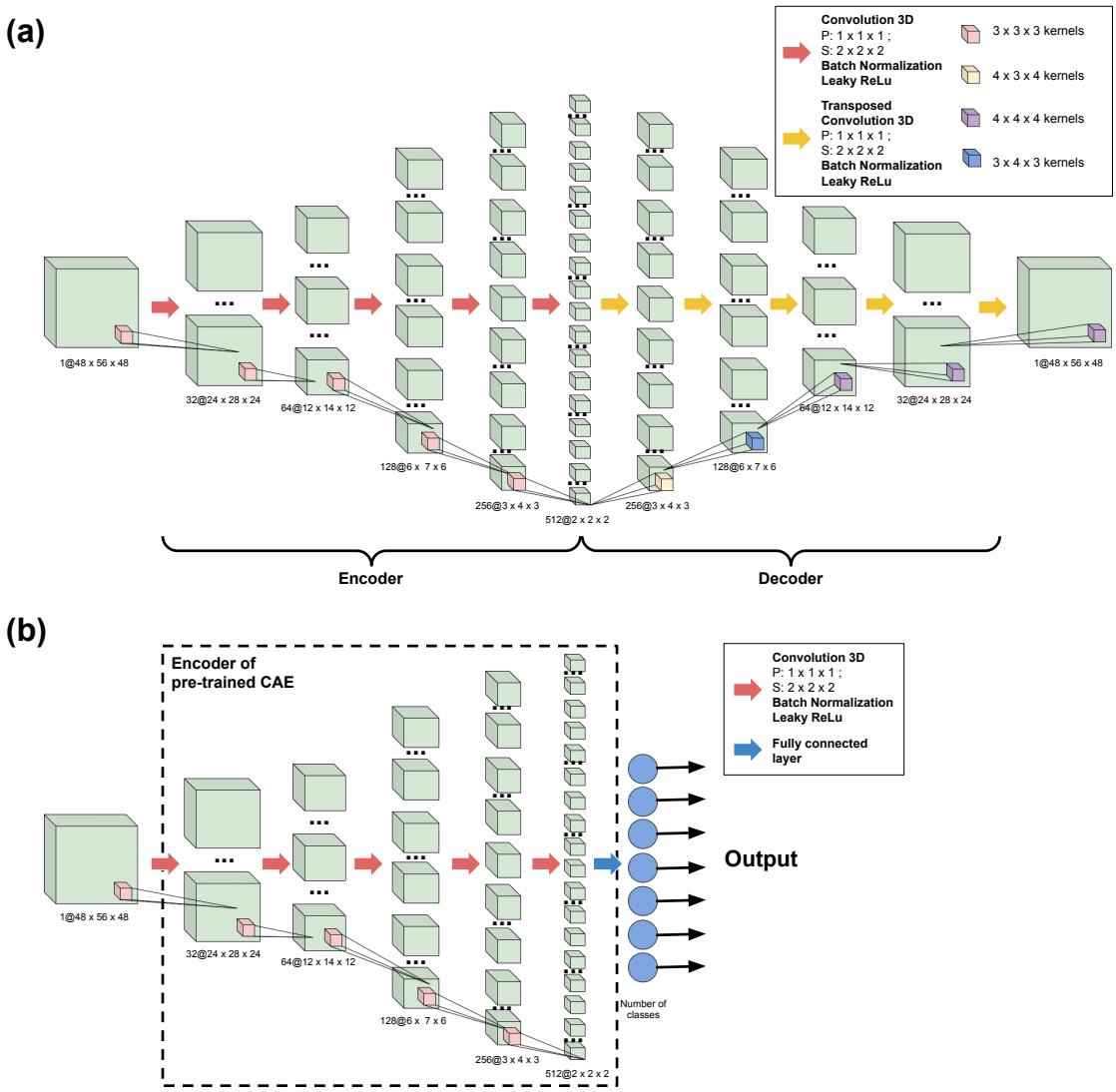

Figure D.2 – Schematic visualisation of the architectures of the CAE (a) and CNN (b) with 5 layers. The CAE is composed of an encoder and a decoder with respectively 5 convolutional and transposed convolutional layers. The size of the latent space is 512 * 2 * 2 * 2. The CNN has the same architecture as the encoder of the CAE with a fully-connected layer added at the end of the network with different numbers of output node depending on the dataset and the classification performed.

the output one, for which a sigmoid function was used in order to obtain output values between -1 and 1. The latent space for this model was of size 512 x 3 x 4 x 3. A schematic representation of this architecture can be found in Figure D.1(a).

In the 5-layer model, one convolutional layer was added at the beginning of the encoder with 32 channels and similar parameters as the other layers of the encoder. A transposed convolutional layer was also added at the end of the decoder with 32 channels. The kernel sizes in the decoder were also modified to maintain the feature map sizes: the first and second layers of the decoder had kernel sizes of 3 x 4 x 3 and 4 x 3 x 4 respectively. All other parameters, batch normalization layers and activation functions were the same. The latent space for this model was of size 512 x 2 x 2 x 2. A schematic representation of this architecture can be found in Figure D.2.

D.1.0.2 Convolutional Neural Network

The 3-dimensional CNN used for classification followed the architecture of the encoder part of the CAE. In the same way as for the CAE, two CNN architectures were derived. For each one, we took the corresponding architecture of the encoder (4 or 5 layers) and added a fully connected layer at the end. The number of nodes in this layer varied depending on the number of classes. A softmax activation function was used for this output layer. Visual representation of the CNN are available in Figure D.1(b) and Figure D.2.

Appendix E

Supplementary materials for Chapter 5

E.1 Implementation settings

The neural network used in CCDPM to predict the noise follows a simple U-Net architecture (Ronneberger et al., 2015) with two downsampling and upsampling blocks with 3D convolutions layers and skip connections. The hyperparameters of the DDPM are the following: $t = 500$ diffusion steps; linear noise schedule with variances in the range of $\beta_1 = 10^4$ and $\beta_t = 0.02$; batch size of 8 and learning rate of 1e-4. The weight w used to control the conditional guidance is optimized on the validation set by comparing $w = 0, w = 0.5$ and $w = 2$ and a value of 0.5 was found to give the best results in terms of Pearson's correlation coefficient between the target ground-truth and the generated image on this set. The model is implemented using PyTorch Paszke et al., 2019 and trained for 200 epochs on 1 GPU NVIDIA Tesla V100.

The CNN used to extract class conditional features is composed of five 3-dimensional convolution layers with 3-dimensional batch normalization and leaky rectified linear units (ReLU) activation functions, followed by a fully connected layer. The latent space corresponds to a $4,096$ flatten vector which is injected as conditioning to the U-Net. It is trained for 150 epochs using a learning rate of 1e-4 and a batch size of 64 on 1 GPU NVIDIA Tesla V100.

E.2 Supplementary Materials

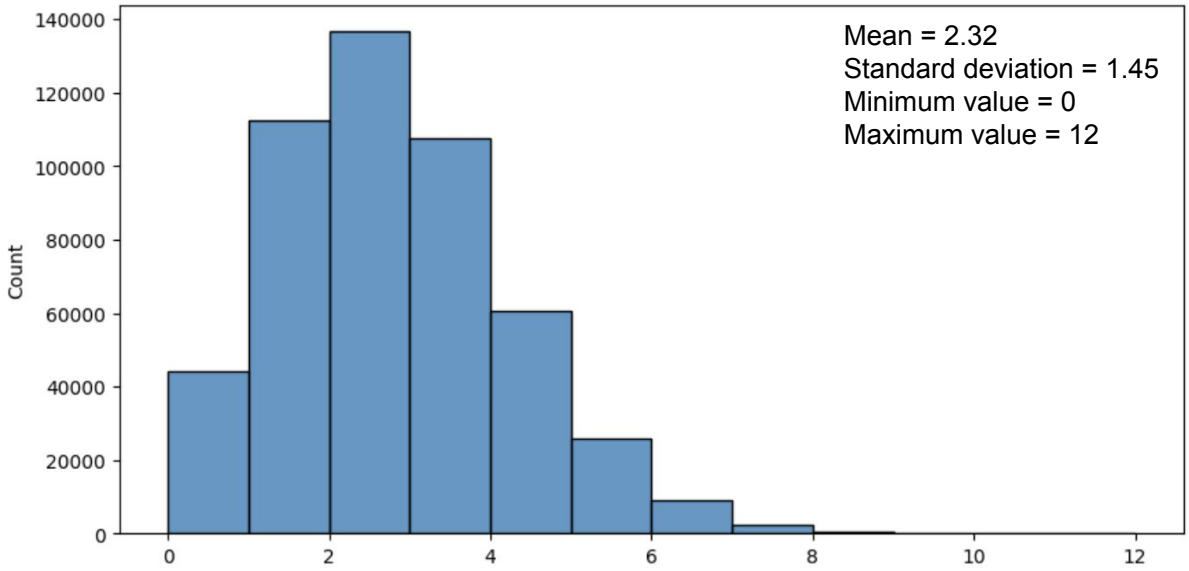

Figure E.1 – Histogram of the number of shared participants between two groups for each pair of groups across the whole dataset. While shared participants between groups can impact our results by slightly over-estimating our performance, the impact is likely to be low due to the small number of shared participants (2.3 on average). In addition, this has no impact the conclusions of our study on the comparison of performance between different models as all models are trained and evaluated on the same sets of groups.

	fsl-1 → spm-0	spm-0 → fsl-1	fsl-1 → spm-1	fsl-1 → fsl-0
Right-hand (included in training set)				
Initial	*76.2*	*76.2*	*82.6*	*91.0*
StarGAN (Choi et al., 2018)	90.6	87.1	87.7	91.8
Right-foot (**NOT** included in training set)				
Initial	*86.5*	*86.5*	*85.7*	*96.2*
StarGAN (Choi et al., 2018)	71.5	71.5	63.4	82.6

Table E.1 – Performance associated with four transfers for StarGAN applied on statistic maps of a task seen during training (*right-hand*), versus a task non-seen during training (*right-foot*). *Initial* represents the metrics between the source image (before transfer) and the ground-truth target image.

Appendix F

Supplementary materials for Chapter 7

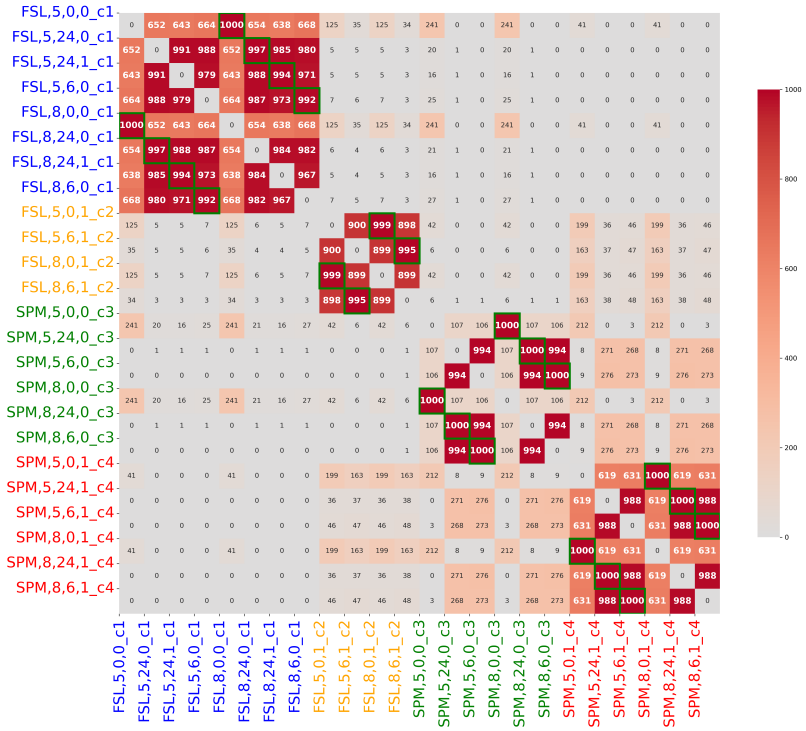

Figure F.1 – Adjacency matrix representing the number of times each pair pipelines belong to the same community across different group-level statistic maps of the contrast *left-hand*

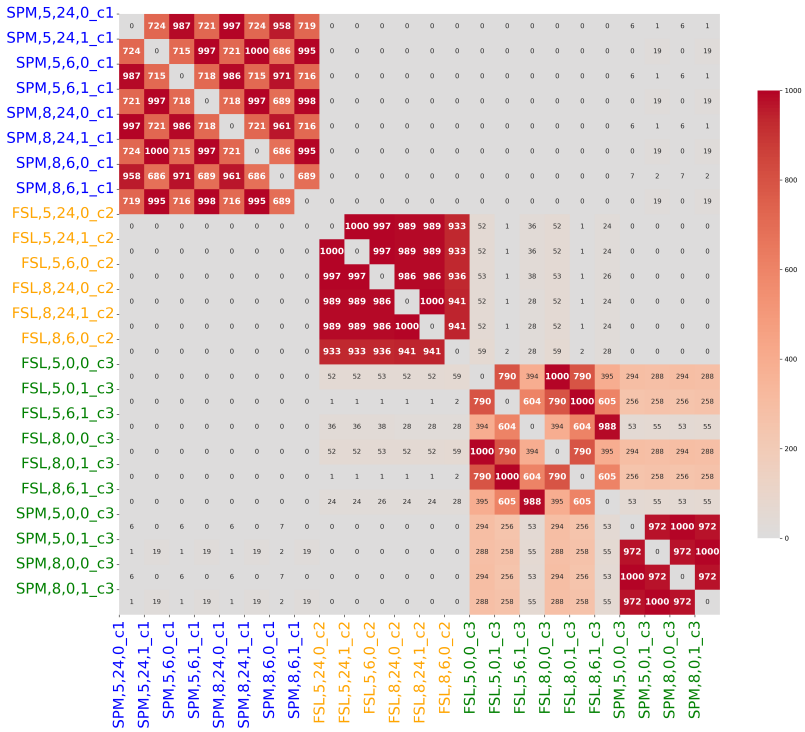

Figure F.2 – Adjacency matrix representing the number of times each pair pipelines belong to the same community across different group-level statistic maps of the contrast *left-foot*

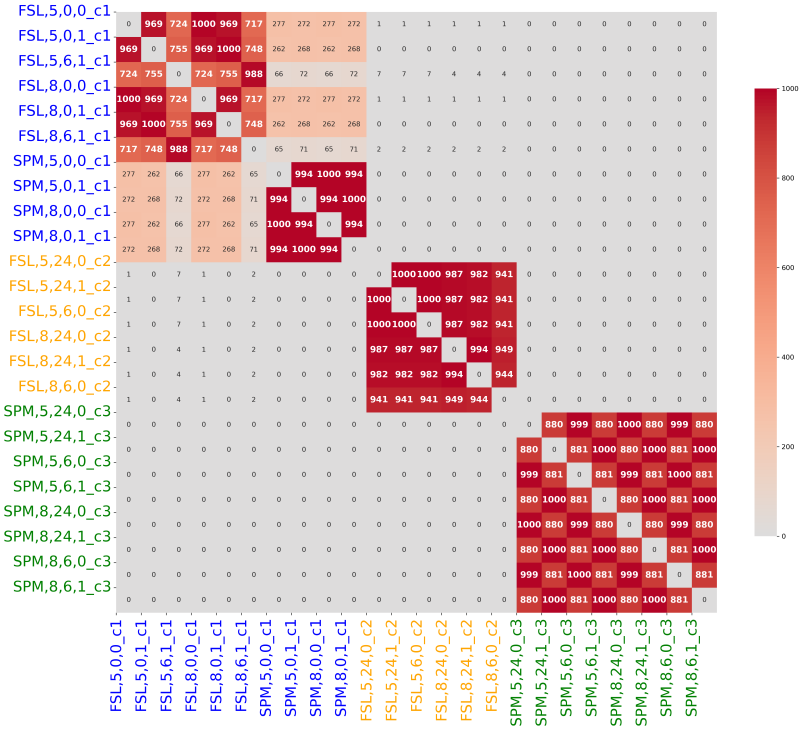

Figure F.3 – Adjacency matrix representing the number of times each pair pipelines belong to the same community across different group-level statistic maps of the contrast *tongue*

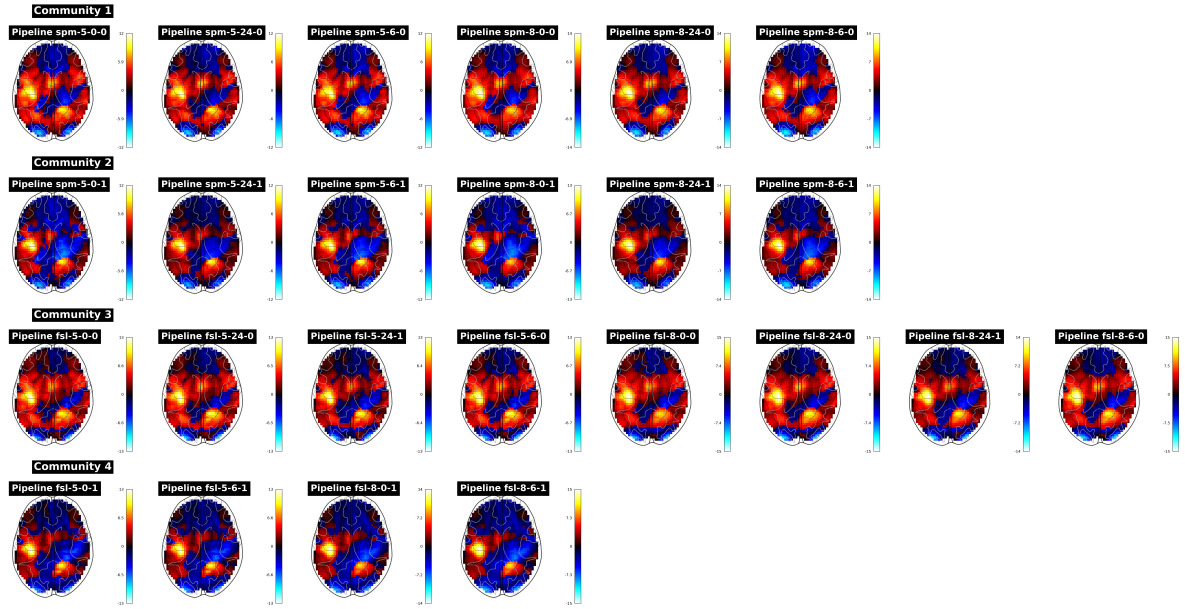

Figure F.4 – Mean statistic map for the contrast *right-hand* across groups (of participants) for a representative pipeline of each community.

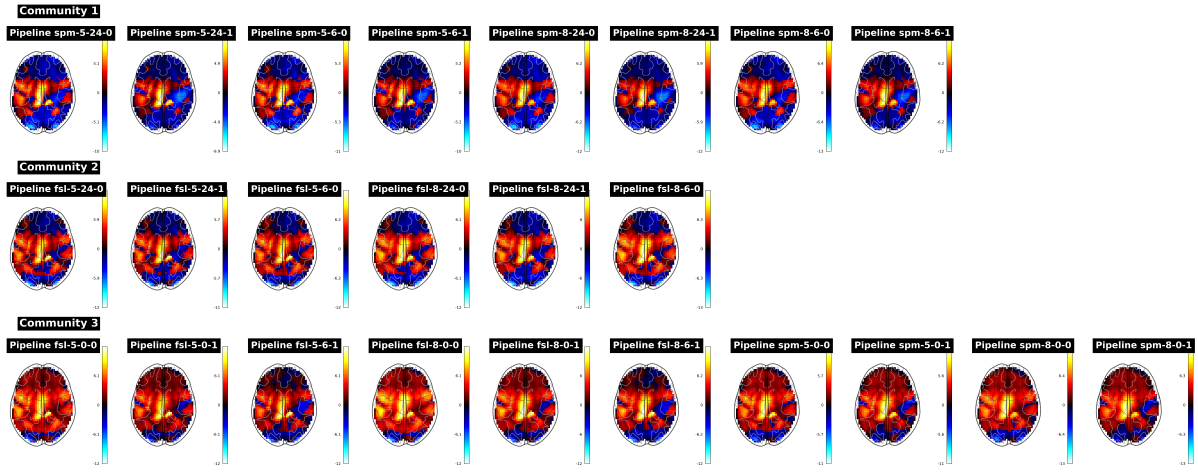

Figure F.5 – Mean statistic map for the contrast *right-foot* across groups (of participants) for a representative pipeline of each community.

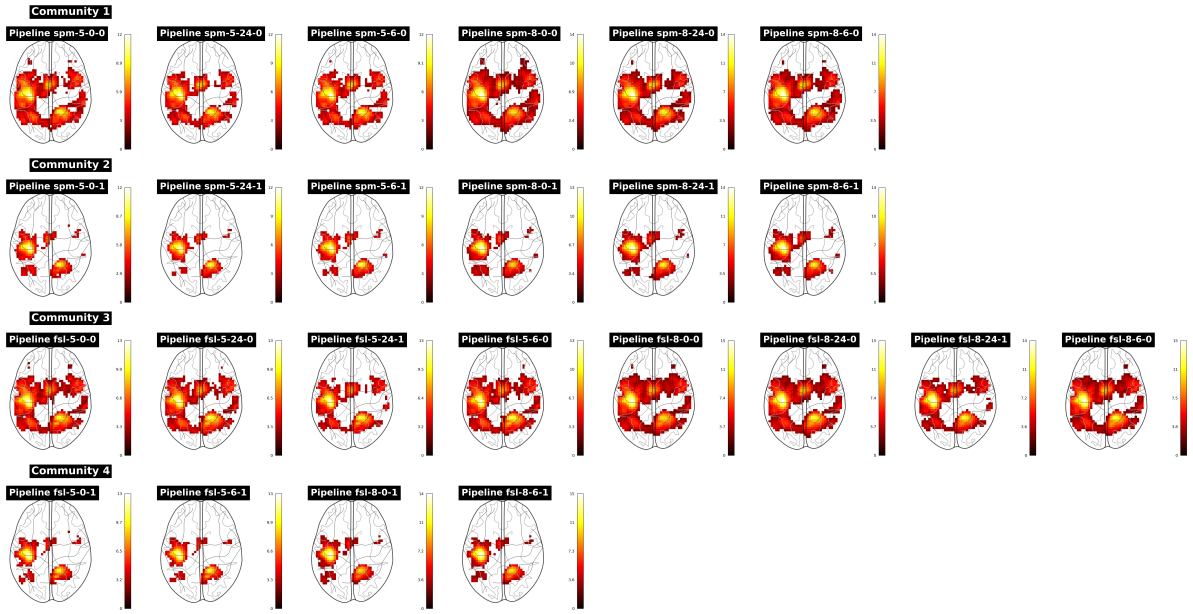

Figure F.6 – Mean statistic map for the contrast *right-hand* across groups (of participants) for a representative pipeline of each community. Unthresholded maps (upper) and thresholded maps (lower) with voxelwise FDR-corrected $p < 0.05$

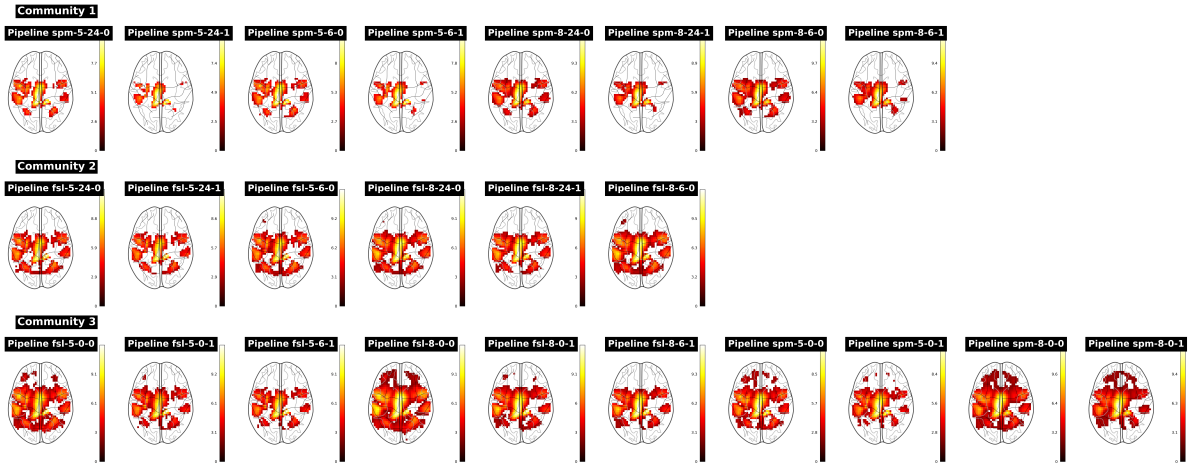

Figure F.7 – Mean statistic map for the contrast *right-foot* across groups (of participants) for a representative pipeline of each community. Unthresholded maps (upper) and thresholded maps (lower) with voxelwise FDR-corrected $p < 0.05$

Appendix G

Supplementary Materials for Chapter 8

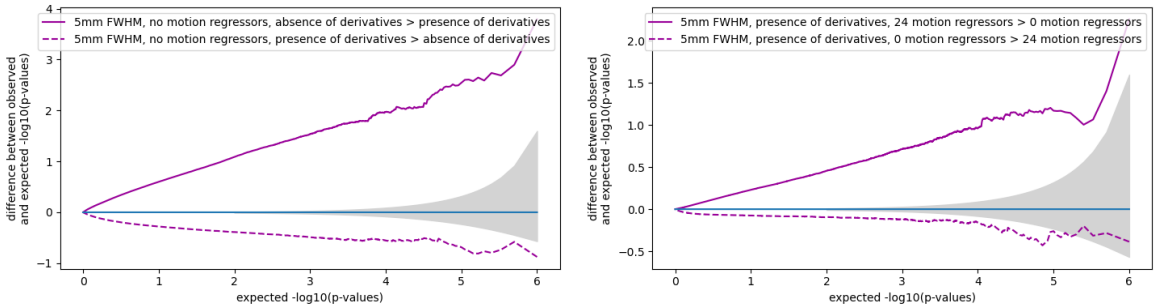

Figure G.1 – Bland-Altman P-P plots for pipelines with two different (right column) parameters and with the same (left column) parameters within SPM. The grey shade corresponds to the 0.95 confidence interval. A curve above (respectively below) the confidence interval indicates invalidity (respectively conservativeness). Default parameters values were modified to 5 mm smoothing, 0 motion regressors and no HRF derivatives to explore the distribution of p-values with different fixed parameters.

SPM

	Smooth 5 mm		Smooth 8 mm	
	No derivatives	Derivatives	No derivatives	Derivatives
0 motion regressors	0.014	0.019	0.025	0.019
6 motion regressors	0.013	0.015	0.021	0.025
24 motion regressors	0.021	0.015	0.018	0.019

FSL

	Smooth 5 mm		Smooth 8 mm	
	No derivatives	Derivatives	No derivatives	Derivatives
No motion regressors	0.01	0.013	0.014	0.014
6 motion regressors	0.015	0.017	0.017	0.022
24 motion regressors	0.017	0.02	0.014	0.012

Table G.1 – False positive rates for between-groups analyses using contrast maps without post-processing with the same pipeline in both groups, with SPM and FSL and for all possible sets of parameters (number of motion regressors, smoothing kernel FWHM and presence or absence of HRF temporal derivatives).

BIBLIOGRAPHY

[1] Martín Abadi et al., « Deep Learning with Differential Privacy », *in*: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318, DOI: 10.1145/2976749.2978318.

[2] Alexandre Abraham et al., « Machine learning for neuroimaging with scikit-learn », *in*: *Frontiers in Neuroinformatics* (2014), DOI: 10.3389/fninf.2014.00014.

[3] Alexandre Abraham et al., « Machine learning for neuroimaging with scikit-learn », *in*: *Frontiers in Neuroinformatics* 8 (2014), p. 14, ISSN: 1662-5196, DOI: 10.3389/fninf.2014.00014.

[4] Jean-François Abramatic, Roberto Di Cosmo, and Stefano Zacchiroli, « Building the Universal Archive of Source Code », *in*: *Communications of the ACM* 61.*10* (2018), ed. by ACM, pp. 29–31, ISSN: 0001-0782, DOI: 10.1145/3183558.

[5] Anees Abrol et al., « Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning », *in*: *Nature Communications* 12.*1* (2021), p. 353, ISSN: 2041-1723, DOI: 10.1038/s41467-020-20655-6.

[6] Soroosh Afyouni and Thomas E. Nichols, « Insight and inference for DVARS », *in*: *NeuroImage* 172 (2018), pp. 291–312, ISSN: 1095-9572, DOI: 10.1016/j.neuroimage.2017.12.098.

[7] Erikson J. de Aguiar, Caetano Traina, and Agma J. M. Traina, « Security and Privacy in Machine Learning for Health Systems: Strategies and Challenges », *in*: *Yearbook of Medical Informatics* 32.*1* (2023), pp. 269–281, ISSN: 0943-4747, DOI: 10.1055/s-0043-1768731.

[8] Mauricio Alférez et al., « Modeling variability in the video domain: language and experience report », *in*: *Software Quality Journal* 27.*1* (2019), pp. 307–347, ISSN: 1573-1367, DOI: 10.1007/s11219-017-9400-8.

[9] Katrin Amunts et al., « Julich-Brain: A 3D probabilistic atlas of the human brain's cytoarchitecture », *in*: *Science* 369.*6506* (2020), pp. 988–992, DOI: 10.1126/science.abb4588.

[10] *Anaconda Software Distribution*, version 2-2.4.0, https://docs.anaconda.com/, 2020.

[11] Jesper L.R. Andersson et al., « Modeling Geometric Deformations in EPI Time Series », *in*: *NeuroImage* 13.*5* (2001), pp. 903–919, ISSN: 1053-8119, DOI: 10.1006/nimg.2001.0746.

[12] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay, « Face aging with conditional generative adversarial networks », *in*: *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 2089–2093, DOI: 10.1109/ICIP.2017.8296650.

[13] Karim Armanious et al., « MedGAN: Medical image translation using GANs », *in*: *Computerized Medical Imaging and Graphics* 79 (2020), p. 101684, ISSN: 0895-6111, DOI: 10.1016/j.compmedimag.2019.101684.

[14] Andrew Arnold, Ramesh Nallapati, and William W. Cohen, « A Comparative Study of Methods for Transductive Transfer Learning », *in*: *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, 2007, pp. 77–82, DOI: 10.1109/ICDMW.2007.109.

[15] Adam R. Aron, Mark A. Gluck, and Russell A. Poldrack, « Long-term test-retest reliability of functional MRI in a classification learning task », *in*: *NeuroImage* 29.*3* (2006), pp. 1000–1006, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2005.08.010.

[16] John Ashburner and Karl J. Friston, « Unified segmentation », *in*: *NeuroImage* 26.*3* (2005), pp. 839–851, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2005.02.018.

[17] Brian B. Avants et al., « A reproducible evaluation of ANTs similarity metric performance in brain image registration », *in*: *NeuroImage* 54.*3* (2011), pp. 2033–2044, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2010.09.025.

[18] Lea Baecker et al., « Machine learning for brain age prediction: Introduction to methods and clinical applications », *in*: *eBioMedicine* 72 (2021), ISSN: 2352-3964, DOI: 10.1016/j.ebiom.2021.103600.

[19] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov, « Differential Privacy Has Disparate Impact on Model Accuracy », *in*: *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

[20] Monya Baker, « 1,500 scientists lift the lid on reproducibility », *in*: *Nature* 533.*7604* (2016), pp. 452–454, ISSN: 1476-4687, DOI: 10.1038/533452a.

[21] Peter A. Bandettini et al., « Processing strategies for time-course data sets in functional mri of the human brain », *in*: *Magnetic Resonance in Medicine* 30.*2* (1993), Publisher: John Wiley & Sons, Ltd, pp. 161–173, ISSN: 0740-3194, DOI: 10.1002/mrm.1910300204.

[22] Lorena A. Barba, *Terminologies for Reproducible Research*, 2018, DOI: 10.48550/arXiv.1802.03311.

[23] Christian Barillot et al., « Shanoir: Applying the Software as a Service Distribution Model to Manage Brain Imaging Research Repositories », *in*: *Frontiers in information and communication technologies* (2016), DOI: 10.3389/fict.2016.00025.

[24] Earl T. Barr et al., « The Oracle Problem in Software Testing: A Survey », *in*: *IEEE Transactions on Software Engineering* 41.*5* (2015), pp. 507–525, ISSN: 1939-3520, DOI: 10.1109/TSE.2014.2372785.

[25] Vishnu M. Bashyam et al., « Deep Generative Medical Image Harmonization for Improving Cross-Site Generalization in Deep Learning Predictors », *in*: *Journal of Magnetic Resonance Imaging* 55.*3* (2022), pp. 908–916, ISSN: 1522-2586, DOI: 10.1002/jmri.27908.

[26] Joanne C Beer et al., « Longitudinal combat: A method for harmonizing longitudinal multi-scanner imaging data », *in*: *Neuroimage* 220 (2020), p. 117129, DOI: 10.1016/j.neuroimage.2020.117129.

[27] C. Glenn Begley and John P.A. Ioannidis, « Reproducibility in Science », *in*: *Circulation Research* 116.*1* (2015), pp. 116–126, DOI: 10.1161/CIRCRESAHA.114.303819.

[28] Yashar Behzadi et al., « A component based noise correction method (CompCor) for BOLD and perfusion based fMRI », *in*: *NeuroImage* 37.*1* (2007), pp. 90–101, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2007.04.042.

[29] Pierre Bellec et al., « Multi-level bootstrap analysis of stable clusters in resting-state fMRI », *in*: *NeuroImage* 51.*3* (2010), pp. 1126–1139, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2010.02.082.

[30] Yoshua Bengio, « Deep Learning of Representations for Unsupervised and Transfer Learning », *in*: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, JMLR Workshop and Conference Proceedings, 2012, pp. 17–36.

[31] Yoshua Bengio, Aaron Courville, and Pascal Vincent, « Representation Learning: A Review and New Perspectives », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.*8* (2013), pp. 1798–1828, ISSN: 0162-8828, DOI: 10.1109/TPAMI.2013.50.

[32] Joseph Berkson, « Application of the Logistic Function to Bio-Assay », *in*: *Journal of the American Statistical Association* 39.*227* (1944), pp. 357–365, ISSN: 0162-1459, DOI: 10.1080/01621459.1944.10500699.

[33] Nikhil Bhagwat et al., « Understanding the impact of preprocessing pipelines on neuroimaging cortical surface analyses », *in*: *GigaScience* 10.*1* (2021), giaa155, ISSN: 2047-217X, DOI: 10.1093/gigascience/giaa155.

[34] Battista Biggio and Fabio Roli, « Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning », *in*: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, Association for Computing Machinery, 2018, pp. 2154–2156, ISBN: 978-1-4503-5693-0, DOI: 10.1145/3243734.3264418.

[35] Bastiaan R Bloem, Michael S Okun, and Christine Klein, « Parkinson's disease », *in*: *The Lancet* 397.*10291* (2021), pp. 2284–2303, ISSN: 0140-6736, DOI: 10.1016/S0140-6736(21)00218-X.

[36] Vincent D Blondel et al., « Fast unfolding of communities in large networks », *in*: *Journal of Statistical Mechanics: Theory and Experiment* 2008.*10* (2008), P10008, DOI: 10.1088/1742-5468/2008/10/P10008.

[37] Myriam Bontonou et al., « Few-Shot Decoding of Brain Activation Maps », *in*: *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1326–1330, DOI: 10.23919/EUSIPCO54536.2021.9616158.

[38] Rotem Botvinik-Nezer and Tor D. Wager, « Reproducibility in Neuroimaging Analysis: Challenges and Solutions », *in*: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, Reliability of Neurocognitive Measures for Mental Health 8.*8* (2023), pp. 780–788, ISSN: 2451-9022, DOI: 10.1016/j.bpsc.2022.12.006.

[39] Rotem Botvinik-Nezer et al., « fMRI data of mixed gambles from the Neuroimaging Analysis Replication and Prediction Study », *in*: *Scientific Data* 6.*1* (2019), p. 106, ISSN: 2052-4463, DOI: 10.1038/s41597-019-0113-7.

[40] Rotem Botvinik-Nezer et al., « Variability in the analysis of a single neuroimaging dataset by many teams », *in*: *Nature* 582.*7810* (2020), pp. 84–88, ISSN: 1476-4687, DOI: 10.1038/s41586-020-2314-9.

[41] Alexander Bowring, Camille Maumet, and Thomas E. Nichols, « Exploring the impact of analysis software on task fMRI results », *in*: *Human Brain Mapping* 40.*11* (2019), pp. 3362–3384, ISSN: 1097-0193, DOI: 10.1002/hbm.24603.

[42] Alexander Bowring, Thomas E. Nichols, and Camille Maumet, « Isolating the sources of pipeline-variability in group-level task-fMRI results », *in*: *Human Brain Mapping* 43.*3* (2022), pp. 1112–1128, ISSN: 1097-0193, DOI: 10.1002/hbm.25713.

[43] Matthew Brett et al., *nipy/nibabel: 3.2.1*, 2020, DOI: 10.5281/zenodo.4295521, URL: https://doi.org/10.5281/zenodo.4295521.

[44] Pete Bridge et al., « Intraobserver Variability: Should We Worry? », *in*: *Journal of Medical Imaging and Radiation Sciences* 47.*3* (2016), pp. 217–220, ISSN: 1939-8654, DOI: 10.1016/j.jmir.2016.06.004.

[45] Buolamwini, Joy, « Artificial Intelligence Has a Problem With Gender and Racial Bias », *in*: *TIME* (2019), (visited on 03/19/2024).

[46] Katherine S. Button et al., « Power failure: why small sample size undermines the reliability of neuroscience », *in*: *Nature Reviews Neuroscience* 14.*5* (2013), pp. 365–376, DOI: 10.1038/nrn3475.

[47] Vince D. Calhoun et al., « The impact of T1 versus EPI spatial normalization templates for fMRI data analyses », *in*: *Human Brain Mapping* 38.*11* (2017), pp. 5331–5342, ISSN: 1097-0193, DOI: 10.1002/hbm.23737.

[48] Craddock Cameron et al., « Towards Automated Analysis of Connectomes: The Configurable Pipeline for the Analysis of Connectomes (C-PAC) », *in*: *Frontiers in Neuroinformatics* 7 (2013), ISSN: 1662-5196, DOI: 10.3389/conf.fninf.2013.09.00042.

[49] Joshua Carp, « On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments », *in*: *Frontiers in Neuroscience* 6 (2012), ISSN: 1662-453X, DOI: 10.3389/fnins.2012.00149.

[50] Joshua Carp, « The secret lives of experiments: Methods reporting in the fMRI literature », *in*: *NeuroImage* 63.*1* (2012), pp. 289–300, ISSN: 10538119, DOI: 10.1016/j.neuroimage.2012.07.004.

[51] Scott Chacon and Ben Straub, *Pro git*, Apress, 2014.

[52] Christopher D. Chambers et al., « *Registered Reports*: Realigning incentives in scientific publishing », *in*: *Cortex* 66 (2015), A1–A2, ISSN: 0010-9452, DOI: 10.1016/j.cortex.2015.03.022.

[53] Adam M. Chekroud et al., « Illusory generalizability of clinical prediction models », *in*: *Science* 383.*6679* (2024), pp. 164–167, ISSN: 0036-8075, 1095-9203, DOI: 10.1126/science.adg8538.

[54] Bing Chen et al., « Individual Variability and Test-Retest Reliability Revealed by Ten Repeated Resting-State Brain Scans over One Month », *in*: *PLOS ONE* 10.*12* (2016), e0144963, DOI: 10.1371/journal.pone.0144963.

[55] Min Chen et al., « Image Registration: Fundamentals and Recent Advances Based on Deep Learning », *in*: *Machine Learning for Brain Disorders*, ed. by Olivier Colliot, Humana, 2023.

[56] Yibei Chen et al., « Reproducing FSL's fMRI data analysis via Nipype: Relevance, challenges, and solutions », *in*: *Frontiers in Neuroimaging* 1 (2022), ISSN: 2813-1193, DOI: 10.3389/fnimg.2022.953215.

[57] Veronika Cheplygina, Marleen de Bruijne, and Josien P. W. Pluim, « Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis », *in*: *Medical Image Analysis* 54 (2019), pp. 280–296, ISSN: 1361-8415, DOI: 10.1016/j.media.2019.03.009.

[58] Junghwan Cho et al., *How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?*, 2016, arXiv: 1511.06348.

[59] Jooyoung Choi et al., « ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models », *in*: *2021 IEEE/CVF International Conference on Computer Vision, ICCV*, IEEE, 2021, pp. 14347–14356, DOI: 10.1109/ICCV48922.2021.01410.

[60] Yunjey Choi et al., « StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation », *in*: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2018, pp. 8789–8797, DOI: 10.1109/CVPR.2018.00916.

[61] Fabien Cignetti et al., « Pros and Cons of Using the Informed Basis Set to Account for Hemodynamic Response Variability with Developmental Data », *in*: *Frontiers in Neuroscience* 10 (2016), ISSN: 1662-453X, DOI: 10.3389/fnins.2016.00322.

[62] Jon F. Claerbout and Martin Karrenbach, « Electronic documents give reproducible research a new meaning », *in*: *SEG Technical Program Expanded Abstracts 1992*, SEG Technical Program Expanded Abstracts, Society of Exploration Geophysicists, 1992, pp. 601–604, DOI: 10.1190/1.1822162.

[63] Mark S. Cohen, « Parametric Analysis of fMRI Data Using Linear Systems Methods », *in*: *NeuroImage* 6.*2* (1997), pp. 93–103, ISSN: 1053-8119, DOI: 10.1006/nimg.1997.0278.

[64] Collection n°1952, *NeuroVault Collection n°1952*, https://identifiers.org/neurovault.collection:1952, Accessed: 2022-01-19, 2016.

[65] Collection n°4337, *NeuroVault Collection n°4337*, https://identifiers.org/neurovault.collection:4337, Accessed: 2022-01-19.

[66] Collection n°457, *NeuroVault Collection n°457*, https://identifiers.org/neurovault.collection:457, Accessed: 2023-05-20, 2015.

[67] The Turing Way Community et al., *The Turing Way: A Handbook for Reproducible Data Science*, 2019, DOI: 10.5281/zenodo.3233986, URL: https://zenodo.org/records/3233986.

[68] Shane Cook, *CUDA Programming: A Developer's Guide to Parallel Computing with GPUs*, 2012, DOI: 10.1016/C2011-0-00029-7.

[69] Roberto Di Cosmo and Stefano Zacchiroli, « Software Heritage: Why and How to Preserve Software Source Code », *in*: *iPRES 2017: 14th International Conference on Digital Preservation*, 2017.

[70] Sergi G. Costafreda, « Pooling fMRI Data: Meta-Analysis, Mega-Analysis and Multi-Center Studies », *in*: *Frontiers in Neuroinformatics* 3 (2009), p. 33, ISSN: 1662-5196, DOI: 10.3389/neuro.11.033.2009.

[71] Robert W. Cox, « AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages », *in*: *Computers and Biomedical Research* 29.*3* (1996), pp. 162–173, ISSN: 0010-4809, DOI: 10.1006/cbmr.1996.0014.

[72] Jessica Dafflon et al., « A guided multiverse study of neuroimaging analyses », *in*: *Nature Communications* 13.*1* (2022), p. 3758, ISSN: 2041-1723, DOI: 10.1038/s41467-022-31347-8.

[73] Saswat Das and Subhankar Mishra, « Advances in Differential Privacy and Differentially Private Machine Learning », *in*: *Information Technology Security: Modern Trends and Challenges*, ed. by Debasis Gountia, Dilip Kumar Dalei, and Subhankar Mishra, Springer Nature, 2024, pp. 147–188, ISBN: 978-981-9704-07-1, DOI: 10.1007/978-981-97-0407-1_7.

[74] Jonas Denck et al., « MR-contrast-aware image-to-image translations with generative adversarial networks », *in*: *International Journal of Computer Assisted Radiology and Surgery* 16.*12* (2021), pp. 2069–2078, ISSN: 1861-6429.

[75] Jia Deng et al., « ImageNet: A large-scale hierarchical image database », *in*: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[76] Rahul S. Desikan et al., « An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest », *in*: *NeuroImage* 31.*3* (2006), pp. 968–980, ISSN: 1053-8119, DOI: https://doi.org/10.1016/j.neuroimage.2006.01.021.

[77] Prafulla Dhariwal and Alexander Quinn Nichol, *Diffusion Models Beat GANs on Image Synthesis*, ed. by A. Beygelzimer et al., 2021.

[78] Adriana Di Martino et al., « The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism », *in*: *Molecular psychiatry* 19.*6* (2014), pp. 659–667, DOI: 10.1038/mp.2013.78.

[79] Jörn Diedrichsen et al., « A probabilistic MR atlas of the human cerebellum », *in*: *NeuroImage* 46.*1* (2009), pp. 39–46, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2009.01.045.

[80] Jérôme Dockès, Gaël Varoquaux, and Jean-Baptiste Poline, « Preventing dataset shift from breaking machine-learning biomarkers », *in*: *GigaScience* 10.*9* (2021), giab055, ISSN: 2047-217X.

[81] Carl Doersch and Andrew Zisserman, « Multi-Task Self-Supervised Visual Learning », *in*: *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2051–2060, DOI: 10.1109/ICCV.2017.226.

[82] Suyu Dong et al., « VoxelAtlasGAN: 3D Left Ventricle Segmentation on Echocardiography with Atlas Guided Generation and Voxel-to-Voxel Discrimination », *in*: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*,

ed. by Alejandro F. Frangi et al., Springer International Publishing, 2018, pp. 622–629, ISBN: 978-3-030-00937-3, DOI: 10.1007/978-3-030-00937-3_71.

[83] Zolnamar Dorjsembe et al., « Conditional Diffusion Models for Semantic 3D Brain MRI Synthesis », *in*: *IEEE Journal of Biomedical and Health Informatics* (2024), pp. 1–10, ISSN: 2168-2208, DOI: 10.1109/JBHI.2024.3385504.

[84] Elizabeth DuPre et al., « Beyond advertising: New infrastructures for publishing integrated research objects », *in*: *PLOS Computational Biology* 18.*1* (2022), e1009651, DOI: 10.1371/journal.pcbi.1009651.

[85] Juan Manuel Durán and Karin Rolanda Jongsma, « Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI », *in*: *Journal of Medical Ethics* (2021), medethics–2020–106820, ISSN: 1473-4257, DOI: 10.1136/medethics-2020-106820.

[86] Cynthia Dwork and Aaron Roth, « The Algorithmic Foundations of Differential Privacy », *in*: *Found. Trends Theor. Comput. Sci.* 9.*3–4* (2014), pp. 211–407, ISSN: 1551-305X, DOI: 10.1561/0400000042.

[87] H. Henrik Ehrsson, Stefan Geyer, and Eiichi Naito, « Imagery of Voluntary Movement of Fingers, Toes, and Tongue Activates Corresponding Body-Part-Specific Motor Representations », *in*: *Journal of Neurophysiology* 90.*5* (2003), pp. 3304–3316, ISSN: 0022-3077, DOI: 10.1152/jn.01113.2002.

[88] Anders Eklund, Thomas E. Nichols, and Hans Knutsson, « Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates », *in*: *Proceedings of the National Academy of Sciences* 113.*28* (2016), pp. 7900–7905, DOI: 10.1073/pnas.1602413113.

[89] Maxwell L. Elliott et al., « What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis », *in*: *Psychological Science* 31.*7* (2020), pp. 792–806, ISSN: 0956-7976, DOI: 10.1177/0956797620916786.

[90] Dumitru Erhan et al., « Why Does Unsupervised Pre-Training Help Deep Learning? », *in*: *The Journal of Machine Learning Research* 11.*19* (2010), pp. 625–660.

[91] Oscar Esteban et al., « fMRIPrep: a robust preprocessing pipeline for functional MRI », *in*: *Nature Methods* 16.*1* (2019), pp. 111–116, ISSN: 1548-7105, DOI: 10.1038/s41592-018-0235-4.

[92] European Organization For Nuclear Research and OpenAIRE, *Zenodo*, en, 2013, DOI: 10.25495/7GXK-RD71, URL: https://www.zenodo.org/.

[93] Alan C. Evans et al., « Brain templates and atlases », *in*: *NeuroImage* 62.*2* (2012), pp. 911–922, ISSN: 10538119, DOI: 10.1016/j.neuroimage.2012.01.024.

[94] David A. Feinberg et al., « Multiplexed Echo Planar Imaging for Sub-Second Whole Brain FMRI and Fast Diffusion Imaging », *in*: *PLOS ONE* 5.*12* (2010), pp. 1–11, DOI: https://doi.org/10.1371/journal.pone.0015710.

[95] Samuel G. Finlayson et al., *Adversarial Attacks Against Medical Deep Learning Systems*, 2019, arXiv: 1804.05296 [cs.CR].

[96] Orhan Firat, Like Oztekin, and Fatos T. Yarman Vural, « Deep learning for brain decoding », *in*: *2014 IEEE International Conference on Image Processing (ICIP)*, Paris, France: IEEE, 2014, pp. 2784–2788, DOI: 10.1109/ICIP.2014.7025563.

[97] Bruce Fischl, « FreeSurfer », *in*: *NeuroImage* 62.*2* (2012), pp. 774–781, ISSN: 1095-9572, DOI: 10.1016/j.neuroimage.2012.01.021.

[98] VS Fonov et al., « Unbiased nonlinear average age-appropriate brain templates from birth to adulthood », *in*: *Organization for Human Brain Mapping 2009 Annual Meeting* 47 (2009), S102, ISSN: 1053-8119, DOI: 10.1016/S1053-8119(09)70884-5.

[99] Jean-Philippe Fortin et al., « Removing inter-subject technical variability in magnetic resonance imaging studies », *in*: *NeuroImage* 132 (2016), pp. 198–212, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2016.02.036.

[100] Alexander L. Fradkov, « Early History of Machine Learning », *in*: *IFAC-PapersOnLine*, 21st IFAC World Congress 53.*2* (2020), pp. 1385–1390, ISSN: 2405-8963, DOI: 10.1016/j.ifacol.2020.12.1888.

[101] Lee Friedman et al., « Reducing inter-scanner variability of activation in a multicenter fMRI study: Role of smoothness equalization », *in*: *NeuroImage* 32.*4* (2006), pp. 1656–1668, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2006.03.062.

[102] K. J. Friston, P. Jezzard, and R. Turner, « Analysis of functional MRI time-series », *in*: *Human Brain Mapping* 1.*2* (1994), pp. 153–171, ISSN: 1097-0193, DOI: 10.1002/hbm.460010207.

[103] K. J. Friston et al., « Event-Related fMRI: Characterizing Differential Responses », *in*: *NeuroImage* 7.*1* (1998), pp. 30–40, ISSN: 1053-8119, DOI: 10.1006/nimg.1997.0306.

[104] Karl J. Friston et al., « Movement-Related effects in fMRI time-series », *in*: *Magnetic Resonance in Medicine* 35.*3* (1996), Publisher: John Wiley & Sons, Ltd, pp. 346–355, ISSN: 0740-3194, DOI: 10.1002/mrm.1910350312.

[105] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge, « Image Style Transfer Using Convolutional Neural Networks », *in*: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2414–2423, DOI: 10.1109/CVPR.2016.265.

[106] Andrew Gelman and Eric Loken, « The garden of forking paths : Why multiple comparisons can be a problem, even when there is no fishing expedition or p-hacking and the research hypothesis was posited ahead of time », *in*: *Department of Statistics, Columbia University* (2019).

[107] Elodie Germani, *Docker image "open_pipeline"*, 2021, URL: \url{https://hub.docker.com/r/elodiegermani/open_pipeline}.

[108] Elodie Germani, *Image processing*, version elodiegermani/nguyen-etal-2021:latest nguyen- etal-2021:latest, 2023, DOI: 10.5281/zenodo.10298335, URL: https://doi.org/10.5281/zenodo.10298335.

[109] Elodie Germani, *Trainer*, version elodiegermani/nguyen-etal-2021:latest nguyen-etal-2021:latest, 2023, DOI: 10.5281/zenodo.10298359, URL: https://doi.org/10.5281/zenodo.10298359.

[110] Elodie Germani, Elisa Fromont, and Camille Maumet, « Supporting data for "On the benefits of self-taught learning for brain decoding" », *in*: *GigaScience Database* (2023), DOI: 10.5524/102377.

[111] Davide Giavarina, « Understanding Bland Altman analysis », *in*: *Biochemia Medica* 25.*2* (2015), pp. 141–151, ISSN: 18467482, DOI: 10.11613/BM.2015.015.

[112] Matthew F Glasser et al., « The Human Connectome Project's neuroimaging approach », *in*: *Nature Neuroscience* (2016), DOI: 10.1038/nn.4361.

[113] Tristan Glatard et al., *Boutiques: a flexible framework for automated application integration in computing platforms*, 2017, eprint: 1711.09713.

[114] Tristan Glatard et al., « Reproducibility of neuroimaging analyses across operating systems », *in*: *Frontiers in Neuroinformatics* 9 (2015), ISSN: 1662-5196, DOI: 10.3389/fninf.2015.00012.

[115] Dylan G. E. Gomes et al., « Why don't we share data and code? Perceived barriers and benefits to public archiving practices », *in*: *Proceedings of the Royal Society B: Biological Sciences* 289.*1987* (2022), p. 20221113, DOI: 10.1098/rspb.2022.1113.

[116] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, http://www.deeplearningbook.org, MIT Press, 2016.

[117] Ian Goodfellow et al., « Generative Adversarial Nets », *in*: *Advances in Neural Information Processing Systems*, ed. by Z. Ghahramani et al., vol. 27, Curran Associates, Inc., 2014.

[118] Krzysztof Gorgolewski, « Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python », *in*: *Frontiers in Neuroinformatics* (2017), p. 15, DOI: https://www.doi.org/10.5281/zenodo.581704.

[119] Krzysztof J Gorgolewski et al., « BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods », *in*: *PLoS computational biology* 13.*3* (2017), e1005209.

[120] Krzysztof J. Gorgolewski et al., « NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain », *in*: *Frontiers in Neuroinformatics* 9 (2015), ISSN: 1662-5196, DOI: 10.3389/fninf.2015.00008.

[121] Krzysztof J. Gorgolewski et al., « The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments », *in*: *Scientific Data* 3.*1* (2016), p. 160044, ISSN: 2052-4463, DOI: 10.1038/sdata.2016.44, URL: http://www.nature.com/articles/sdata201644.

[122] C. Goutte, F.A. Nielsen, and K.H. Hansen, « Modeling the hemodynamic response in fMRI using smooth FIR filters », *in*: *IEEE Transactions on Medical Imaging* 19.*12* (2000), pp. 1188–1201, ISSN: 1558-254X, DOI: 10.1109/42.897811.

[123] Ed H. B. M. Gronenschild et al., « The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements », *in*: *PLOS ONE* 7.*6* (2012), pp. 1–13, DOI: 10.1371/journal.pone.0038234.

[124] Katrina Gwinn et al., « Parkinson's disease biomarkers: perspective from the NINDS Parkinson's Disease Biomarkers Program », *in*: *Biomarkers in Medicine* 11.*6* (2017), pp. 451–473, ISSN: 1752-0363, DOI: 10.2217/bmm-2016-0370.

[125] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart, « Exploring Network Structure, Dynamics, and Function using NetworkX », *in*: *Proceedings of the 7th Python in Science Conference*, ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman, 2008, pp. 11 –15.

[126] Yaroslav Halchenko and Michael Hanke, « Open is Not Enough. Let's Take the Next Step: An Integrated, Community-Driven Computing Platform for Neuroscience », *in*: *Frontiers in Neuroinformatics* 6 (2012), DOI: https://www.doi.org/10.3389/fninf.2012.00022.

[127] Yaroslav Halchenko et al., *nipy/heudiconv: v1.1.1*, 2024, DOI: 10.5281/zenodo.11100373.

[128] Yaroslav O. Halchenko et al., « DataLad: distributed system for joint management of code, data, and their relationship », *in*: *Journal of Open Source Software* 6.*63* (2021), p. 3262, ISSN: 2475-9066, DOI: 10.21105/joss.03262.

[129] Daniel A. Handwerker, John M. Ollinger, and Mark D'Esposito, « Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses », *in*: *NeuroImage* 21.*4* (2004), pp. 1639–1651, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2003.11.029.

[130] Tom E. Hardwicke and John P. A. Ioannidis, « Populating the Data Ark: An attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles », *in*: *PLOS ONE* 13.*8* (2018), ed. by Jelte M. Wicherts, e0201856, ISSN: 1932-6203, DOI: 10.1371/journal.pone.0201856.

[131] Tom E. Hardwicke et al., « Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal Cognition », *in*: *Royal Society Open Science* 5.*8* (2018), p. 180448, DOI: 10.1098/rsos.180448.

[132] Charles R. Harris et al., « Array programming with NumPy », *in*: *Nature* 585 (2020), 357–362, DOI: 10.1038/s41586-020-2649-2.

[133] Satoru Hayasaka and Thomas E. Nichols, « Validating cluster size inference: random field and permutation methods », *in*: *NeuroImage* 20.*4* (2003), pp. 2343–2356, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2003.08.003.

[134] Simon S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, 1999, ISBN: 978-0-13-273350-2.

[135] Kaiming He et al., « Deep Residual Learning for Image Recognition », *in*: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, 2016, pp. 770–778, ISBN: 978-1-4673-8851-1, DOI: 10.1109/CVPR.2016.90.

[136] Kaiming He et al., « Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification », *in*: *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE Computer Society, 2015, pp. 1026–1034, ISBN: 978-1-4673-8391-2, DOI: 10.1109/ICCV.2015.123.

[137] Geoffrey E Hinton, « Deep belief networks », *in*: *Scholarpedia* 4.*5* (2009), p. 5947.

[138] Geoffrey E Hinton and Terrence J Sejnowski, « Optimal perceptual inference », *in*: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, vol. 448, 1983, pp. 448–453.

[139] Jonathan Ho, Ajay Jain, and Pieter Abbeel, « Denoising Diffusion Probabilistic Models », *in*: *Advances in Neural Information Processing Systems*, ed. by H. Larochelle et al., vol. 33, Curran Associates, Inc., 2020, pp. 6840–6851.

[140] Jonathan Ho and Tim Salimans, « Classifier-Free Diffusion Guidance », *in*: *"Deep Generative Models and Downstream Applications" Workshop@NeurIPS'21*, 2021.

[141] Sabine Hoffmann et al., « The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines », *in*: *Royal Society Open Science* 8.*4* (2021), p. 201925, DOI: 10.1098/rsos.201925.

[142] A. P. Holmes and K. J. Friston, « Generalisability, Random Effects & Population Inference », *in*: *NeuroImage* 7.*4, Part 2* (1998), S754, ISSN: 1053-8119, DOI: 10.1016/S1053-8119(18)31587-8.

[143] Sara Hooker, « Moving beyond "algorithmic bias is a data problem" », *in*: *Patterns* 2.*4* (2021), p. 100241, ISSN: 2666-3899, DOI: 10.1016/j.patter.2021.100241.

[144] Yanbing Hou and Huifang Shang, « Magnetic Resonance Imaging Markers for Cognitive Impairment in Parkinson's Disease: Current View », *in*: *Frontiers in Aging Neuroscience* 14 (2022), ISSN: 1663-4365, DOI: 10.3389/fnagi.2022.788846.

[145] YanBing Hou et al., « Prediction of individual clinical scores in patients with Parkinson's disease using resting-state functional magnetic resonance imaging », *in*: *Journal of the Neurological Sciences* 366 (2016), pp. 27–32, ISSN: 0022-510X, DOI: 10.1016/j.jns.2016.04.030.

[146] Jinlong Hu et al., « A Multichannel 2D Convolutional Neural Network Model for Task-Evoked fMRI Data Classification », *in*: *Computational Intelligence and Neuroscience* 2019 (2019), ed. by Laura Marzetti, p. 5065214, ISSN: 1687-5265, DOI: 10.1155/2019/5065214.

[147] Xiao-Fei Hu et al., « Amplitude of Low-frequency Oscillations in Parkinson's Disease: A 2-year Longitudinal Resting-state Functional Magnetic Resonance Imaging Study », *in*: *Chinese Medical Journal* 128.*05* (2015), pp. 593–601, DOI: 10.4103/0366-6999.151652.

[148] Xiaojie Huang, Jun Xiao, and Chao Wu, « Design of Deep Learning Model for Task-Evoked fMRI Data Classification », *in*: *Computational Intelligence and Neuroscience* 2021 (2021), ed. by António Dourado, pp. 1–10, ISSN: 1687-5273, 1687-5265, DOI: 10.1155/2021/6660866.

[149] S.A. Huettel, A.W. Song, and G. McCarthy, *Functional Magnetic Resonance Imaging*, Functional Magnetic Resonance Imaging vol. 1, Sinauer Associates, 2004, ISBN: 978-0-87893-288-7.

[150] *Human Connectome Project: Data Usage Agreement*, https://www.humanconnectome.org/study/hcp-young-adult/document/wu-minn-hcp-consortium-open-access-data-use-terms, 2013.

[151] Chloe Hutton et al., « Image Distortion Correction in fMRI: A Quantitative Evaluation », *in*: *NeuroImage* 16.*1* (2002), pp. 217–240, ISSN: 1053-8119, DOI: 10.1006/nimg.2001.1054.

[152] *ICA-AROMA & fmriprep using child template - fmriprep - Neurostars*, 2019, URL: https://neurostars.org/t/ica-aroma-fmriprep-using-child-template/5139 (visited on 10/17/2023).

[153] John P. A. Ioannidis, « Why Most Discovered True Associations Are Inflated: » *in*: *Epidemiology* 19.*5* (2008), pp. 640–648, ISSN: 1044-3983, DOI: 10.1097/EDE.0b013e31818131e7.

[154] John P. A. Ioannidis, « Why Most Published Research Findings Are False », *in*: *PLoS Medicine* 2.*8* (2005), e124, ISSN: 1549-1676, DOI: 10.1371/journal.pmed.0020124, (visited on 03/18/2024).

[155] John P. A. Ioannidis et al., « Increasing value and reducing waste in research design, conduct, and analysis », *in*: *Lancet (London, England)* 383.*9912* (2014), pp. 166–175, ISSN: 1474-547X, DOI: 10.1016/S0140-6736(13)62227-8.

[156] John Pa Ioannidis, « Effectiveness of antidepressants: an evidence myth constructed from a thousand randomized trials? », *in*: *Philosophy, Ethics, and Humanities in Medicine* 3.*1* (2008), p. 14, ISSN: 1747-5341, DOI: 10.1186/1747-5341-3-14.

[157] Sergey Ioffe and Christian Szegedy, « Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift », *in*: *CoRR* abs/1502.03167 (2015), DOI: 10.48550/arXiv.1502.03167.

[158] Phillip Isola et al., « Image-to-Image Translation with Conditional Adversarial Networks », *in*: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.

[159] Clifford R Jack Jr et al., « The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods », *in*: *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 27.*4* (2008), pp. 685–691, DOI: 10.1002/jmri.21049.

[160] Arman Jahanpour et al., *Neurobagel Query Tool: Web app for cohort searches across Neurobagel graphs.* 2023, DOI: 10.5281/zenodo.8088224, URL: https://doi.org/10.5281/zenodo.8088224.

[161] Mark Jenkinson et al., « FSL », *in*: *20 YEARS OF fMRI* 62.*2* (2012), pp. 782–790, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2011.09.015.

[162] Lan Jiang et al., « CoLa-Diff: Conditional Latent Diffusion Model for Multi-modal MRI Synthesis », *in*: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, ed. by Hayit Greenspan et al., Springer Nature Switzerland, 2023, pp. 398–408, ISBN: 978-3-031-43999-5, DOI: 10.1007/978-3-031-43999-5_38.

[163] Ziheng Jiang et al., « Characterizing Structural Regularities of Labeled Data in Overparameterized Models », *in*: *Proceedings of the 38th International Conference on Machine Learning*, PMLR, 2021, pp. 5034–5044.

[164] Cheng-Bin Jin et al., « Deep CT to MR Synthesis Using Paired and Unpaired Data », *in*: *Sensors* 19.*10* (2019), ISSN: 1424-8220, DOI: 10.3390/s19102361.

[165] Jakub Kaczmarzyk et al., *kaczmarj/neurodocker: Version 0.4.2*, 2018, DOI: 10.5281/zenodo.1477094, URL: https://doi.org/10.5281/zenodo.1477094.

[166] Shizuo Kaji and Satoshi Kida, « Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging », *in*: *Radiological Physics and Technology* 12.*3* (2019), pp. 235–248, ISSN: 1865-0341, DOI: 10.1007/s12194-019-00520-y.

[167] Sayash Kapoor and Arvind Narayanan, « Leakage and the reproducibility crisis in machine-learning-based science », *in*: *Patterns* 4.*9* (2023), p. 100804, ISSN: 2666-3899, DOI: https://doi.org/10.1016/j.patter.2023.100804.

[168] Davood Karimi et al., « Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis », *in*: *Medical Image Analysis* 65 (2020), p. 101759, ISSN: 1361-8415, DOI: 10.1016/j.media.2020.101759.

[169] Will Kay et al., *The Kinetics Human Action Video Dataset*, 2017, URL: http://arxiv.org/abs/1705.06950.

[170] Shahrzad Kharabian Masouleh et al., « Influence of Processing Pipeline on Cortical Thickness Measurement », *in*: *Cerebral Cortex* 30.*9* (2020), pp. 5014–5027, ISSN: 1047-3211, 1460-2199, DOI: 10.1093/cercor/bhaa097.

[171] Gregory Kiar et al., « Numerical uncertainty in analytical pipelines lead to impactful variability in brain networks », *in*: *PLOS ONE* 16.*11* (2021), pp. 1–16, DOI: 10.1371/journal.pone.0250755.

[172] Hee E. Kim et al., « Transfer learning for medical image classification: a literature review », *in*: *BMC Medical Imaging* 22.*1* (2022), p. 69, ISSN: 1471-2342, DOI: 10.1186/s12880-022-00793-7.

[173] Diederik P. Kingma and Jimmy Ba, *Adam: A Method for Stochastic Optimization*, 2017, arXiv: 1412.6980 [cs.LG].

[174] Diederik P Kingma and Max Welling, *Auto-Encoding Variational Bayes*, 2022, arXiv: 1312.6114 [stat.ML].

[175] Simon Klau et al., *Comparing the vibration of effects due to model, data preprocessing and sampling uncertainty on a large data set in personality psychology*, tech. rep., 2020, DOI: 10.5282/ubm/epub.70485.

[176] Arno Klein et al., « Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration », *in*: *NeuroImage* 46.*3* (2009), pp. 786–802, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2008.12.037.

[177] Thomas Kluyver et al., « Jupyter Notebooks – a publishing format for reproducible computational workflows », *in*: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, ed. by F. Loizides and B. Schmidt, IOS Press, 2016, pp. 87 –90.

[178] Lingke Kong et al., « Breaking the Dilemma of Medical Image-to-image Translation », *in*: *Advances in Neural Information Processing Systems*, ed. by A. Beygelzimer et al., 2021.

[179] Max Korbmacher, Lars T. Westlye, and Ivan I. Maximov, *FreeSurfer version-shuffling can boost brain age predictions*, 2024, DOI: 10.1101/2024.06.14.599070.

[180] Sotetsu Koyamada et al., *Deep learning of fMRI big data: a novel approach to subject-transfer decoding*, 2015, arXiv: 1502.00093 [stat.ML].

[181] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, « ImageNet Classification with Deep Convolutional Neural Networks », *in*: *Advances in Neural Information Processing Systems*, vol. 25, Curran Associates, Inc., 2012.

[182] Gregory M. Kurtzer, Vanessa Sochat, and Michael W. Bauer, « Singularity: Scientific containers for mobility of compute », *in*: *PLOS ONE* 12.*5* (2017), e0177459, ISSN: 1932-6203, DOI: 10.1371/journal.pone.0177459.

[183] Stephen LaConte et al., « The Evaluation of Preprocessing Choices in Single-Subject BOLD fMRI Using NPAIRS Performance Metrics », *in*: *NeuroImage* 18.*1* (2003), pp. 10–27, ISSN: 1053-8119, DOI: 10.1006/nimg.2002.1300.

[184] Nicholas Lange and Scott L. Zeger, « Non-linear Fourier Time Series Analysis for Human Brain Mapping by Functional Magnetic Resonance Imaging », *in*: *Journal of the Royal Statistical Society Series C: Applied Statistics* 46.*1* (2002), pp. 1–29, ISSN: 0035-9254, DOI: 10.1111/1467-9876.00046.

[185] Agostina J. Larrazabal et al., « Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis », *in*: *Proceedings of the National Academy of Sciences* 117.*23* (2020), pp. 12592–12594, DOI: 10.1073/pnas.1919012117.

[186] Anna Laurinavichyute, Himanshu Yadav, and Shravan Vasishth, « Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy », *in*: *Journal of Memory and Language* 125 (2022), p. 104332, ISSN: 0749-596X, DOI: 10.1016/j.jml.2022.104332, (visited on 03/19/2024).

[187] Maël Lebreton et al., « Assessing inter-individual differences with task-related functional neuroimaging », *in*: *Nature Human Behaviour* 3.*9* (2019), pp. 897–905, ISSN: 2397-3374, DOI: 10.1038/s41562-019-0681-8.

[188] Y. Lecun et al., « Gradient-based learning applied to document recognition », *in*: *Proceedings of the IEEE* 86.*11* (1998), pp. 2278–2324, DOI: 10.1109/5.726791.

[189] Yann Lecun, « Une procedure d'apprentissage pour reseau a seuil asymmetrique (A learning scheme for asymmetric threshold networks) », *in*: *Proceedings of Cognitiva 85, Paris, France* (1985), pp. 599–604.

[190] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, « Deep learning », *in*: *Nature* 521.*7553* (2015), pp. 436–444, ISSN: 1476-4687, DOI: 10.1038/nature14539.

[191] Yann LeCun et al., « Handwritten Digit Recognition with a Back-Propagation Network », *in*: *Advances in Neural Information Processing Systems*, vol. 2, Morgan-Kaufmann, 1989.

[192] Karim Lekadir et al., *FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging*, 2023, arXiv: 2109.09658 [cs.CV].

[193] Douglas B. Lenat et al., « Cyc: toward programs with common sense », *in*: *Communications of the ACM* 33.*8* (1990), pp. 30–49, ISSN: 0001-0782, DOI: 10.1145/79173.79176, (visited on 06/10/2024).

[194] Haoying Li et al., « SRDiff: Single image super-resolution with diffusion probabilistic models », *in*: *Neurocomputing* 479 (2022), pp. 47–59, ISSN: 0925-2312, DOI: 10.1016/j.neucom.2022.01.029.

[195] Xiangrui Li et al., « The first step for neuroimaging data analysis: DICOM to NIfTI conversion », *in*: *Journal of Neuroscience Methods* 264 (2016), pp. 47–56, ISSN: 0165-0270, DOI: 10.1016/j.jneumeth.2016.03.001, (visited on 03/21/2024).

[196] Xiaoxiao Li et al., « Multi-site fMRI Analysis Using Privacy-preserving Federated Learning and Domain Adaptation: ABIDE Results », *in*: *Medical image analysis* 65 (2020), p. 101765, ISSN: 1361-8415, DOI: 10.1016/j.media.2020.101765.

[197] Xinhui Li et al., « Learning pipeline-invariant representation for robust brain phenotype prediction », *in*: *Proceedings of the Data-centric Machine Learning Research (DMLR) Workshop at the 40 th International Conference on Machine Learning (ICML)*, Honolulu, Hawaii, USA, 2023.

[198] Xinhui Li et al., *Moving Beyond Processing and Analysis-Related Variation in Neuroscience*, 2021, DOI: 10.1101/2021.12.01.470790.

[199] Yanghao Li et al., « Demystifying Neural Style Transfer », *in*: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 2230–2236, DOI: 10.24963/ijcai.2017/310.

[200] Martin A. Lindquist et al., « Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling », *in*: *NeuroImage* 45.*1, Supplement 1* (2009), S187–S198, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2008.10.065.

[201] Geert Litjens et al., « A survey on deep learning in medical image analysis », *in*: *Medical Image Analysis* 42 (2017), pp. 60–88, ISSN: 1361-8415, DOI: 10.1016/j.media.2017.07.005.

[202] Mengting Liu et al., « Style Transfer Using Generative Adversarial Networks for Multi-site MRI Harmonization », *in*: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, ed. by Marleen de Bruijne et al., Cham: Springer International Publishing, 2021, pp. 313–322, ISBN: 978-3-030-87199-4, DOI: 10.1007/978-3-030-87199-4_30.

[203] Thomas T. Liu, « The Development of Event-Related fMRI Designs », *in*: *Neuroimage* 62.*2* (2012), pp. 1157–1162, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2011.10.008.

[204] Nikos K. Logothetis et al., « Neurophysiological investigation of the basis of the fMRI signal », *in*: *Nature* 412.*6843* (2001), pp. 150–157, ISSN: 1476-4687, DOI: 10.1038/35084005.

[205] Eric Loken and Andrew Gelman, « Measurement error and the replication crisis », *in*: *Science* 355.*6325* (2017), pp. 584–585, DOI: 10.1126/science.aal3618.

[206] Lingjuan Lyu et al., « Privacy and Robustness in Federated Learning: Attacks and Defenses », *in*: *IEEE Transactions on Neural Networks and Learning Systems* (2024), pp. 1–21, ISSN: 2162-237X, 2162-2388, DOI: 10.1109/TNNLS.2022.3216981.

[207] Qing Lyu and Ge Wang, *Conversion Between CT and MRI Images Using Diffusion and Score-Matching Models*, 2022, arXiv: 2209.12104 [eess.IV].

[208] Dennis Mackin et al., « Measuring Computed Tomography Scanner Variability of Radiomics Features », *in*: *Investigative Radiology* 50.*11* (2015), pp. 757–765, ISSN: 1536-0210, DOI: 10.1097/RLI.0000000000000180.

[209] J. MacQueen, « Some methods for classification and analysis of multivariate observations », *in*: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, vol. 5.1, 1967, pp. 281–298.

[210] Nahiyan Malik and Danilo Bzdok, « From YouTube to the brain: Transfer learning can improve brain-imaging predictions with deep learning », *in*: *Neural Networks* 153 (2022), pp. 325–338, ISSN: 0893-6080, DOI: 10.1016/j.neunet.2022.06.014.

[211] Kenneth Marek et al., « The Parkinson's progression markers initiative (PPMI) - establishing a PD biomarker cohort », eng, *in*: *Annals of clinical and translational neurology* 5.*12* (2018), pp. 1460–1477, ISSN: 2328-9503, DOI: 10.1002/acn3.644.

[212] Christopher J Markiewicz et al., « The OpenNeuro resource for sharing of neuroscience data », *in*: *eLife* 10 (2021), e71774, ISSN: 2050-084X, DOI: 10.7554/eLife.71774.

[213] Christos Matsoukas et al., « What Makes Transfer Learning Work for Medical Images: Feature Reuse & Other Factors », *in*: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2022, pp. 9215–9224, ISBN: 978-1-66546-946-3, DOI: 10.1109/CVPR52688.2022.00901.

[214] Arthur Mensch et al., « Extracting representations of cognition across neuroimaging studies improves brain decoding », *in*: *PLOS Computational Biology* (2014), ed. by Daniele Marinazzo, DOI: 10.1371/journal.pcbi.1008795.

[215] Romuald Menuet et al., « Comprehensive decoding mental processes from Web repositories of functional brain images », *in*: *Scientific Reports* 12.*1* (2022), p. 7050, DOI: 10.1038/s41598-022-10710-1.

[216] Dirk Merkel, « Docker: lightweight linux containers for consistent development and deployment », *in*: *Linux journal* 2014.*239* (2014), p. 2, DOI: 10.5555/2600239.2600241.

[217] Shun Miao, Z. Jane Wang, and Rui Liao, « A CNN Regression Approach for Real-Time 2D/3D Registration », *in*: *IEEE Transactions on Medical Imaging* 35.*5* (2016), pp. 1352–1363, ISSN: 1558-254X, DOI: 10.1109/TMI.2016.2521800.

[218] Karla L. Miller et al., « Multimodal population brain imaging in the UK Biobank prospective epidemiological study », *in*: *Nature Neuroscience* 19.*11* (2016), pp. 1523–1536, ISSN: 1546-1726, DOI: 10.1038/nn.4393.

[219] Marvin Minsky and Seymour Papert, *Perceptrons; an Introduction to Computational Geometry*, MIT Press, 1969, ISBN: 978-0-262-13043-1.

[220] Mehdi Mirza and Simon Osindero, *Conditional Generative Adversarial Nets*, 2014, arXiv: 1411.1784 [cs.LG].

[221] Trina Mitchell et al., « Emerging Neuroimaging Biomarkers Across Disease Stage in Parkinson Disease: A Review », *in*: *JAMA Neurology* 78.*10* (2021), pp. 1262–1272, ISSN: 2168-6149, DOI: 10.1001/jamaneurol.2021.1312.

[222] Steen Moeller et al., « Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI », *in*: *Magnetic Resonance in Medicine* 63.*5* (2010), pp. 1144–1153, DOI: https://doi.org/10.1002/mrm.22361.

[223] Sameera V. Mohd Sagheer and Sudhish N. George, « A review on medical image denoising algorithms », *in*: *Biomedical Signal Processing and Control* 61 (2020), p. 102036, ISSN: 1746-8094, DOI: 10.1016/j.bspc.2020.102036.

[224] Stefano Moia et al., « Proceedings of the OHBM Brainhack 2022 », *in*: *Aperture Neuro* 4 (2024), DOI: 10.52294/001c.92760.

[225] Alicja Moskal et al., « Artifact Detection on X-ray of Lung with COVID-19 Symptoms », *in*: *Information Technology in Biomedicine*, ed. by Ewa Pietka et al., Cham: Springer International Publishing, 2022, pp. 234–245, ISBN: 978-3-031-09135-3, DOI: 10.1007/978-3-031-09135-3_20.

[226] Antonio Mucherino, Petraq J. Papajorgji, and Panos M. Pardalos, « k-Nearest Neighbor Classification », *in*: *Data Mining in Agriculture*, Springer, 2009, pp. 83–106.

[227] J.A. Mumford and T. Nichols, « Modeling and inference of multisubject fMRI data », *in*: *IEEE Engineering in Medicine and Biology Magazine* 25.*2* (2006), pp. 42–51, ISSN: 1937-4186, DOI: 10.1109/MEMB.2006.1607668.

[228] Jeanette A. Mumford and Thomas Nichols, « Simple group fMRI modeling and inference », *in*: *NeuroImage* 47.*4* (2009), pp. 1469–1475, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2009.05.034.

[229] Jeanette A. Mumford et al., « The response time paradox in functional magnetic resonance imaging analyses », *in*: *Nature Human Behaviour* 8.*2* (2024), pp. 349–360, ISSN: 2397-3374, DOI: 10.1038/s41562-023-01760-0.

[230] M. R. Munafò, G. Stothart, and J. Flint, « Bias in genetic association studies and impact factor », *in*: *Molecular Psychiatry* 14.*2* (2009), pp. 119–120, ISSN: 1476-5578, DOI: 10.1038/mp.2008.77.

[231] Marcus R. Munafò, Angela S. Attwood, and Jonathan Flint, « Bias in genetic association studies: effects of research location and resources », *in*: *Psychological Medicine* 38.*8* (2008), pp. 1213–1214, ISSN: 0033-2917, DOI: 10.1017/S003329170800353X.

[232] Gustav Müller-Franzes et al., « A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis », *in*: *Scientific Reports* 13.*1* (2023), p. 12098, ISSN: 2045-2322, DOI: 10.1038/s41598-023-39278-0.

[233] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang, « What is being transferred in transfer learning? », *in*: *Advances in neural information processing systems* 33 (2020), pp. 512–523.

[234] Kevin P. Nguyen et al., « Predicting Parkinson's disease trajectory using clinical and neuroimaging baseline measures », *in*: *Parkinsonism & Related Disorders* 85 (2021), pp. 44–51, ISSN: 1353-8020, DOI: 10.1016/j.parkreldis.2021.02.026.

[235] Alexander Quinn Nichol and Prafulla Dhariwal, « Improved Denoising Diffusion Probabilistic Models », *in*: *38th International Conference on Machine Learning*, PMLR, 2021, pp. 8162–8171.

[236] Thomas Nichols, *SPM plot units*, 2012, URL: https://web.archive.org/web/20230606094719/https://blog.nisox.org/2012/07/31/spm-plot-units (visited on 2012).

[237] Thomas E Nichols et al., « Best practices in data analysis and sharing in neuroimaging using MRI », *in*: *Nature Neuroscience* 20.*3* (2017), pp. 299–303, ISSN: 1097-6256, 1546-1726, DOI: 10.1038/nn.4500.

[238] Dong Nie et al., « Medical Image Synthesis with Deep Convolutional Adversarial Networks », *in*: *IEEE Transactions on Biomedical Engineering* 65.*12* (2018), pp. 2720–2730, ISSN: 1558-2531, DOI: 10.1109/TBME.2018.2814538.

[239] Guy Nir et al., « Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts », *in*: *Medical Image Analysis* 50 (2018), pp. 167–180, ISSN: 1361-8415, DOI: 10.1016/j.media.2018.09.005.

[240] Guiomar Niso et al., « Open and reproducible neuroimaging: From study inception to publication », *in*: *NeuroImage* 263 (2022), p. 119623, ISSN: 1095-9572, DOI: 10.1016/j.neuroimage.2022.119623.

[241] Stephanie Noble, Dustin Scheinost, and R. Todd Constable, « A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis », *in*: *NeuroImage* 203 (2019), p. 116157, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2019.116157.

[242] Martin Nørgaard et al., « Different preprocessing strategies lead to different conclusions: A [^{11}C]DASB-PET reproducibility study », *in*: *Journal of Cerebral Blood Flow & Metabolism* 40.*9* (2020), pp. 1902–1911, ISSN: 0271-678X, 1559-7016, DOI: 10.1177/0271678X19880450.

[243] Luke Oakden-Rayner et al., « Hidden stratification causes clinically meaningful failures in machine learning for medical imaging », *in*: *Proceedings of the ACM Conference on Health, Inference, and Learning*, Association for Computing Machinery, 2020, pp. 151–159, ISBN: 978-1-4503-7046-2, DOI: 10.1145/3368555.3384468.

[244] T. R. Oakes et al., « Comparison of fMRI motion correction software tools », *in*: *NeuroImage* 28.*3* (2005), pp. 529–543, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2005.05.058.

[245] Augustus Odena, Christopher Olah, and Jonathon Shlens, « Conditional Image Synthesis with Auxiliary Classifier GANs », *in*: *Proceedings of the 34th International Conference on Machine Learning*, ed. by Doina Precup and Yee Whye Teh, vol. 70, Proceedings of Machine Learning Research, PMLR, 2017, pp. 2642–2651.

[246] Kanghan Oh et al., « Classification of schizophrenia and normal controls using 3D convolutional neural network and outcome visualization », *in*: *Schizophrenia Research* 212 (2019), pp. 186–195, DOI: 10.1016/j.schres.2019.07.034.

[247] Wiktor Olszowy et al., « Accurate autocorrelation modeling substantially improves fMRI reliability », *in*: *Nature Communications* 10 (2019), p. 1220, ISSN: 2041-1723, DOI: 10.1038/s41467-019-09230-w.

[248] Open Science Collaboration, « Estimating the reproducibility of psychological science », *in*: *Science* 349.*6251* (2015), aac4716–aac4716, ISSN: 0036-8075, 1095-9203, DOI: 10.1126/science.aac4716.

[249] Seyedmehdi Orouji et al., *"Task-relevant autoencoding" enhances machine learning for human neuroscience*, 2023, arXiv: 2208.08478 [q-bio.NC].

[250] Muzaffer Ozbey et al., « Unsupervised Medical Image Translation With Adversarial Diffusion Models », *in*: *IEEE transactions on medical imaging* 42.*12* (2023), pp. 3524–3539, ISSN: 1558-254X, DOI: 10.1109/tmi.2023.3290149.

[251] L. Palumbo et al., « Evaluation of the intra- and inter-method agreement of brain MRI segmentation software packages: A comparison between SPM12 and FreeSurfer v6.0 », *in*: *Physica Medica* 64 (2019), pp. 261–272, ISSN: 1120-1797, DOI: 10.1016/j.ejmp.2019.07.016.

[252] Shaoyan Pan et al., *Cycle-guided Denoising Diffusion Probability Model for 3D Cross-modality MRI Synthesis*, 2023, arXiv: 2305.00042 [eess.IV].

[253] Sinno Jialin Pan and Qiang Yang, « A Survey on Transfer Learning », *in*: *IEEE Transactions on Knowledge and Data Engineering* 22.*10* (2010), pp. 1345–1359, ISSN: 1558-2191, DOI: 10.1109/TKDE.2009.191.

[254] HuiZe Pang et al., « Use of machine learning method on automatic classification of motor subtype of Parkinson's disease based on multilevel indices of rs-fMRI », *in*: *Parkinsonism & Related Disorders* 90 (2021), pp. 65–72, ISSN: 1353-8020, DOI: 10.1016/j.parkreldis.2021.08.003.

[255] David B. Parker and Qolamreza R. Razlighi, « The Benefit of Slice Timing Correction in Common fMRI Preprocessing Pipelines », *in*: *Frontiers in Neuroscience* 13 (2019), ISSN: 1662-453X, DOI: 10.3389/fnins.2019.00821.

[256] Linden Parkes et al., « An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI », *in*: *NeuroImage* 171 (2018), pp. 415–436, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2017.12.073.

[257] Lance Parsons, Ehtesham Haque, and Huan Liu, « Subspace clustering for high dimensional data: a review », *in*: *ACM SIGKDD Explorations Newsletter* 6.*1* (2004), pp. 90–105, ISSN: 1931-0145, DOI: 10.1145/1007730.1007731, URL: https://doi.org/10.1145/1007730.1007731.

[258] Adam Paszke et al., « PyTorch: An Imperative Style, High-Performance Deep Learning Library », *in*: *Advances in Neural Information Processing Systems 32* (2019), pp. 8024–8035, DOI: 10.48550/arXiv.1912.01703.

[259] W.D. Penny et al., *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, Elsevier, 2011.

[260] Jeffrey M. Perkel, « Challenge to scientists: does your ten-year-old code still run? », *in*: *Nature* 584.*7822* (2020), pp. 656–658, DOI: 10.1038/d41586-020-02462-7.

[261] Steven E. Petersen and Joseph W. Dubis, « The mixed block/event-related design », *in*: *NeuroImage* 62.*2* (2012), pp. 1177–1184, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2011.09.084.

[262] Russell A Poldrack and Krzysztof J Gorgolewski, « Making big data open: data sharing in neuroimaging », *in*: *Nature Neuroscience* 17.*11* (2014), pp. 1510–1517, ISSN: 1546-1726, DOI: 10.1038/nn.3818.

[263] Russell A. Poldrack, Jeanette A. Mumford, and Thomas E. Nichols, *Handbook of Functional MRI Data Analysis*, 1st ed., Cambridge University Press, 2011, DOI: 10.1017/CBO9780511895029.

[264] Russell A. Poldrack et al., « Scanning the horizon: towards transparent and reproducible neuroimaging research », *in*: *Nature Reviews Neuroscience* (2017), DOI: 10.1038/nrn.2016.167.

[265] Russell A. Poldrack et al., « The Cognitive Atlas: Toward a Knowledge Foundation for Cognitive Neuroscience », *in*: *Frontiers in Neuroinformatics* 5 (2011), ISSN: 1662-5196, DOI: 10.3389/fninf.2011.00017.

[266] Jean-Baptiste Poline et al., « Data sharing in neuroimaging research », *in*: *Frontiers in Neuroinformatics* 6 (2012), ISSN: 1662-5196, DOI: 10.3389/fninf.2012.00009.

[267] Jonathan D. Power et al., « Methods to detect, characterize, and remove motion artifact in resting state fMRI », *in*: *NeuroImage* 84 (2014), pp. 320–341, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2013.08.048.

[268] Jonathan D. Power et al., « Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion », *in*: *NeuroImage* 59.*3* (2012), pp. 2142–2154, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2011.10.018.

[269] Jonathan D. Power et al., « Steps toward optimizing motion artifact removal in functional connectivity MRI; a reply to Carp », *in*: *NeuroImage* 76 (2013), pp. 439–441, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2012.03.017.

[270] Konpat Preechakul et al., « Diffusion Autoencoders: Toward a Meaningful and Decodable Representation », *in*: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10609–10619, DOI: 10.1109/CVPR52688.2022.01036.

[271] Florian Prinz, Thomas Schlange, and Khusru Asadullah, « Believe it or not: how much can we rely on published data on potential drug targets? », *in*: *Nature Reviews Drug Discovery* 10.*9* (2011), pp. 712–712, ISSN: 1474-1784, DOI: 10.1038/nrd3439-c1.

[272] Raimon H.R. Pruim et al., « ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data », *in*: *NeuroImage* 112 (2015), pp. 267–277, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2015.02.064.

[273] *Public nEUro*, 2020, URL: https://public-neuro.github.io/index.html (visited on 12/20/2023).

[274] Zhiwei Qin et al., « Style transfer in conditional GANs for cross-modality synthesis of brain magnetic resonance images », *in*: *Computers in Biology and Medicine* 148 (2022), p. 105928, ISSN: 0010-4825, DOI: 10.1016/j.compbiomed.2022.105928.

[275] J. Rademacher et al., « Topographical Variation of the Human Primary Cortices: Implications for Neuroimaging, Brain Mapping, and Neurobiology », *in*: *Cerebral Cortex* 3.*4* (1993), pp. 313–329, ISSN: 1047-3211, DOI: 10.1093/cercor/3.4.313.

[276] Maithra Raghu et al., « Transfusion: Understanding Transfer Learning for Medical Imaging », *in*: *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

[277] Rajat Raina et al., « Self-taught learning: transfer learning from unlabeled data », *in*: *Proceedings of the 24th international conference on Machine learning*, Corvalis Oregon USA: ACM, 2007, pp. 759–766, ISBN: 978-1-59593-793-3, DOI: 10.1145/1273496.1273592.

[278] Limor Raviv, Gary Lupyan, and Shawn C. Green, « How variability shapes learning and generalization », *in*: *Trends in Cognitive Sciences* 26.*6* (2022), pp. 462–483, ISSN: 1364-6613, DOI: 10.1016/j.tics.2022.03.007.

[279] Muhammad Habib ur Rehman et al., « Federated learning for medical imaging radiology », *in*: *The British Journal of Radiology* 96.*1150* (2023), p. 20220890, ISSN: 0007-1285, DOI: 10.1259/bjr.20220890.

[280] María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante, « Addressing fairness in artificial intelligence for medical imaging », *in*: *Nature Communications* 13.*1* (2022), p. 4581, ISSN: 2041-1723, DOI: 10.1038/s41467-022-32186-3.

[281] Nicola Rieke et al., « The future of digital health with federated learning », *in*: *npj Digital Medicine* 3.*1* (2020), pp. 1–7, ISSN: 2398-6352, DOI: 10.1038/s41746-020-00323-1.

[282] Xavier Rolland, « Impact of analytical variability on data compatibility in functional Magnetic Resonance Imaging studies », PhD thesis, 2022.

[283] Xavier Rolland, Pierre Maurel, and Camille Maumet, « Towards efficient fmri data re-use: can we run between-group analyses with datasets processed differently with spm ? », *in*: *ISBI 2022 - IEEE International Symposium on Biomedical Imaging*, 2022, pp. 1–4.

[284] Edmund T. Rolls et al., « Automated anatomical labelling atlas 3 », *in*: *NeuroImage* 206 (2020), p. 116189, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2019.116189, URL: https://www.sciencedirect.com/science/article/pii/S1053811919307803.

[285] Robin Rombach et al., « High-Resolution Image Synthesis with Latent Diffusion Models », *in*: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2022, pp. 10674–10685, ISBN: 978-1-66546-946-3, DOI: 10.1109/CVPR52688.2022.01042.

[286] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, « U-Net: Convolutional Networks for Biomedical Image Segmentation », *in*: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ed. by Nassir Navab et al., Cham: Springer International Publishing, 2015, pp. 234–241, ISBN: 978-3-319-24574-4, DOI: 10.1007/978-3-319-24574-4_28.

[287] F. Rosenblatt, « The perceptron: A probabilistic model for information storage and organization in the brain », *in*: *Psychological Review* 65.*6* (1958), pp. 386–408, ISSN: 1939-1471, DOI: 10.1037/h0042519.

[288] Alaleh Sadraee, Martin Paulus, and Hamed Ekhtiari, « fMRI as an outcome measure in clinical trials: A systematic review in clinicaltrials.gov », *in*: *Brain and Behavior* 11.*5* (2021), e02089, ISSN: 2162-3279, DOI: 10.1002/brb3.2089.

[289] Chitwan Saharia et al., « Palette: Image-to-Image Diffusion Models », *in*: *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, New York, NY, USA: Association for Computing Machinery, 2022, pp. 1–10, ISBN: 978-1-4503-9337-9, DOI: 10.1145/3528233.3530757.

[290] Zohaib Salahuddin et al., « Transparency of deep neural networks for medical image analysis: A review of interpretability methods », *in*: *Computers in Biology and Medicine* 140 (2022), p. 105111, ISSN: 0010-4825, DOI: 10.1016/j.compbiomed.2021.105111.

[291] Tim Salimans et al., « Improved Techniques for Training GANs », *in*: *Advances in Neural Information Processing Systems*, ed. by D. Lee et al., vol. 29, Curran Associates, Inc., 2016.

[292] Gholamreza Salimi-Khorshidi et al., « Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies », *in*: *Neuroimage* 45.*3* (2009), pp. 810–823, DOI: 10.1016/j.neuroimage.2008.12.039.

[293] Jacob Sanz-Robinson et al., « NeuroCI: Continuous Integration of Neuroimaging Results Across Software Pipelines and Datasets », *in*: *2022 IEEE 18th Interna-*

tional Conference on e-Science (e-Science), 2022, pp. 105–116, DOI: 10.1109/eScience55777.2022.00025.

[294] Hiroshi Sasaki, Chris G. Willcocks, and Toby P. Breckon, *UNIT-DDPM: UNpaired Image Translation with Denoising Diffusion Probabilistic Models*, 2021, arXiv: 2104.05358 [cs.CV].

[295] Theodore D. Satterthwaite et al., « An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data », *in*: *NeuroImage* 64 (2013), pp. 240–256, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2012.08.052.

[296] Alexander Schaefer et al., « Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI », *in*: *Cerebral Cortex* 28.*9* (Sept. 2018), pp. 3095–3114, ISSN: 1047-3211, DOI: 10.1093/cercor/bhx179, (visited on 10/18/2023).

[297] GD. Schott, « Penfield's homunculus: a note on cerebral cartography », *in*: *J Neurol Neurosurg Psychiatry* 56.*4* (1993), pp. 329–333, DOI: 10.1136/jnnp.56.4.329..

[298] Kawin Setsompop et al., « Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty », *in*: *Magnetic Resonance in Medicine* 67.*5* (2012), pp. 1210–1224, DOI: https://doi.org/10.1002/mrm.23097.

[299] M. Tarek Shaban et al., « Staingan: Stain Style Transfer for Digital Histological Images », *in*: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019, pp. 953–956, DOI: 10.1109/ISBI.2019.8759152.

[300] Wei Shen et al., « Multi-scale Convolutional Neural Networks for Lung Nodule Classification », *in*: *Information Processing in Medical Imaging*, ed. by Sebastien Ourselin et al., Cham: Springer International Publishing, 2015, pp. 588–599, ISBN: 978-3-319-19992-4, DOI: 10.1007/978-3-319-19992-4_46.

[301] R. Silberzahn et al., « Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results », *in*: *Advances in Methods and Practices in Psychological Science* 1.*3* (2018), pp. 337–356, ISSN: 2515-2459, 2515-2467, DOI: 10.1177/2515245917747646.

[302] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn, « False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant », *in*: *Psychological Science* 22.*11* (2011), pp. 1359–1366, ISSN: 0956-7976, DOI: 10.1177/0956797611417632.

[303] K. Simonyan and A. Zisserman, « Very deep convolutional networks for large-scale image recognition », *in*: *3rd International Conference on Learning Representations (ICLR 2015)* (2015), (visited on 04/25/2024).

[304] Paul Smolensky, « Information processing in dynamical systems: Foundations of harmony theory », *in*: *Parallel Distributed Process* 1 (1986).

[305] Jiaming Song, Chenlin Meng, and Stefano Ermon, « Denoising Diffusion Implicit Models », *in*: *International Conference on Learning Representations*, 2021.

[306] Sara Steegen et al., « Increasing Transparency Through a Multiverse Analysis », *in*: *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 11.*5* (2016), pp. 702–712, ISSN: 1745-6924, DOI: 10.1177/1745691616658637.

[307] Stephen Strother et al., « Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis », *in*: *NeuroImage*, Mathematics in Brain Imaging 23 (2004), S196–S207, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2004.07.022.

[308] Stephen C. Strother et al., « The Quantitative Evaluation of Functional Neuroimaging Experiments: The NPAIRS Data Analysis Framework », *in*: *NeuroImage* 15.*4* (2002), pp. 747–771, ISSN: 1053-8119, DOI: 10.1006/nimg.2001.1034.

[309] Cathie Sudlow et al., « UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age », *in*: *PLoS Medicine* 12.*3* (2015), e1001779, ISSN: 1549-1277, DOI: 10.1371/journal.pmed.1001779.

[310] Roger Sun, Eric Deutsch, and Laure Fournier, « Intelligence artificielle et imagerie médicale », *in*: *Bulletin du Cancer* 109.*1* (2022), pp. 83–88, ISSN: 0007-4551, DOI: 10.1016/j.bulcan.2021.09.009.

[311] Michele Svanera et al., « Transfer learning of deep neural network representations for fMRI decoding », *in*: *Journal of Neuroscience Methods* (2019), DOI: 10.1016/j.jneumeth.2019.108319.

[312] Christian Szegedy et al., « Going deeper with convolutions », *in*: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA: IEEE, 2015, pp. 1–9, ISBN: 978-1-4673-6964-0, DOI: 10.1109/CVPR.2015.7298594.

[313] Haiming Tang, Nanfei Sun, and Steven Shen, « Improving Generalization of Deep Learning Models for Diagnostic Pathology by Increasing Variability in Training Data: Experiments on Osteosarcoma Subtypes », *in*: *Journal of Pathology Informatics* 12 (2021), p. 30, ISSN: 2229-5089, DOI: 10.4103/jpi.jpi_78_20.

[314] Leho Tedersoo et al., « Data sharing practices and data availability upon request differ across scientific disciplines », *in*: *Scientific Data* 8.*1* (2021), p. 192, ISSN: 2052-4463, DOI: 10.1038/s41597-021-00981-0.

[315] Merle Temme, « Algorithms and Transparency in View of the New General Data Protection Regulation », *in*: *European Data Protection Law Review* 3.*4* (2017), pp. 473–485, ISSN: 2364284X, DOI: 10.21552/edpl/2017/4/9.

[316] Thijs Kooi, *Deep learning: From natural to medical images*, 2018, URL: https://medium.com/merantix/deep-learning-from-natural-to-medical-images-74827bf51d6b.

[317] Armin Thomas, Christopher Ré, and Russell Poldrack, « Self-Supervised Learning of Brain Dynamics from Broad Neuroimaging Data », *in*: *Advances in Neural Information Processing Systems* 35 (2022), pp. 21255–21269.

[318] Armin W. Thomas, Christopher Ré, and Russell A. Poldrack, *Challenges for cognitive decoding using deep learning methods*, 2021, arXiv: 2108.06896 [cs.LG].

[319] Armin W. Thomas et al., « Evaluating deep transfer learning for whole-brain cognitive decoding », *in*: *Journal of the Franklin Institute* 360.*13* (2023), pp. 9754–9787, ISSN: 0016-0032.

[320] Paul M. Thompson et al., « Three-Dimensional Statistical Analysis of Sulcal Variability in the Human Brain », *in*: *Journal of Neuroscience* 16.*13* (1996), pp. 4261–4274, ISSN: 0270-6474, 1529-2401, DOI: 10.1523/JNEUROSCI.16-13-04261.1996.

[321] Zhigang Tu et al., « Multi-stream CNN: Learning representations based on human-related regions for action recognition », *in*: *Pattern Recognition* 79 (2018), pp. 32–43, ISSN: 00313203.

[322] Andri C. Tziortzi et al., « Connectivity-Based Functional Analysis of Dopamine Release in the Striatum Using Diffusion-Weighted MRI and Positron Emission Tomography », *in*: *Cerebral Cortex* 24.*5* (2014), pp. 1165–1177, ISSN: 1047-3211, DOI: 10.1093/cercor/bhs397.

[323] Seyed Abolfazl Valizadeh et al., « Identification of individual subjects on the basis of their brain anatomical features », *in*: *Scientific Reports* 8.*1* (2018), p. 5611, ISSN: 2045-2322, DOI: 10.1016/j.patcog.2018.01.020.

[324] David C. Van Essen et al., « The WU-Minn Human Connectome Project: An overview », *in*: *NeuroImage* 80 (2013), pp. 62–79, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2013.05.041.

[325] Quentin Vanderbecq et al., « Comparison and validation of seven white matter hyperintensities segmentation software in elderly patients », *in*: *NeuroImage: Clinical* 27 (2020), p. 102357, ISSN: 2213-1582, DOI: 10.1016/j.nicl.2020.102357.

[326] Gael Varoquaux and Olivier Colliot, « Evaluating Machine Learning Models and Their Diagnostic Value », *in*: *Machine Learning for Brain Disorders*, Neuromethods, Springer US, 2023, pp. 601–630, ISBN: 978-1-07-163195-9, DOI: 10.1007/978-1-0716-3195-9_20.

[327] Gaël Varoquaux and Veronika Cheplygina, « Machine learning for medical imaging: methodological failures and recommendations for the future », *in*: *npj Digital Medicine* (2022), DOI: 10.1038/s41746-022-00592-y.

[328] Gaël Varoquaux et al., « Atlases of cognition with large-scale human brain mapping », *in*: *PLOS Computational Biology* (2018), DOI: 10.1371/journal.pcbi.1006565.

[329] Hanna V Vesterinen et al., « Systematic Survey of the Design, Statistical Analysis, and Reporting of Studies Published in the 2008 Volume of the Journal of Cerebral Blood Flow and Metabolism », *in*: *Journal of Cerebral Blood Flow & Metabolism* 31.*4* (2011), pp. 1064–1072, ISSN: 0271-678X, DOI: 10.1038/jcbfm.2010.217.

[330] Gaël Vila et al., « The Impact of Hardware Variability on Applications Packaged with Docker and Guix: a Case Study in Neuroimaging », 2024.

[331] Pascal Vincent et al., « Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion », *in*: *Journal of Machine Learning Research* 11.*110* (2010), pp. 3371–3408, ISSN: 1533-7928, DOI: 10.5555/1756006.1953039.

[332] Hanh Vu, Hyun-Chul Kim, and Jong-Hwan Lee, « 3D convolutional neural network for feature extraction and classification of fMRI volumes », *in*: *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 2018, pp. 1–4, DOI: 10.1109/PRNI.2018.8423964.

[333] Hanh Vu et al., « fMRI volume classification using a 3D convolutional neural network robust to shifted and scaled neuronal activations », *in*: *NeuroImage* (2020), DOI: 10.1016/j.neuroimage.2020.117328.

[334] Christian Wachinger et al., « Detect and correct bias in multi-site neuroimaging datasets », *in*: *Medical Image Analysis* 67 (2021), p. 101879, ISSN: 1361-8423, DOI: 10.1016/j.media.2020.101879.

[335] Jacques Wainer and Gavin Cawley, *Nested cross-validation when selecting classifiers is overzealous for most practical applications*, 2018, URL: http://arxiv.org/abs/1809.09446.

[336] Jonathon Walters et al., « Predicting brain activation maps for arbitrary tasks with cognitive encoding models », *in*: *NeuroImage* 263 (2022), p. 119610, DOI: 10.1016/j.neuroimage.2022.119610.

[337] Hua Wang, Feiping Nie, and Heng Huang, « Robust and discriminative self-taught learning », *in*: *Proceedings of the 30th International Conference on Machine Learning* (2013).

[338] Risheng Wang et al., « Medical image segmentation using deep learning: A survey », *in*: *IET Image Processing 5* (2022), pp. 1243–1267, ISSN: 1751-9667, DOI: 10.1049/ipr2.12419.

[339] Xiaoxiao Wang et al., « Decoding and mapping task states of the human brain via deep learning », *in*: *Human Brain Mapping* (2020), DOI: 10.1002/hbm.24891.

[340] Xin-Di Wang, Chao-Gan Yan, and Yu-Feng Zang, « Linear trend of resting-state fMRI time series », 2014.

[341] Xuan Wang et al., « A Review of GAN-Based Super-Resolution Reconstruction for Optical Remote Sensing Images », *in*: *Remote Sensing* 15.*20* (2023), p. 5062, ISSN: 2072-4292, DOI: 10.3390/rs15205062.

[342] Zheng Wang, Yangqiu Song, and Changshui Zhang, « Transferred dimensionality reduction », *in*: *Proceedings of the 2008th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*, ECMLPKDD'08, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 550–565, ISBN: 978-3-540-87480-5.

[343] Jason D. Warren et al., « Molecular nexopathies: a new paradigm of neurodegenerative disease », *in*: *Trends in Neurosciences* 36.*10* (2013), pp. 561–569, ISSN: 0166-2236, DOI: 10.1016/j.tins.2013.06.007.

[344] Joseph P. Weir, « Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM », *in*: *Journal of Strength and Conditioning Research* 19.*1* (2005), pp. 231–240, ISSN: 1064-8011, DOI: 10.1519/15184.1.

[345] Andreas Weissenbacher et al., « Correlations and anticorrelations in resting-state functional connectivity MRI: A quantitative comparison of preprocessing strategies », *in*: *NeuroImage* 47.*4* (2009), pp. 1408–1416, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2009.05.005.

[346] Junhao Wen et al., « Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation », *in*: *Medical Image Analysis* 63 (2020), p. 101694, ISSN: 1361-8415, DOI: 10.1016/j.media.2020.101694, URL: https://www.sciencedirect.com/science/article/pii/S1361841520300591.

[347] Mark D. Wilkinson et al., « The FAIR Guiding Principles for scientific data management and stewardship », *in*: *Scientific Data* 3.*1* (2016), p. 160018, ISSN: 2052-4463, DOI: 10.1038/sdata.2016.18.

[348] Martin J. Willemink et al., « Preparing Medical Imaging Data for Machine Learning », *in*: *Radiology* 295 (2020), pp. 4–15, ISSN: 0033-8419, DOI: 10.1148/radiol.2020192224.

[349] Mandy Melissa Jane Wittens et al., « Inter- and Intra-Scanner Variability of Automated Brain Volumetry on Three Magnetic Resonance Imaging Systems in Alzheimer's Disease and Controls », *in*: *Frontiers in Aging Neuroscience* 13 (2021), ISSN: 1663-4365, DOI: 10.3389/fnagi.2021.746982.

[350] Jelmer M. Wolterink et al., « Deep MR to CT Synthesis Using Unpaired Data », *in*: *Simulation and Synthesis in Medical Imaging*, ed. by Sotirios A. Tsaftaris et al., Cham: Springer International Publishing, 2017, pp. 14–23, ISBN: 978-3-319-68127-6, DOI: 10.1007/978-3-319-68127-6_2.

[351] Jelmer M. Wolterink et al., « Generative Adversarial Networks for Noise Reduction in Low-Dose CT », *in*: *IEEE Transactions on Medical Imaging* 36.*12* (2017), pp. 2536–2545, ISSN: 1558-254X, DOI: 10.1109/TMI.2017.2708987.

[352] Mark W. Woolrich, Timothy E. J. Behrens, and Stephen M. Smith, « Constrained linear basis sets for HRF modelling using Variational Bayes », *in*: *NeuroImage* 21.*4* (2004), pp. 1748–1761, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2003.12.024.

[353] Mark W. Woolrich et al., « Temporal Autocorrelation in Univariate Linear Modeling of FMRI Data », *in*: *NeuroImage* 14.*6* (2001), pp. 1370–1386, DOI: https://doi.org/10.1006/nimg.2001.0931.

[354] K. J. Worsley et al., « A General Statistical Analysis for fMRI Data », *in*: *NeuroImage* 15.*1* (2002), pp. 1–15, ISSN: 1053-8119, DOI: 10.1006/nimg.2001.0933.

[355] Tao Wu et al., « Regional homogeneity changes in patients with Parkinson's disease », *in*: *Human Brain Mapping* 30.*5* (2009), pp. 1502–1510, ISSN: 1097-0193, DOI: 10.1002/hbm.20622.

[356] Ullrich Wüllner et al., « The heterogeneity of Parkinson's disease », *in*: *Journal of Neural Transmission* 130.*6* (2023), pp. 827–838, ISSN: 0300-9564, DOI: 10.1007/s00702-023-02635-4.

[357] Junqian Xu et al., « Highly accelerated whole brain imaging using aligned-blipped-controlled-aliasing multiband EPI », *in*: *Proceedings of the 20th Annual Meeting of ISMRM*, vol. 2306, 2012, pp. 1907–1913.

[358] Ting Xu et al., « ReX: an integrative tool for quantifying and optimizing measurement reliability for the study of individual differences », *in*: *Nature Methods* 20.*7* (2023), pp. 1025–1028, ISSN: 1548-7105, DOI: 10.1038/s41592-023-01901-3.

[359] Y. Gao et al., « Decoding Behavior Tasks From Brain Activity Using Deep Transfer Learning », *in*: *IEEE Access* 7 (2019), pp. 43222–43232, ISSN: 2169-3536, DOI: 10.1109/ACCESS.2019.2907040.

[360] Chao-Gan Yan et al., « A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics », *in*: *NeuroImage* 76 (2013), pp. 183–201, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2013.03.004.

[361] Chao-Gan Yan et al., « Addressing head motion dependencies for small-world topologies in functional connectomics », *in*: *Frontiers in human neuroscience* 7 (2013), p. 910, ISSN: 1662-5161, DOI: 10.3389/fnhum.2013.00910.

[362] Qianye Yang et al., « MRI Cross-Modality Image-to-Image Translation », *in*: *Scientific Reports* 10.*1* (2020), p. 3753, ISSN: 2045-2322, DOI: 10.1038/s41598-020-60520-6.

[363] Qingsong Yang et al., « Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss », *in*: *IEEE Transactions on Medical Imaging* 37.*6* (2018), pp. 1348–1357, ISSN: 1558-254X, DOI: 10.1109/TMI.2018.2827462.

[364] Ruixin Yang and Yingyan Yu, « Artificial Convolutional Neural Network in Object Detection and Semantic Segmentation for Medical Imaging Analysis », *in*: *Frontiers in Oncology* 11 (2021), ISSN: 2234-943X, DOI: 10.3389/fonc.2021.638182.

[365] Xiao Yang, Roland Kwitt, and Marc Niethammer, « Fast Predictive Image Registration », *in*: *Deep Learning and Data Labeling for Medical Applications*, ed. by Gustavo Carneiro et al., Springer International Publishing, 2016, pp. 48–57, ISBN: 978-3-319-46976-8, DOI: 10.1007/978-3-319-46976-8_6.

[366] Tal Yarkoni et al., « Large-scale automated synthesis of human functional neuroimaging data », *in*: *Nature Methods* 8.*8* (2011), pp. 665–670, DOI: 10.1038/nmeth.1635.

[367] Wutao Yin, Longhai Li, and Fang-Xiang Wu, « Deep learning for brain disorder diagnosis based on fMRI images », *in*: *Neurocomputing* 469 (2022), pp. 332–345, DOI: 10.1016/j.neucom.2020.05.113.

[368] Jinsung Yoon, Lydia N. Drumright, and Mihaela van der Schaar, « Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN) », *in*: *IEEE Journal of Biomedical and Health Informatics* 24.*8* (2020), pp. 2378–2388, DOI: 10.1109/JBHI.2020.2980262.

[369] Sunao Yotsutsuji, Miaomei Lei, and Hiroyuki Akama, « Evaluation of Task fMRI Decoding With Deep Learning on a Small Sample Dataset », *in*: *Frontiers in neuroinformatics* 15 (2021), pp. 577451–577451, DOI: 10.3389/fninf.2021.577451.

[370] Yumei Yue et al., « ALFF and ReHo Mapping Reveals Different Functional Patterns in Early- and Late-Onset Parkinson's Disease », *in*: *Frontiers in Neuroscience* 14 (2020), ISSN: 1662-453X, DOI: doi.org/10.3389/fnins.2020.00141.

[371] Yufeng Zang et al., « Regional homogeneity approach to fMRI data analysis », *in*: *NeuroImage* 22.*1* (2004), pp. 394–400, ISSN: 1053-8119, DOI: 10.1016/j.neuroimage.2003.12.030.

[372] Oliver Zendel et al., « How Good Is My Test Data? Introducing Safety Analysis for Computer Vision », *in*: *International Journal of Computer Vision* 125.*1* (2017), pp. 95–109, ISSN: 1573-1405, DOI: 10.1007/s11263-017-1020-z.

[373] Xu Zheng et al., « STaDA: Style Transfer as Data Augmentation », *in*: *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019) - Volume 4: VISAPP*, vol. 2, SCITEPRESS, 2019, pp. 107–114, ISBN: 978-989-758-354-4, DOI: 10.5220/0007353401070114.

[374] Jun-Yan Zhu et al., « Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks », *in*: *IEEE International Conference on Computer Vision, ICCV*, IEEE Computer Society, 2017, pp. 2242–2251.

[375] Peiye Zhuang, Alexander G Schwing, and Oluwasanmi Koyejo, « Fmri data augmentation via synthesis », *in*: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE, 2019, pp. 1783–1787.

[376] Qi-Hong Zou et al., « An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: Fractional ALFF », *in*: *Journal of Neuroscience Methods* 172.*1* (2008), pp. 137–141, ISSN: 0165-0270, DOI: 10.1016/j.jneumeth.2008.04.012.

COLLEGE DOCTORAL BRETAGNE
MATHS, TELECOMS
INFORMATIQUE, SIGNAL
SYSTEMES, ELECTRONIQUE

Université de Rennes

Titre : Exploration et atténuation de la variabilité analytique en IRM fonctionnelle par apprentissage de représentations

Mot clés : Imagerie cérébrale, variabilité analytique, apprentissage de représentations

Résumé : Les études d'imagerie cérébrale sont soumises à un grand nombre de sources de variabilité, à différents niveaux. Dans cette thèse, nous nous intéressons aux variations induites par différentes méthodes d'analyse, également appelé *variabilité analytique*. Ce phénomène est désormais connu dans la communauté, l'objectif est maintenant de mieux comprendre les facteurs menant à cette variabilité et de trouver des solutions pour mieux la prendre en compte. Ici, nous apprenons des représentations des résultats d'IRMf, une technique d'imagerie qui permet d'étudier l'activité cérébrale, pour répondre aux défis liés à la variabilité analytique. Tout d'abord, nous proposons deux solutions pour faciliter la ré-utilisation des nombreuses cartes statistiques disponibles dans les bases de données publiques. Ensuite, nous explorons l'espace analytique et présentons un ensemble de données multi-pipeline que nous avons utilisé pour explorer la stabilité des relations entre les méthodes d'analyse et la validité des études combinant des données traitées avec différentes méthodes. Nos résultats montrent que nos solutions utilisant l'apprentissage non supervisé, associées à une meilleure connaissance de l'espace analytique, permettent le développement d'études robustes avec des données plus nombreuses et diversifiées provenant des données publiques.

Title: Exploring and mitigating analytical variability in fMRI results using representation learning

Keywords: Brain imaging, analytical variability, representation learning

Abstract: Brain imaging studies are subjected to a large number of sources of variability, arising at different levels. In this thesis, we focus on the variations in the results induced by different pipeline implementations, also known as analytical variability. While this phenomenon is now well known in the community, there is a need for a better understanding of the factors leading to this variability and for solutions to take it into account when building studies. Here, we aim at building comprehensible and meaningful representations of fMRI results, a brain imaging technique that explores brain activity under different contexts, to answer different challenges related to analytical variability. In a first set of contributions, we propose two solutions to facilitate the reuse of the large amount of statistic maps available in public databases. In a second set of contributions, we dive into the fMRI analytical space and start by presenting a multi-pipeline dataset that we used to explore the stability of pipelines relationships and the validity of studies combining data processed with different pipelines. Our results show that our methods based on unsupervised learning, coupled with a better knowledge of the analytical space, could facilitate the development of studies with larger and more diverse data by re-using public data.